



**Desarrollo de Técnica
Reconstructiva y Contrastiva para
el análisis de Palabras en Lenguajes
de Señas mediante Espacios
Latentes con Datos y Recursos
Limitados**

DANIEL CAMILO BERNAL TERNERA

UNIVERSIDAD SERGIO ARBOLEDA
ESCUELA DE CIENCIAS EXACTAS E INGENIERÍA
CIENCIAS DE LA COMPUTACIÓN E INTELIGENCIA ARTIFICIAL
BOGOTÁ D.C.
2025

**Desarrollo de Técnica
Reconstructiva y Contrastiva para
el análisis de Palabras en Lenguajes
de Señas mediante Espacios
Latentes con Datos y Recursos
Limitados**

DANIEL CAMILO BERNAL TERNERA

DOCUMENTO FINAL

Asesor
Juan Pablo Ospina Lopez

UNIVERSIDAD SERGIO ARBOLEDA
ESCUELA DE CIENCIAS EXACTAS E INGENIERÍA
CIENCIAS DE LA COMPUTACIÓN E INTELIGENCIA ARTIFICIAL
BOGOTÁ D.C.
2025

Nota de aceptación:

Firma del presidente del jurado

Firma del jurado

Firma del jurado



UNIVERSIDAD
SERGIO ARBOLEDA

Dedicatoria

A mi tierra, a la fuerza de mi raíz,
al profundo e inmenso cielo azul del llano
colombiano, donde pude encomendar mis
ideas a un gavilán pescador, para que las
dejara volar con la libertad del viento,
y a la selva, donde la colosal ceiba, como re-
gio soporte, permita florecer de forma incon-
mensurable los posibles proyectos que pue-
dan surgir de este estudio.



Agradecimientos

En primer lugar, me gustaría expresar mi agradecimiento a la generosidad de la Universidad Sergio Arboleda, que me brindó la oportunidad de cursar toda mi carrera con una beca que materializó mis sueños y metas. Además, agradezco por concederme las diferentes condecoraciones Honores Rodrigo Noguera que porto con orgullo y me animan a dar mi mejor esfuerzo. De igual manera, agradezco la oportunidad de permitirme expandir mis horizontes con un increíble intercambio con la Escuela de Ingeniería Julio Garavito. También quiero expresar mi agradecimiento a Juan Sebastián Malagón, que gracias a su apoyo inicial, pude desarrollar un punto de vista innovador al problema planteado en este proyecto. Por otro lado, a mi asesor Juan Pablo Ospina, por su valiosa ayuda y orientación que, a pesar de no estar presente en todo el desarrollo del proyecto, lo acogió con mucho cariño como si fuera propio. En última instancia, pero no menos importante, agradezco a mi familia y amigos que siempre me brindaron su apoyo incondicional en todo momento que lo necesité. Quiero hacer una mención especial a mi hermana Alejandra Bernal, quien dedicó considerable parte de su tiempo para ayudarme con la producción escrita de este documento y ser mi apoyo emocional en momentos complicados.

Contenido

1. Introducción	9
2. Problema de investigación	11
3. Justificación	29
4. Objetivos	32
1. Objetivo general	32
2. Objetivos específicos	32
5. Marco Teórico	33
1. Marco Histórico	33
1.1. Historia y Reconocimiento del Lenguaje de Señas	33
1.2. Evolución de la Tecnología de Reconocimiento del Lenguaje de Señas	37
2. Marco Referencial	40
2.1. Panorama general del reconocimiento de lenguaje de señas	40
2.2. Investigaciones particulares relevantes	42
2.3. Research Gap dentro del campo	44
3. Marco Conceptual	46
3.1. Fundamentos Lingüísticos y Culturales del Lenguaje de Señas	46
3.2. Perspectivas sobre Discapacidad, Accesibilidad e Inclusión	48
3.3. Barreras de Accesibilidad e Impacto de la Falta de Conciencia	49
3.4. Contexto Histórico de la Educación de Sordos	49
4. Marco Teórico	49
4.1. Teoría del Aprendizaje por Representación (Pilar Central)	49
4.2. El Principio del Information Bottleneck y la Teoría de Modelos Generativos	50
4.3. Teoría del Aprendizaje por Transferencia (Transfer Learning)	50
4.4. Teorías de Aprendizaje Auto-Supervisado (SSL) y Métrico (Metric Learning)	51
4.5. Modelado de Características Espacio-Temporales y la Hipótesis del Múltiple	51
5. Marco Tecnológico y científico	52
5.1. Tecnología para el Procesamiento del Lenguaje de Señas	52
5.2. Tecnología para el Procesamiento del Lenguaje de Señas	53
5.3. Estrategias de Aprendizaje, Optimización y Evaluación	54

6. Capítulo 1: Preprocesamiento y organización de los datos	56
1. Estado Inicial de los Datos	56
1.1. Origen y propósito de los Datos	56
1.2. Estructuración Original de los Datos	58
2. Preprocesamiento de los datos	58
2.1. Reestructuración de los Datos	58
2.2. Selección de las etiquetas	61
2.3. Redimensionamiento de tamaño y duración de las secuencias	62
2.4. Recorte del fondo de las secuencias	63
2.5. Recuento de los datos disponibles	64
7. Capítulo 2: Desarrollo de la técnica propuesta	71
1. Creación de los Datos de Entrada	71
1.1. Preparación de datos	71
1.2. Construcción de las variantes	72
2. Implementación del Modelo	72
2.1. Construcción	72
2.2. Entrenamiento	76
2.3. Evaluación	78
2.4. Obtención de Resultados	80
8. Capítulo 3: Evaluación los resultados obtenidos	81
1. Análisis general	81
2. Análisis de las gráficas y tablas por experimento	82
2.1. Experimentos con el dataset WSL	82
2.2. Experimentos con el dataset ISL	90
2.3. Experimentos con el dataset SLOVO	98
2.4. Experimentos con los tres datasets	106
3. Comparación entre experimentos	114
3.1. Experimentos con WSL	114
3.2. Experimentos con ISL	115
3.3. Experimentos con SLOVO	115
3.4. Experimentos con los tres datasets	116
9. Capítulo 4: Conclusiones y trabajos futuros	117
1. Conclusiones	117
2. Trabajos futuros y recomendaciones	118
10. Anexos	120
1. Gráficas de los Experimentos Realizados	120
1.1. Experimentos del dataset de WSL	120
1.2. Con el dataset de ISL	164
1.3. Con el dataset de SLOVO	208

Contenido

1.4. Con los tres conjuntos de datos	252
--	-----



UNIVERSIDAD
SERGIO ARBOLEDA

Introducción

En esta investigación se establece que el lenguaje de señas constituye un campo de estudio de gran relevancia, abordado desde múltiples disciplinas que reconocen su naturaleza como un sistema lingüístico completo y complejo, que tiene su propia gramática, vocabulario y estructura sintáctica. Donde la importancia de su estudio radica en la necesidad de superar las barreras de comunicación que enfrenta la comunidad con discapacidad auditiva en todo el mundo. Superar esta barrera es algo muy complicado por la enorme diversidad lingüística, con más de 200 lenguajes de señas distintos, sumado a la falta de recursos accesibles, lo cual incrementa las desigualdades en áreas críticas como la atención médica y el desarrollo profesional.

El origen de esta investigación está en el problema práctico y teórico que enfrenta el Reconocimiento de Lenguaje de Señas (SLR) con inteligencia artificial. Donde a pesar de que el estado del arte ha explorado diferentes metodologías, desde enfoques tradicionales hasta redes neuronales profundas, el avance se ve frenado por limitaciones significativas. Siendo una de las principales limitaciones identificadas la carencia de datasets grandes, diversos y estandarizados que tengan palabras completas en video. Donde la mayoría conjuntos de datos que hay son pequeños, diversos en sus condiciones de grabación, y carecen de una estructuración uniforme, lo que dificulta el entrenamiento de los modelos. Y la falta de recursos computacionales suficientes para poder entrenar modelos complejos con estos grandes datasets.

Ante este panorama, los objetivos de este proyecto se centran en el diseño y análisis de una nueva técnica de aprendizaje profundo. Con un objetivo principal que no es la clasificación directa, sino la creación de un espacio latente estructurado, entrenando específicamente para entender y codificar la información temporal que hay en las secuencias de video del lenguaje de señas y su significado semántico. Donde se busca explorar un camino diferente que pueda manejar la variabilidad de los datos existentes de una manera más robusta.

El alcance de este estudio está dentro del campo del Reconocimiento de Lenguaje de Señas Continuo (CSLR) a nivel de palabra, aplicando conceptos del aprendizaje profundo espacio temporal y el aprendizaje métrico. La metodología empleada para alcanzar los objetivos propuestos se divide en tres fases fundamentales. La primera se basa en un cuidadoso proceso de preprocesamiento y organización de los datos, diseñado para unificar y normalizar tres datasets de diferentes idiomas (WLASL, ISL y SLOVO). La segunda se encarga de describir el desarrollo de la arquitectura del modelo y la lógica

de su entrenamiento. Y por último, la tercera se encarga de definir como se hará la evaluación, lo cual incluye diferentes gráficas y tablas para poder medir la sensibilidad temporal, verificar la correcta codificación de la estructura secuencial y analizar si se separan de manera correcta las etiquetas de cada seña.

La importancia de este estudio es buscar aportar al conocimiento del campo respectivo utilizando un nuevo enfoque centrado en el aprendizaje de la dinámica temporal, representada por el lenguaje corporal, como un elemento importante que a menudo se omite. Esta investigación pretende sentar una base metodológica que sirva como un aporte para futuras investigaciones orientadas al desarrollo de soluciones más complejas y robustas en el área, demostrando la importancia de generar contribuciones significativas mediante investigaciones organizadas e innovadoras.

Problema de investigación

Panorama del lenguaje de señas

Mediante una investigación, se pudo evidenciar que el lenguaje de señas abarca un campo de estudio de gran relevancia y tamaño que ha sido abordado a lo largo de los años desde múltiples disciplinas, las cuales incluyen la lingüística, la sociología, la antropología y las ciencias de la computación, entre otras más. Esta aproximación multidisciplinaria va acorde a su naturaleza como un sistema lingüístico bastante completo y complejo, que va más allá de la percepción que usualmente se tiene de ser únicamente una serie de gestos manuales y sistemáticos.¹ Esta profundidad surge en parte por la diversidad y contexto en el que se desarrolla «Hay muchos lenguajes de señas diferentes en el mundo, cada uno con sus características únicas. Por lo tanto, traducir de una lengua nacional a una lengua de signos requiere no sólo conocimiento de la lengua de signos, sino también comprensión de la cultura y el contexto social de las personas que la utilizan.».² Además, como sugiere Probierz, como sistema lingüístico natural, el lenguaje de señas posee su propia gramática, vocabulario y estructura sintáctica, lo que requiere un enfoque interdisciplinario que integre la lingüística con el análisis cultural para poder entender toda su complejidad.³

No obstante, esta complejidad aumenta bastante al considerar que, a pesar de lo que las personas piensan comúnmente, no existe un lenguaje de señas universal. Tan solo en Estados Unidos, de los 30 millones de personas con pérdida auditiva, únicamente 6 millones utilizan el ASL, y entre esos, solo 2 millones son usuarios activos.⁴ A nivel mundial, la situación se vuelve más complicada y fragmentada, teniendo más de 200 lenguajes de señas distintos. Esta diversidad, si bien es un reflejo de la riqueza cultural, representa un gran obstáculo significativo para la comunicación global. Además, la falta de recursos accesibles se convierte en una barrera sistemática que contribuye a que las personas con discapacidades queden atrasadas respecto a sus pares sin discapacidades

¹B. Probierz et al. «Sign language interpreting - relationships between research in different areas - overview». En: *Annals of Computer Science and Information Systems* 35 (2023), págs. 213-223. DOI: [10.15439/2023f2503](https://doi.org/10.15439/2023f2503), p. 216.

²Ibid., p. 213.

³Ibid., p. 213.

⁴H. J. Adler. «Language complexities for deaf and hard of hearing individuals in their pursuit of a career in science, technology, engineering, mathematics, and medicine: Perspectives from an LSL/ASL user». En: *Ear And Hearing* (2025), págs. 851-855. DOI: [10.1097/aud.0000000000001637](https://doi.org/10.1097/aud.0000000000001637), p. 851.

en la búsqueda de una carrera profesional, como ocurre en STEMM.⁵

Estableciendo que el lenguaje es un constructo profundo, con múltiples capas en su origen y por ende en su significado, donde el componente cultural desempeña un papel fundamental como uno de los pilares centrales en su constitución. Se puede analizar que en el caso de las lenguas de señas, estas no solo funcionan como un sistema de comunicación, sino que también encarnan una identidad y una cultura propias, enriquecidas por diferentes elementos históricos, sociales y educativos que complementan y construyen su complejidad.⁶ Además, para entender el lenguaje de señas, también es importante analizar la sordera desde sus diferentes componentes «La sordera como fenómeno histórico y social. El fenómeno se analiza a través de un enfoque interdisciplinario, que incluye la historia de las personas sordas, El lenguaje de Señas, Cultura sorda, Identidad sorda, Educación para sordos, etc.».⁷ Con relación a lo anterior, el lenguaje de señas se conecta y puede abordar desde diferentes puntos de vista que lo componen para desentrañar su complejidad, como cualquier otra forma de comunicación entre humanos, van más allá de la sola funcionalidad de comunicar para ser en un pilar de la identidad cultural de las comunidades sordas a través del mundo.⁸

Siguiendo ese orden de ideas, las lenguas de señas no son muy diferentes de las orales, porque ambas están ligadas a variaciones dialécticas que son un reflejo de la diversidad en la cultura de las diferentes regiones de los hablantes. Cuando se habla de variaciones, se hace referencia a acentos, regionalismos y otras expresiones lingüísticas que enriquecen la diversidad del lenguaje.⁹ Con lo anterior claro, ya no es descabellado pensar que el Lenguaje de Señas Colombiano (LSC) presenta diferencias entre regiones, como entre la región del pacífico y de la capital, siguiendo la lógica del español que se habla allí, donde se pueden notar grandes diferencias en los acentos de los habitantes de estas áreas. Este concepto antes descrito se conoce como la diversidad dialectal en las lenguas de señas, este se alinea con el reconocimiento de la multilingualidad en los distintos estudios dirigidos a personas con discapacidad auditiva, en los cuales se ha podido observar que está ocurriendo un cambio dentro el estudio de múltiples lenguas, los cuales tienen en cuenta las lenguas de señas heredadas.¹⁰ Este cambio se ha podido observar en diferentes estudios «Los académicos a menudo realizaban sus estudios en un solo par de idiomas (por ejemplo, inglés y ASL). El interés por el bilingüismo bimodal se ha desplazado recientemente hacia el multilingüismo (Swanwick 2016), incluidas las

⁵Adler, «Language complexities for deaf and hard of hearing individuals in their pursuit of a career in science, technology, engineering, mathematics, and medicine: Perspectives from an LSL/ASL user», p. 851.

⁶D. Bedoin. «Exploring identity building, language transmission and educational strategies for immigrant d/Deaf multilingual learners». En: *Journal of Multilingual and Multicultural Development* (2024), págs. 162-175. DOI: [10.1080/01434632.2024.2390570](https://doi.org/10.1080/01434632.2024.2390570), p. 166.

⁷Ibid., p. 163.

⁸Ibid., p. 163.

⁹Ibid., p. 162.

¹⁰Ibid., p. 166.

lenguas patrimoniales habladas y de signos.».¹¹ Lo que implica que cada vez es más fácil, al tener más estudios en esta área y enfoque, sustentar que las lenguas de señas no son homogéneas, sino que, por el contrario, estas tienen ciertas variaciones que, como se establece con anterioridad, son un reflejo de la riqueza lingüística y cultural de las comunidades con alguna discapacidad de tipo auditivo. Es por eso que también se pueden plantear posibles conflictos relacionados con problemas sociales actuales, como lo puede ser el incremento de la migración, donde las lenguas heredadas pueden jugar un papel esencial, al significar choques culturales entre la lengua establecida en el territorio y la que ha tenido que desplazarse e integrarse al mismo.

Por lo cual, cuando se empieza a ver las lenguas de señas desde otra perspectiva y no solo como herramientas comunicativas, se puede ver que constituyen un componente muy importante de la identidad cultural de las diferentes comunidades con discapacidad al rededor del mundo. Entonces, cuando se tiene esta identidad, que va más allá de las limitaciones físicas y se construye a partir de una comunidad que históricamente se ha desarrollado como ninguna otra por sus diferentes condiciones de vida y contextos, se puede ver que en realidad abarca un área inmensamente diversa de experiencias culturales y lingüísticas. Como indica Bedoin «Las personas sordas han sido consideradas durante mucho tiempo miembros de una comunidad única y uniforme, con una lengua de signos, una cultura y una identidad sordas. Esto conduce a la afirmación de una cultura sorda homogénea o incluso universal, en la que una fuerte identidad sorda juega un papel positivo en el reconocimiento de las comunidades sordas e».¹²

El problema con el lenguaje de señas

Teniendo en cuenta los estudios de las experiencias de las personas que tienen esta discapacidad, que muestran cómo los malentendidos, en su forma de comunicarse usando el lenguaje de señas, no solo impulsan el uso de etiquetas erróneas, sino que también, como se aborda en la justificación mas adelante, a potenciar desigualdades en la atención médica, afectando la calidad de vida de las PDHL. Además de revisar la importancia cultural y significativa que puede tener algo tan profundo a la comunidad como lo es el lenguaje, entre otros motivos que se irán explorando a lo largo del documento, se decide abordar el problema de la barrera del idioma en el lenguaje de señas. Con la intención de poder aportar a futuras investigaciones que a través de ciertas técnicas de inteligencia artificial puedan llegar a crear soluciones efectivas que impulse la unificación del lenguaje de señas a nivel global.

¹¹Ibid., p. 163.

¹²Ibid., p. 163.

Indagación de estudios previos con el lenguaje de señas

Ahora bien, con un problema definido que se quiere solucionar, se realiza una investigación al estado del arte relacionado con la integración de tecnologías avanzadas, como la inteligencia artificial, el aprendizaje automático y el procesamiento del lenguaje natural con el lenguaje de señas para poder encontrar un campo que no haya sido trabajo de manera exhaustiva, identificando lo que se realizó y cuál es la importancia que cada estudio puede aportar. Es importante precisar que no se mencionarán todos los estudios consultados en este apartado, pues además de ser bastante extenso y perder la cohesión entre los argumentos, solo se quiere mencionar a los principales estudios que más adelante moldearían el curso de la investigación y el cómo se llegó a una pregunta de investigación bien delimitada, los demás estudios serán referenciados en el marco teórico.

Al revisar el estudio «A Survey on Chinese Sign Language Recognition: From Traditional Methods to Artificial Intelligence»¹³ podemos ver que esencialmente se trata de una revisión exhaustiva de las metodologías y tecnologías empleadas en el reconocimiento de la Lengua de Señas China (CSL) a lo largo de los últimos 20 años. Lo realiza haciendo especial énfasis en la evolución de los métodos, desde enfoques tradicionales hasta innovaciones basadas en inteligencia artificial.

Inicialmente, durante la realización del estudio se utilizaron Modelos Ocultos de Markov (HMM) y Máquinas de Vectores de Soporte (SVM), las cuales se centraban en la clasificación de gestos a partir de determinadas características extraídas, así como la estrategia Dynamic Time Warping (DTW), la cual permitía alinear secuencias de gestos con variaciones en la velocidad. Sin embargo, en la última década, la investigación cambió su punto de vista, ahora hacia Redes Neuronales Profundas (DNN), que han demostrado ser muy importantes para mejorar la precisión y robustez del reconocimiento de gestos, por otro lado, la implementación de modelos híbridos que combinan diferentes arquitecturas para optimizar el rendimiento también han demostrado ser de gran utilidad.

Además, el estudio también abarca la exploración de la integración de múltiples modalidades, lo anterior quiere decir que, busca fusionar características tanto de gestos como con expresiones faciales y con movimientos de labios, lo que demostró ser crucial para una interpretación más precisa y fiel del lenguaje de señas. Teniendo todo lo anterior en cuenta, los objetivos de esta investigación no solo van hasta mejorar la precisión y robustez de los sistemas de reconocimiento, sino que también desarrollar conjuntos

¹³X. Jiang et al. «A survey on Chinese sign language recognition: From traditional methods to Artificial Intelligence». En: *Computer Modeling in Engineering & Sciences* 140.1 (2024), págs. 1-40. DOI: [10.32604/cmes.2024.047649](https://doi.org/10.32604/cmes.2024.047649), p. 1-40.

de datos más grandes y estandarizados para que se pueda facilitar el entrenamiento de modelos, así como promover aplicaciones prácticas que hagan accesible esta tecnología al público general. No obstante, a pesar de los avances significativos que se mencionan, el documento también registra que hay desafíos que no se han podido superar aún, como lo es la fusión de características de lengua de señas continuas y la coordinación de gestos con expresiones faciales, incluso señalando la necesidad de mejorar la robustez y el rendimiento de este tipo de algoritmos en tiempo real. En conclusión, se puede prever que el continuo desarrollo de nuevas tecnologías y la integración de diferentes campos científicos impulsarán aún más de forma positiva el reconocimiento de la lengua de señas china, con un énfasis determinado en modelos híbridos y diferentes técnicas de aprendizaje profundo, lo que promete generar un futuro mucho más accesible y efectivo para las todas aquellas personas con discapacidades auditivas.

Por otro lado, al revisar el estudio «A Systematic Review of Hand Gesture Recognition: An Update From 2018 to 2024»¹⁴ se puede ver que este presenta una revisión sistemática de la literatura sobre el reconocimiento de lenguaje de señas, haciendo énfasis especial en los métodos de visión, sensores y enfoques híbridos que se han utilizado entre 2018 y 2024.

Cuando se revisa en profundidad, este muestra que la metodología del estudio se basa en un proceso estructurado que incluye primeramente la identificación del dominio de investigación, luego la selección de estudios a través de bases de datos como IEEE Xplore, ScienceDirect, Scopus y Web of Science, y por último la aplicación de criterios de inclusión y exclusión para filtrar un total de 1,316 artículos, de los cuales solo se seleccionaron 256 para un análisis exhaustivo. Los objetivos principales del estudio se centran en evaluar la representación de lenguaje de señas, la adquisición de datos y la precisión de los métodos de reconocimiento. Así mismo, se concluye que la precisión del reconocimiento puede variar en gran medida, alcanzando de esta manera entre 64 % y 98 % en casos donde se puede conocer la identidad del intérprete, y entre 52 % y 98 % en situaciones donde no es relevante, con un promedio de 87.9 % y 79 % respectivamente.

Al revisar lo anterior con detenimiento, se pueden identificar desafíos en la caracterización de gestos continuos y de igual manera se destaca la necesidad de mejorar la viabilidad práctica de los sistemas de reconocimiento de gestos basados en visión, así como resaltar la importancia de integrar otros enfoques interdisciplinarios que consideren aspectos de interacción que estén enfocados en la interacción humano-computadora sumada a la complejidad del lenguaje de señas, que al igual que el estudio anterior, se concluye que comprende no solo movimientos de las manos, sino que también expresiones faciales y lenguaje corporal.

¹⁴A. O. Hashi, S. Z. M. Hashim y A. B. Asamah. «A Systematic Review of Hand Gesture Recognition: An update from 2018 to 2024». En: *IEEE Access* 12 (2024), págs. 1-35. DOI: [10.1109/access.2024.3421992](https://doi.org/10.1109/access.2024.3421992), p. 1-35.

El documento de la investigación menciona varios métodos utilizados en el reconocimiento de gestos, específicamente en el contexto del reconocimiento de lenguaje de señas. Estos métodos se pueden clasificar en tres categorías principales, siendo los primeros los métodos Basados en Visión. Estos métodos se caracterizan por utilizar información visual para reconocer gestos, esto quiere decir que, incluyen varias etapas, como la recopilación de datos, el preprocesamiento, la segmentación, la extracción de características y la clasificación. Estos se enfocan de manera principal en el análisis de imágenes o secuencias de video para identificar gestos estáticos y dinámicos.

En otra instancia, están los métodos Basados en Sensores, los cuales se enfocan en utilizar sensores, como indica su nombre, que usualmente están integrados en guantes o dispositivos portátiles, que son utilizados para capturar datos sobre los movimientos de las manos. Estos sensores son capaces de medir parámetros como lo son la flexión, la orientación y la rotación de la mano. Esto hace que este método sea menos susceptible a las condiciones que se relacionan al entorno, lo que permite una captura de datos más precisa, aunque puede ser percibido como incómodo debido a la necesidad de usar múltiples dispositivos.

Por último, se tiene a los métodos Híbridos, estos métodos se caracterizan por combinar técnicas de visión además de sensores para aprovechar las ventajas de ambos enfoques. Al integrar datos visuales y de sensores, se busca mejorar la precisión y la robustez del reconocimiento de gestos. Es importante recalcar que el documento también destaca la importancia de emplear algoritmos de aprendizaje profundo, como lo pueden ser las redes neuronales convolucionales (CNN), para mejorar la eficacia del reconocimiento de gestos, así como la necesidad de abordar aspectos no manuales del lenguaje de señas, como las expresiones faciales y el lenguaje corporal, lo cual puede significar en la omisión de datos que son cruciales para una interpretación fiel del significado en el lenguaje de señas.

Siguiendo con la idea del uso de sensores, se analiza el estudio «American Sign Language Recognition and Translation Using Perception Neuron Wearable Inertial Motion Capture System»¹⁵ el cual trata del reconocimiento y la traducción de la Lengua de Señas Americana (ASL) a través de un sistema de captura de movimiento inercial conocido como «Perception Neuron».

A lo largo del estudio se recopila un conjunto de datos en los cuales se incluyen 300 oraciones comúnmente usadas en ASL, compuestas por 455 gestos diferentes, a los que se les asignan dos tipos de etiquetas. La primera de ellas se basa en la gramática de la lengua de señas, mientras que la otra se asigna desacuerdo a la gramática del lenguaje hablado. Así mismo, se desarrollan dos modelos de procesamiento de lenguaje natural

¹⁵Y. Gu, H. Oku y M. Todoh. «American Sign Language Recognition and translation using Perception Neuron Wearable Inertial Motion Capture System». En: *Sensors* 24.2 (2024), págs. 1-15. doi: [10.3390/s24020453](https://doi.org/10.3390/s24020453), p. 1-15.

(NLP) para el reconocimiento de secuencias y la traducción de extremo a extremo.

La metodología de los procesos antes descritos van desde la segmentación manual de palabras para su validación, es decir, se seleccionan 20 palabras para evaluar la precisión del clasificador, que combina un extractor de características CNN y una capa de clasificación totalmente conectada, donde los resultados muestran una alta precisión en la clasificación con un promedio de alrededor del 88 %.

A pesar de que se identifican confusiones entre palabras con gestos similares, las conclusiones del estudio muestran que, aunque el modelo de reconocimiento logra una alta precisión, se encontró que la traducción de extremo a extremo presenta una mayor cantidad de errores debido a la falta de un conocimiento gramatical. Aparte de eso, se observa que la validación a nivel de oración se vuelve más difícil con un vocabulario más extenso, lo que termina reduciendo la tasa de precisión. Dentro del estudio se subraya de gran manera la importancia de las diferencias individuales dentro de cada uno de los gestos y la necesidad de mejorar la comprensión gramatical para optimizar la traducción de ASL.

Para tener un enfoque diverso en la metodología para hallar una solución, se revisa el estudio «Sign language recognition using the fusion of image and hand landmarks through multi-headed convolutional neural network»¹⁶ el cual se centra en la fusión de técnicas de procesamiento de imágenes y aprendizaje profundo, más específicamente a través de la utilización de una red neuronal convolucional multi-cabeza (CNN). Para lograrlo se hace uso de un conjunto de datos de imágenes denominado «Finger Spelling, A» para entrenar el modelo, esto con el objetivo de mejorar la precisión en la detección de gestos de la lengua de señas americana (ASL).

Las metodologías incluyen la extracción de características utilizando transformadas de características invariantes a escala como lo es (SIFT) y la clasificación mediante el uso de redes neuronales, así también como la implementación de técnicas de detección de bordes como lo pueden ser Canny y ORB. Así mismo, como se ve en otros estudios, se aplican redes neuronales convolucionales 3D (3DRCNN) para capturar tanto la información espacial como temporal de los gestos.

Para finalizar, este estudio concluye en que la combinación del procesamiento de imágenes tradicional sumado con la extracción de puntos de referencia de la mano más el uso de CNN multi-cabeza puede permitir una tasa de detección mucho mejor, logrando unos resultados positivos incluso en condiciones donde hay mucho ruido y variabilidad en las imágenes. A través de este enfoque no solo se mejora la precisión del reconocimiento, sino que también reduce la necesidad de hardware costoso, lo que

¹⁶R. K. Pathan et al. «Sign language recognition using the fusion of image and hand landmarks through multi-headed convolutional neural network». En: *Scientific Reports* 13.1 (2023), págs. 1-11. DOI: [10.1038/s41598-023-43852-x](https://doi.org/10.1038/s41598-023-43852-x), p. 1-11.

lo hace accesible.

Para analizar en mejor medida como se pueden usar tanto videos como imágenes, se introduce «Sign Language Recognition System for Communicating to People with Disabilities»¹⁷ Este estudio se caracteriza por usar una metodología que se basa en redes neuronales convolucionales (CNN) que para poder facilitar la comunicación entre personas con discapacidades auditivas y la sociedad en general, realiza una recolección de un conjunto de datos de señas en lenguaje americano (ASL) de Kaggle, que incluye 24 clases de gestos, aplicando un filtro de desenfoque gaussiano para mejorar la calidad de las imágenes.

Posteriormente, para poder lograr el objetivo, se implementan diversas técnicas de procesamiento de imágenes, como lo es la detección de contornos, la extracción de características KAZE y la conversión de espacios de color, para poder preparar los datos para su respectiva clasificación. Esta clasificación se realiza utilizando algoritmos como el de vecinos más cercanos y, principalmente, el de CNN, que ha demostrado ser el método más preciso, siendo capas de alcanzar una precisión del 100 % en imágenes y del 73 % en videos.

En conclusión, lo valioso de este sistema es que está diseñado para recibir entradas de video en tiempo real, procesar las imágenes y convertir los gestos en texto que puede leer cualquier persona, lo cual permite una interacción más fluida y efectiva. También se destaca la importancia de este tipo de tecnologías en la mejora de la comunicación para aquellas personas con discapacidad, subrayando la necesidad de herramientas accesibles y precisas que faciliten la interacción en situaciones cotidianas y críticas.

Para profundizar aún más sobre los videos y su relación con los autoencoders, se trae a colación el documento «VTAN: A Novel Video Transformer Attention-Based Network for Dynamic Sign Language Recognition»¹⁸ el cual se caracteriza por utilizar un modelo innovador para el reconocimiento de lenguaje de señas de una manera dinámica, el modelo combina un autoencoder convolucional (CAE) y un transformador que se enfoca en atención suave.

Ahora bien, en este documento se abordan dos problemas muy importantes, siendo el primero la redundancia de fotogramas importantes en videos de lenguaje de señas y el segundo, la necesidad de enfocarse principalmente en las regiones de las manos, las cuales se considera que son muy importantes para la traducción de gestos. Para lograrlo se usa un módulo de agregación de características visuales, a través de un CAE

¹⁷Y. Obi et al. «Sign language recognition system for communicating to people with disabilities». En: *Procedia Computer Science* 216 (2023), págs. 13-20. DOI: [10.1016/j.procs.2022.12.106](https://doi.org/10.1016/j.procs.2022.12.106), p. 13-20.

¹⁸Z. Deng et al. «VTAN: A novel Video Transformer Attention-Based Network for Dynamic Sign Language Recognition». En: *Computers, Materials & Continua* 82.2 (2024), págs. 2793-2812. DOI: [10.32604/cmc.2024.057456](https://doi.org/10.32604/cmc.2024.057456), p. 2793-2812.

para extraer fotogramas clave mediante clustering K-means, reduciendo de esta manera los datos redundantes y mejorando de manera considerable la eficiencia computacional. Por otro lado, este mismo módulo de mejora de características espacio-temporales (STHE) emplea un transformador para poder priorizar las características de las manos, capturando de esta manera dinámicas existentes entre el espacio y su temporalidad.

Para probar lo anterior, los experimentos se realizaron en los conjuntos de datos AUTSL y SLR500, los cuales muestran mejoras significativas en precisión, alcanzando la cifra de 93.6 % y 91.3 %, respectivamente, destacando la efectividad que tiene VTAN frente a otros métodos previos.

Por último, para finalizar esta breve compilación de los documentos que principalmente moldearon la idea principal, se tiene el trabajo «G2P-DDM: Generating Sign Pose Sequence from Gloss Sequence with Discrete Diffusion Model»¹⁹ el cual propone un modelo innovador, este modelo se enfoca en la producción de lenguaje de señas (SLP), centrándose en la transformación de secuencias de etiquetas, es decir anotaciones o etiquetas textuales de los signos, en secuencias de poses (G2P) En pocas palabras, una traducción de texto a señas.

Para poder lograrlo, el estudio utiliza un enfoque que consta de dos etapas, la primera de estas se basa en que el modelo llamado «Pose-VQVAE» convierte secuencias continuas de poses en códigos latentes discretos, el modelo realiza esta tarea dividiendo el esqueleto humano en tres partes principales siendo estas el cuerpo, la mano derecha y la mano izquierda, empleando múltiples estructuras en códigos para mejorar la reconstrucción. Mientras que en la segunda etapa, la cual se llama «G2P-DDM», se basa principalmente en un modelo de difusión discreta, es decir, el modelo genera secuencias de poses a partir de las diferentes etiquetas usando la herramienta CodeUnet. Posteriormente, se utiliza un algoritmo de clustering secuencial-KNN el cual se encarga de predecir longitudes variables en las secuencias.

En otras palabras, este trabajo centra su investigación en la producción automática de lenguaje de señas al discretizar el espacio de poses y emplear modelos de difusión, para luego abordar desafíos como lo pueden ser la variabilidad en la longitud de las secuencias y la complejidad de los gestos dinámicos. A través del uso de representaciones latentes discretas, las cuales permiten explorar el espacio latente para tareas como la generación de datos sintéticos hasta la interpolación de gestos.

¹⁹P. Xie et al. «G2P-DDM: Generating Sign Pose Sequence from Gloss Sequence with Discrete Diffusion Model». En: *Proceedings of the AAAI Conference on Artificial Intelligence* 38.6 (2024), págs. 6234-6242. doi: [10.1609/aaai.v38i6.28441](https://doi.org/10.1609/aaai.v38i6.28441), p. 6234-6242.

Identificación de los conceptos principales del problema

Teniendo en cuenta los estudios que se seleccionaron previamente, se realiza un recuento secuencial de cómo cada uno influenció y moldeó el planteamiento del problema. Bajo ese orden de ideas, se comienza con la necesidad de realizar un avance en un campo, de preferencia uno aún no investigado en gran medida, para poder contribuir en el mismo realizando una investigación que sirva de base.

El primer estudio, «A Survey on Chinese Sign Language Recognition: From Traditional Methods to Artificial Intelligence»,²⁰ el cual mostró una revisión exhaustiva durante dos décadas de la evolución en el reconocimiento de la Lengua de Señas China (CSL), con la cual se destacó la transición de métodos tradicionales hacia enfoques más recientes basados en Redes Neuronales Profundas (DNN). Con este enfoque se llegó a una conclusión muy importante, la cual fue resaltar la importancia de capturar la información completa del intérprete, es decir, gestos manuales, expresiones faciales y movimientos corporales, en lugar de limitarse solamente a las manos. Además de mostrar los beneficios de centrarse en la búsqueda de una solución avanzada, alejándose de enfoques tradicionales y justificando la evolución hacia técnicas de aprendizaje profundo. Por estos motivos se comenzó a indagar en la posibilidad de emplear un autoencoder capaz de procesar toda la información que se le es transmitida, asegurando de esta manera una representación más completa de los datos.

Por otro lado, el segundo documento, «A Systematic Review of Hand Gesture Recognition: An Update From 2018 to 2024»,²¹ fundamento el uso de la herramienta que se planteó en el anterior documento al revisar 1,300 estudios, de los cuales la mayoría perteneció a métodos de visión, sensores e híbridos. Esto quiere decir que los autoencoders no han sido explorados de manera sustancial en comparación con otras técnicas. Esta investigación también llegó a la conclusión que es importante considerar elementos aparte de las manos para el reconocimiento del lenguaje de señas, como las expresiones faciales y otro lenguaje corporal, para llegar a una interpretación precisa. Por lo tanto, lo encontrado solo reforzó aún más la elección de un autoencoder que capture el contexto completo.

Por el contrario, el tercer estudio, «American Sign Language Recognition and Translation Using Perception Neuron Wearable Inertial Motion Capture System»,²² ayudó a delimitar la idea, pues con su investigación, respecto a la evaluación del uso de sensores

²⁰Jiang et al., «A survey on Chinese sign language recognition: From traditional methods to Artificial Intelligence», p. 1-40.

²¹Hashi, Hashim y Asamah, «A Systematic Review of Hand Gesture Recognition: An update from 2018 to 2024», p. 1-35.

²²Gu, Oku y Todoh, «American Sign Language Recognition and translation using Perception Neuron Wearable Inertial Motion Capture System», p. 1-15.

iniciales para el reconocimiento de la Lengua de Señas Americana (ASL), expuso las limitaciones de los métodos basados en sensores. Siendo estas su gran costo, la complejidad de implementación y la incomodidad de los usuarios al llevar puestos diferentes dispositivos portátiles. Con lo anterior claro, se pudo descartar esta aproximación para poder llegar a soluciones más accesibles y escalables. Sin embargo, el estudio también resaltó la importancia de distinguir gestos similares, lo que inspiró la idea teórica de utilizar un espacio latente para diferenciarlos según sus distancias representacionales. Además, esta idea es muy compatible con el funcionamiento de los autoencoders, los cuales son capaces de generar representaciones compactas y significativas, sentando las bases para su posible uso no solo en reconocimiento, sino que también en una potencial traducción.

Teniendo la idea un poco más clara, se pudo explorar una alternativa para poder hacer funcionar lo anterior, así es como en el cuarto estudio, «Sign language recognition using the fusion of image and hand landmarks through multi-headed convolutional neural network»,²³ se demostró el potencial del uso de técnicas que involucran redes 3D convolucionales (3DRCNN) para capturar información espacial y temporal, lo cual sugirió la posibilidad de desarrollar arquitecturas que sean capaces de detectar la temporalidad de los gestos, lo cual es un aspecto muy importante del lenguaje de señas.

Ahora bien, para sustentar la utilización de los datos, el quinto estudio, «Sign Language Recognition System for Communicating to People with Disabilities»,²⁴ presentó una gran idea en el uso de CNN para procesar secuencias de video de ASL, lo cual se puede unir con el anterior estudio demostrando que es posible realizar este proceso, en este caso particular, se logró realizar con técnicas de preprocesamiento como detección de contornos y extracción de características. Este documento, además de aportar un ejemplo concreto de éxito en el procesamiento de video, reforzando la confianza en las capacidades de los autoencoders para este propósito, también reveló pasos importantes dentro del preprocesamiento de los videos para tener un mayor margen de éxito.

De igual manera, para sentar las últimas bases de las generalidades del problema, se revisó el sexto estudio, «VTAN: A Novel Video Transformer Attention-Based Network for Dynamic Sign Language Recognition»,²⁵ que introdujo un modelo que ya se centraba completamente en un autoencoder convolucional (CAE) con un transformador, lo cual deja en evidencia su posible potencial que se estableció teóricamente con los primeros estudios analizados. Además, este enfoque abordó un problema clave cuando se trata con videos, siendo este mismo, la redundancia de frames en videos de lenguaje de señas

²³Pathan et al., «Sign language recognition using the fusion of image and hand landmarks through multi-headed convolutional neural network», p. 1-11.

²⁴Obi et al., «Sign language recognition system for communicating to people with disabilities», p. 13-20.

²⁵Deng et al., «VTAN: A novel Video Transformer Attention-Based Network for Dynamic Sign Language Recognition», p. 2793-2812.

para seleccionar características relevantes y reducir datos innecesarios.

En última instancia, el séptimo documento, «G2P-DDM: Generating Sign Pose Sequence from Gloss Sequence with Discrete Diffusion Model»,²⁶ aunque tiene una aproximación muy diferente, sirvió de inspiración para constatar cómo sería evaluado el proyecto, al proporcionar un punto teórico que refuerza lo visto en previos estudios y es la implementación e importancia de representaciones latentes discretas dentro de un espacio para realizar traducciones. Esta idea fue clave para definir el enfoque final, pues sirve como puente para sustentar el concepto de utilizar autoencoders no solo para identificar patrones, sino también para traducir palabras en lenguajes de señas a través de un espacio latente. Esta traducción se basa en el comportamiento de diferentes lenguajes de señas en un mismo espacio latente donde se teoriza que, luego de un proceso determinado, es posible que se agrupen las mismas palabras de distintos lenguajes, indicando que existe alguna relación que pueda permitir la asociación de etiquetas al ingresar una nueva palabra de un lenguaje que no ha sido entrenado, esto al ubicarse cerca de sus palabras equivalentes en otros idiomas.

Delimitación del problema

Dejando en claro cómo surge la base del problema a través de ciertos documentos claves, ahora se realiza una acotación para poder enfocar el proyecto de investigación a alcances realistas y obtener resultados determinados, para hacer recomendaciones certeras sobre cómo se puede enfocar una investigación de este tipo en un campo en concreto. Es por esto que se opta por hacer la mención de «recursos determinados» como una delimitación en la investigación, esto con el fin de responder a una decisión consciente del alcance del proyecto, basada en las herramientas y capacidades accesibles en el contexto de un estudiante de pregrado de la Universidad Sergio Arboleda, sin incurrir en costos monetarios significativos.

Teniendo lo anterior en cuenta, primero se realiza una búsqueda preliminar de los datos a utilizar. Esta búsqueda es muy importante, pues desde la teoría y la práctica se ha establecido como un principio fundamental que es muy valioso en el aprendizaje automático el señalar que la calidad y el rendimiento de un modelo de IA son directamente proporcionales a la calidad y representatividad de sus datos de entrenamiento. Este principio es conocido como «Garbage In, Garbage Out».²⁷

Por eso, se establecen los siguientes parámetros de búsqueda basándose en los es-

²⁶Xie et al., «G2P-DDM: Generating Sign Pose Sequence from Gloss Sequence with Discrete Diffusion Model», p. 6234-6242.

²⁷EBSCO Information Services, Inc. *Garbage in, garbage out (GIGO)*. s.f. EBSCO. URL: <https://www.ebsco.com/research-starters/computer-science/garbage-garbage-out-gigo> (visitado 04-07-2025).

tudios previamente mencionados, comenzando por la característica que el dataset esté compuesto por videos de intérpretes realizando palabras continuas, no por deletreo de señas. Se toma esta decisión para poder detectar la temporalidad de la secuencia que compone a una palabra, como se indica en el estudio «VTAN: A Novel Video Transformer Attention-Based Network for Dynamic Sign Language Recognition»²⁸ Además, este parámetro se respalda con los estudios anteriores que denotan la importancia de buscar toda la información que se omite cuando solo se concentra en una seña en particular, al solo tener un deletreo, no se hace uso de otras partes del cuerpo que pueden dar información importante adicional.

Consecuentemente con lo anterior, se busca que los videos contengan vistas completas del cuerpo para capturar toda la información posible, de igual manera que haya cierta diversidad de intérpretes y de videos por palabra es importante, como se menciona en «A Systematic Review of Hand Gesture Recognition: An Update From 2018 to 2024»²⁹ «Surge una necesidad crítica para el desarrollo de bases de datos más completas y diversas para facilitar una investigación más exhaustiva de los sistemas SLR.» donde luego se justifica que al cumplir esta diversidad se lograría que el modelo generalizara de mejor manera.

Siguiendo con los requerimientos, unos muy importantes son los que están relacionados con su calidad, es decir, fondo limpio, condiciones buenas de iluminación, anotaciones y etiquetas consistentes y tamaños normalizados. Esto es muy importante para simplificar pasos de preprocesamiento y limpieza de datos como los que se mencionan en la sección de preprocesamiento de imágenes en el estudio «A Systematic Review of Hand Gesture Recognition: An Update From 2018 to 2024»³⁰ donde se repasa desde métodos para la eliminación de objetos no deseados en el fondo hasta la configuración del tamaño de los videos.

Por último, para lograr una traducción parcial consistente de palabras por medio de sus representaciones latentes, se necesita que se tengan las mismas etiquetas, esto implica que cuando se menciona anteriormente la necesidad de tener variedad en las palabras, es importante que se logre sin caer en regionalismos para tener una buena equivalencia entre diferentes idiomas. Consecuentemente, se necesita tener las mismas palabras en diferentes idiomas para lograr dicha equivalencia.

Después de una búsqueda exhaustiva tanto en los documentos como en la red, se llegó a la conclusión de que cuando se delimita la búsqueda según los anteriores parámetros establecidos, existe la escasez de datasets públicos, a gran escala y de alta

²⁸Deng et al., «VTAN: A novel Video Transformer Attention-Based Network for Dynamic Sign Language Recognition», p. 2793-2812.

²⁹Hashi, Hashim y Asamah, «A Systematic Review of Hand Gesture Recognition: An update from 2018 to 2024», p. 1-35.

³⁰Ibid., p. 1-35.

calidad. Los recursos que existen actualmente tiene la particularidad de no cumplir frecuentemente con las características necesarias para un entrenamiento ideal, lo cual no resulta ser una sorpresa, pues ya se han mencionado en este documento, con varios estudios, que a partir de este problema se enfocaron en la creación de datasets de calidad, como lo es el ejemplo de «A Survey on Chinese Sign Language Recognition: From Traditional Methods to Artificial Intelligence»,³¹ «A Systematic Review of Hand Gesture Recognition: An Update From 2018 to 2024»³² y «VTAN: A Novel Video Transformer Attention-Based Network for Dynamic Sign Language Recognition»³³ el cual se caracteriza por utilizar un modelo innovador para el reconocimiento de lenguaje de señas de una manera dinámica, por mencionar unos pocos.

A pesar de esto, se pudieron identificar tres datasets que cumplen con la mayoría de lo anteriormente descrito, sin embargo, tienen ciertas limitaciones. El primer dataset seleccionado consiste en videos de palabras del lenguaje de señas americano. Este dataset cuenta con alrededor de 2000 palabras, siendo el dataset más grande que hay hasta la fecha de su creación del lenguaje americano, la versión que se emplea del dataset es la 03.³⁴

El segundo dataset que se eligió fue de lenguaje de señas en indio, se escogió este porque aparte de cumplir con la mayoría de requisitos antes descritos, tiene 4292 videos de diferentes palabras, siendo éstas muy variadas.³⁵

El último dataset que se decidió utilizar consiste en 20400 videos grabados por 194 diferentes intérpretes, el problema principal con este dataset es que no están grabados en espacios uniformes y su captura fue con diferentes cámaras.³⁶

Se decidió utilizar solo estos tres datasets, pues son los que más videos etiquetados contienen, además de estar abiertos al público general, permitiendo una descarga accesible. También son los que presentan la mayor calidad en la presentación de sus datos, haciéndolos la mejor opción. Sin embargo, las etiquetas y la manera en que los datos están estructurados son totalmente diferentes entre sí, lo que complica el preprocesa-

³¹Jiang et al., «A survey on Chinese sign language recognition: From traditional methods to Artificial Intelligence», p. 1-40.

³²Hashi, Hashim y Asamah, «A Systematic Review of Hand Gesture Recognition: An update from 2018 to 2024», p. 1-35.

³³Deng et al., «VTAN: A novel Video Transformer Attention-Based Network for Dynamic Sign Language Recognition», p. 2793-2812.

³⁴Dongxu Li et al. *WLASL: Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset*. s.f. WLASL. URL: <https://dxli94.github.io/WLASL/> (visitado 05-07-2025).

³⁵A. Sridhar et al. *INCLUDE: A Large Scale Dataset for Indian Sign Language Recognition*. Data set, ACM Multimedia 2020 (ACMMM2020). Zenodo. 2020. DOI: [10.5281/zenodo.3413528](https://doi.org/10.5281/zenodo.3413528). URL: <https://zenodo.org/record/4010759> (visitado 05-07-2025).

³⁶Alexander Kapitanov et al. *Slovo: Russian Sign Language Dataset*. arXiv:2305.14527. 2023. URL: <https://arxiv.org/abs/2305.14527> (visitado 05-07-2025).

miento y unificación de un formato de los datos. Por las razones anteriores, se decidió no incluir más datasets, los diferentes costos que representan en tiempo de investigación el incluir más conjuntos de datos que no están estructurados de la misma manera no es viable. Además, el experimento funciona teóricamente de buena manera con solo dos lenguajes de señas.

Los recursos que se utilizaron para este proyecto de investigación fueron brindados por la universidad gracias a las cuentas de prueba académicas de AWS academy learner lab, se contó con 50 dólares, con los cuales se pudo acceder a diferentes cuadernos del servicio Amazon SageMaker AI, que se alojaron en la nube para ejecutar todo el código. Las mejores instancias a las que se pudieron acceder son las ml.c5.xlarge, estas cuentan con optimización para computación, no son de inicio rápido, 4 vCPU y 8GiB de memoria. No se contó con memoria GPU para la realización del proyecto, por lo cual no se pudieron realizar diferentes estrategias de cómputo con CUDA o aceleración con gráficas.³⁷ Es importante recalcar que cada sesión del laboratorio tiene un máximo de 4 horas para trabajar antes de que se reinicie, una vez ocurre esto, es necesario volver a cargar todos los recursos, pues se borra toda la información que tenía el kernel, además de tener que encender todos los servicios que se estaban utilizando antes del apagado.

La última delimitación que se realiza antes de tener el problema completamente formulado es la teórica, se toma la decisión de enfocar el trabajo de investigación en una técnica de aprendizaje autosupervisado (Self-Supervised Learning) para aprender representaciones de video. Se piensa abordar esta estrategia con un modelo que se compone de un autoencoder basado en convoluciones 3D (3D-CNN) para extraer características espaciotemporales, y una red neuronal recurrente (GRU) bidireccional para modelar las secuencias temporales. Posteriormente, para estructurar el espacio latente, se hace uso de una función de pérdida compuesta que incluye la reconstrucción (MSE), múltiples pérdidas triplet para la estructura temporal y una pérdida KL. Por último, se ve el resultado de la estructuración de las representaciones por medio de PCA y UMAP.

La decisión de utilizar un enfoque de aprendizaje autosupervisado nace del estudio «Advancing video self-supervised learning via image foundation models»³⁸ el cual señala que hay una manera de realizar de manera eficiente el aprendizaje de un modelo cuando no se dispone de datos etiquetados a nivel de frame. Además, presenta un concepto que se utilizará en el proyecto de investigación, las tareas de pretexto, según la anterior investigación, estas son las que se centran en crear desafíos que permitan a los modelos aprender relaciones espaciales y temporales en los datos sin etiquetas.

Así mismo, la arquitectura 3D se justifica al leer el documento «Batch feature stan-

³⁷Amazon Web Services. *Tipos de instancias disponibles para su uso con Studio Classic*. Publicado el 24 de octubre de 2022. Amazon SageMaker. 2022. URL: https://docs.aws.amazon.com/es_es/sagemaker/latest/dg/notebooks-available-instance-types.html (visitado 05-07-2025).

³⁸J. Wu, Z. Huang y C. Liu. «Advancing video self-supervised learning via image foundation models».

dardization network with triplet loss for weakly-supervised video anomaly detection»³⁹ en el cual se expone en la sección 2.1 la idea de que los autoencoders, especialmente los 3D, son muy buenos cuando se emplean para aprender representaciones en videos, siendo capaces de reconstruir, con un buen rendimiento, eventos normales y potencialmente anomalías debido a su capacidad de generalización en los patrones espacio-temporales que capturan.

En acompañamiento a esta arquitectura, se establece que la mejor opción sería incluir una capa GRU bidireccional, esto gracias a las afirmaciones de la investigación «Entwicklung und Evaluation eines Deep-Learning Algorithmus für die Worterkennung aus Lippenbewegungen für die deutsche Sprache»⁴⁰ en la cual se establece que una red neuronal recurrente (GRU) bidireccional para modelar las secuencias temporales es una arquitectura más que adecuada para tareas de reconocimiento de acciones y lenguaje de señas en formato de video. Esto lo explica con diferentes aproximaciones, mencionando que las CNN 3D extraen información espacial y temporal de los videos, mientras que las GRU modelan las secuencias temporales, haciéndolo un modelo híbrido que es capaz de captar tanto información espacial como temporal en videos.

En cuanto a las funciones de pérdida, se escoge la aproximación antes descrita en el primer párrafo, basándose en el documento «ATCM-Net: A deep learning method for phase unwrapping based on perception optimization and learning enhancement»⁴¹ en donde se describe la formulación de una función de pérdida compuesta y cómo esta combina diferentes componentes para mejorar el entrenamiento y la capacidad del modelo. Al realizar una buena combinación, se permite enfocar el aprendizaje en múltiples objetivos, siguiendo prácticas recomendadas para mejorar la precisión y la robustez del modelo. Sin embargo, el método que se expone es uno limitado, es por eso que se decidió adoptar el concepto de la función compuesta, pero empleando triplet loss. Se tomó esta decisión luego de analizar el documento «A novel triplet loss architecture with visual explanation for detecting the unwanted rotation of bolts in safety-critical environments»⁴² donde se expone que el triplet loss está diseñado para ser robusto frente a variaciones en las condiciones de captura. Esto es muy bueno porque permite que el

En: *Pattern Recognition Letters* 192 (2025), págs. 22-28. doi: [10.1016/j.patrec.2025.03.015](https://doi.org/10.1016/j.patrec.2025.03.015).

³⁹S. Yi, Z. Fan y D. Wu. «Batch feature standardization network with triplet loss for weakly-supervised video anomaly detection». En: *Image and Vision Computing* 120 (2022), págs. 1-9. doi: [10.1016/j.imavis.2022.104397](https://doi.org/10.1016/j.imavis.2022.104397).

⁴⁰D. N. Pham y T. Rahne. «Entwicklung und Evaluation eines Deep-Learning-Algorithmus für die Worterkennung aus Lippenbewegungen für die deutsche Sprache». En: *HNO* 70.6 (2022), págs. 456-465. doi: [10.1007/s00106-021-01143-9](https://doi.org/10.1007/s00106-021-01143-9).

⁴¹M. Xu et al. «ATCM-Net: A deep learning method for phase unwrapping based on perception optimization and learning enhancement». En: *Optics & Laser Technology* 190 (2025), págs. 1-20. doi: [10.1016/j.optlastec.2025.113185](https://doi.org/10.1016/j.optlastec.2025.113185).

⁴²T. Bolton et al. «A novel triplet loss architecture with visual explanation for detecting the unwanted rotation of bolts in safety-critical environments». En: *Engineering Applications of Artificial Intelligence* 156 (2025), págs. 1-16. doi: [10.1016/j.engappai.2025.111097](https://doi.org/10.1016/j.engappai.2025.111097).

modelo se vuelva invariante a ciertos tipos de ruido, lo que es crítico en el contexto del proyecto de investigación donde, como se explicó anteriormente, los datasets no son los mejores y las imágenes pueden no ser perfectas.

De igual manera, para poder garantizar que el modelo pueda tener una sólida representación óptima de un dato, se integra el uso combinado de la pérdida de Divergencia de Kullback-Leibler (KL) y el Error Cuadrático Medio (MSE) respaldándose en el documento «TB-Net: Intra- and inter-video correlation learning for continuous sign language recognition»⁴³ el cual muestra que reconocen la relevancia del principio del IB para el aprendizaje de representaciones. Para ahondar más en este concepto y cómo afecta al aprendizaje, se refirió a la bibliografía de este documento para encontrar que en «PAC-BAYES INFORMATION BOTTLENECK»⁴⁴ se evalúa la diferencia entre la distribución de las representaciones latentes aprendidas y una distribución previa simple, por medio de la pérdida de divergencia KL y MSE. Donde se busca minimizar esta divergencia para que de esta forma el espacio latente sea estructurado y evitar que el modelo memorice los datos, lo que se alinea con el objetivo de mantener una representación de mínima complejidad.

Por último, se decide emplear las herramientas UMAP y TSNE para la visualización de los espacios latentes, puesto que, luego de consultar «Comparison of dimensionality reduction techniques for the visualisation of chemical space in organometallic catalysis»⁴⁵ se puede definir que para tener un panorama más completo de los datos, se usarán las dos herramientas donde TSNE puede ser utilizada para obtener una representación inicial de los datos, mientras que UMAP puede afinar esa representación, proporcionando una mejor separación y visualización.

Definición final del problema

Luego de la revisión documental completa, que se plantea anteriormente como registro de los antecedentes de la creación y diseño de la solución al problema de investigación, finalmente se llega a su definición completa que se usará para el resto del proyecto de investigación, la cual, considerando todo lo anteriormente dicho, contempla que este proyecto de investigación se enfoca en sentar las bases para un contexto debidamente limitado, el análisis de la representación de palabras individuales en un espacio latente a través de unas técnicas y arquitecturas específicas. Donde, principalmente, se busca aportar conocimiento sobre el uso de autoencoders con estrategias específicas para

⁴³J. Liu et al. «TB-Net: Intra- and inter-video correlation learning for continuous sign language recognition». En: *Information Fusion* 109 (2024), págs. 1-10. DOI: [10.1016/j.inffus.2024.102438](https://doi.org/10.1016/j.inffus.2024.102438).

⁴⁴Z. Wang et al. «PAC-Bayes information bottleneck». En: *Proceedings of the International Conference on Learning Representations*. 2022. URL: <https://openreview.net/forum?id=ilHOIDsPv1P>.

⁴⁵M. Villares, C. M. Saunders y N. Fey. «Comparison of dimensionality reduction techniques for the visualisation of chemical space in organometallic catalysis». En: *Artificial Intelligence Chemistry* 2.1 (2024), págs. 1-10. DOI: [10.1016/j.aichem.2024.100055](https://doi.org/10.1016/j.aichem.2024.100055).

subsanar el problema que se encontró previamente dentro del campo del lenguaje de señas. En otras palabras, el problema de investigación se centra en el comportamiento y la organización de las palabras dentro de un determinado espacio vectorial de baja dimensionalidad, para facilitar su visualización, con el fin de plantear recomendaciones a trabajos futuros que puedan resultar en una posible herramienta que sea capaz de unificar el lenguaje de señas. Teniendo en cuenta esto, surge la siguiente pregunta de investigación «¿En qué medida puede un autoencoder, a través de determinadas técnicas, entrenado con tres lenguajes de señas y evaluado con recursos determinados, ayudar a identificar patrones útiles para representar palabras en un espacio latente?».

Justificación

Los problemas del lenguaje de señas, como se indicó en el apartado anterior, no solo generan barreras y rompen la comunicación, sino que también limitan el acceso a la información y a las oportunidades para las comunidades que tienen esta discapacidad. Por ejemplo, como se menciona anteriormente, se estima que de 30 millones de estadounidenses, de 12 años o más, con pérdida auditiva, 6 millones usan ASL como modo de comunicación, de los cuales 2 millones son usuarios activos.¹ Es por esto que se considera que una herramienta que pueda comprender y, con el paso del tiempo y su desarrollo, ayudar a unificar las variantes del lenguaje de señas, como una solución de alto impacto. Al poder reducir costos en traducciones, personalizar la atención, entre otros beneficios.

Se hace este señalamiento porque, a pesar de todas las investigaciones que se han realizado desde diferentes perspectivas, aún no se ha llegado a una solución efectiva del problema, esto solo genera una mayor falta de sensibilización y el surgimiento de etiquetas equivocadas que se le asignan tanto al lenguaje como a las personas involucradas, siendo consecuencias mayormente significativas para las PDHL (People Disabling Hearing Loss, traducido al español como Personas con discapacidad auditiva). Esto se puede ver en el acceso a servicios de salud, donde un estudio en el Reino Unido reveló que el 64.4 % de los pacientes PDHL informaron perderse el 50 % o más de la información importante durante sus citas médicas.² Además, el 81 % de los pacientes reportó que sus necesidades de comunicación no fueron satisfechas y solo el 11 % de los pacientes cubiertos por el Estándar de Información Accesible (AIS) tuvo un acceso equitativo a los servicios del NHS.³ Esto resulta contradictorio, pues debería ser el área de la salud la que tuviera este problema cubierto en mayor medida.

La falta de conciencia sobre la pérdida de audición puede resultar en experiencias negativas que afectan la independencia, la confianza y el bienestar psicológico de las personas que poseen esta discapacidad.⁴ Un informe de RNID encontró que el 72 % de

¹Adler, «Language complexities for deaf and hard of hearing individuals in their pursuit of a career in science, technology, engineering, mathematics, and medicine: Perspectives from an LSL/ASL user», p. 851.

²B. Parmar et al. «I always feel like I'm the first deaf person they have ever met:"Deaf Awareness, Accessibility and Communication in the United Kingdom's National Health Service (NHS): How can we do better?» En: *PLoS ONE* 20.5 (2025), págs. 1-18. doi: [10.1371/journal.pone.0322850](https://doi.org/10.1371/journal.pone.0322850), p. 2.

³Ibid., p. 3.

⁴Ibid., p. 7.

los usuarios de BSLS (Lenguaje de Señas Británico) encontraron actitudes y comportamientos negativos por parte del personal médico.⁵ Así mismo, es importante destacar cómo la falta de ciertos ajustes al sistema y la necesidad constante de justificarse por sí mismas generan frustración y agotamiento emocional en las personas con esta discapacidad, donde solo el 32 % de los pacientes PDHL se mostró satisfecho con las habilidades de comunicación del personal de salud.⁶ «Los encuestados en este estudio detallaron sus experiencias durante las consultas clínicas del NHS, destacando cómo la falta de conciencia sobre las personas sordas ha afectado negativamente su acceso a los servicios, su independencia, su confianza y su bienestar psicológico. Además, informaron sentir una responsabilidad percibida de autodefensa para asegurar un mejor nivel de atención.»⁷

Así mismo, estos problemas se identifican en el área académica y profesional, incluso en las interacciones informales, o en espacios profesionales para fomentar el networking y el aprendizaje colaborativo, donde frecuentemente excluyen a las personas con discapacidad, ya que las ayudas visuales propuestas, como las autocapturas o la escritura resultan insuficientes ante la rapidez y complejidad de estas conversaciones.⁸ De igual manera, un científico discapacitado puede encontrarse constantemente en una dinámica de ponerse al día, en la cual, cuando logra entender toda la información y desatrasarse, sus compañeros de trabajo con la capacidad de escuchar ya han avanzado bastante en la discusión, empeorando una brecha de conocimiento que afecta directamente su competitividad. Esta brecha tiene consecuencias grandes, como que el salario promedio de un científico con discapacidades es entre \$10,000 y \$15,000 más bajo que el de un científico sin discapacidades,⁹ y las tasas de éxito en la obtención de subvenciones entre los investigadores principales que reportan discapacidades son más bajas que las de aquellos que no reportan ninguna.¹⁰ De igual manera, la falta de intérpretes de ASL, y posiblemente cualquier lenguaje de señas, con conocimientos técnicos en áreas especializadas hace mucho peor esta situación, donde se sufre el riesgo posible de caer en malas interpretaciones o traducciones que pueden ser críticas en un contexto científico.¹¹

Igualmente, las personas con discapacidad mencionaron con urgencia que tienen sentimientos de ansiedad, humillación y miedo a no poder transmitir información vital debido a barreras del lenguaje. Un preocupante 18.3 % de los pacientes PDHL reportó no haber entendido nada de la información proporcionada durante sus citas médicas en

⁵Parmar et al., «I always feel like I'm the first deaf person they have ever met:"Deaf Awareness, Accessibility and Communication in the United Kingdom's National Health Service (NHS): How can we do better?», p. 3.

⁶Ibid., p. 2.

⁷Ibid., p. 14.

⁸Adler, «Language complexities for deaf and hard of hearing individuals in their pursuit of a career in science, technology, engineering, mathematics, and medicine: Perspectives from an LSL/ASL user», p. 852.

⁹Ibid., p. 854.

¹⁰Ibid., p. 854.

¹¹Ibid., p. 853.

el NHS, y un 17.3 % adicional entendió solo el 25 %.¹² Esto, en los peores escenarios, lleva a las personas a evitar la atención médica. «La manera en que me han tratado cada vez que he necesitado acudir a un hospital o a una cita con el médico me ha hecho dar miedo a ir por mi cuenta y tiendo a evitar contactar con los servicios de salud incluso cuando es probable que los necesite.»¹³

Con esas ideas en mente, se es consciente de la gran amplitud y magnitud de dicho desafío, sin olvidar las limitaciones de recursos que rodean a muchos proyectos académicos de pregrado. Al tener ciertos recursos computacionales explícitamente definidos y delimitados, este proyecto de investigación no pretende desarrollar una herramienta de unificación lingüística final. Es por esto que su impacto se busca desde determinados puntos, con la intención final de poder sentar las bases y generar nuevo conocimiento de técnicas innovadoras para la traducción parcial de palabras en lenguaje de señas, mostrando a la comunidad investigativa el gran potencial que se oculta detrás de estas herramientas.

¹²Parmar et al., «I always feel like I'm the first deaf person they have ever met: "Deaf Awareness, Accessibility and Communication in the United Kingdom's National Health Service (NHS): How can we do better?",» p. 8.

¹³Ibid., p. 11.

Objetivos

1. Objetivo general

- Desarrollar una técnica reconstructiva y contrastiva, que a través de patrones en el espacio latente de un autoencoder entrenado con lenguajes de señas, evalúe una posible traducción de palabras.

2. Objetivos específicos

- Establecer e implementar los pasos adecuados de preprocesamiento y organización a los datos para un buen uso en la técnica seleccionada.
- Implementar una técnica de manera sistemática con respaldo en investigaciones previas para la obtención de datos.
- Aplicar y evaluar los resultados obtenidos para poder establecer una base técnica para futuras investigaciones en representaciones latentes de señas y su posible equivalencia interlingüística.

Marco Teórico

1. Marco Histórico

1.1. Historia y Reconocimiento del Lenguaje de Señas

Cuando se hace un recuento histórico, se puede evidenciar que antes del siglo XVII, hay pocos registros escritos sobre el lenguaje de señas o la educación de personas sordas. Es por esto que es difícil hacer un recuento histórico de fechas que daten antes de esto, sin embargo, es importante anotar que las señas como una forma de comunicación que surge primero que la hablada en la historia de la humanidad. Siguiendo con la documentación histórica del lenguaje de señas, usualmente se reconocen figuras determinadas que demostraron el potencial intelectual de las personas sordas. Como por ejemplo, en Francia, una de las figuras más antiguas e importantes es Etienne de Fay (1669 - 1750). El cual fue sordo de nacimiento, además de ser educado en la abadía de Amiens, donde no solo aprendió a leer y escribir, sino que también fue capaz de convertirse en un muy buen arquitecto, escultor y maestro para otros niños sordos.¹ De Fay era capaz de comunicarse con buenas habilidades a través del lenguaje de señas y también de educar a sus estudiantes utilizando tanto las señas como la escritura, demostrando que la educación podía llevar a la autonomía y la responsabilidad, es considerado el primer profesor para personas con discapacidad auditiva.² Además, se considera que su vida representa una gran diferencia para la historia de los sordos en Francia, incluso siendo anterior a otra de las figuras más conocidas en esta parte de la historia, siendo L'Epée.³ El cual planteaba métodos diferentes de enseñanza en estas comunidades en particular. Por otro lado, la comunidad sorda siempre ha sufrido de ser marginada en varios aspectos desde períodos mucho más antiguos de lo que datan los registros y esta época no es la excepción, la mayoría de las personas sordas vivían en el aislamiento donde casi siempre eran confundidas con personas sin capacidad intelectual, y su acceso a la educación era prácticamente inexistente, a menos que pertenecieran a familias con muchos recursos que pudieran permitirse buenos tutores, marcando la separación entre la élite y las masas.⁴

Luego se realiza un salto en la historia a principios del siglo XIX, donde se puede

¹R. Fischer y H. L. Lane. *Looking back: a reader on the history of deaf communities and their sign languages*. 1993. URL: <http://ci.nii.ac.jp/ncid/BA29032250>, p. 13-15.

²Ibíd., p. 18.

³Ibíd., p. 14.

⁴Ibíd., p. 26.

1. Marco Histórico

evidenciar que el lenguaje de señas americano (ASL) está intrínsecamente ligado a la herencia francesa que se revisó anteriormente. Porque en 1817, gracias a la colaboración del reverendo estadounidense Thomas Hopkins Gallaudet y de Laurent Clerc, un muy buen profesor sordo del instituto de París, se pudo fundar el Asilo Americano para la Educación de Sordomudos (ahora conocida como la Escuela Americana para Sordos o ASD) en Hartford, Connecticut. El profesor Clerc no solo fue cofundador, sino que también el principal profesor, importando el Lenguaje de Señas Francés (LSF) al continente americano.⁵

El hecho de traer este lenguaje a un nuevo lugar provocó un cambio cultural en la ASD, donde el LSF de Clerc se logró mezclar con diversas formas de comunicación que ya utilizaban los estudiantes allí, quienes provenían de diferentes partes del país. Así mismo, entre estas formas de comunicación se encontraba el lenguaje de señas que ya existía en la isla de Martha's Vineyard. En este lugar, el lenguaje de señas se volvió una necesidad por una alta incidencia de sordera hereditaria que originó un lenguaje de señas utilizado tanto por personas sordas como oyentes. La combinación de estas diferentes corrientes, con sus bases predominantes en el LSF, dio origen a lo que hoy se conoce como ASL. Desde Hartford, el modelo educativo y el nuevo lenguaje de señas se expandieron rápidamente por todo Estados Unidos, donde los graduados de la ASD fundaban nuevas escuelas residenciales, asegurando de esta manera una notable uniformidad del ASL en las primeras generaciones de los hablantes.⁶

Posteriormente, hubo un problema dentro de la comunidad hablante de señas debido al congreso de Milán y la resistencia americana. Lo que sucedió, consistió en que hacia finales del siglo XIX, la educación de la gente sorda se vio envuelta en un complicado debate filosófico que se originó entre los defensores del método manual, es decir los que usaban las señas, y los del método oral, que peleaban por enseñar a los sordos a hablar y leer los labios. Este problema alcanzó su punto más alto en el Segundo Congreso Internacional sobre la Educación de las personas sordas, el cual tuvo su origen en Milán en 1880. De esta manera, en el congreso, la mayoría de educadores, que cabe aclarar no poseían ninguna discapacidad, votó a favor de la prohibición del lenguaje de señas en la educación, lo que ocasionó la prevalencia del método oral sobre toda Europa.⁷

Este suceso ocurrido en el edicto de Milán tuvo consecuencias devastadoras en Europa, llevando a la supresión casi en su totalidad del lenguaje de señas en las escuelas, además de la marginación de los profesores sordos y los estudiantes que usaran las señas, donde los vigilaban constantemente para evitar que lo hicieran, esto fue muy malo, pues era la manera en la cual podían enseñar y comunicarse. En ese momento,

⁵E. Shaw e Y. Delaporte. *A Historical and Etymological Dictionary of American Sign Language: the origin and evolution of more than 500 signs.* 2015. URL: <http://ci.nii.ac.jp/ncid/BB1918069X>, p. X.

⁶Ibíd., p. X-XI.

⁷Ibíd., p. XI.

en Francia, la comunidad sorda tuvo que ver con desesperación cómo su lengua era borrada de las aulas durante décadas.⁸ Sin embargo, en Estados Unidos la reacción fue bastante diferente. Aunque el oralismo ganó mucho terreno, la comunidad sorda de América, gracias a que estaba fortalecida por varias escuelas que enseñaban las señas por varios años, además de tener muchos exalumnos educados y organizaciones como la Asociación Nacional de Sordos (NAD) que se dedicaban a las señas, pudo crear una muy buena resistencia. Durante esta época de resistencia, en los próximos años 1904 al 1910, la NAD generó varias películas que mostraban la indignación de las personas ante los "falsos profetas" que impartían el oralismo, unificando a la comunidad sorda de América..⁹ Por otro lado, factores como la distancia cultural a Europa, además de la autonomía que tenían las escuelas estatales, sin olvidar el apoyo de grupos religiosos que veían en las señas un medio eficaz para la evangelización permitieron que el ASL sobreviviera, aunque algunas veces de forma clandestina dentro de las escuelas que oficialmente adoptaban el oralismo. Pero prevaleciendo el hecho que en América aumentó el uso de las señas comparando con Francia, esto se puede ver también en la manera en la que cambiaron los lenguajes a través de la historia.¹⁰

A pesar de este problema, a principios del siglo XX, en plena pelea que estaba en contra del oralismo, la comunidad sorda estadounidense empezó medidas bastante buenas para preservar su lengua. Como se mencionó antes, la NAD, bajo el liderazgo de su presidente George Veditz, emprendió un proyecto pionero, el cual consistía en grabar una serie de películas que luego se popularizarían entre los años 1910 y 1920 para documentar y preservar "la belleza y la gracia del lenguaje de señas" para las generaciones futuras.¹¹ Estas películas no solo son un testimonio de la resistencia de la comunidad, sino también una gran oportunidad para admirar y apreciar la evolución del ASL, esto porque se dice porque en estas aparecen "maestros de las señas" de diferentes generaciones, mostrando las variaciones diacrónicas y sincrónicas del lenguaje a lo largo de los años. Por ejemplo, se puede observar cómo señas que originalmente eran compuestos de dos movimientos, como la seña de PADRE (una combinación de "hombrez .engendrar"), se fueron reduciendo y simplificando con el tiempo hasta llegar finalmente a la forma moderna de un solo toque en la frente.¹²

Es en este mismo momento en que se logran publicar los primeros diccionarios de ASL, como los de J. Schuyler Long (1910) y Daniel D. Higgins (1923), con el objetivo de llegar a estandarizar y preservar la "pureza original" de la lengua frente a la creciente amenaza del oralismo. Este periodo, aunque difícil y conocido como la época oscura, demostró la resiliencia de la comunidad sorda y su profundo arraigo con su identidad

⁸Ibid., p. XI.

⁹Ibid., p. XIII.

¹⁰Ibid., p. XIV.

¹¹T. Supalla y P. Clark. *Sign Language Archaeology*. 2015. DOI: [10.2307/j.ctv2rcng45](https://doi.org/10.2307/j.ctv2rcng45), p. 3.

¹²Ibid., p. 4-5.

1. Marco Histórico

lingüística y cultural.¹³

Así mismo, esta resiliencia dio frutos porque pesar de la opresión, la comunidad sorda logró resistir. Es importante señalar, que el punto de inflexión más significativo del siglo XX llegó desde el punto de vista académico. Esto gracias a William C. Stokoe, el cual fue un lingüista estadounidense que en 1960, fue el primero en demostrar científicamente que el ASL no era algo simple y superficial o un código gestual, sino más bien una lengua verdadera y compleja, con su propia fonología, morfología y sintaxis.¹⁴ Con este análisis lingüístico estableció la riqueza del ASL e impulso en diferentes círculos académicos su reintroducción en la educación y sentó las bases para su reconocimiento académico.

Gracias al trabajo de Stokoe, junto con el impulso que tuvo el movimiento por los derechos civiles en Estados Unidos, liderado por Martin Luther King, generó un cambio de percepción, donde las personas sordas comenzaron a verse a sí mismas como una minoría lingüística y cultural haciendo que su lucha por los derechos se intensificara.¹⁵ Por otro lado, en Francia, este surgimiento de espíritu de lucha se conoció como el Réveil sourd (El despertar sordo), el cual consistió en un movimiento para reafirmar el Lenguaje de Señas Francés (LSF) después de un siglo de represión desde el Congreso de Milán.¹⁶

Sin embargo, también estaban ocurriendo otros movimientos a nivel internacional, el movimiento por el reconocimiento de los lenguajes de señas ganó impulso en otros lados del mundo. Por ejemplo, en Alemania, un logro bastante grande fue en el Congreso de Hamburgo de 1985, donde se declaró por primera vez que el Lenguaje de Señas Alemán era un sistema lingüístico independiente y completo. Además, este evento fue un antes y después en la historia de la comunidad sorda alemana y significó después en la fundación del Centro de Lenguaje de Señas Alemán en Hamburgo en 1987. Por otro lado, también significó que la lucha política se intensificara, dando como resultado a los líderes sordos comenzando a exigir una participación activa en la educación de los niños sordos.¹⁷

Por último, la década de 1980 se conoció por culminar con dos eventos muy importantes. El primero de ellos, En 1988, fue que el Parlamento Europeo adoptó una resolución comunicando a los países miembros a reconocer legalmente sus respectivos lenguajes de señas nacionales.¹⁸ Y la segunda fue la victoria en la universidad Gallaudet

¹³Shaw y Delaporte, *A Historical and Etymological Dictionary of American Sign Language: the origin and evolution of more than 500 signs*, p. XIV.

¹⁴Ibid., p. XIV.

¹⁵Fischer y Lane, *Looking back: a reader on the history of deaf communities and their sign languages*, p. 517.

¹⁶Shaw y Delaporte, *A Historical and Etymological Dictionary of American Sign Language: the origin and evolution of more than 500 signs*, p. XII.

¹⁷Fischer y Lane, *Looking back: a reader on the history of deaf communities and their sign languages*, p. 188-189.

¹⁸Ibid., p. 204.

en Rusia, donde los estudiantes protestaron con éxito para que se nombrara al primer presidente sordo de la universidad, este hecho se supo mucho después, luego de 75 años, por culpa de la dificultad que tenía ese país en dicha época, donde la información era bastante limitada y lo que le pasaba a la comunidad sorda de Rusia estaba totalmente aislado.¹⁹ De estos hechos a la actualidad, la lucha por los derechos de las personas sordas sigue en pie, donde a través de la tecnología se ha tratado de dar una solución, sin embargo, aún no se ha encontrado una respuesta definitiva como se puede apreciar en el siguiente apartado con la revisión histórica de la detección de señas.

1.2. Evolución de la Tecnología de Reconocimiento del Lenguaje de Señas

Cuando se analiza la evolución de la tecnología se pueden reconocer diferentes etapas, que se diferencian por los avances en inteligencia artificial y visión por computador. Cuando se habla de la primera década de los últimos 20 años, se ve que en el campo del reconocimiento del lenguaje de señas, y también el del Lenguaje de Señas Chino (CSLR), se centró en su mayoría en aplicaciones y sistemas basados en sensores. Además, durante este período de experimentación, los investigadores tenían de métodos que requerían contacto físico o dispositivos que eran muy especializados. Donde la recolección de datos se realizaba a través de guantes de datos (data gloves) y sensores de movimiento como Kinect y Leap Motion.²⁰

Siguiendo este orden de ideas, las tecnologías de clasificación que se veían de manera más común en esta etapa eran los Modelos Ocultos de Márkov (HMM), las Máquinas de Vectores de Soporte (SVM) y el Almacenamiento de Contraste Dinámico (DTW).²¹ Donde estos métodos clásicos se aplicaban principalmente al reconocimiento de señas aisladas o al deletreo manual (dactilografía), la particularidad de estos gestos es que se realizan de uno en uno en un entorno controlado, lo cual representaría un futuro problema donde hasta los métodos actuales se esfuerzan por poder reconocer los gestos en espacios continuos.²² No obstante, esta primera generación de sistemas presentaba limitaciones muy grandes, las cuales eran, el costo del reconocimiento, donde este era relativamente alto y la precisión era comparativamente baja. Sin olvidar que, la dependencia de tener sensores hacía que los sistemas fueran casi siempre invasivos y poco prácticos para el uso del día a día.²³

¹⁹Ibid., p. 195.

²⁰Jiang et al., «A survey on Chinese sign language recognition: From traditional methods to Artificial Intelligence», p. 3.

²¹Ibid., p. 3.

²²Hashi, Hashim y Asamah, «A Systematic Review of Hand Gesture Recognition: An update from 2018 to 2024», p. 23.

²³Jiang et al., «A survey on Chinese sign language recognition: From traditional methods to Artificial Intelligence», p. 3.

1. Marco Histórico

Por otro lado, se provocó un cambio en las dinámicas, permitiendo que surgiera en mayor medida las técnicas por visión por computador y de aprendizaje profundo. Este cambio se originó en la última década, impulsado por el rápido desarrollo de la visión por computadora y las tecnologías de inteligencia artificial.²⁴ Este período tiene la peculiaridad de haber marcado un cambio de paradigma desde los métodos basados en sensores hacia técnicas basadas en la visión, que utilizan cámaras para capturar el movimiento de las manos y el cuerpo sin necesidad de contacto para medir las distancias de los dedos con respecto a la posición de la mano.²⁵

Es en esta transición que ocurre algo importante, que sería la publicación del trabajo de Su et al. en 2016, quienes propusieron un método no visual para el reconocimiento del lenguaje de señas basado en acelerometría (ACC) y electromiografía de superficie (sEMG), utilizando un algoritmo de Random Forest que alcanzó una notable precisión del 98.25 % en la clasificación de 121 palabras del CSL.²⁶ Casi al mismo tiempo de este hecho, el potencial del aprendizaje profundo se hizo mucho más evidente, gracias a que en 2017, Yang et al. utilizaron una Red Neuronal Convolutacional (CNN) que junto a una segmentación de manos para verificar 40 palabras del vocabulario de señas, pudo lograr una tasa de reconocimiento del 99.00 %.²⁷

Consecuentemente con lo anterior, estos avances demostraron que los enfoques basados en visión y aprendizaje profundo no solo podían igualar, sino que también superar la precisión de los anteriores sistemas basados en sensores, dándole paso a una nueva era de investigación. La introducción de arquitecturas innovadoras como el modelo Transformer por Vaswani et al. en 2017 y BERT por Devlin et al. en 2018 ayudaron a sentar las bases para modelos de lenguaje y reconocimiento mucho más potentes y contextuales..²⁸

Ahora bien, con las bases sentadas, se puede ver que desde 2018 hasta el presente, la investigación en SLR (reconocimiento de lenguaje de señas) ha entrado en una fase que se puede considerar de perfeccionamiento, con un enfoque particular en la aplicación de técnicas de aprendizaje profundo para abordar los aspectos más complejos del lenguaje de señas.²⁹ Cuando se menciona el uso de la tecnología CNN, se evidencia que se ha generalizado desde 2019, y además ha sido complementado por arquitecturas más avanzadas como las 3D-CNN, Redes Neuronales Recurrentes (entre otras de sus variantes como LSTM), y los modelos Transformer, los cuales son más adecuados para

²⁴Jiang et al., «A survey on Chinese sign language recognition: From traditional methods to Artificial Intelligence», p. 2.

²⁵Ibid., p. 3.

²⁶Ibid., p. 12.

²⁷Ibid., p. 13.

²⁸Ibid., p. 19.

²⁹Hashi, Hashim y Asamah, «A Systematic Review of Hand Gesture Recognition: An update from 2018 to 2024», p. 3.

el reconocimiento de lenguaje de señas continuo debido a su capacidad para procesar la dinámica temporal y la información contextual.³⁰

Así mismo, los enfoques de investigación actuales se centran en superar determinados desafíos clave. Como por ejemplo, uno de ellos es lograr un reconocimiento de lenguaje de señas que pueda ser independiente del intérprete, donde los marcos de aprendizaje profundo pudieron mostrar resultados buenos.³¹ De igual manera, se está reconociendo cada vez más la importancia de los componentes no manuales del lenguaje de señas, es decir, las expresiones faciales y la postura corporal, además de trabajar en la integración de estos elementos en las tecnologías para mejorar la precisión y robustez de los sistemas.³² Gracias a esto, se pudo potenciar a 2021 las investigaciones que destacan la integración de redes neuronales y de sensores equipables, como por ejemplo el uso de guantes de datos para el reconocimiento en tiempo real de movimientos dinámicos de los dedos.³³

No obstante, a pesar de los avances que se han hecho hasta el día de hoy, todavía permanecen obstáculos significativos. Como por ejemplo, siendo este de los más importantes, la limitación de las bases de datos existentes, que casi siempre son pequeñas y carecen de diversidad, conteniendo solo alfabetos, números o un conjunto reducido de palabras.³⁴ Gracias a la falta de datasets a gran escala y que sean diversos, especialmente enfocados para el lenguaje de señas continuo, es decir, en formato de videos o secuencias, dificulta el desarrollo de sistemas que puedan ser verdaderamente prácticos y, más aún, robustos.³⁵ De igual manera, otros desafíos incluyen la necesidad de desarrollar mejores métodos híbridos de extracción de características para poder reducir la dimensionalidad de los datos brutos y la también capacidad que tienen los sistemas para poder manejar condiciones enfocadas al mundo real, como por ejemplo lo pueden ser occlusiones (cuando un objeto bloquea la vista de la mano) y variaciones en la iluminación o escena.³⁶

³⁰ Jiang et al., «A survey on Chinese sign language recognition: From traditional methods to Artificial Intelligence», p. 28.

³¹ Hashi, Hashim y Asamah, «A Systematic Review of Hand Gesture Recognition: An update from 2018 to 2024», p. 3.

³² Ibíd., p. 3.

³³ Ibíd., p. 11.

³⁴ Ibíd., p. 1.

³⁵ Ibíd., p. 23.

³⁶ Ibíd., p. 5.

2. Marco Referencial

2. Marco Referencial

2.1. Panorama general del reconocimiento de lenguaje de señas

Como se ha podido evidenciar, la investigación en SLR (reconocimiento de lenguaje de señas) ha ido evolucionando de manera significativa, pasando de métodos tradicionales a enfoques avanzados basados en inteligencia artificial, lo que muestra un progreso avanzado en la interacción entre el humano y el computador. Como se pudo apreciar en el apartado anterior con la evolución del campo, se puede segmentar en tres áreas principales de investigación que representan diferentes niveles de complejidad y desarrollo, siendo estas el reconocimiento de señas aisladas, el reconocimiento de señas continuas y por último la traducción del lenguaje de señas.

En primera instancia, al evaluar las señas aisladas, se ve que esta es la categoría más fundamental del SLR y se enfoca meramente en la identificación de señas individuales que son ejecutadas de manera planeada y con pausas claras entre ellas. Cuando se aborda esta aproximación, se necesita que cada video o secuencia de datos contenga una única seña que debe ser clasificada en una categoría que ya está definida. Por esto es que este tipo de reconocimiento se considera análogo al reconocimiento de gestos estáticos, donde la información sobre la manera en que se presenta la mano es de suma importancia, o en su defecto a los gestos dinámicos simples cuyo movimiento sigue un patrón bien definido y establecido.³⁷ Como se pudo ver con anterioridad, a lo largo de los años, esta tarea se pudo lograr con éxito usando métodos de aprendizaje automático tradicionales como los Modelos Ocultos de Márkov (HMM), las Máquinas de Vectores de Soporte (SVM) y la Deformación Dinámica del Tiempo (DTW).³⁸ Hay que tener en cuenta, a pesar de que el reconocimiento de señas aisladas pudo alcanzar altos niveles de precisión en vocabularios controlados, a la hora de la verdad, su aplicación es muy limitada en la comunicación natural y fluida, porque las personas hablantes no suelen hacer pausas entre cada seña y mucho menos deletrearlas en una conversación normal.

Por otro lado, cuando se revisa el reconocimiento de señas continuas (CSLR), se puede ver en la revisión documental que representa un salto significativo en complejidad y representa el enfoque de la mayoría de las investigaciones actuales. Cuando se habla de este enfoque, se hace referencia a poder identificar y transcribir una secuencia de señas a partir de un video sin tener que realizarle algún corte, por este motivo, no hay delimitadores marcados entre una seña y la siguiente.³⁹ Por lo anterior, se considera

³⁷ Hashi, Hashim y Asamah, «A Systematic Review of Hand Gesture Recognition: An update from 2018 to 2024», p. 2.

³⁸ Jiang et al., «A survey on Chinese sign language recognition: From traditional methods to Artificial Intelligence», p. 1.

³⁹ A. Khan et al. «Deep Learning Approaches for Continuous Sign Language Recognition: A Comprehensive review». En: *IEEE Access* 13 (2025), págs. 1-21. DOI: [10.1109/access.2025.3554046](https://doi.org/10.1109/access.2025.3554046), p. 2.

que esta tarea es mucho más complicada y también se pueden identificar dos desafíos principales, siendo el primero de estos, la segmentación temporal, es difícil determinar solo usando código dónde termina una seña y dónde comienza la siguiente en parte por los movimientos de transición que no hacen el cambio de manera evidente.⁴⁰ Por otro lado, el segundo, es el efecto de co-articulación, donde la apariencia y ejecución de una seña se ven influenciadas por las señas que se hicieron antes y las que se harán después, esto añade una capa extra de variabilidad y complejidad.⁴¹ Con esto en mente, para poder abordar estos desafíos, se ha identificado la utilización de los enfoques que están fundamentados en aprendizaje profundo, que utilizan arquitecturas como las Redes Neuronales Convolucionales (CNN) para la extracción de características espaciales y Redes Neuronales Recurrentes (RNN) o Transformers para el modelado temporal^{42, 43}.

Ahora bien, cuando se habla de la última categoría, que corresponde a la Traducción del Lenguaje de Señas, se dice que es la más complicada, porque diferencia del CSLR, que solo busca transcribir la secuencia de señas, la SLT tiene como objetivo generar una oración gramaticalmente correcta además de tener sentido dentro del contexto de un lenguaje hablado. Teniendo en cuenta lo anterior, se puede deducir que este proceso no solo requiere el reconocimiento preciso de las señas, sino también un profundo entendimiento de la gramática, la sintaxis y la estructura lingüística del lenguaje de señas de origen, así como del lenguaje hablado al que se quiera hacer la traducción. Como se ha podido evidenciar en la investigación, traducir de un lenguaje hablado a uno de señas no solo exige conocimiento del lenguaje, sino también una comprensión de la cultura y el contexto social de sus usuarios. Por esto mismo es que la SLT debe manejar tanto el hecho de que el orden de las señas no siempre corresponde directamente con el orden de las palabras en la oración traducida, como con el hecho de incorporar información crucial transmitida a través de todo el cuerpo del intérprete, como las expresiones faciales y el movimiento del cuerpo, que son parte integral del significado en el lenguaje de señas. En esta área, los casi todos los trabajos de la actualidad, plantean lo anterior como investigaciones futuras o como pasos que ya se han realizado para poder llegar a este ideal, como lo puede ser el reconocimiento de la importancia de capturar todo el intérprete en lugar de solo las manos o la integración de nuevas tecnológicas para encontrar la manera de capturar la temporalidad.^{44, 45, 46}

Ahora bien, con el panorama general analizado, se puede ubicar este proyecto de

⁴⁰Ibid., p. 3.

⁴¹Ibid., p. 14.

⁴²Ibid., p. 9.

⁴³Ibid., p. 11.

⁴⁴A. Akarsh et al. «Sign Language Recognition: Advancing Human-Computer Interaction through Machine Learning». En: *IEEE Xplore* 2 (2024), págs. 1-4. DOI: [10.1109/icetcs61022.2024.10544114](https://doi.org/10.1109/icetcs61022.2024.10544114), p. 2.

⁴⁵Hashi, Hashim y Asamah, «A Systematic Review of Hand Gesture Recognition: An update from 2018 to 2024», p. 23.

⁴⁶Khan et al., «Deep Learning Approaches for Continuous Sign Language Recognition: A Com-

2. Marco Referencial

investigación dentro de la categoría de Reconocimiento de Señas Continuas (CSLR), con una posible extensión hacia los fundamentos de la Traducción del Lenguaje de Señas (SLT). Esto se puede decir por qué al plantear la naturaleza del lenguaje de señas como un «sistema lingüístico bastante completo y complejo» que posee «su propia gramática, vocabulario y estructura sintáctica», este trabajo de investigación reconoce desde el principio los desafíos que van más allá de la simple clasificación de gestos.⁴⁷ Además, la investigación se alinea con la corriente principal del CSLR, que se basa en buscar el desarrollo de sistemas capaces de interpretar secuencias de señas en un flujo continuo con videos, enfrentando directamente los problemas de segmentación y co-articulación que caracterizan a esta área.⁴⁸ Sin embargo, al enfocarse en la estructura del lenguaje, el proyecto sienta las bases para una futura transición hacia la SLT, usando nuevas técnicas en un área que requiere más atención para poder desentrañar el potencial de las herramientas avanzadas de inteligencia artificial.

2.2. Investigaciones particulares relevantes

Como se pudo ver en la definición del problema, se presentaron los trabajos más influyentes que inspiraron y moldearon el desarrollo y elección de la técnica que se plantea en esta investigación. No obstante, también hubo otros estudios que de igual manera se enfocan en áreas de interés más particulares para el estudio que sirvieron para probar diferentes teorías más adelante en los análisis de resultados. Esta revisión se inicia con las primeras y más comunes arquitecturas particulares en el reconocimiento de lenguaje de señas y tareas relacionadas, como lo puede ser por ejemplo la lectura de labios, estas se basan en la combinación de Redes Neuronales Convolucionales (CNN) para la extracción de características espaciales y Redes Neuronales Recurrentes (RNN) para el modelado de la dependencia temporal.

Un ejemplo claro de esta aproximación es el trabajo de Dey et al., quienes proponen una red para el reconocimiento de gestos de preguntas "Wh." en lenguaje de señas americano. En ese estudio, se optó por emplear una arquitectura de Red Convolutional 3D para poder capturar las características espaciales y temporales de bajo nivel directamente de los fotogramas de un video.⁴⁹ Posteriormente, se emplea una BiLSTM (Long Short-Term Memory Bidireccional) la cual procesa estas características para entender la secuencia del gesto en ambos sentidos, es decir, hacia adelante y hacia atrás. De igual manera, se incorpora un mecanismo de atención multi-cabeza (Multi-head Attention) el cual permite al modelo establecer la importancia de diferentes partes de la

⁴⁷Khan et al., «Deep Learning Approaches for Continuous Sign Language Recognition: A Comprehensive review», p. 11.

⁴⁸Ibid., p. 14.

⁴⁹A. Dey, S. Biswas y D. Le. «Recognition of Wh-Question Sign Gestures in Video Streams using an Attention Driven C3D-BiLSTM Network». En: *Procedia Computer Science* 235 (2024), págs. 2920-2931. doi: [10.1016/j.procs.2024.04.276](https://doi.org/10.1016/j.procs.2024.04.276), p. 2920.

secuencia, mejorando de esta forma, la capacidad de capturar características complejas y relevantes para un reconocimiento mucho más preciso.⁵⁰

Por otro lado, en otro dominio, siendo más específico en el de la lectura de labios, con el estudio de Inamdar et al. se propone un modelo que también combina convoluciones 3D y una LSTM. Donde se aprovechaba este enfoque híbrido de las CNN 3D para aprender representaciones de video y las LSTM para modelar las secuencias temporales de los movimientos de los labios.⁵¹ En otra mano, el estudio de Innocente et al. tiene la similitud de seguir esta misma línea para desarrollar un sistema de lectura de labios en italiano con el fin de asistir a pacientes con patologías en las cuerdas vocales.⁵² Sin embargo, su modelo consta de una CNN espacio-temporal, la cual se enfoca en un vocabulario específico y cuidadosamente seleccionado para cubrir necesidades de comunicación esenciales y de emergencia, demostrando la aplicabilidad de estas arquitecturas en diferentes entornos con éxito. Además, este estudio también resalta un desafío muy importante en el reconocimiento visual del habla, siendo este la ambigüedad visual donde movimientos de labios similares pueden corresponder a fonemas diferentes (homofenos).⁵³

También se han analizado enfoques más complejos que buscan superar las limitaciones de los modelos tradicionales con el fin de mejorar la calidad de las representaciones de video. Este cambio de perspectiva se puede ver en la propuesta de Geng et al., quienes abordan el Reconocimiento Continuo de Lenguaje de Señas (CSLR) no como un problema de alineamiento, sino como una tarea de generación de video a texto, teniendo en cuenta que el alineamiento se refiere al categorizar ciertos conjuntos de frames en señas específicas.⁵⁴ Argumentan que el alineamiento entre los fotogramas del video y las glosas (unidades léxicas del lenguaje de señas) es inherentemente propenso a errores debido a la naturaleza débilmente supervisada de los datos. Para evitar este paso, utilizan un modelo de difusión para generar la secuencia de glosas a partir de las características visuales extraídas. Este enfoque generativo, combinado con aprendizaje contrastivo y mecanismos de atención cruzada (cross-attention), permite al modelo aprender una relación más directa y robusta entre el video y el texto, obteniendo resultados competitivos.⁵⁵

⁵⁰Ibid., p. 2921.

⁵¹R. Inamdar et al. «Lips Reading Using 3D Convolution and LSTM». En: *Vellore Institute of Technology* (2023), págs. 1-6. DOI: [10.20944/preprints202312.0928.v1](https://doi.org/10.20944/preprints202312.0928.v1), p. 1.

⁵²C. Innocente et al. «Deep Learning-Based Lip-Reading for Vocal Impaired Patient Rehabilitation». En: *Computer Modeling in Engineering & Sciences* 143.2 (2025), págs. 1355-1379. DOI: [10.32604/cmes.2025.063186](https://doi.org/10.32604/cmes.2025.063186), p. 1.

⁵³Ibid., p. 1.

⁵⁴X. Geng et al. «No blind alignment but generation: A different view of continuous sign language recognition based on diffusion». En: *Pattern Recognition* (2025), págs. 1-11. DOI: [10.1016/j.patcog.2025.111960](https://doi.org/10.1016/j.patcog.2025.111960), p. 1.

⁵⁵Ibid., p. 3.

2. Marco Referencial

Finalmente, cuando se evalúa el aprendizaje auto-supervisado, se presenta como una solución prometedora para que ya no sea drástica la dependencia a los datos etiquetados de manera masiva. Esto se puede ver con el estudio de Dave et al. el cual se enfoca en desarrollar un TCLR (Temporal Contrastive Learning for Video Representation), a través de un marco de aprendizaje contrastivo diseñado específicamente para datos de video.⁵⁶ A diferencia de los métodos anteriores, el de TCLR se conoce por introducir explícitamente pérdidas que hacen que el modelo pueda discriminar no solo entre diferentes videos, sino también entre clips no superpuestos dentro del mismo video. Además, esto hace que las características aprendidas sean diversas a lo largo de la dimensión temporal.⁵⁷ Como resultado, se tienen mejores representaciones de video que impulsan significativamente el rendimiento en tareas posteriores como el reconocimiento de acciones, incluso con etiquetas limitadas. Este enfoque es particularmente importante para el lenguaje de señas, donde se ve que la dinámica temporal y el poder diferenciar entre señas que son visualmente similares es fundamental.⁵⁸

2.3. Research Gap dentro del campo

Luego de realizar la investigación del campo, se puede evidenciar que a pesar de los avances significativos en el reconocimiento del lenguaje de señas (SLR), aún hay brechas críticas que no permiten el desarrollo de soluciones robustas y generalizables. Lo anterior se puede sustentar en que una de las principales carencias identificadas es el enfoque común en las tareas de reconocimiento en lugar de representación semántica, lo que se quiere decir con esto es que la mayoría de los trabajos revisados, desde modelos híbridos como C3D-BiLSTM con atención hasta enfoques generativos basados en difusión, están optimizados para la transcripción continua y una posible traducción en el futuro, pero no para el análisis existente de las relaciones semánticas entre señas⁵⁹.⁶⁰ Lo que se ha evidenciado también es que estas representaciones internas son utilizadas en pro del entrenamiento y no como objetivo principal, es decir, están diseñadas para maximizar la precisión en clasificación, no para organizar el espacio latente según similitudes. Por otro lado, lo que se quiere lograr con este proyecto, se centra explícitamente en la estructuración semántica de palabras en un espacio de baja dimensionalidad, un cambio de paradigma desde la traducción hacia la representación, creando nuevos posibles campos y aproximaciones que se pueden estudiar en el futuro.

Además, otra brecha detectada es la limitación de los trabajos a contextos monolingües. Es decir, los estudios como los de Jiang (CSL) o Gu (ASL) operan exclusivamente

⁵⁶I. Dave et al. «TCLR: Temporal contrastive learning for video representation». En: *Computer Vision and Image Understanding* 219 (2022), págs. 1-9. doi: [10.1016/j.cviu.2022.103406](https://doi.org/10.1016/j.cviu.2022.103406), p. 1.

⁵⁷Ibíd., p. 6.

⁵⁸Ibíd., p. 7.

⁵⁹Dey, Biswas y Le, «Recognition of Wh-Question Sign Gestures in Video Streams using an Attention Driven C3D-BiLSTM Network», p. 2921.

⁶⁰Geng et al., «No blind alignment but generation: A different view of continuous sign language recognition based on diffusion», p. 3.

sobre datasets específicos (RWTH-PHOENIX-Weather 2014T, AQSVd), ignorando la diversidad dialectal y la falta de estándares universales en lenguajes de señas al rededor del mundo⁶¹.⁶² Así mismo, como señala Bedoin, los lenguajes de señas varían regionalmente, y su tratamiento como sistemas homogéneos limita su aplicabilidad real a diferentes contextos.⁶³ Por eso mismo es que este proyecto aborda este vacío al explorar un espacio latente compartido para tres lenguajes (ASL, indio y ruso), donde señas equivalentes puedan agruparse semánticamente, sentando bases para una futura unificación.

Siguiendo con lo detectado durante el análisis, resulta llamativo encontrar campos de otras disciplinas que han implementado con éxito soluciones de inteligencia artificial que pueden ser muy valiosas y aún no se han probado en el de lenguaje de señas. Como por ejemplo, la lectura de labios o la rehabilitación motriz donde se han desarrollado técnicas innovadoras que podrían adaptarse, como por ejemplo el trabajo de Inamdar et al. el cual combina CNN 3D y LSTM para modelar movimientos labiales en pacientes con patologías vocales,⁶⁴ mientras que, por otro lado, Innocente et al. emplean arquitecturas espacio-temporales con capas bidireccionales para vocabularios médicos esenciales.⁶⁵ A lo que se quiere llegar es que estas aproximaciones centradas en capturar dinámicas temporales y ambigüedades visuales (homofenismos), tienen la particularidad de ser directamente aplicables al SLR, donde la co-articulación y las expresiones faciales son obstáculos que han sido detectados en múltiples escenarios.

Por último, se detectó un vacío arquitectónico y metodológico en técnicas determinadas. Por esto es que el núcleo de esta investigación reside en una combinación de técnicas que no ha sido explorada anteriormente, esta se compone en primer medida de un Autoencoder Convolucional 3D, el cual es ideal para comprimir videos de señas en representaciones latentes, preservando información espacio-temporal. El cual es evaluado por una función de pérdida compuesta, que se compone de MSE, el cual garantiza reconstrucción confiable de las señas. Que está acompañada de una función Triplet Loss, que su funcionamiento es estructurar el espacio latente mediante distancias semánticas (anclas, positivos y negativos). Acompañada también por una divergencia KL, la cual se encarga de regularizar la distribución latente para evitar sobreajuste y permitir interpolación entre señas. Se dice que no ha sido explorada, porque esta arquitectura difiere de enfoques como TCLR (aprendizaje contrastivo temporal), que diferencian clips pe-

⁶¹ Jiang et al., «A survey on Chinese sign language recognition: From traditional methods to Artificial Intelligence», p. 1-40.

⁶² Gu, Oku y Todoh, «American Sign Language Recognition and translation using Perception Neuron Wearable Inertial Motion Capture System», p. 1-15.

⁶³ Bedoin, «Exploring identity building, language transmission and educational strategies for immigrant d/Deaf multilingual learners», p. 166.

⁶⁴ Inamdar et al., «Lips Reading Using 3D Convolution and LSTM», p. 1.

⁶⁵ Innocente et al., «Deep Learning-Based Lip-Reading for Vocal Impaired Patient Rehabilitation», p. 1.

3. Marco Conceptual

ro no organizan el espacio semánticamente.⁶⁶ Además, de integrar investigaciones con otros objetivos como de modelos médicos, como puede ser la lectura de labios, o de producción de señas, como el modelo G2P-DDM para generación de poses,⁶⁷ pero organizada bajo un único objetivo, el cual es poder mapear señas multilingües en un espacio estructurado.

En conclusión, se pudo identificar que las brechas observadas revelan que el SLR está lejos de ser perfeccionado o estar en una etapa final. Esto por los diferentes enfoques centrados en representación, no solo reconocimiento. Además de no tener modelos multilingües destacables que capturen diversidad dialectal. Sin mencionar que faltan arquitecturas híbridas que combinen técnicas de aprendizaje autosupervisado y regularización semántica bajo funciones de perdida determinadas. Sin duda alguna, este proyecto llena un vacío crítico al proponer un marco para analizar y visualizar relaciones entre señas de distintos lenguajes, utilizando recursos limitados pero estratégicos (datasets ASL, indio y ruso). Donde los resultados podrían guiar futuras investigaciones hacia sistemas de traducción universal más inclusivos, alineados con las necesidades reales de las comunidades sordas.⁶⁸

3. Marco Conceptual

3.1. Fundamentos Lingüísticos y Culturales del Lenguaje de Señas

En esta subsección se aborda la naturaleza del lenguaje de señas visto como un sistema lingüístico complejo y además el pilar de una cultura compleja.

- Lenguaje de Señas como Sistema Lingüístico.

Es un sistema de comunicación natural, completo y complejo, utilizado por las comunidades sordas. No obstante, no consiste en una simple serie de gestos, sino que más bien en poseer todos los componentes de una lengua, siendo estos la gramática, sintaxis y un léxico rico. Además, su estudio requiere un enfoque interdisciplinario que integre la lingüística con el análisis cultural para su mejor comprensión.

- Parámetros Constituyentes.

⁶⁶Dave et al., «TCLR: Temporal contrastive learning for video representation», p. 6.

⁶⁷Xie et al., «G2P-DDM: Generating Sign Pose Sequence from Gloss Sequence with Discrete Diffusion Model», p. 6234.

⁶⁸Adler, «Language complexities for deaf and hard of hearing individuals in their pursuit of a career in science, technology, engineering, mathematics, and medicine: Perspectives from an LSL/ASL user», p. 851.

Son los cinco componentes fundamentales que constituyen una seña individual. Es decir, las características que dan información respecto a ella, que cuando se da una alteración de cualquiera de estos parámetros, se puede cambiar por completo el significado de la seña.

Configuración Manual: La forma específica que adopta una mano al realizar una seña, como por ejemplo, el puño cerrado o los dedos extendidos.

Ubicación: El lugar en el cuerpo o, en su defecto, en el espacio donde se ejecuta la seña, como por ejemplo en la frente o en el pecho.

Movimiento: La acción o trayectoria que realizan las manos o partes del cuerpo durante la ejecución de la seña, como lo pueden ser el movimiento circular o la línea recta.

Orientación de la Palma: La dirección hacia la cual apunta la palma o los dedos de la mano, como por ejemplo, hacia arriba, hacia abajo o incluso hacia el intérprete.

Componentes no Manuales: Incluyen expresiones faciales, como lo puede ser una ceja levantada para una pregunta, movimientos de la cabeza, la boca o el torso. Se evidencia que estos componentes son muy importantes, ya que son capaces de cumplir un papel importante en la interpretación, pudiendo añadir matices emocionales o hasta intensificar el significado.

- Diversidad Lingüística del Lenguaje de Señas.

Mito del Lenguaje de Señas Universal: Se refiere a la creencia errónea que se tiene de que existe un único lenguaje de señas para todas las personas sordas del mundo. De hecho, en realidad, existen cientos de lenguajes de señas distintos, siendo más de 200 documentados como se ha visto en la investigación, donde cada uno cuenta con su propia historia y evolución, como por ejemplo la evolución histórica que se estudió del Lenguaje de Señas Americano con relación al Lenguaje de Señas Francés.

Dialectos y Regionalismos: Al igual que los idiomas hablados, los lenguajes de señas presentan variaciones en su dialecto. Esto quiere decir que en una misma seña se puede tener formas ligeramente diferentes o pueden existir señas distintas y que estas tengan un mismo concepto dependiendo de la región geográfica, como por ejemplo las diferencias en el lenguaje entre Bogotá y la costa.

- Cultura Sorda

Esta es el conjunto de creencias, comportamientos, arte, tradiciones literarias, historia y valores compartidos por las comunidades de personas que se han visto afectadas por una discapacidad auditiva. Esto quiere decir que el lenguaje de señas no es solo una herramienta de comunicación, sino el pilar central que porta

3. Marco Conceptual

y preserva esta identidad cultural.

- **Bilingüismo Bimodal y Multilingüismo**

Bilingüismo Bimodal: Se refiere al dominio y uso de un lenguaje de señas, es decir, en la modalidad de gestos, y un lenguaje hablado o escrito, es decir, en modalidad oral o auditiva, por parte de un usuario.

Multilingüismo en Comunidades Sordas: Esto quiere decir que se reconoce, particularmente en contextos multiculturales o en casos como los migratorios, donde las personas sordas pueden llegar a dominar múltiples lenguajes de señas o, en su defecto, combinaciones de estos con varias lenguas orales.

3.2. Perspectivas sobre Discapacidad, Accesibilidad e Inclusión

En esta subsección se enmarca el proyecto dentro de un contexto social, destacando de esta manera las barreras que enfrentan las personas sordas y también la importancia de las soluciones tecnológicas.

- **Modelo Social de la Discapacidad**

Es una perspectiva que entiende lo que es la discapacidad no como una deficiencia inherente al individuo, sino más bien como el resultado de las barreras sociales, culturales, actitudinales y físicas que pueden impedir la participación plena de las personas, por ejemplo, la exclusión de una persona sorda no se debe a su sordera, sino más bien a la falta de intérpretes, de tecnologías accesibles o de conciencia social.

- **Accesibilidad y Diseño Universal**

Accesibilidad: Se refiere al diseño de productos, servicios o entornos para que puedan ser utilizados por el mayor número posible de personas, independientemente de sus capacidades. Sin embargo, en el contexto tecnológico, esto implica crear herramientas que eliminan las barreras de comunicación.

Diseño Universal: Va un paso más allá de la accesibilidad, donde realmente se enfoca en la creación de soluciones que atiendan a la diversidad humana desde su creación, sin la necesidad de adaptaciones posteriores, como lo puede ser una app que está diseñada desde el inicio con opciones de visualización bastante flexibles que no requieran de más actualizaciones.

3.3. Barreras de Accesibilidad e Impacto de la Falta de Conciencia

Esta subsección menciona la falta de recursos accesibles y de conciencia sobre la cultura sorda, lo cual genera barreras sistemáticas que impactan negativamente la vida de las personas sordas en ámbitos académicos, profesionales y de salud. Esto significa en su exclusión, falta de independencia, y puede generar sentimientos de frustración, ansiedad y agotamiento emocional.

3.4. Contexto Histórico de la Educación de Sordos

Esta subsección se encarga de añadir una perspectiva histórica muy importante para entender los conflictos y valores dentro de la comunidad sorda.

- Oralismo vs. Método Manual

Oralismo: Es un enfoque educativo, históricamente dominante tras el Congreso de Milán de 1880, el cual prioriza la enseñanza del habla y la lectura de labios, frecuentemente prohibiendo el uso del lenguaje de señas.

Método Manual: Es el que defiende el uso del lenguaje de señas y lo establece como el método principal y más natural para la educación de las personas con alguna discapacidad.

- Resiliencia Comunitaria

Se refiere a las estrategias de las comunidades sordas para preservar su lengua y cultura frente a la opresión, como la ocurrida frente al oralismo, como se puede apreciar en las películas de la National Association of the Deaf en EE.UU., que se encargaron de documentar el ASL para futuras generaciones.

4. Marco Teórico

4.1. Teoría del Aprendizaje por Representación (Pilar Central)

Cuando se habla del aprendizaje por representación, se hace referencia a que es el paradigma fundamental de esta investigación. Esto se dice porque su objetivo no es solo predecir, sino más bien aprender transformaciones de los datos que extraigan información útil para la realización de tareas. Cuando se toma esta aproximación, ya no se enfoca en operar sobre los píxeles brutos de un video, es decir, que el modelo aprende una nueva representación en un espacio vectorial de menor dimensionalidad, conocido como espacio latente.

4. Marco Teórico

En esta teoría se establece la piedra angular del proyecto, ya que el problema se enfoca explícitamente en «el comportamiento y la organización de las palabras dentro de un determinado espacio vectorial de baja dimensionalidad». Así mismo, la elección de un autoencoder como núcleo de la arquitectura es una consecuencia directamente a esta teoría, se toma esta decisión también teniendo en cuenta que el modelo se entrena no para clasificar, sino más bien en comprimir una señal en una representación latente y luego reconstruirla. Por ende, la calidad y estructura de esta representación es el verdadero objetivo, pues se teoriza que en este espacio las relaciones semánticas entre señas, incluso de diferentes idiomas, se harán evidentes.

4.2. El Principio del Information Bottleneck y la Teoría de Modelos Generativos

En un principio, el principio del Information Bottleneck (IB) muestra que una representación ideal debe ser un cuello de botella para la información, esto quiere decir que se debe comprimir al máximo la entrada reteniendo solo la información que es relevante. Cuando se transporta al contexto del autoencoder, esto se puede traducir en un equilibrio entre compresión, una representación simple, y preservación de la información, una reconstrucción precisa. Por lo tanto, este proyecto integra el IB a través de la función de pérdida compuesta, es decir que el Error Cuadrático Medio (MSE) asegura la fidelidad de la reconstrucción, mientras que la Divergencia de Kullback-Leibler (KL) la fuerza a que las representaciones latentes sigan una distribución simple, actuando como un regularizador que estructura el espacio y evita la memorización.

Por otro lado, la teoría de los Modelos Generativos, muestra que al utilizar un Autoencoder Variacional, el modelo no es solo de compresión, sino más bien generativo. Esto significa que no solo aprende a codificar y decodificar, sino que también aprende la distribución de probabilidad subyacente de los datos. Con esta perspectiva se tiene un enfoque más poderoso, ya que este implica que el modelo entiende todo lo que compone de una señal, pudiendo teóricamente generar nuevas instancias de señales visualmente coherentes. Además, esta capacidad generativa sirve como una validación para mostrar que el espacio latente ha sido capaz de capturar la estructura fundamental de los datos.

4.3. Teoría del Aprendizaje por Transferencia (Transfer Learning)

Cuando se habla de esta teoría, se dice que es una adición fundamental para poder justificar el objetivo de la unificación global. Esto se puede afirmar, porque el Transfer Learning se enfoca en aprovechar el conocimiento adquirido en una tarea o dominio para mejorar el rendimiento en otro.

Con lo anterior claro, se puede traer a colación el problema de la escasez de datasets

etiquetados a gran escala, siendo esta una limitación crítica en el campo. Por otro lado, no es viable crear un dataset masivo para cada uno de los 200+ lenguajes de señas, por este motivo, la teoría del aprendizaje por transferencia, es ideal al entrenar el autoencoder en un lenguaje de señas de gran escala usando aprendizaje auto-supervisado. Donde la importancia reside en esta fase, en la cual el modelo aprenderá características universales sobre el movimiento humano y la gesticulación.

4.4. Teorías de Aprendizaje Auto-Supervisado (SSL) y Métrico (Metric Learning)

Cuando se habla de estas dos teorías, se puede ver que ambas definen la estrategia de entrenamiento para aprender representaciones semánticas sin necesidad de etiquetas masivas.

Con más detalle, se ve que el aprendizaje Auto-Supervisado (SSL) es una respuesta estratégica a la escasez de datos etiquetados a nivel de frame. Donde el propio dato de entrada proporciona la supervisión, en cambio, la elección de la reconstrucción de video como tarea de pretexto es una implementación directa de SSL. Al utilizarlo, esto permite que el modelo aprenda características espaciotemporales de manera autónoma.

Por otro lado, el aprendizaje Métrico y la Función de Pérdida Triplet es el mecanismo clave para poder organizar el espacio latente por su significado. Mientras que el SSL aprende el cómo se ve una seña, el aprendizaje métrico le enseña al modelo, qué significa. Como se puede ver, la Pérdida Triplet estructura el espacio de una manera que las representaciones de señas semánticamente similares, siendo estas el ancla y el positivo, por ejemplo, gracias en ASL y gracias en LSC, estén más cerca que las de señas no relacionadas, siendo estas el ancla y el negativo, por ejemplo gracias y lunes en ASL. Por otra parte, esta teoría es muy importante para poder superar las variaciones o ruido superficial, como lo puede ser diferente, iluminación, ropa del intérprete, o incluso el idioma, para poder agrupar las señas por su equivalencia semántica.

4.5. Modelado de Características Espacio-Temporales y la Hipótesis del Múltiple

Por otro lado, estas teorías justifican la arquitectura de red neuronal específica que se eligió para el proyecto. Siendo esta el modelado de características espaciotemporales, donde el lenguaje de señas se caracteriza por ser intrínsecamente dinámico y para poder abordar estas características se emplea una arquitectura híbrida. La cual se conforma por un autoencoder convolucional 3D (3D-CNN), el cual extrae características locales que son capaces de fusionar el espacio y tiempo a corto plazo, es decir la forma de la mano mientras se mueve. Y una red neuronal recurrente (GRU) bidireccional, la cual se encarga de modelar las dependencias temporales a largo plazo, capturando de esta

5. Marco Tecnológico y científico

manera el contexto completo de la seña al procesar la secuencia de características en ambas direcciones. Teniendo en cuenta lo anterior, esta arquitectura determinada está teóricamente diseñada para ser capas de poder capturar desde los movimientos más pequeños hasta la estructura narrativa más completa de un gesto.

Por otro lado, también se tiene la hipótesis del múltiple (Manifold Hypothesis), donde esta hipótesis señala que los datos de alta dimensionalidad del mundo real, como lo pueden ser los píxeles de un video en este caso, en realidad se encuentran en una estructura subyacente de baja dimensionalidad (un manifold). Este proyecto se basa en la presunción de que todos los videos posibles de la seña «aprender», sin importar sus muchos píxeles que la componen, se pueden agrupar en una pequeña región de este manifold. Donde, en este caso, el objetivo del autoencoder es precisamente descubrir y parametrizar este manifold, utilizando las herramientas de visualización como UMAP y t-SNE. Por tanto, estas últimas herramientas, cuentan como métodos para validar empíricamente esta hipótesis, permitiendo analizar si el modelo ha sido capaz de aprender una estructura coherente donde las señas se organizan de manera lógica.

5. Marco Tecnológico y científico

5.1. Tecnología para el Procesamiento del Lenguaje de Señas

- Reconocimiento de Lenguaje de Señas por Computadora

Reconocimiento de Señas Aisladas (ISLR): Es la tarea de identificar señas individuales que se realizan con pausas claras y marcadas entre ellas. Además, es una aproximación más sencilla, pero limitada a determinados entornos controlados.

Reconocimiento Continuo de Lenguaje de Señas (CSLR): Esta es la tarea de transcribir una secuencia continua y fluida de señas, la cual puede ser una frase, una conversación o inclusive, como en el contexto de este proyecto, palabras que están compuestas por varios movimientos, a partir de un video. Esta es mucho más desafiante debido a la falta de pausas y la co-articulación.

Traducción del Lenguaje de Señas (SLT): Por otro lado, esta representa el objetivo final de muchos trabajos, que no solo reconoce las señas, sino que las convierte a texto, voz u otro medio de comunicación, en un idioma hablado, respetando las diferencias gramaticales y sintácticas entre ambas lenguas.

- Desafíos Fundamentales en CSLR

Alineamiento Débilmente Supervisado: Este representa un problema bastante común donde se dispone de un video y su transcripción textual completa, pero no de una correspondencia exacta y minuciosa, siendo este dicho el alineamiento

entre cada fotograma y la seña específica, dificultando el entrenamiento de los modelos.

Segmentación Temporal: Esta es la dificultad de identificar dónde termina una seña y comienza la siguiente en un flujo continuo.

Co-articulación: Este es el fenómeno donde la ejecución de una seña es influenciada por las señas que van antes y después.

Limitaciones de Datasets: Es la escasez de conjuntos de datos públicos, grandes, diversos y de alta calidad, que representen todos dialectos y variaciones culturales.

Robustez en Condiciones Reales: Es la dificultad de los sistemas para manejar occlusiones, es decir, manos bloqueadas por algún otro elemento, variaciones de iluminación, fondos complejos y diferencias individuales en la manera de realizar las señas.

5.2. Tecnología para el Procesamiento del Lenguaje de Señas

- Redes Neuronales para Extracción de Características

Red Neuronal Convolutacional (CNN): Está diseñada para procesar datos en formato de rejilla como las imágenes. También se utiliza para la extracción de características espaciales, como formas o texturas de los fotogramas de un video.

Red Neuronal Convolutacional 3D (3D-CNN o C3D): Esta es una configuración de la CNN que tiene la particularidad de operar sobre volúmenes de datos que tienen el formato alto x ancho x tiempo, que además tiene la capacidad de capturar simultáneamente características espaciales y temporales, haciéndola ideal para el movimiento entre fotogramas y para el análisis de gestos dinámicos.

- Redes Neuronales para el Modelado de Secuencias

Red Neuronal Recurrente (RNN): Esta está diseñada para procesar datos secuenciales al tener una memoria que le permite usar información de pasos anteriores para cambiar la salida actual.

Red Neuronal Long Short-Term Memory (LSTM) y BiLSTM: Esta red tiene la particularidad de estar diseñada para procesar datos secuenciales, que como sucede con la RNN, tiene una memoria que le permite usar información de pasos anteriores para cambiar la salida actual. Sin embargo, lo que diferencia a la LSTM es que es una RNN avanzada que puede resolver el problema de aprender dependencias a largo plazo. Por otro lado, una BiLSTM es aquella que procesa la secuencia en ambas direcciones, es decir, hacia adelante y hacia atrás, proporcionando de esta manera un contexto temporal más completo.

- Arquitecturas de Aprendizaje de Representación

5. Marco Tecnológico y científico

Autoencoder: Es una arquitectura no supervisada que aprende a comprimir datos, es decir, una codificación en una representación de baja dimensionalidad llamada espacio latente, para luego reconstruir la entrada original, es decir una decodificación. Donde su objetivo es el aprendizaje de características y la reducción de dimensionalidad.

Autoencoder Convolucional 3D (3D-CAE): Es una implementación específica del autoencoder que usa capas convolucionales 3D para poder comprimir videos, capturando de esta manera patrones espaciotemporales de una manera eficiente.

Transformer: Esta es una arquitectura avanzada que utiliza mecanismos de atención para poder procesar contextos largos y capturar de esta manera relaciones complejas entre elementos de una secuencia, superando algunas limitaciones de las RNNs.

- Mecanismo de Atención (Attention Mechanism)

Este es un componente importante que permite a un modelo neuronal ponderar de manera dinámica la importancia de diferentes partes de una secuencia de entrada al generar una salida determinada. Es decir, en lugar de tratar todos los fotogramas o características por igual, el modelo aprende a prestar atención a los segmentos que tienen más relevancia.

5.3. Estrategias de Aprendizaje, Optimización y Evaluación

- Paradigmas de Aprendizaje

Aprendizaje Auto-Supervisado (Self-Supervised Learning): Este es un paradigma donde el modelo puede aprender representaciones significativas directamente de los datos sin la necesidad de etiquetas manuales. Además, esta tarea se logra mediante la creación de tareas pretexto, como por ejemplo predecir un frame determinado que está más adelante en la secuencia.

Aprendizaje Contrastivo (Contrastive Learning): Este es un enfoque del aprendizaje auto-supervisado que se encarga de enseñar al modelo la creación un espacio de representación donde las muestras similares, como lo pueden ser dos clips de la misma seña, están juntas, y las muestras diferentes, como lo pueden ser videos de señas diferentes, están separadas.

Temporal Contrastive Learning (TCLR): Esta es una implementación específica para videos que se encarga de asegurarse que el modelo pueda aprender características diversas a lo largo del tiempo de entrenamiento.

- Función de Pérdida Compuesta (Composite Loss Function)

Error Cuadrático Medio (MSE): Esta métrica, es la que mide la diferencia promedio al cuadrado entre la entrada original y la reconstruida por el autoencoder, forzando una reconstrucción fiel.

Pérdida Triplet (Triplet Loss): Esta es una función de pérdida que opera sobre un triplete, el cual está constituido de una ancla, un elemento positivo y uno negativo para estructurar semánticamente el espacio latente, minimizando de esta manera la distancia entre señas de la misma clase y maximizando la distancia entre señas de clases diferentes.

Divergencia de Kullback-Leibler (KL): Esta se encarga de desempeñar un papel como un término de regularización para forzar que el espacio latente se ajuste a una distribución de probabilidad conocida como lo puede ser la normal, resultando en un espacio más suave y bien organizado, evitando el sobreajuste.

- Técnicas de Reducción de Dimensionalidad y Visualización

UMAP / t-SNE: Estos son algoritmos utilizados para poder visualizar el espacio latente de alta dimensionalidad en un gráfico 2D o 3D. Se usan en el proyecto, ya que tienen la particularidad de permitir la inspección visualmente de si el modelo ha logrado agrupar señas similares.

Capítulo 1: Preprocesamiento y organización de los datos

1. Estado Inicial de los Datos

1.1. Origen y propósito de los Datos

Luego de realizar la búsqueda exhaustiva que se menciona en la delimitación del problema, se encuentra que los candidatos ideales se redujeron a 3. Estos datasets se encuentran representados en los lenguajes de señas, inglés, indio y ruso.

En primera instancia, con el desarrollo del proyecto se utilizó el conjunto de datos en inglés, WLASL (por sus siglas Word-Level American Sign Language). Así mismo, este conjunto de datos se considera que es actualmente el mayor repositorio de videos disponible públicamente para el reconocimiento de palabras individuales en el Lenguaje de Señas Americano (ASL) además de contener un vocabulario con 2,000 señas comunes.¹

Por otro lado, el propósito más importante detrás de la creación del Dataset WLASL fue el facilitar la investigación en el campo de la comprensión del lenguaje de señas. Teniendo en cuenta lo anterior, se buscó desarrollar tecnologías que eventualmente pudieran mejorar la comunicación entre las comunidades sordas y también las oyentes, queriendo responder a la necesidad de herramientas más precisas para el reconocimiento y traducción de señas.

A la hora de realizar la recolección de los datos, estos fueron recopilados de dos fuentes principales de internet, siendo la primera de estas, sitios web educativos, donde se trajeron videos especializados en lenguaje de señas, como ASLU y ASL-LEX. La segunda fuente consistió en videos tutoriales de la plataforma YouTube, seleccionando solo los videos con títulos que describían de manera explícita y clara la seña mostrada.²

Por último, en la recolección de los datos, para poder asegurar que el conjunto de datos se enfocara exclusivamente en palabras individuales, se aplicó un criterio de filtrado estricto. El cual consistió en el descarte de todos los videos en los que la etiqueta

¹Li et al., *WLASL: Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset*.

²Risang Baskoro. *WLASL Processed Dataset*. s.f. WLASL. URL: <https://www.kaggle.com/datasets/risangbaskoro/wlasl-processed> (visitado 05-05-2025).

o anotación de la seña estuviera compuesta por más de dos palabras en inglés.

Es importante hacer la aclaración de que el conjunto de datos WLDSL fue desarrollado por Dongxu Li y Hongdong Li. Se especifica que su uso está restringido a fines académicos y computacionales, prohibiéndose de esta manera cualquier tipo de explotación comercial, el dataset se distribuye bajo la licencia Computational Use of Data Agreement (C-UDA).

Por otro lado, el siguiente se denomina INCLUDE Indian Lexicon Sign Language Dataset. Este conjunto de datos fue desarrollado para abordar la carencia de un dataset público y estandarizado de Lenguaje de Señas Indio (ISL), lengua que es utilizada por más de 5 millones de personas sordas en India.³ Donde el objetivo principal de su creación fue proporcionar un recurso robusto para poder entrenar y evaluar diferentes modelos de Reconocimiento de Lenguaje de Señas (SLR).

Además, el dataset fue creado influenciado por la iniciativa AI4Bharat y también se detalla en su publicación académica que para hacer la recopilación de los videos, se contó con la colaboración de estudiantes sordos de la St. Louis School for the Deaf en Adyar, Chennai, quienes son intérpretes experimentados, garantizando de esta manera que las señas grabadas se asemejen a las condiciones de comunicación que se tienen en el día a día.

No obstante, el conjunto de datos completo está conformado por un total de 4,292 videos, los cuales abarcan 263 señas de palabras distintas que pertenecen a 15 categorías diferentes, donde cada video muestra la acción de una única seña.

En última instancia se tiene el dataset SLOVO, el cual es un repositorio de videos a gran escala que está enfocado en el Lenguaje de Señas Ruso (RSL). Por otro lado, sus creadores alegan que el origen del dataset surge de la necesidad de contar con recursos específicos para cada lengua de señas, ya que estas varían significativamente entre países, y a la dificultad general de recopilar este tipo de datos.⁴

Por esto mismo es que el dataset contiene 20,400 videos que representan 1,000 gestos o señas distintas, las cuales están interpretadas por un total de 194 intérpretes. Por esta razón, el tamaño total del conjunto de datos es de aproximadamente 16 GB, con una duración acumulada de video de 9.2 horas donde la calidad de las grabaciones es alta, al tener un 65 % de los videos en resolución FullHD y el resto en HD.

³Sridhar et al., *INCLUDE: A Large Scale Dataset for Indian Sign Language Recognition*.

⁴Kapitanov et al., *Slovo: Russian Sign Language Dataset*.

2. Preprocesamiento de los datos

1.2. Estructuración Original de los Datos

Teniendo más claro el origen y motivo detrás de cada conjunto de datos, se analiza a mayor escala la estructura de cada uno de estos, al estar almacenados, grabados y etiquetados de diferentes maneras.

El dataset WLASL viene organizado de la siguiente manera, hay una carpeta que contiene todos los videos, donde cada video tiene su id en el nombre. Un archivo JSON, el cual contiene toda la información relacionada de cada video, como lo puede ser la etiqueta, número de frames, entre otros datos. Un archivo de texto que contiene los IDs de los videos que no están asociados a alguna etiqueta. Otro archivo TXT que contiene todas las etiquetas que existen en el dataset.

Sin embargo, el dataset INCLUDE - ISL está estructurado de una manera diferente, está compuesto por diferentes carpetas anidadas donde se divide según el tipo de palabra o la categoría a la que puede pertenecer. Por ejemplo, la carpeta «adjetivos_1de5» contiene otra carpeta llamada «Adjetivos» que a su vez tiene diferentes carpetas de adjetivos como «1. Ruidoso» que contienen todos los videos que tienen esa etiqueta.

Por último, se tiene el dataset SLOVO que contiene la carpeta de los videos, donde cada video está guardado con un serial específico y además un CSV, el cual contiene la asociación de cada video con diferentes datos relevantes como su etiqueta en ruso, la duración, tamaño, entre otras.

2. Preprocesamiento de los datos

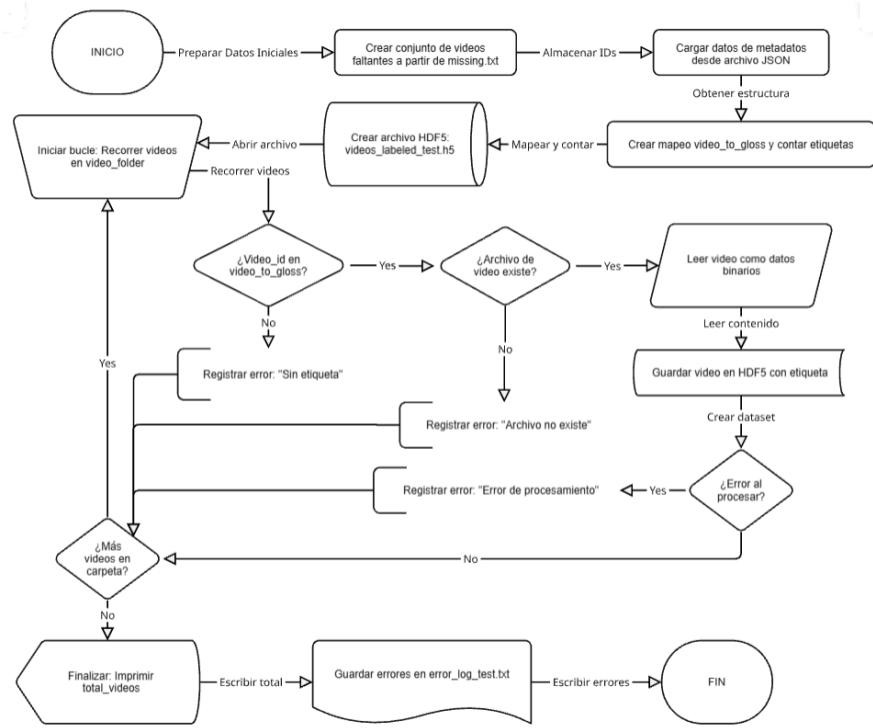
2.1. Reestructuración de los Datos

Debido a que los datos se organizan de maneras diferentes, se utilizan distintas aproximaciones para poder extraer los videos, asociarles una etiqueta y llegar a un mismo formato, para de esta manera poder almacenar en nuevos archivos H5 los diferentes datasets con una misma estructura. Con esto en mente, se define que la estructura que se utilizará es que cada elemento en el archivo H5 será un video (conjunto de frames) y asociado a cada elemento se tendrá un atributo que se llama «gloss» el cual contendrá la etiqueta en inglés de cada video.

Para empezar, el proceso que se le realizó a los datos de WLASL constó de diferentes pasos, iniciando con la lectura del archivo de texto «missing.txt». Este paso filtra videos problemáticos desde el inicio, evitando procesar datos que no existen o no son válidos, lo que ahorra tiempo y previene errores. Luego se lee el archivo «WLASL_v0.3.json», el cual contiene metadatos, como los IDs de los videos y sus etiquetas, se carga este archivo para asociar cada video con su significado. Es por esto que el siguiente paso consiste

en crear un mapeo de video a etiqueta, procesarla y contarla para verificar que estén completas, cuando se procesa, se asegura que no existan caracteres que no se puedan imprimir o que la etiqueta esté vacía. Posteriormente, se crea un archivo HDF5 que servirá de contenedor estructurado para almacenar videos y sus etiquetas de una forma determinada, facilitando su uso en análisis posteriores. Luego se verifica la existencia del archivo, evitando errores al intentar leer un archivo que no está presente, también se filtran videos sin etiqueta, asegurando que solo se procesen videos válidos. Después de asegurar la integridad de los archivos, se leen y guardan en el archivo HDF5 para, por último registrar y guardar errores. El anterior procedimiento se resume y define en el siguiente diagrama de flujo.

Figura 6.1: Diagrama de flujo con el procesamiento de los datos del dataset de WLSL.03

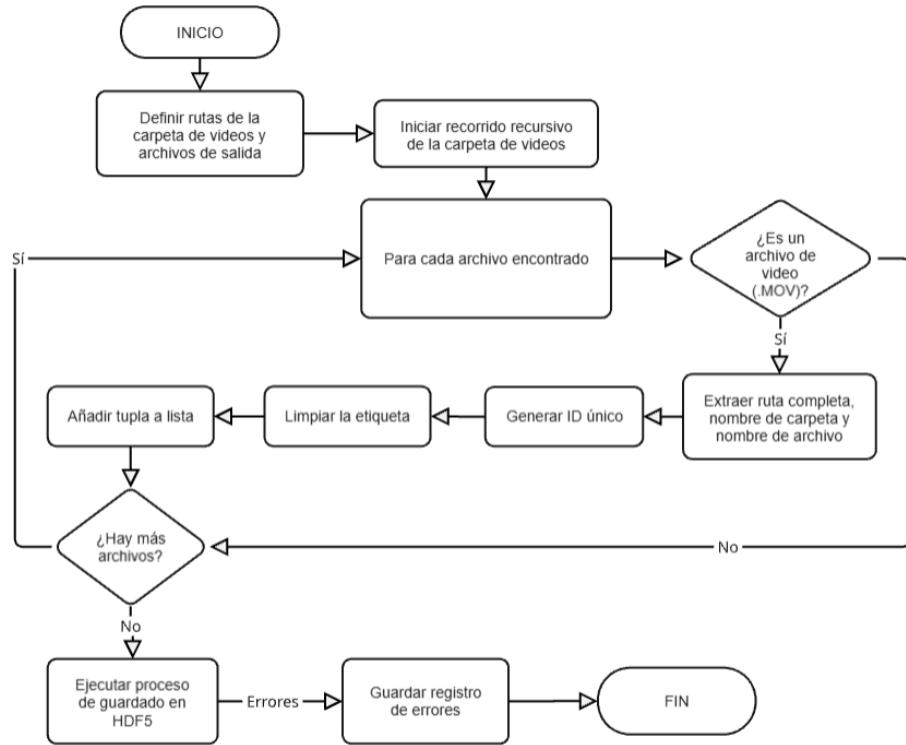


En cuanto al dataset de ISL, se comienza haciendo un recorrido de manera recursiva a la carpeta donde se encuentran los videos y, cuando se encuentra un video, se extrae la ruta completa del mismo, combinando la carpeta y el nombre del archivo, para lograr obtener el nombre de la carpeta donde está almacenado y el nombre de una manera fácil. Despues se crea un identificador único por cada archivo que consta de la combinación anterior entre la carpeta y nombre, por ejemplo, «Adjectives_1of81.loud.MOV». Luego se realiza una limpieza a las etiquetas, se realiza de esta manera porque en su versión original los nombres de los videos vienen con prefijos como «1. loud» y convertirlos en «loud». Posteriormente, se crea una tupla que contiene la ruta del video, el nombre completo y su etiqueta limpia. Ya teniendo los videos y su respectiva etiqueta procesada,

2. Preprocesamiento de los datos

se prosigue a realizar el guardado en el formato indicado. Se realiza el mismo proceso de guardado que en el caso de WSL, por este motivo se va a obviar el proceso, teniendo el siguiente diagrama de flujo.

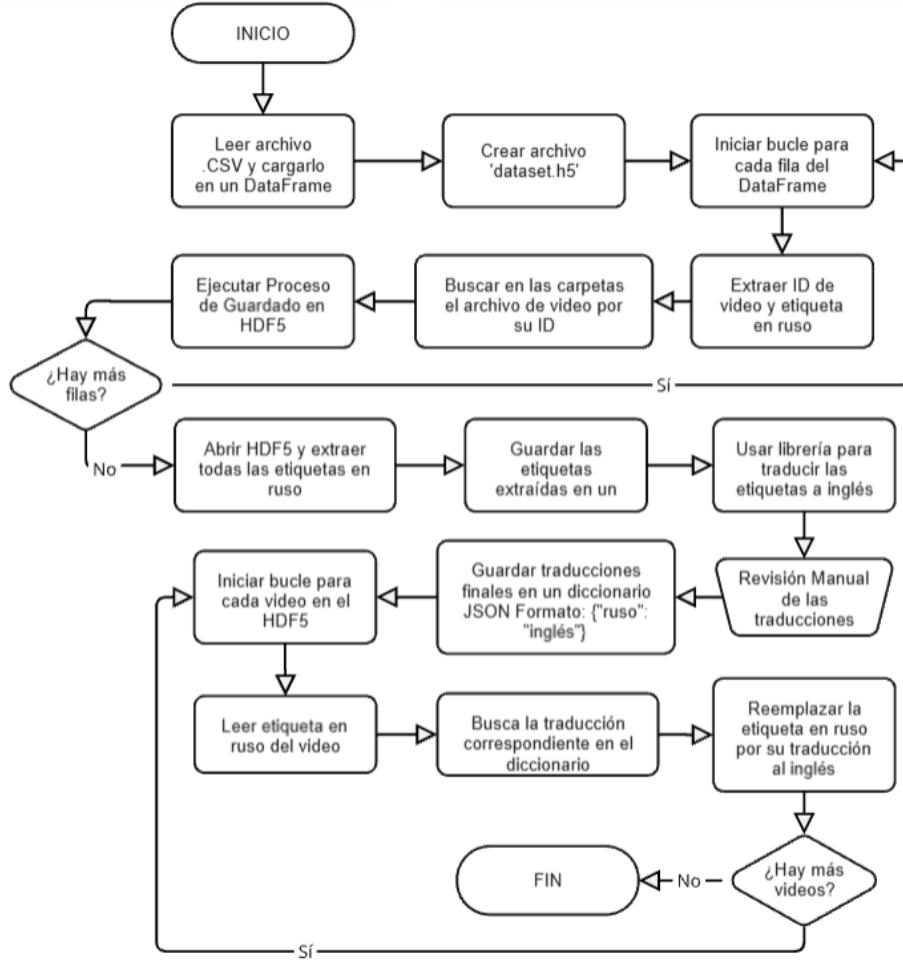
Figura 6.2: Diagrama de flujo con el procesamiento de los datos del dataset de ISL



Por último, para el dataset SLOVO se tuvo que emplear pasos extra, como se menciona en la descripción del conjunto de datos, las etiquetas, aparte de estar organizadas de otra manera, están en ruso, sin embargo, la traducción se realiza al final. Por esto es que se comienza con la lectura del CSV, pasándolo a un dataframe para realizarle operaciones más fácilmente. Se crea un archivo H5 para poder almacenar los nuevos datos organizados, luego se lee el df, extrayendo el id del video y su respectiva etiqueta para realizar una búsqueda por id en las subcarpetas donde se almacenan los videos. Con el nombre del video (id) y su etiqueta, se realiza el guardado en el archivo H5 de la misma manera que en los casos anteriores, por este motivo se obvia este proceso. Por último, se realiza la traducción de este nuevo archivo H5 comenzando por la extracción de las etiquetas y dejándolas en un archivo TXT, luego se utiliza la librería «googletrans import Translator» para traducir todas las etiquetas del TXT a inglés, seguido a esto, se revisa manualmente que cada etiqueta esté traducida correctamente al inglés y que la palabra exista. Esta traducción se guarda en un archivo JSON en modo diccionario, donde está la clave original en ruso asignada con su traducción, para luego abrir el archivo H5 y empezar a reemplazar las etiquetas en ruso por las traducidas, teniendo

como resultado el siguiente diagrama.

Figura 6.3: Diagrama de flujo con el procesamiento de los datos del dataset de SLOVO



2.2. Selección de las etiquetas

Para las diferentes pruebas realizadas, se utiliza una herramienta que permite comparar todas las etiquetas procesadas que hay en común en los tres sets de datos, esto para poder asegurar que hay una equivalencia interlingüística entre etiquetas. Con esta herramienta, luego de realizar un análisis a todas las etiquetas existentes de los 3 datasets, se encontró lo siguiente, en total se tienen 2000 etiquetas en WSL, 263 etiquetas en ISL y 927 etiquetas en SLOVO. Las diferentes cantidades de etiquetas en común son estas, 204 etiquetas comunes entre WSL e ISL, 456 etiquetas comunes entre WSL y SLOVO, 99 etiquetas comunes entre ISL y SLOVO. Teniendo 93 etiquetas comunes entre los tres diferentes datasets, las cuales se pueden representar con la siguiente lista de palabras.

2. Preprocesamiento de los datos

Etiquetas en común

Etiquetas en común		
animal	bad	beautiful
bird	black	blue
book	boy	brother
brown	cat	cheap
child	clean	cold
cow	daughter	dog
dry	evening	expensive
family	famous	father
fish	friend	girl
good	green	happy
hard	heavy	high
hot	hour	house
i	key	light
long	man	mean
minute	monday	money
month	morning	mother
mouse	new	newspaper
nice	night	old
orange	paper	pencil
pink	poor	price
red	religion	restaurant
saturday	school	second
short	sister	slow
soft	son	spring
strong	summer	sunday
thursday	time	today
tomorrow	train	tuesday
waiter	warm	week
white	wife	winter
woman	year	yellow
yesterday	you	young

2.3. Redimensionamiento de tamaño y duración de las secuencias

También se le realiza un preprocesamiento especial a los datos seleccionados, el cual consiste en un recorte específico para uniformar la duración de las secuencias, un ajuste de las dimensiones de los videos para igualar los tamaños, un cambio a blanco y negro para que las secuencias se puedan procesar de manera más eficiente y una separación

del intérprete del fondo, dejando un solo fondo para las secuencias.

Se decide optar por la utilización de estos cambios para mejorar la tasa de aprendizaje y su rendimiento. Para el mejor resultado del proyecto, se necesita el empleo de varias etiquetas y de múltiples videos para que el tiempo de ejecución no sea muy grande, se consideran todas las opciones que hay disponibles para recortar la mayor cantidad de información, asegurando un resultado satisfactorio. Es por esto que se decide igualar todas las secuencias con una estrategia específica, el primer paso de esta es calcular la mediana de la cantidad de frames de todas las secuencias dadas. Luego, teniendo como referencia esta métrica, se realiza una comprobación para saber si la secuencia que se está procesando es menor, mayor o igual a la mediana.

Si es menor a la mediana, se crea una nueva secuencia que tiene el número de frames específico, esto se logra usando una interpolación lineal para generar una nueva secuencia con una duración determinada, por ejemplo, si la mediana fuera 8 se distribuyen los frames (originales e interpolados) uniformemente en la nueva secuencia. Los nuevos frames se crean como combinaciones ponderadas de los frames originales en posiciones intermedias, determinadas por los índices generados por la función «`np.linspace`». Siguiendo el ejemplo, se añaden 3 frames a una secuencia de 5, donde los nuevos frames estarán distribuidos aproximadamente entre los frames originales, con pesos calculados según su posición relativa.

Si es mayor a la mediana, se reduce la secuencia de video seleccionando un subconjunto de frames distribuidos uniformemente, eliminando de esta manera los frames sobrantes, cabe aclarar que en este caso no se realiza interpolación ni transformación, solo selecciona frames ya existentes. Por ejemplo, si la duración actual del clip son 10 frames y la mediana es 7, se utiliza la función «`np.linspace(0, 9, num=7, dtype=int)`», dando como resultado algo como [0, 1, 3, 4, 6, 7, 9]. Esta función selecciona 7 frames de los 10 originales, distribuidos lo más uniformemente posible para evitar errores como que se eliminen de manera seguida y perder fragmentos importantes de video.

Por último, si la secuencia resulta tener la misma duración de la mediana, se deja igual, realizando una copia de los frames de la secuencia que se esté procesando en el momento. Posteriormente a este paso, se redimensionan los frames a 120 x 160 y se pasa a blanco y negro, ya que no se pierde la información de lo que está realizando el intérprete, y es necesario para poder procesar la mayor cantidad de secuencias posible con los recursos delimitados.

2.4. Recorte del fondo de las secuencias

Para asegurarse de que el modelo se encargue de aprender las características temporales, se decide quitar el fondo y dejar solamente a los intérpretes. Se refuerza esta

2. Preprocesamiento de los datos

decisión teniendo en cuenta que es bastante probable que el modelo empiece a agrupar secuencias de video en el espacio latente según el fondo en el que se es grabado. Como, por ejemplo, que agrupe todas las secuencias de ISL que tengan un tablero en el fondo, o que agrupe todas las secuencias de WSL que tengan un fondo blanco. Para lograr esto, se diseña un algoritmo que tiene como base la utilización del modelo «yolov8m-seg.pt». No se emplea alguna librería preentrenada de OpenCV o de Google porque se busca un enfoque que sea muy flexible y personalizable, se busca esta aproximación en mayor medida porque hay algunos casos específicos que no se detectan de manera correcta con estas librerías. Como lo puede ser cuando se tiene la interpolación de frames, se tienen dos o una figura translúcida, pues es un frame de transición donde el intérprete está en medio de dos señas.

En el caso del dataset de ISL y WSL, coincidieron en que la mejor variación de hiperparámetros para hallar el mejor recorte es la utilización de un método híbrido. Este método es la combinación de dilatación de la máscara, con el fin de adaptarse mejor a la forma del intérprete, con un margen adicional alrededor del contorno, ofreciendo un enfoque más robusto. Sin embargo, para el dataset SLOVO, la mejor opción fue el método de dilatación con un kernel 15x15.

La elección de estos métodos, y que se emplee un margen alrededor de todos los intérpretes, se debe principalmente a que es muy complicado generar un algoritmo que se adapte a todas las diferentes formas que puede producir un intérprete con pocos videos. Además, hay muchos intérpretes diferentes que, como en el caso del dataset SLOVO, todos estaban grabados de maneras diferentes, cambiando tanto el fondo como la iluminación bruscamente. Por este motivo es que no se llega a recortar enteramente la silueta de los videos, puede que funcione en la mayoría de los videos, pero si hay varias secuencias de video donde se recortan los dedos de la seña que se está haciendo, el método se considera insuficiente.

Luego de realizar los recortes, al revisar todos los frames para corroborar que se habían recortado correctamente, se pudo evidenciar que en contados momentos, muy rara vez se producía un error donde dejaba un frame en negro, también algunas secuencias tenían frames en negro desde el inicio. En vista de que es un problema muy poco frecuente donde se recorta todo el contenido, se decidió que cuando se encuentre un frame de este tipo, fuera reemplazado por la combinación del que tiene antes y después.

2.5. Recuento de los datos disponibles

Con los datos preprocesados se hacen subconjuntos de los tres datasets que se trabajan. Estos subconjuntos están delimitados por sus etiquetas y se asegura que estas sean las mismas para cada subconjunto de cada uno de los lenguajes de señas, se crean subconjuntos para las siguientes cantidades de etiquetas; 5, 10, 20, 35, 50 y 72 etiquetas. Como se señaló, se crea un subconjunto por cada uno de los datasets, resultando en 18

subconjuntos de datos, que son equivalentes entre sí en cuanto a su denominación, en otras palabras, el subconjunto de 5 etiquetas de ISL tiene las mismas etiquetas que el dataset de 5 etiquetas en SLOVO.

Esto se realiza para poder realizar comparaciones que estén en las mismas condiciones y para que los conjuntos de datos estén alineados con el objetivo del proyecto, el cual es encontrar una equivalencia entre lenguajes. Además de poder realizar diferentes simulaciones sin tener el problema que no permite utilizar todas las etiquetas al mismo tiempo debido a falta de recursos y poder computacional en la plataforma de AWS. Cabe aclarar que no se tiene la misma cantidad de videos por los tres idiomas en cada etiqueta, lo cual significa que se tiene un gran desbalance, donde en el peor caso posible un idioma podría superar a otro por más de 50 videos. Para poder detectar cuáles son las etiquetas que tienen un mayor desbalance, se decidió realizar una tabla, la cual contiene la etiqueta, la cantidad de videos que tiene por lenguaje, el mínimo de videos que tienen en algún lenguaje y, por último, el total de videos por etiqueta.

RANKING DE ETIQUETAS POR BALANCE Y COMPLETITUD

(Ordenado de la etiqueta más equilibrada a la menos equilibrada)

#	ETIQUETA	ISL	SLOVO	WLSL_V03	MÍNIMO (Balance)	TOTAL
1	short	21	40	13	13	74
2	cold	20	20	12	12	52
3	man	19	20	12	12	51
4	brother	20	20	11	11	51
5	mother	19	20	11	11	50
6	woman	19	20	11	11	50
7	dog	18	20	11	11	49
8	family	16	20	11	11	47
9	thursday	11	20	11	11	42
10	good	21	60	10	10	91
11	black	19	40	10	10	69
12	bad	21	20	10	10	51
13	hot	21	20	10	10	51
14	daughter	19	20	10	10	49
15	orange	19	20	10	10	49
16	son	19	20	10	10	49
17	white	19	20	10	10	49
18	bird	18	20	10	10	48
19	fish	18	20	10	10	48
20	year	11	20	10	10	41
21	yesterday	11	20	10	10	41
22	dry	21	20	9	9	50
23	new	21	20	9	9	50

2. Preprocesamiento de los datos

RANKING DE GLOSAS POR BALANCE Y COMPLETITUD

(Ordenado de la glosa más equilibrada a la menos equilibrada)

#	GLOSA	ISL	SLOVO	WSLV_V03	MÍNIMO (Balance)	TOTAL
24	school	20	20	9	9	49
25	child	19	20	9	9	48
26	cow	19	20	9	9	48
27	girl	19	20	9	9	48
28	pink	19	20	9	9	48
29	train	19	20	9	9	48
30	animal	18	20	9	9	47
31	cat	18	20	9	9	47
32	minute	11	20	9	9	40
33	today	11	20	9	9	40
34	week	11	20	9	9	40
35	you	21	80	8	8	109
36	yellow	19	40	8	8	67
37	happy	21	20	8	8	49
38	slow	21	20	8	8	49
39	boy	20	20	8	8	48
40	brown	20	20	8	8	48
41	blue	19	20	8	8	47
42	wife	19	20	8	8	47
43	sunday	11	20	8	8	39
44	time	11	20	8	8	39
45	expensive	8	20	8	8	36
46	light	8	20	8	8	36
47	mean	8	20	8	8	36
48	soft	8	20	8	8	36
49	strong	8	20	8	8	36
50	beautiful	8	40	7	7	55
51	old	21	20	7	7	48
52	house	20	20	7	7	47
53	restaurant	20	20	7	7	47
54	father	19	20	7	7	46
55	friend	19	20	7	7	46
56	green	19	20	7	7	46
57	red	19	20	7	7	46
58	sister	19	20	7	7	46
59	hour	11	20	7	7	38
60	month	11	20	7	7	38
61	saturday	11	20	7	7	38
62	summer	11	20	7	7	38



RANKING DE GLOSAS POR BALANCE Y COMPLETITUD

(Ordenado de la glosa más equilibrada a la menos equilibrada)

#	GLOSA	ISL	SLOVO	WSL_V03	MÍNIMO (Balance)	TOTAL
63	tomorrow	11	20	7	7	38
64	tuesday	11	20	7	7	38
65	cheap	8	20	7	7	35
66	clean	8	20	7	7	35
67	paper	7	20	8	7	35
68	religion	7	20	7	7	34
69	spring	11	40	6	6	57
70	hard	8	40	6	6	54
71	monday	11	20	6	6	37
72	morning	11	20	6	6	37
73	winter	11	20	6	6	37
74	famous	8	20	6	6	34
75	high	8	20	6	6	34
76	poor	8	20	6	6	34
77	book	7	20	6	6	33
78	key	7	20	6	6	33
79	money	7	20	6	6	33
80	price	7	20	6	6	33
81	long	21	20	5	5	46
82	warm	21	20	5	5	46
83	young	21	20	5	5	46
84	mouse	18	20	5	5	43
85	night	11	20	5	5	36
86	heavy	8	20	5	5	33
87	pencil	7	20	5	5	32
88	i	21	60	4	4	85
89	evening	11	20	4	4	35
90	newspaper	7	20	4	4	31
91	nice	4	20	6	4	30
92	waiter	11	20	3	3	34
93	second	3	20	6	3	29

Teniendo en cuenta la tabla anterior, los subconjuntos que se mencionaron anteriormente van a estar definidos por el orden de la tabla, en otras palabras, los subconjuntos con 5 etiquetas tendrán las que aparecen en el top 5 de la tabla y así sucesivamente. Se realiza esto para asegurarse de que se usarán las mejores etiquetas posibles en los diferentes experimentos.

Siguiendo ese orden de ideas, el trabajo de investigación se trabajará con los siguientes subconjuntos de los tres datasets principales en la medida en que lo permitan los

2. Preprocesamiento de los datos

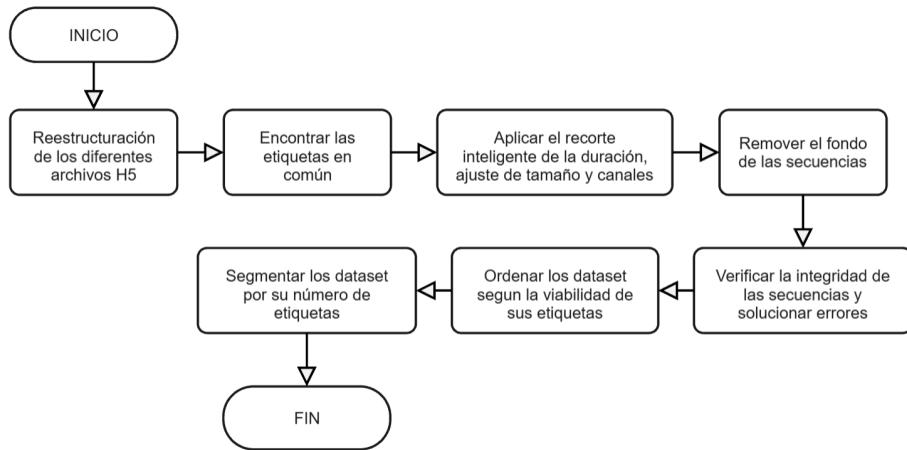
recursos computacionales:

- **Datasets que tienen las primeras 5 etiquetas:** En primera instancia se tiene «ISL_5gloss.h5» con 99 videos, luego «SLOVO_5gloss.h5» con 120 videos y por último a «WLSL_5gloss.h5» conteniendo un total de 59 videos.
- **Datasets que tienen las primeras 10 etiquetas:** También se tiene «ISL_5gloss.h5» con 184 videos, luego «SLOVO_5gloss.h5» con 260 videos y por último a «WLSL_5gloss.h5» conteniendo un total de 113 videos.
- **Datasets que tienen las primeras 20 etiquetas:** En esta división se tiene «ISL_5gloss.h5» con 368 videos, luego «SLOVO_5gloss.h5» con 480 videos y por último a «WLSL_5gloss.h5» conteniendo un total de 213 videos.
- **Datasets que tienen las primeras 35 etiquetas:** Por otro lado, se tiene «ISL_5gloss.h5» con 626 videos, luego «SLOVO_5gloss.h5» con 840 videos y por último a «WLSL_5gloss.h5» conteniendo un total de 348 videos.
- **Datasets que tienen las primeras 50 etiquetas:** Además, en esta subdivisión se tiene «ISL_5gloss.h5» con 835 videos, luego «SLOVO_5gloss.h5» con 1180 videos y por último a «WLSL_5gloss.h5» conteniendo un total de 467 videos.
- **Datasets que tienen las primeras 72 etiquetas:** En este conjunto se tiene «ISL_5gloss.h5» con 1128 videos, luego «SLOVO_5gloss.h5» con 1660 videos y por último a «WLSL_5gloss.h5» conteniendo un total de 618 videos.
- **Datasets que tienen todas las etiquetas:** Y finalmente, se tiene «ISL_5gloss.h5» con 1355 videos, luego «SLOVO_5gloss.h5» con 2120 videos y por último a «WLSL_5gloss.h5» conteniendo un total de 728 videos.

Para resumir todo el proceso y que quede más claro todo el proceso de preprocesado de los datos, se propone este pequeño diagrama.



Figura 6.4: Diagrama de flujo con el resumen de todo el procesamiento de los datos de los tres datasets



Además, se presentan tres ejemplos de cómo se transforman las secuencias de video durante el proceso. En la primera fila se tienen ocho frames distribuidos de la secuencia original, luego en la segunda fila, se tiene la secuencia luego del recorte inteligente, por último, la secuencia final pertenece al resultado que se le pasa al modelo sin fondo ni frames en negro. En el caso de la figura que tiene el ejemplo, WSLT presenta algunos errores que contiene el dataset en sus videos, donde al final se presentan algunos frames en negro. Este error se soluciona con el recorte inteligente, al ser pocos frames en negro, no se tienen en cuenta. Sin embargo, en algunas secuencias los frames en negro alcanzan procesos finales donde ya se eliminó el fondo. Por este motivo está la última etapa de verificación y depuración que se mencionó con anterioridad, es muy importante, aunque no aplique para todos los casos.

Figura 6.5: Evolución de las secuencias del dataset ISL durante el procesamiento.



2. Preprocesamiento de los datos

Figura 6.6: Evolución de las secuencias del dataset SLOVO durante el procesamiento.

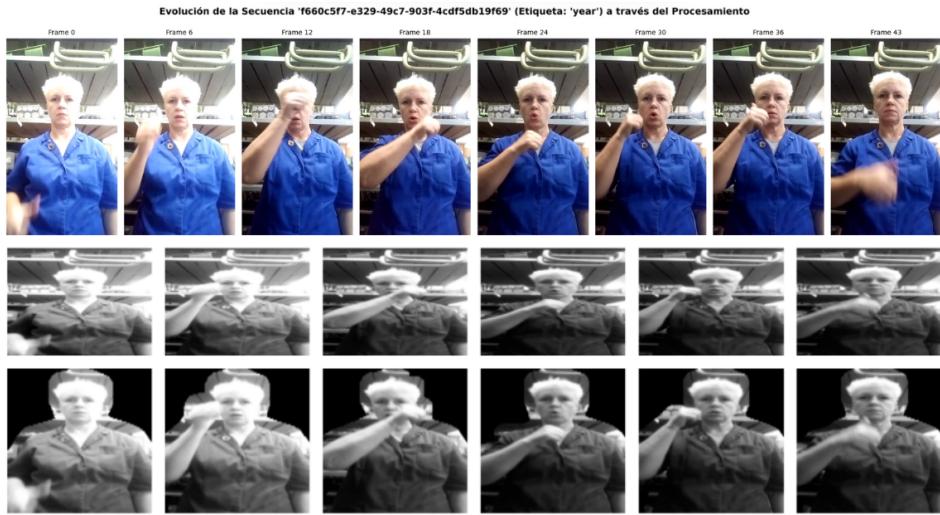


Figura 6.7: Evolución de las secuencias del dataset WSLSL durante el procesamiento.



Capítulo 2: Desarrollo de la técnica propuesta

Se dividirá esta sección en las diferentes partes de la técnica desarrollada. El orden de las subsecciones se da a partir del flujo de datos a través del programa para poder cumplir el objetivo principal de la misma. El cual es entrenar un modelo de aprendizaje profundo para aprender representaciones vectoriales, conocidas como «embeddings», de videos de lenguaje de señas. La meta no solo es clasificar las señas, sino crear un «espacio latente» donde los videos con la misma seña estén agrupados y los videos con señas diferentes estén separados.

1. Creación de los Datos de Entrada

1.1. Preparación de datos

Para preparar los datos, se comienza con una división de los mismos. Por ejemplo, si se tiene un conjunto de 60 videos, este se divide en un conjunto de entrenamiento, con 48 videos, y otro de validación, con 12 videos. Es importante hacer una división estratificada para que la proporción de cada glosario sea la misma en ambos conjuntos, esta observación se realiza porque la cantidad de videos por etiqueta varía fuertemente dependiendo de la palabra que se esté utilizando. De no realizar la estratificación correctamente, puede que todos los videos de una etiqueta resulten en train, pero no en test, haciendo que los conjuntos estén desbalanceados y no sean representativos.

Luego se realiza una normalización donde los valores de los píxeles, que inicialmente van de 0 a 255, se escalan a un rango de 0 a 1. Es importante destacar que esta normalización se calcula usando solo los valores mínimos y máximos del conjunto de entrenamiento para evitar la filtración de información del conjunto de validación al modelo, lo que podría llegar a sesgar el rendimiento del modelo al exponerlo a datos que no debería saber durante la fase de entrenamiento. Además, antes de la normalización, los datos se convierten al tipo float32 para garantizar precisión en los cálculos. Posteriormente, se crean mapas de etiquetas a índices para ambos conjuntos, lo que facilita la organización de los datos por categorías para tareas posteriores, como lo es el aprendizaje con la técnica triplet. Por último, se liberan los datos originales de la memoria mediante la eliminación de variables y la recolección de basura, optimizando el uso de recursos computacionales.

2. Implementación del Modelo

1.2. Construcción de las variantes

El siguiente paso es muy importante para todo el entrenamiento. Esto se dice, pues se hace toda la transformación a cada video que se encuentre en los conjuntos de entrenamiento y validación. Esta transformación es la generación de quintetos basada en la técnica de pérdida de triplet, es decir, que para cada video de entrada, en otras palabras, el ancla, se generan cuatro variantes adicionales, generando de esta forma un grupo de 5. En este grupo se tiene el Ancla, la cual es el video sin cambios. Luego se tiene el Positivo Desplazado, El cual es el mismo video, pero con los fotogramas desplazados temporalmente en una posición, esta se considera una variación buena por su similitud temporal. Posteriormente, se tiene el Negativo de una etiqueta, este es un video aleatorio de una señal diferente. Es una variación que se considera mala. También se tiene el Negativo Invertido, este es el video original, pero reproducido hacia atrás. Este también se considera una variación mala por su estructura temporal. Por último, se tiene el Negativo Permutado, el cual consiste en que los fotogramas del video original están en orden aleatorio. Esta es la última variación temporalmente mala.

Este nuevo formato de datos de entrada ahora tiene una dimensión extra de tamaño 5 para contener estas variantes. Por ejemplo, la forma de los datos de entrenamiento pasa a ser (48, 5, 7, 120, 160, 1). Donde 48 es el número de muestras, 5 el número de variantes, 7 el número de frames de las secuencias, 120 y 160 es el tamaño de cada frame y, por último, el 1 son los canales.

El propósito de la inclusión de estas variantes es que el modelo pueda aprender de la mejor manera el «embedding» que representa cada señal. El modelo aprende esta representación al ser forzado a resolver una tarea pretexto, que en este caso es diferenciar entre variaciones «buenas» y «malas» de un mismo video. Se afirma que es capaz de aprender a capturar las características temporales y espaciales porque, al tener el Ancla y el Positivo Desplazado con representaciones similares, el modelo aprende a ser invariantes a pequeños cambios temporales. Por otro lado, al también forzar a que las representaciones del Ancla y los tres Negativos sean diferentes, el modelo aprende a ser más sensible a lo que hace única una señal, es decir, el orden temporal correcto y la dirección del movimiento.

2. Implementación del Modelo

2.1. Construcción

En el siguiente paso se construye lo que sería la base del modelo, un autoencoder Conv3D. Este modelo primero redimensiona los fotogramas de entrada, pasando de 120x160 a 80x60 para reducir la carga computacional. Luego, codifica la secuencia de video en una representación latente más pequeña y finalmente la decodifica para reconstruir el video original.

La codificación se realiza con cuello de botella, este se logra mediante diferentes capas convolucionales y de agrupación. Se utilizan capas Conv3D que, a diferencia de las capas Conv2D que analizan imágenes que están estáticas, estas utilizan filtros tridimensionales. Esto quiere decir que el filtro no solo se desliza sobre la altura y el ancho del fotograma, sino que también a través del eje del tiempo. Esto es algo muy importante, pues es lo que hace que el modelo aprenda características espaciotemporales, como lo pueden ser patrones de movimiento, gestos y la dinámica de la señal como tal, en lugar de solo formas estáticas^{1,2}. El número de filtros, primero 16 y luego 32, aumenta la capacidad del modelo para aprender características más complejas en cada nivel. También se usan capas MaxPooling3D después de cada convolución. Por ejemplo, una capa con `pool_size=(1, 2, 2)` se encarga de reducir la dimensionalidad espacial, dividiendo la altura y ancho por 2, pero mantiene intacta la dimensión temporal. Esto hace que la representación sea más abstracta, robusta a movimientos cortos y previene el overfitting,³ cabe aclarar que también esto hace que la longitud original de la secuencia de la señal sea igual. Por otro lado, también se usan capas BatchNormalization, las cuales se utilizan después de las convoluciones para poder estabilizar y acelerar el proceso de entrenamiento, normalizando de esta manera las activaciones de la capa anterior. Por último, se emplea el Cuello de Botella, `bottleneck_sequence`, esta se considera la salida final del codificador, es una secuencia de tensores que representa la versión más comprimida y abstracta del video original. Esta también se considera la representación latente.

En cuanto a la decodificación, se realiza el proceso inverso, esto quiere decir que se toma la representación latente compacta y se expande para reconstruir el video original. Teniendo esto en mente, se usan capas Conv3DTranspose, las cuales son el complemento de las capas Conv3D y MaxPooling3D. Estas, mediante el uso de `strides=(1, 2, 2)`, pueden realizar una «deconvolución» o también conocida como «upsampling», duplicando las dimensiones espaciales mientras aprenden a llenar los detalles perdidos durante la codificación. Además, se tiene una capa de salida, esta se considera la capa final, es una Conv3D con un número de filtros igual a los canales del video original. Esta hace uso de una función de activación sigmoid, se escoge esta función porque los píxeles de la imagen de entrada se normalizan típicamente al rango [0, 1], permitiendo que la función sigmoid se asegure de que los píxeles del video reconstruido también se encuentren en este mismo rango, haciendo de esta manera que la función de pérdida sea más efectiva.

El modelo se encarga de devolver dos valores muy importantes, siendo el primero la reconstrucción. Aunque el objetivo del proyecto no es tener el video final reconstruido,

¹Dey, Biswas y Le, «Recognition of Wh-Question Sign Gestures in Video Streams using an Attention Driven C3D-BiLSTM Network», p. 4.

²Innocente et al., «Deep Learning-Based Lip-Reading for Vocal Impaired Patient Rehabilitation», p. 1355, 1369.

³Dey, Biswas y Le, «Recognition of Wh-Question Sign Gestures in Video Streams using an Attention Driven C3D-BiLSTM Network», p. 2924.

2. Implementación del Modelo

juega un papel relevante porque es la manera en la que se calculará la pérdida de reconstrucción para obligar al autoencoder a aprender y mantener la información esencial del video, como lo pueden ser la cantidad de dedos levantados, entre otros indicadores. Mientras que la otra salida es la representación latente del cuello de botella. Esta salida es la que se utilizará en la etapa de aprendizaje contrastivo. Esta arquitectura con dos salidas es extremadamente eficiente, esto se puede afirmar porque con una sola pasada hacia adelante a través del codificador, se pueden obtener las representaciones latentes necesarias para la pérdida de triplet y luego, al continuar por el decodificador, también se obtiene la salida para tener la pérdida de reconstrucción. Como se puede ver, esto permite entrenar el modelo para dos tareas simultáneamente, siendo estas la discriminación y reconstrucción, potenciando el aprendizaje de representaciones que son a la vez discriminativas y buenas en contenido.

Sin embargo, el modelo también tiene una particularidad y es que el autoencoder se envuelve en una clase personalizada llamada «TemporalTripletAutoencoder», esta clase le añade dos capas adicionales. La primera es una capa GRU bidireccional para poder analizar la secuencia temporal proveniente del cuello de botella del autoencoder. Mientras que la otra es una capa de pooling final para condensar la salida de la GRU en un único vector de características de dimensión 256 para cada video. Esta es una implementación común y potente para tareas de video, en este caso, en lugar de un autoencoder, generalmente se utiliza una red Conv3D o una CNN espaciotemporal, que actúa como la parte codificadora para extraer características visuales, donde después estas características se pasan a una red recurrente para analizar la secuencia⁴⁵.⁶

Se toma esta aproximación por las ventajas que significa tener cierta encapsulación y lógica compleja, es decir, el wrapper encapsula no solo la arquitectura, sino que también toda la lógica de entrenamiento. Esto permite gestionar de forma ordenada los múltiples componentes de la función de pérdida, en este caso de reconstrucción, varias pérdidas de triplet, etc., y sus respectivos pesos que se definen en el constructor. Esto también significa que hay más control del flujo de datos. Además, proporciona un control explícito sobre el paso hacia adelante, en otras palabras, esto significa que define exactamente cómo los datos fluyen desde la entrada, a través del autoencoder, y luego a través de las nuevas capas temporales para producir el vector final. No obstante, también proporciona flexibilidad en el entrenamiento, pudiendo sobreescribir el método «train_step» de Keras para implementar un ciclo de entrenamiento completamente personalizado, donde se calculan y combinan las diferentes pérdidas, algo que es más complicado en un modelo secuencial simple.

Esta nueva aproximación agrega más capas que procesan la salida del cuello de

⁴Innocente et al., «Deep Learning-Based Lip-Reading for Vocal Impaired Patient Rehabilitation», p. 1367.

⁵Inamdar et al., «Lips Reading Using 3D Convolution and LSTM», p. 1.

⁶Dey, Biswas y Le, «Recognition of Wh-Question Sign Gestures in Video Streams using an Attention

botella del autoencoder. Para poder realizarlo, antes de que la secuencia pueda ser procesada por la GRU, se prepara porque el cuello de botella del autoencoder produce una secuencia de mapas de características 3D, con la siguiente forma «(Tiempo, Alto, Ancho, Canales)». Sin embargo, la capa GRU espera una secuencia de vectores 1D, con forma «(Tiempo, Características)^{7,8}.

Con este paso previo realizado, se puede comenzar con la agregación temporal usando una GRU bidireccional. Se puede considerar que esta es la capa más importante para entender la dinámica temporal de una señal. Esto porque una Unidad Recurrente Cerrada (GRU) es un tipo de Red Neuronal Recurrente (RNN) que está diseñada para procesar secuencias. Esto quiere decir que posee filtros internos que le permiten decidir qué información de los pasos anteriores es relevante, mantener y cuál se puede olvidar, permitiéndole de esta manera capturar detalles a lo largo del tiempo. Por ejemplo, se puede decir que es como leer un dicho, para poder entenderlo, la GRU tiene la capacidad de recordar el contexto inicial para entender las palabras finales.

Ahora bien, aplicado al proyecto, una GRU estándar procesa la secuencia en un solo sentido, pero con un envoltorio bidireccional duplica la capa GRU. Mientras una procesa la secuencia hacia adelante, la otra la procesa hacia atrás para luego juntar ambas salidas. Esto es muy importante para el reconocimiento de señas, ya que el significado de un gesto depende no solo de lo que el intérprete ya realizó, sino también de lo que hará después. Esto le da al modelo un contexto completo de toda la secuencia en cada punto de tiempo.⁹

Ahora bien, teniendo en cuenta lo anterior, la GRU bidireccional tiene como resultado una secuencia de vectores Enriquecidos con contexto temporal. Sin embargo, para la tarea de comparación que serán las distintas pérdidas, se necesita de un único vector de características que represente toda una secuencia de video. Para esto, la capa GlobalAveragePooling1D toma la secuencia de salida de la GRU y calcula el promedio de todos los vectores a lo largo de la dimensión temporal. Dando como resultado una condensación de una secuencia de longitud variable en un único vector de tamaño fijo.

A continuación, se muestra un diagrama que tiene toda la arquitectura del modelo antes descrita y las relaciones que tiene cada parte entre sí. Además, se muestran las interacciones de las diferentes pérdidas triplet, donde la flecha verde indica que se atraen, mientras que la roja indica que hay cierto grado de repulsión. Se verá con más detalle adelante cómo funcionan estas pérdidas en el espacio latente de visualización.

Driven C3D-BiLSTM Network», p. 2923.

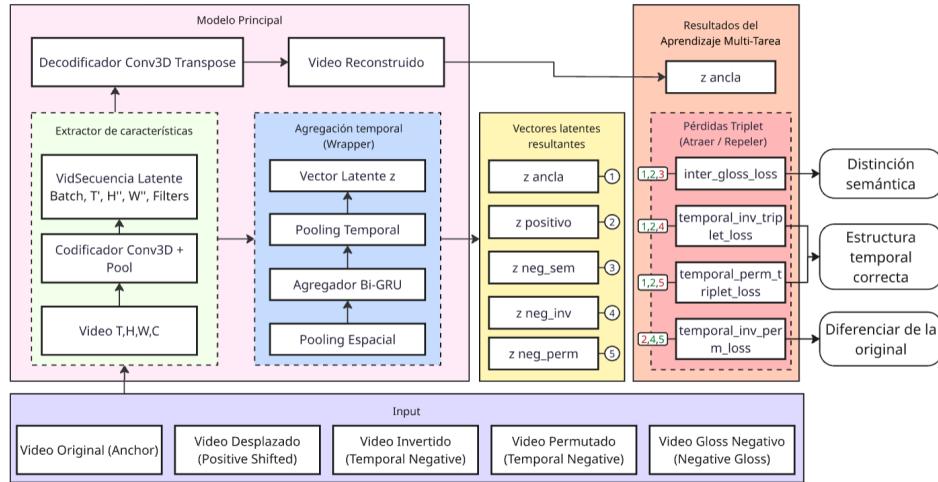
⁷Inamdar et al., «Lips Reading Using 3D Convolution and LSTM», p. 3.

⁸Dey, Biswas y Le, «Recognition of Wh-Question Sign Gestures in Video Streams using an Attention Driven C3D-BiLSTM Network», p. 2923.

⁹Innocente et al., «Deep Learning-Based Lip-Reading for Vocal Impaired Patient Rehabilitation», p. 1371.

2. Implementación del Modelo

Figura 7.1: Diagrama de la construcción del modelo

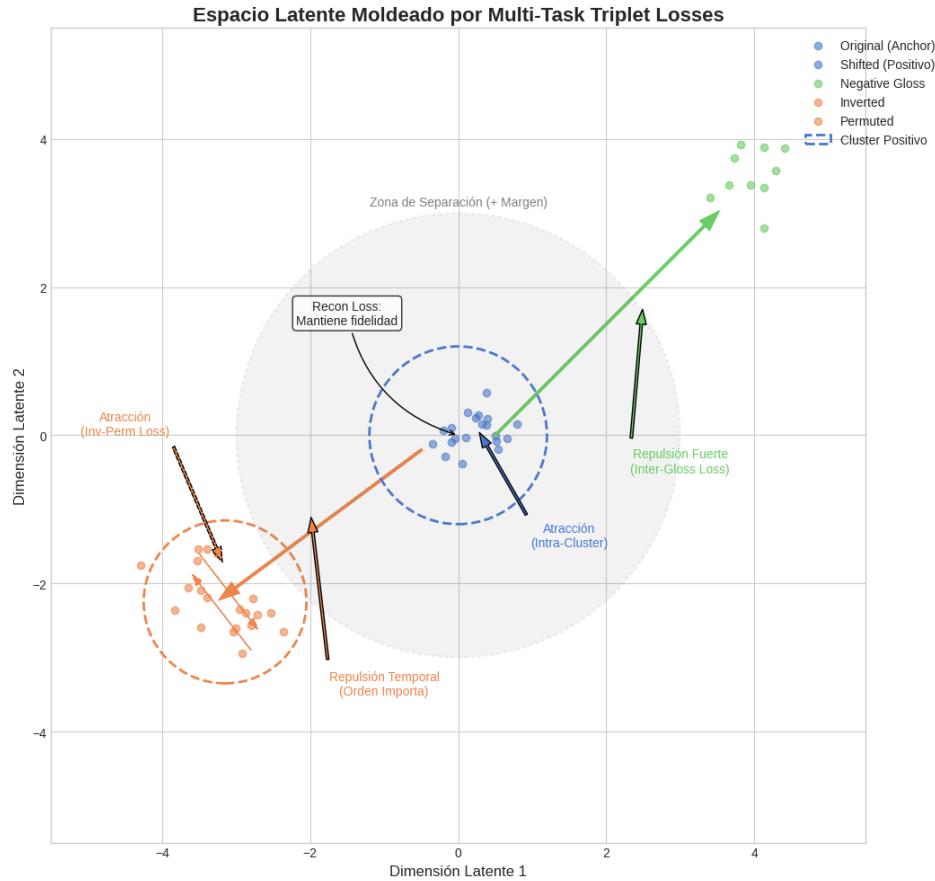


2.2. Entrenamiento

Como se mencionó con anterioridad, se tiene un modelo wrapper que facilita la personalización del entrenamiento. Por este motivo se tiene su propio «train_step», que es el núcleo de la lógica de aprendizaje. Con esta modificación, cada paso de entrenamiento calcula una pérdida combinada que se encarga de incluir las diferentes pérdidas que se mencionaron con anterioridad en la construcción de las variantes.

Durante el entrenamiento, las triplet losses, al usar las distancias cuadradas, crean un espacio latente métrico donde la geometría puede llegar a reflejar similitudes semánticas y temporales. Al diferenciar shifts leves, es decir, los positivos, de cambios drásticos, como lo pueden ser negativos como gloss diferentes o alteraciones temporales, el modelo puede llegar a generalizar mejor los datos ruidosos.

Figura 7.2: Comportamiento esperado para el Triplet Loss entre secuencias



Por ejemplo, puede ignorar las variaciones menores, pero detecta cambios en el significado o la secuencia. Por otro lado, usar distancias cuadradas es computacionalmente barato y, junto a márgenes grandes, permite separaciones fuertes sin sobrepenalizar. Esta combinación de triplets múltiples permite multi-task learning, es decir, reconstrucción para fidelidad, inter-gloss para discriminación, y temporales para invariancia al orden.

En cuanto a su compilación, el modelo se compila con el optimizador Adam con una tasa de aprendizaje inicial baja. Por otro lado, utiliza una serie de configuraciones de callbacks y entrenamiento específicas, estas son EarlyStopping, que sirve para detener el entrenamiento si el rendimiento en el conjunto de validación deja de mejorar, y la otra es ReduceLROnPlateau, que funciona para reducir la tasa de aprendizaje si el entrenamiento se estanca. Para que se pueda realizar el entrenamiento de manera efectiva y completa en la plataforma de AWS, se limita el entrenamiento durante un máximo de 100 épocas, pero se detiene antes si se llega a activar el EarlyStopping, restaurando los mejores pesos de la época.

2. Implementación del Modelo

2.3. Evaluación

En cuanto a la evaluación después del entrenamiento, se tienen en cuenta las métricas finales donde se mide el modelo final con los mejores pesos en el conjunto de validación completo para obtener las métricas de rendimiento finales. Esto lo hace por medio de una prueba de sensibilidad temporal, la cual se ejecuta con una función que verifica si el modelo aprendió correctamente la estructura temporal, midiendo si los vectores de los videos desplazados están, en promedio, más cerca de los originales que los vectores de los videos invertidos y permutados. Si el resultado se cumple, se puede confirmar que el aprendizaje fue exitoso.

Comparación con Baselines

Por otro lado, para poder determinar si hay una mejoría o para simplemente establecer un punto de comparación para la interpretación de los resultados, se opta por la utilización de diferentes aproximaciones para definir modelos «baselines» como el estándar de lo que se debe superar para considerar alguna mejoría. Teniendo claro lo anterior, se definen tres aproximaciones para estos «baselines», la primera es uno muy efectivo y fácil de implementar que compara el modelo entrenado contra una versión de sí mismo con los pesos inicializados de forma aleatoria, en otras palabras, sin entrenar. El segundo se construye para poder comparar la evolución de la pérdida total, consiste en un modelo con una arquitectura más simple. Se toma esta aproximación porque si el modelo actual que es más complejo obtiene una pérdida final más baja justifica su complejidad. Y por último se tiene una aproximación clásica con PCA, que consta de una técnica estándar para la reducción de dimensionalidad, este es un «baseline» lineal que permite ver el valor añadido de las transformaciones no lineales de tu red neuronal.

En primera instancia, para poder demostrar que un modelo sin entrenar es valioso, se puede partir del hecho de que funciona para confirmar que el modelo ha aprendido patrones significativos de los datos y que su rendimiento no se debe a una casualidad o a algún sesgo que tenga en su arquitectura^{10,11}. Además, los estudios revisados comparan sus propuestas con otros modelos de última generación para establecer una jerarquía de rendimiento, donde se puede analizar que un modelo que no está entrenado representa el punto de inicio, es decir, el nivel de rendimiento de pura casualidad. Donde superar este «baseline» significa que cualquier mejora en la precisión o disminución en la pérdida es un resultado directo del proceso de aprendizaje y no un hecho aleatorio.¹²

Siguiendo con ese orden de ideas, para justificar el uso de un modelo simplificado, se puede ver que esta práctica es común en la investigación de reconocimiento de ges-

¹⁰Inamdar et al., «Lips Reading Using 3D Convolution and LSTM», p. 5.

¹¹Xu et al., «ATCM-Net: A deep learning method for phase unwrapping based on perception optimization and learning enhancement», p. 7.

¹²Innocente et al., «Deep Learning-Based Lip-Reading for Vocal Impaired Patient Rehabilitation», p. 1368.

tos para comprobar el rendimiento de nuevas arquitecturas. Como por ejemplo, en el desarrollo del modelo «C3D-BiLSTM» «MHAttention», los autores lo evaluaron comparándolo con otros modelos establecidos como «Two-Stream CNN», «CNN+LSTM» y «3DCNN», donde decidieron tener un enfoque de comparar rendimiento con la complejidad.¹³ Por otro lado, es igual de importante justificar la complejidad del modelo que se está proponiendo, porque si un modelo más complejo no puede ofrecer una mejora significativa en el rendimiento cuando se compara con uno más simple, su complejidad adicional no se puede llegar a justificar.¹⁴

Por último, se tiene que el análisis de componentes principales (PCA) puede llegar a ser un «baseline» para la linealidad, esto debido a que las redes neuronales son capaces de aprender transformaciones no lineales complejas y, al compararlas con un método lineal estándar como PCA, permite cuantificar el valor agregado de esta no linealidad. Esta técnica de reducción de dimensionalidad es ampliamente reconocida y utilizada por su función de transformar un conjunto de variables posiblemente correlacionadas en un conjunto de valores linealmente no correlacionados llamados componentes principales.¹⁵ Cuando se evalúa la participación de esta técnica dentro de la investigación, se puede ver desde el reconocimiento de gestos hasta la química organometálica.¹⁶ Gracias a que su naturaleza lineal hace que la interpretación de sus resultados sea relativamente directa, es fácil establecer un «baseline» basado en PCA como una estrategia robusta para demostrar que las características aprendidas por la red neuronal capturen relaciones no lineales.

Elección de las métricas

Para poder medir qué tanto el modelo está aprendiendo las representaciones latentes, se opta por la utilización de dos métricas que son la distancia l2 y la distancia coseno. Esto se decide realizar porque, al proponer una función de pérdida que tenga ambas métricas de manera conjunta, es posible proporcionar un mejor resultado durante el entrenamiento del modelo.¹⁷ Siguiendo ese orden de ideas, la distancia euclíadiana se entiende como el poder aplicar la magnitud absoluta entre las muestras, donde resulta sensible a la escala de los vectores, mientras que la distancia coseno se encarga de resaltar el ángulo entre los vectores diferentes, independientemente de la magnitud que tengan. Por este motivo es que cuando se utilizan juntas, se pueden complementar sus limitaciones individuales, proporcionando de esta forma una evaluación más completa

¹³Dey, Biswas y Le, «Recognition of Wh-Question Sign Gestures in Video Streams using an Attention Driven C3D-BiLSTM Network», p. 2928.

¹⁴Ibíd., p. 2923.

¹⁵Hashi, Hashim y Asamah, «A Systematic Review of Hand Gesture Recognition: An update from 2018 to 2024», p. 8.

¹⁶Villares, Saunders y Fey, «Comparison of dimensionality reduction techniques for the visualisation of chemical space in organometallic catalysis», p. 3.

¹⁷W. Zeng et al. «Clustering-Guided Pairwise Metric Triplet Loss for Person Reidentification». En: *IEEE Internet of Things Journal* 9.16 (2022), págs. 1-11. DOI: [10.1109/jiot.2022.3147950](https://doi.org/10.1109/jiot.2022.3147950), p. 1.

2. Implementación del Modelo

de las representaciones.¹⁸

Ahora bien, para aplicar lo anterior al proyecto, se utiliza la siguiente comprobación de distancias entre secuencias, tanto para la distancia l2 como la distancia coseno, «Dist Shift <Dist Inv» y «Dist Shift <Dist Perm» como una metodología sólida para categorizar de una buena manera la importancia de una «disrupción» en la secuencia. Se puede afirmar esto, pues demuestra que no solo las transformaciones más pequeñas resultan en representaciones más cercanas en términos de magnitud, siendo esta la distancia L2. Sino que también la dirección semántica de la representación se guarda de mejor manera en transformaciones pequeñas, como la distancia coseno, validando finalmente que el modelo está aprendiendo un espacio de características que tenga sentido y sea significativo.¹⁹

2.4. Obtención de Resultados

Para poder facilitar la recolección y documentación de los datos, se definen ciertos elementos gráficos, además de las métricas claves señaladas en la subsección anterior, que serán los mismos durante todos los experimentos para tener un punto sólido de comparación. Estos elementos son gráficas que comparan los diferentes valores de entrenamiento en train y test, como lo son las distintas pérdidas. También, las gráficas que comparan el modelo en general con los diferentes «baselines» y, por último, la evolución del espacio latente a través de las épocas, tanto en «UMAP» como en «PCA». Resultando en un total de 4 gráficas de comparación con el «baseline», 6 gráficas de la evolución de las pérdidas y rendimiento en general y 14 gráficas de la evolución del espacio latente, haciendo un total de 24 gráficas por experimento. Estas gráficas se encuentran en los anexos al final del documento, para facilitar la lectura del mismo.

¹⁸Zeng et al., «Clustering-Guided Pairwise Metric Triplet Loss for Person Reidentification», p. 4.

¹⁹Ibid., p. 4.

Capítulo 3: Evaluación los resultados obtenidos

1. Análisis general

En cuanto a los datasets que se encontraron, se puede evidenciar que el dataset más grande a nivel de palabras es el WSL con el idioma en inglés, lo cual se puede respaldar con la investigación realizada en el marco histórico, donde se evidencia que el lenguaje de señas en este idioma fue el que tuvo más tiempo para ser investigado y desarrollado, sufriendo menos censura y restricciones que otros lenguajes de señas. Sin embargo, es importante resaltar que este dataset es más grande en el contexto de su riqueza léxica, pero no en cantidad de videos, ya que son los otros datasets los que cuentan con más videos que el WSL.

Siguiendo ese orden de ideas, se puede afirmar que la misma razón por la que faltan sets de datos grandes en el contexto del lenguaje de señas se debe a la persecución que existió hacia las personas sordas y la falta de apoyo por parte de los diferentes gobiernos hacia estas comunidades, lo que llevó a que en muchos países se prohibiera el uso del lenguaje de señas y se obligara a las personas sordas a aprender a hablar y leer los labios, lo cual terminó en que muchas personas no aprendieran el lenguaje de señas y, por ende, no existiera una base sólida que apoyara la creación de estos datasets en un futuro.

Estos experimentos mostraron resultados tanto positivos como negativos, los mejores resultados se obtuvieron con el dataset WSL, el cual es un dataset más pequeño en cantidad de videos, pero con una mejor calidad y uniformidad en los datos. Los resultados negativos fueron con los tres datasets, los de SLOVO e ISL con 3 etiquetas. En el caso de ISL, el experimento con 2 etiquetas es relativamente bueno, pero mostrando un poco de sobreajuste.

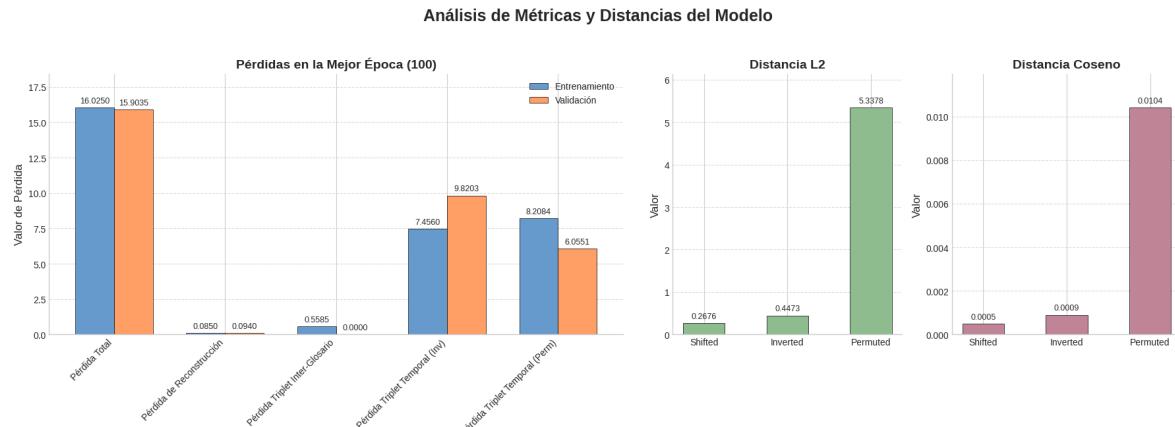
2. Análisis de las gráficas y tablas por experimento

2. Análisis de las gráficas y tablas por experimento

2.1. Experimentos con el dataset WSL

Experimento con 2 etiquetas

Figura 8.1: Estas gráficas muestran los valores de las tablas del resumen de entrenamiento y de la evaluación de la sensibilidad temporal.



Cuadro 8.1: Resumen del entrenamiento

Métrica	Valor en la Mejor Época (100)
Pérdida Total (Entrenamiento)	16.0250
Pérdida de Reconstrucción (Entrenamiento)	0.0850
Pérdida Triplet Inter-Glosario (Entrenamiento)	0.5585
Pérdida Triplet Temporal (Inv) (Entrenamiento)	7.4560
Pérdida Triplet Temporal (Perm) (Entrenamiento)	8.2084
Pérdida Total (Validación)	15.9035
Pérdida de Reconstrucción (Validación)	0.0940
Pérdida Triplet Inter-Glosario (Validación)	0.0000
Pérdida Triplet Temporal (Inv) (Validación)	9.8203
Pérdida Triplet Temporal (Perm) (Validación)	6.0551
Tasa de Aprendizaje (lr)	0.0000

Interpretación de todas las gráficas y tablas

- Separación Semántica y Sensibilidad Temporal.

Cuadro 8.2: Evaluación de Sensibilidad Temporal

Métrica de Distancia	Shifted	Inverted	Permuted
Distancia L2	0.2676	0.4473	5.3378
Distancia Coseno	0.0005	0.0009	0.0104

Evolución del Ratio de Separación Semántica vs. Baseline (Figura 10.1 y Figura 10.4): Estas gráficas miden la capacidad del modelo para poder distinguir entre diferentes señas, esto se representa con la distancia Inter-clase, en relación con la similitud de las señas de la misma clase, representada por la distancia Intra-clase. Por ejemplo, en la Figura 7.3, se puede observar que el modelo entrenado supera al baseline no entrenado, llegando a superar el ratio de 0.77 durante aproximadamente las primeras 20 épocas, alcanzando un máximo de casi 0.79. Sin embargo, después de este punto, es que el ratio semántico del modelo entrenado baja por debajo del baseline, estabilizándose en un valor aproximado de 0.73 hacia el final del entrenamiento. Por otro lado, la Figura 7.6, que incluye una comparación adicional con un baseline de PCA, con un ratio de 0.46, da una perspectiva más a la anterior afirmación. Donde se puede apreciar que, aunque el modelo supera ampliamente a PCA, no logra establecer una ventaja consistente sobre el modelo no entrenado después de las primeras épocas.

Evolución de la Sensibilidad Temporal vs. Baseline (Figura 10.2): Los resultados de esta gráfica muestran que, luego de un cierto periodo inicial, aproximadamente luego de la época 20, las distancias de las tres variantes aumentan. Es importante mencionar que la distancia de la variante permutada crece de forma muy pronunciada, superando de gran manera a las otras dos, lo que indica que el modelo es muy sensible a la desorganización de los fotogramas. Por otro lado, la distancia de la variante invertida también aumenta, pero de una forma más controlada. También se puede apreciar que la distancia con la variante desplazada se mantiene como la más baja de las tres, lo cual es un indicador positivo del aprendizaje de la estructura temporal. Además, todas las distancias del modelo entrenado superan a las de los baselines no entrenados, lo que quiere decir que el entrenamiento está creando un espacio latente que es más estructurado, donde las diferencias temporales son más notorias.

Evolución de la Separación Semántica en el Espacio Latente (Figura 10.5): Esta gráfica se encarga de mostrar por separado la distancia promedio Intra-clase y la Inter-clase. Se puede observar que durante las primeras épocas, ambas distancias disminuyen hasta aproximadamente la época 10, para luego aumentar bastante hasta aproximadamente la época 45. Es después de este punto que las distancias tienden a estabilizarse, manteniendo una separación clara entre la distancia, donde la Inter-clase es mayor a la Intra-clase. Este comportamiento demuestra que el modelo está aprendiendo a agrupar las señas de la misma clase, mientras que también separa las de clases diferentes.

2. Análisis de las gráficas y tablas por experimento

- Comparación del Modelo y Análisis de Pérdidas.

Comparación de Evolución de Pérdida Total en Validación (Figura 10.3): Se puede ver que, al comparar el modelo principal con un modelo más simple, el modelo principal logra una pérdida total mucho menor a lo largo del entrenamiento. Mientras que la pérdida del modelo simple se detiene rápidamente en un valor cercano a 19.5, la del modelo principal continúa bajando hasta un valor menor a 16. Este comportamiento justifica la mayor complejidad de la técnica propuesta, ya que evidencia un mejor rendimiento.

Pérdida Total (Figura 10.6): La gráfica de la pérdida total de validación muestra una clara tendencia a bajar, a pesar de los picos que se observan en la pérdida de entrenamiento. Esto finalmente sugiere que el modelo está generalizando de buena manera los datos no vistos.

Pérdida de Reconstrucción (Figura 10.7): La pérdida de esta gráfica, que se considera asociada al autoencoder, disminuye de forma constante para ambos conjuntos de entrenamiento y validación, indicando que el modelo aprende a reconstruir los videos de entrada de manera efectiva.

Pérdida Triplet Semántica (Figura 10.8): Esta gráfica ayuda a medir si el modelo puede diferenciar entre señas distintas, en este caso siendo «brother» y «cold». Donde la pérdida de validación cae a cero alrededor de la época 20 y se mantiene de esa manera, lo que indica una separación perfecta entre las clases en el conjunto de validación. A simple vista levanta alarmas, pues es muy poco probable que el modelo sea capaz de realizar esta separación de manera tan buena, sin embargo, son elementos como un conjunto pequeño de validación, la probabilidad de que el espacio latente hubiera comenzado con una generación aleatoria favorable, la fácil diferenciación en la forma en que se representan las dos señas y diferentes pistas de las demás gráficas, como la evolución del espacio latente, que muestran que no es fallo del modelo o pura suerte.

Pérdidas Triplet Temporales (Figura 10.9 y Figura 10.10): Estas pérdidas se encargan de evaluar si el modelo puede diferenciar entre la secuencia original y sus versiones invertida y permutada. En las gráficas, se puede apreciar que ambas pérdidas de validación muestran resultados positivos, lo que confirma que el modelo está aprendiendo a ser sensible a la estructura temporal correcta de las señas. Que la línea de validación esté constante en la gráfica 7.11 es un indicador de que, desde etapas tempranas del entrenamiento, el modelo pudo alcanzar un óptimo desempeño para diferenciar este tipo de secuencias, lo cual es positivo y acorde con los demás resultados.

- Evolución del Espacio Latente (PCA y UMAP).

Épocas Iniciales (5-25): Como se observa, en las primeras etapas, los clústeres de

las dos señas, «brother» en azul y «cold» en naranja, evidencian una separación inicial, pero no muy definida, pues los puntos todavía están esparcidos por el plano. Por otro lado, dentro de cada clúster de seña, las variantes temporales, siendo estas la original, la desplazada, la invertida y la permutada, no tienen una estructura clara.

Épocas Intermedias (45-65): A medida que avanza y evoluciona el entrenamiento, la separación entre los clústeres de las dos señas se vuelve un poco más definida y también la distancia entre ellos se agranda en la mayoría de muestras. A su vez, dentro de cada cluster, se puede empezar a observar una especie de sub-estructura. Siendo mas específico, una que hace que los puntos correspondientes a los videos originales, que se representan por un círculo, y los desplazados, que están representados por una cruz, tiendan a acercarse, mientras que las otras variantes tienden a tener en su mayoría una distancia mas pronunciada.

Épocas Finales (85-100): En la etapa final del entrenamiento, se puede apreciar que los clústeres están claramente definidos y compactos. Además, la forma que muestran las secuencias es muy similar en casi su totalidad, donde la agrupación de las variantes originales y desplazadas es muy fuerte, y están claramente separadas de las variantes invertidas y permutadas dentro de su propio clúster de seña. Esto es visible de mejor manera en PCA, aunque UMAP también comparte estas similitudes, esto demuestra visualmente que el modelo ha aprendido con éxito a separar semánticamente las dos señas y entre alteraciones temporales leves y severas, agrupando las que son temporalmente similares.

- Análisis de los Resultados Numéricos Finales.

Resumen del Entrenamiento (Tabla 8.1): En esta tabla, se puede apreciar que las pérdidas de validación son, en general, iguales o hasta mejores que las de entrenamiento, como por ejemplo, la Pérdida Total es 15.9035 en validación contra 16.0250 en entrenamiento. También, la Pérdida Triplet Inter-Glosario de validación es 0.0000, lo que confirma la perfecta separación semántica observada en la gráfica, es probable que no se aprecie tan perfecta por la reducción de dimensiones. Por otro lado, la tasa de aprendizaje llega a un valor muy bajo, lo cual es coherente y esperable con el uso del callback ReduceLROnPlateau.

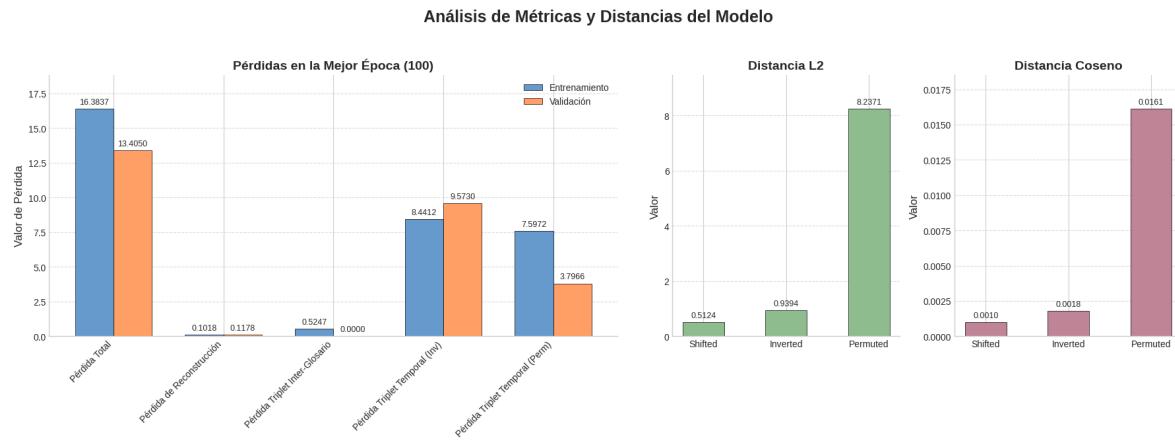
Evaluación de Sensibilidad Temporal (Tabla 8.2): Esta tabla es importante para poder respaldar lo que se observa en la Figura 7.4 y en los espacios latentes. También se puede ver que, tanto para la distancia L2 como para la distancia Coseno, se cumple la condición deseada de las distancias Shifted Inverted Permuted . Por otro lado, la distancia L2 muestra que, en cuanto a magnitud, los videos desplazados, con 0.2676, están más cerca de los originales que los invertidos con 0.4473 y mucho más que los permutados con 5.3378. Así mismo, la distancia Coseno, la cual refleja la similitud en dirección, independientemente de la magnitud. Tiene valores bajos para todas las variantes, lo que indica que conservan cierta

2. Análisis de las gráficas y tablas por experimento

similaridad direccional, sin embargo, la jerarquía de distancias se mantiene, lo que valida que el modelo está aprendiendo un espacio de características significativo y coherente con la estructura temporal de las secuencias.

Experimento con 3 etiquetas

Figura 8.2: Estas gráficas muestran los valores de las tablas del resumen de entrenamiento y de la evaluación de la sensibilidad temporal.



Cuadro 8.3: Resumen del entrenamiento

Métrica	Valor en la Mejor Época (100)
Pérdida Total (Entrenamiento)	16.3837
Pérdida de Reconstrucción (Entrenamiento)	0.1018
Pérdida Triplet Inter-Glosario (Entrenamiento)	0.5247
Pérdida Triplet Temporal (Inv) (Entrenamiento)	8.4412
Pérdida Triplet Temporal (Perm) (Entrenamiento)	7.5972
Pérdida Total (Validación)	13.4050
Pérdida de Reconstrucción (Validación)	0.1178
Pérdida Triplet Inter-Glosario (Validación)	0.0000
Pérdida Triplet Temporal (Inv) (Validación)	9.5730
Pérdida Triplet Temporal (Perm) (Validación)	3.7966
Tasa de Aprendizaje (lr)	0.0000

Interpretación de todas las gráficas y tablas

- Separación Semántica y Sensibilidad Temporal.

Cuadro 8.4: Evaluación de Sensibilidad Temporal

Métrica de Distancia	Shifted	Inverted	Permuted
Distancia L2	0.5124	0.9394	8.2372
Distancia Coseno	0.0010	0.0018	0.0161

Evolución del Ratio de Separación Semántica vs. Baseline (Figura 10.25 y Figura 10.28): Estas gráficas muestran que el ratio semántico del modelo entrenado comienza con un valor bajo alrededor de 0.75, que después de subir y bajar un poco, comienza a subir de manera constante a partir de la época 30. Logra superar al baseline no entrenado aproximadamente en la época 48 y continúa hasta estabilizarse cerca de 0.98 hasta que termina el entrenamiento. En cuanto al baseline de PCA, el modelo lo pudo superar desde el inicio, queriendo indicar que hay un aprendizaje bueno de la separación semántica en los datos que no se han visto.

Evolución de la Sensibilidad Temporal vs. Baseline (Figura 10.26): Esta gráfica evalúa cómo el modelo diferencia las secuencias originales de sus variantes temporales, se puede apreciar que los baselines no entrenados tienen distancias bajas, con 0.17 para la Shifted, 0.24 para la Inverted y 0.34 para la Permuted. Durante el entrenamiento, las distancias aumentan en gran medida, superando los baselines, donde la distancia a la versión permutada crece de forma exponencial, alcanzando un valor superior a 2.511. Las otras distancias también crecen de manera positiva, aunque no tanto en comparación con la permutada. Es importante notar que el modelo mantiene la jerarquía de las distancias Shifted > Inverted > Permuted, lo que demuestra que es más sensible a alteraciones temporales drásticas, como la permutación, que a alteraciones moderadas, como la inversión, o leves, como el desplazamiento.

Evolución de la Separación Semántica en el Espacio Latente (Figura 10.29): Esta gráfica muestra el ratio semántico al mostrar las distancias euclidianas promedio Intra-Clase e Inter-Clase. Se puede ver que ambas distancias experimentan una caída inicial, hasta la época 10, para luego tener un gran aumento hasta la época 38. Algo importante a destacar es que, durante gran parte del entrenamiento, hasta aproximadamente la época 65, la distancia promedio Intra-Clase, es decir, la azul, es mayor que la distancia Inter-Clase, la que está en naranja, y hacia el final del entrenamiento, esta relación parece tener la tendencia a invertirse ligeramente, aunque no pase al terminar las épocas antes de mostrar este comportamiento.

- Comparación del Modelo y Análisis de Pérdidas.

Comparación de Evolución de Pérdida Total en Validación (Figura 10.27): Esta gráfica muestra que, al comparar el modelo principal con un modelo más

2. Análisis de las gráficas y tablas por experimento

simple, se justifica la complejidad del mismo. Esto se puede afirmar porque la pérdida del modelo simple deja de mejorar rápidamente en un valor cercano a 20.0. Mientras que la pérdida del modelo principal, aunque inicialmente sube por aproximadamente 10 épocas, comienza una bajada muy grande después de la época 20, terminando en un valor inferior a 14.0. Esto finalmente demuestra que la arquitectura más compleja es más efectiva para esta tarea.

Pérdida Total (Figura 10.30): Esta gráfica muestra que la pérdida total de validación presenta una curva suave con una tendencia a bajar de manera constante, pasando de aproximadamente un valor de 21 a uno de aproximadamente 13. Por otro lado, se puede ver que la pérdida de entrenamiento es mucho más inestable, pero con una tendencia general descendente mucho menor. Este comportamiento se puede traducir en una buena generalización del modelo.

Pérdida de Reconstrucción (Figura 10.31): Esta gráfica representa la capacidad del autoencoder para poder reconstruir secuencias. Se puede apreciar que la pérdida de validación baja hasta alcanzar su punto más bajo alrededor de la época 35, con un valor aproximado de 0.10. Para luego comenzar a subir ligeramente y volver a tomar un comportamiento descendente, terminando con un valor aproximado de 0.12. Este comportamiento muestra que el modelo, en un entorno de aprendizaje multitarea, puede estar priorizando las pérdidas triplet semánticas y temporales sobre la fidelidad de la reconstrucción en las etapas finales del entrenamiento. Lo cual es positivo, dado que el objetivo principal del modelo no se enfoca en la reconstrucción, sino en la separación de las secuencias.

Pérdida Triplet Semántica (Figura 10.32): Esta gráfica muestra la separación entre las tres señas. En la cual se puede observar que la pérdida de validación muestra una bajada muy agresiva y alcanza el valor de 0.0 alrededor de la época 35. Además, se mantiene en cero desde ese punto hasta el final del entrenamiento, lo que indica que el modelo logró una separación semántica perfecta de las tres clases en el conjunto de validación.

Pérdidas Triplet Temporales (Figura 10.33 y Figura 10.34): Estas gráficas miden la capacidad del modelo para diferenciar la secuencia original de sus versiones invertida y permutada. Donde se puede ver que en ambos casos, las pérdidas de validación muestran una tendencia a bajar, aunque más leve en la gráfica de la invertida comparada con la constante y exponencial bajada de la permutada. Esto finalmente confirma que el modelo está aprendiendo exitosamente a penalizar las secuencias con un orden temporal incorrecto, cumpliendo el objetivo de la tarea pretexto.

- Evolución del Espacio Latente (PCA y UMAP).

Épocas Iniciales (5-25): Se puede ver que en las primeras etapas, los clústeres semánticos están mal definidos y bastante mezclados. Esto quiere decir que no hay una estructura clara que pueda separar las señas ni las variantes temporales.

Épocas Intermedias (45-65): Por otro lado, se encuentra que a mitad del entrenamiento, la separación entre las tres etiquetas de las señas se vuelve más evidente. Indicando que los tres grupos de colores comienzan a distanciarse en el espacio latente.

Épocas Finales (85-100): Finalmente, en las etapas finales del entrenamiento, el resultado de las métricas anteriormente vistas es visualmente más claro, observando cómo se separan las señas diferentes y también cómo se organiza su estructura de las variantes. Como se puede ver en la gráfica de la mejor época representada con UMAP, dentro de cada cluster, se forma una sub-estructura con los videos originales, representados por un círculo, y los desplazados, representados por una cruz, están agrupados más cerca que las variantes Invertida, representada por una equis, y la Permutada, representada por un cuadrado, que están más distanciadas. Esto demuestra que el modelo aprendió exitosamente a agrupar por significado y separar por diferencia temporal, siendo invariante a pequeños desplazamientos pero sensible a inversiones y permutaciones.

- Análisis de los Resultados Numéricos Finales.

Resumen del Entrenamiento (Tabla 8.1Figura 8.1): Esta tabla, que muestra las métricas en la época 100, muestra que la Pérdida Total de Validación, con un valor de 13.4050, fue inferior a la Pérdida Total de Entrenamiento, con un valor de 16.3837, lo que muestra una muy buena generalización y que no hay sobreajuste. Por otro lado, el valor de 0.0000 para la Pérdida Triplet Inter-Glosario en validación confirma de forma numérica la separación semántica que se pudo observar en las gráficas. Así mismo, la tasa de aprendizaje terminando en 0.0000, muestra que el callback ReduceLROnPlateau redujo la tasa al dejar de mejorar.

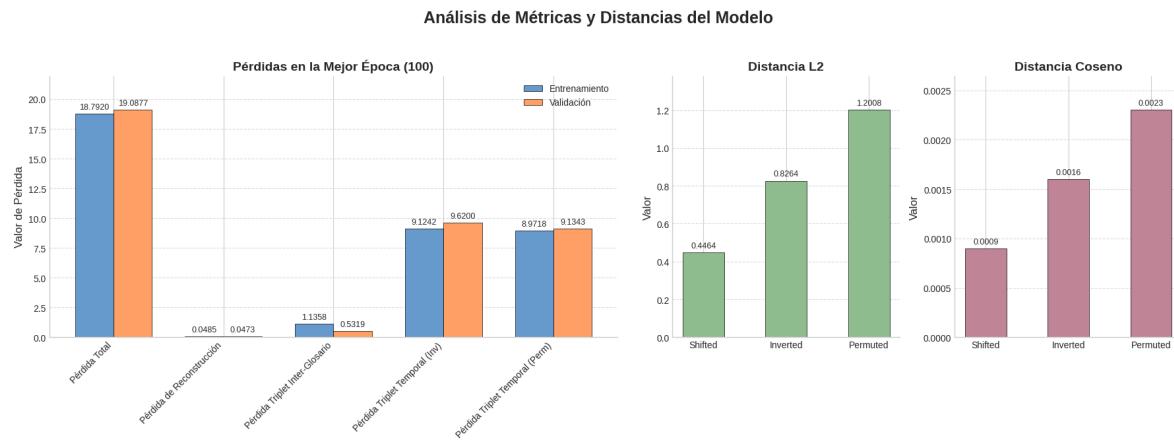
Evaluación de Sensibilidad Temporal (Tabla 8.4): Esta tabla cuantifica el éxito del aprendizaje. Donde los resultados señalan que se cumple con la metodología de evaluación tanto para la Distancia L2 en su magnitud como para la Distancia Coseno en su dirección. Esta jerarquía de distancias demuestra que el modelo aprendió exitosamente a reconocer la secuencia permutada como la más diferente de la original, seguida por la invertida, mientras tiene a la secuencia desplazada como la que más se parece.

2. Análisis de las gráficas y tablas por experimento

2.2. Experimentos con el dataset ISL

Experimento con 2 etiquetas

Figura 8.3: Estas gráficas muestran los valores de las tablas del resumen de entrenamiento y de la evaluación de la sensibilidad temporal.



Cuadro 8.5: Resumen del entrenamiento

Métrica	Valor en la Mejor Época (100)
Pérdida Total (Entrenamiento)	18.7920
Pérdida de Reconstrucción (Entrenamiento)	0.0485
Pérdida Triplet Inter-Glosario (Entrenamiento)	1.1358
Pérdida Triplet Temporal (Inv) (Entrenamiento)	9.1242
Pérdida Triplet Temporal (Perm) (Entrenamiento)	8.9718
Pérdida Total (Validación)	19.0877
Pérdida de Reconstrucción (Validación)	0.0473
Pérdida Triplet Inter-Glosario (Validación)	0.5319
Pérdida Triplet Temporal (Inv) (Validación)	9.6200
Pérdida Triplet Temporal (Perm) (Validación)	9.1343
Tasa de Aprendizaje (lr)	0.0000

Cuadro 8.6: Evaluación de Sensibilidad Temporal

Métrica de Distancia	Shifted	Inverted	Permuted
Distancia L2	0.4464	0.8264	1.2008
Distancia Coseno	0.0009	0.0016	0.0023

Interpretación de todas las gráficas y tablas

- Separación Semántica y Sensibilidad Temporal.

Evolución del Ratio de Separación Semántica vs. Baseline (Figura 10.49 y Figura 10.52): Esta gráfica mide el ratio entre la distancia Inter-clase, es decir, entre señas diferentes, y la distancia Intra-clase, que es entre la misma seña. Se puede ver que el modelo entrenado se compara con un baseline no entrenado, que tiene un ratio de 1.22, el modelo entrenado supera este baseline desde la primera época. Mostrando un pico inicial alcanzando aproximadamente un valor de 1.49, para luego bajar a un valor de aproximadamente 1.37 hasta la época 35 para empezar un crecimiento constante y suave. Finalmente, el valor termina en 1.50 al terminar el entrenamiento, queriendo indicar que el proceso de entrenamiento mejoró significativamente la separación semántica en comparación con la inicialización aleatoria. Por otro lado, cuando se hace la comparación con un baseline de PCA, que tiene un ratio de 0.60. Se puede ver que el modelo entrenado se mantiene muy por encima durante las 100 épocas. Queriendo demostrar que el aprendizaje no lineal del modelo captura relaciones más valiosas que una simple reducción de dimensionalidad lineal de PCA y que el entrenamiento aporta un valor de separación muy grande.

Evolución de la Sensibilidad Temporal vs. Baseline (Figura 10.50): Esta gráfica evalúa cómo el modelo diferencia las secuencias originales de sus variantes temporales. Se puede ver que los baselines no entrenados tienen distancias muy bajas y cercanas entre sí, con valores de 0.10 para la Shifted, 0.12 para la Inverted y 0.16 para la Permuted. A diferencia de las distancias del modelo entrenado, pues después de una pequeña bajada al comienzo, estas aumentan de forma constante, teniendo que la distancia permutada es la que más crece, superando el valor de 1.0, seguida por la distancia de la invertida, terminando en un valor de aproximadamente 0.85 y, por último, la distancia de la desplazada con un valor de 0.63. El modelo mantiene la jerarquía de distancias deseada Shifted Inverted Permuted, demostrando que el modelo ha aprendido a ser más sensible a alteraciones temporales drásticas, como la permutación, que a alteraciones leves como el desplazamiento. Además, se puede ver que el entrenamiento ha logrado estructurar el espacio latente para hacer estas diferencias temporales mucho más notorias que en el modelo no entrenado.

Evolución de la Separación Semántica en el Espacio Latente (Figura 10.53): Esta gráfica muestra cómo los componentes del ratio semántico, es decir la distancia promedio Inter-Clase y la distancia promedio Intra-Clase, presentan una bajada inicial en las primeras épocas, que luego se convierte en un incremento rápido para ambas distancias para luego disminuir su crecimiento. Es importante anotar que, la distancia Inter-clase se mantiene consistentemente por encima de la distancia Intra-clase durante todo el entrenamiento y presentando una tendencia

2. Análisis de las gráficas y tablas por experimento

más grande al crecimiento. Esto es un resultado que es muy bueno, porque muestra que el modelo está logrando el objetivo principal de alejar las representaciones de señas diferentes, es decir, la distancia Inter-clase, mientras mantiene juntas las representaciones de la misma seña, en otras palabras, la distancia Intra-clase.

- Comparación del Modelo y Análisis de Pérdidas.

Comparación de Evolución de Pérdida Total en Validación (Figura 10.51): Esta gráfica compara el modelo principal con un modelo más simple para poder justificar su complejidad. Lo que se puede observar es que la pérdida del modelo simple desciende, pero se estanca rápidamente en un valor cercano a 20.0, concluyendo su entrenamiento antes de finalizar las 100 épocas gracias a la callback EarlyStopping en la época 82. Por otro lado, el modelo principal, aunque tiene una subida inicial, muestra un descenso mucho más grande y sostenido, alcanzando un valor final mucho más bajo de aproximadamente 19.0. Finalmente, este comportamiento justifica la arquitectura más compleja del modelo principal.

Pérdida Total (Figura 10.54): Esta gráfica muestra que la pérdida total en el conjunto de validación presenta una curva suave y una tendencia descendente constante, a pesar de tener una pequeña subida al comienzo, lo que indica que el modelo tiene un proceso de aprendizaje estable con una buena generalización. Por otro lado, la pérdida de entrenamiento es mucho más inestable con varios picos, pero manteniendo una tendencia general a bajar.

Pérdida de Reconstrucción (Figura 10.55): Esta representa la pérdida del autoencoder, a través de la fidelidad de la reconstrucción. Se puede ver que la pérdida de validación desciende bruscamente, alcanzando un valor muy bajo alrededor de la época 35. Para luego experimentar un ligero aumento antes de volver a bajar hasta el final del entrenamiento. Este comportamiento es esperable en el aprendizaje multitarea, donde el modelo puede priorizar momentáneamente las otras pérdidas triplet sobre la de reconstrucción, antes de encontrar un mejor equilibrio general.

Pérdida Triplet Semántica (Figura 10.56): Esta gráfica muestra cómo evoluciona la separación entre las dos señas, mostrando que la pérdida de validación cae bruscamente a un valor por debajo de 1.0 en la primera época, para luego presentar una pequeña subida hasta recuperar el valor que tenía antes de esta en la época 20, para estabilizarse aproximadamente en un valor de 0.5 durante el resto del entrenamiento. Esto quiere decir que hay una muy buena separación semántica en el conjunto de validación y se hace de manera rápida.

Pérdidas Triplet Temporales (Figura 10.57 y Figura 10.58): Estas gráficas miden la capacidad del modelo para diferenciar la secuencia original de sus versiones invertida y permutada. En ambos casos, la pérdida de validación muestra una clara y constante tendencia a bajar, aunque no es tan grande como las anteriores, este comportamiento confirma que el modelo está aprendiendo exitosamente a

penalizar las secuencias con un orden temporal incorrecto, cumpliendo el objetivo de esta tarea pretexto.

- Evolución del Espacio Latente (PCA y UMAP).

Épocas Iniciales (5-25): En las primeras etapas, se puede observar que los clústeres semánticos están mezclados y la estructura temporal interna no se puede diferenciar. Luego, en la época 25, los colores comienzan a separarse, pero las variantes temporales siguen dispersas dentro de cada grupo.

Épocas Intermedias (45-65): En la mitad del entrenamiento, la separación entre los clústeres semánticos se vuelve mucho más evidente y la distancia entre ellos aumenta. Además, se comienza a ver una sub-estructura temporal, donde las secuencias originales y las Shifted tienden a agruparse, mientras que las Inverted y las Permuted están mas lejos dentro del mismo clúster de color.

Épocas Finales (85-100): Finalizando el entrenamiento, se puede ver que las visualizaciones de las últimas épocas muestran el resultado final del aprendizaje de forma clara. Teniendo los clústeres de las señas bastante separados en regiones diferentes del espacio y con una estructura visible de las diferentes variantes. Esta evolución demuestra que el modelo aprendió exitosamente a organizar el espacio latente según los dos objetivos de separar por significado semántico y por su estructura temporal.

- Análisis de los Resultados Numéricos Finales.

Resumen del Entrenamiento (Tabla 8.5): Esta tabla se encarga de presentar las métricas finales en la época 100. Donde la Pérdida Total de Validación con un valor de 19.0877 es ligeramente más grande que la Pérdida Total de Entrenamiento, la cual alcanzó un valor de 18.7920, lo cual indica un pequeño sobreajuste. Así mismo, la Pérdida de Reconstrucción en Validación con un valor de 0.0473 es menor que la de entrenamiento, con un valor de 0.0485, lo que sugiere una muy buena generalización del autoencoder. Por otro lado, la Pérdida Triplet Inter-Glosario de Validación con un valor de 0.5319 indica que, aunque la separación semántica es fuerte, como se puede ver en el ratio de 1.53 de la figura 7.99, no es perfecta. Además, la tasa de aprendizaje finalizando en 0.0000 confirma que el callback ReduceLROnPlateau se activó para reducir la tasa de aprendizaje al detectar un estancamiento en el rendimiento de validación.

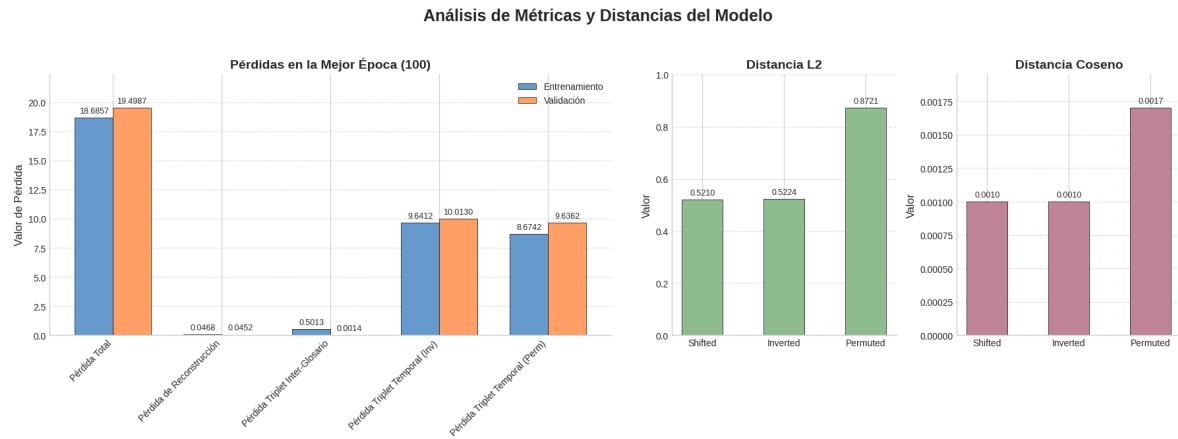
Evaluación de Sensibilidad Temporal (Tabla 8.6): Esta tabla muestra el resultado final del aprendizaje temporal, donde la distancia L2 cumple con la relación esperada con los siguientes valores, 0.4464 para la Shifted, 0.8264 para la Inverted y 1.2008 para la Permuted. De igual manera, la distancia Coseno también cumple con la jerarquía deseada, presentando los siguientes valores, 0.0009 para la Shifted, 0.0016 para la Inverted y 0.0023 para la Permuted. Esto demuestra que el

2. Análisis de las gráficas y tablas por experimento

modelo ha aprendido con éxito la estructura temporal de las secuencias, identificando las permutaciones como el cambio más grande, seguido por las inversiones, y percibiendo los desplazamientos leves como la alteración más similar al video original.

Experimento con 3 etiquetas

Figura 8.4: Estas gráficas muestran los valores de las tablas del resumen de entrenamiento y de la evaluación de la sensibilidad temporal.



Cuadro 8.7: Resumen del entrenamiento

Métrica	Valor en la Mejor Época (100)
Pérdida Total (Entrenamiento)	18.6857
Pérdida de Reconstrucción (Entrenamiento)	0.0468
Pérdida Triplet Inter-Glosario (Entrenamiento)	0.5013
Pérdida Triplet Temporal (Inv) (Entrenamiento)	9.6412
Pérdida Triplet Temporal (Perm) (Entrenamiento)	8.6742
Pérdida Total (Validación)	19.4987
Pérdida de Reconstrucción (Validación)	0.0452
Pérdida Triplet Inter-Glosario (Validación)	0.0014
Pérdida Triplet Temporal (Inv) (Validación)	10.0130
Pérdida Triplet Temporal (Perm) (Validación)	9.6362
Tasa de Aprendizaje (lr)	0.0000

Interpretación de todas las gráficas y tablas

Cuadro 8.8: Evaluación de Sensibilidad Temporal

Métrica de Distancia	Shifted	Inverted	Permuted
Distancia L2	0.5210	0.5224	0.8721
Distancia Coseno	0.0010	0.0010	0.0017

- Separación Semántica y Sensibilidad Temporal.

Evolución del Ratio de Separación Semántica vs. Baseline (Figura 10.73 y Figura 10.76): Esta gráfica mide el ratio entre la distancia Inter-clase, es decir, entre señas diferentes, y la distancia Intra-clase, que es entre la misma seña. Se puede ver que el modelo entrenado se compara con un baseline no entrenado, que tiene un ratio de 1.21. Se puede observar que el modelo entrenado comienza por debajo de este baseline, a pesar de que la gráfica tiene una pequeña subida, luego baja para estabilizarse en un valor de aproximadamente 1.15. Durante las 100 épocas de entrenamiento, el ratio del modelo entrenado no logra superar el baseline no entrenado, sin embargo, sí supera de buena manera el baseline PCA, el cual tiene un valor de 0.62, lo que indica que el aprendizaje no lineal del modelo es superior a una simple reducción de dimensionalidad lineal.

Evolución de la Sensibilidad Temporal vs. Baseline (Figura 10.74): Esta gráfica evalúa la capacidad del modelo para entender el orden temporal. Se puede apreciar que el modelo entrenado supera en gran medida los baselines no entrenados para todas las variantes. La distancia de la Permuted es la que más crece, terminando en un valor aproximado de 0.85. Por otro lado, es importante resaltar que las distancias para la Shifted e Inverted son casi iguales durante todo el entrenamiento, finalizando ambas en un valor cercano a 0.51. Si se tiene solo la información de esta gráfica, el análisis indicaría que el modelo aprendió a ser sensible a las permutaciones, pero falló en aprender la jerarquía entre una alteración leve, como la desplazada, y una severa, como la invertida, tratándolas como igualmente diferentes del original.

Evolución de la Separación Semántica en el Espacio Latente (Figura 10.77): Esta gráfica muestra cómo los componentes del ratio semántico, es decir, la distancia promedio Inter-Clase y la distancia promedio Intra-Clase, evolucionan a través de las épocas. Donde la Inter-clase se mantiene consistentemente por encima de la Intra-Clase. Esto indicando que el modelo está, alejando las clases diferentes entre sí de una manera efectiva. Sin embargo, si se detalla un análisis más profundo, la razón de tener ratio bajo es que la distancia Intra-Clase, es decir, la compactación del clúster, es menor que la distancia Inter-Clase. Esto sugiere que, si bien el modelo separa las clases, no logra compactar eficientemente los miembros de una misma clase. Una posible causa, que luego se analizará más adelante en los espacios latentes, es que la métrica Intra-Clase está siendo dañada por la separación temporal dentro del mismo clúster semántico.

2. Análisis de las gráficas y tablas por experimento

- Comparación del Modelo y Análisis de Pérdidas.

Comparación de Evolución de Pérdida Total en Validación (Figura 10.75):

Esta gráfica compara el modelo principal con un modelo más simple para justificar su complejidad. El Modelo Principal logra una pérdida de validación final de un valor aproximado de 19.5, el cual es significativamente más bajo que el Modelo Simple, que deja de mejorar en un valor de aproximadamente 20.2. Estos resultados justifican el uso de la arquitectura más compleja.

Pérdida Total (Figura 10.78): Esta gráfica muestra que la pérdida total en el conjunto de validación presenta una curva suave y una tendencia descendente constante, a pesar de tener una pequeña subida al comienzo, lo que indica que el modelo tiene un proceso de aprendizaje estable con una buena generalización. Por otro lado, la pérdida de entrenamiento es mucho más inestable con varios picos, pero manteniendo una tendencia general a bajar.

Pérdida de Reconstrucción (Figura 10.79): Esta gráfica muestra que la pérdida de validación del autoencoder disminuye a una gran velocidad, hasta alcanzar un valor de aproximadamente 0.063 alrededor de la época 25, luego desciende de manera más suave hasta llegar a una pérdida final baja de aproximadamente 0.045. Lo que demuestra que el componente de reconstrucción del modelo es funcional.

Pérdida Triplet Semántica (Figura 10.80): Esta gráfica muestra un resultado para determinar la distancia de las señas. La pérdida de validación para la separación de las tres señas inicia muy baja, para tener una subida alrededor de la época 15, para bajar igual de rápido y estabilizarse en 0.0. Esto indica una separación casi perfecta de los tripletes semánticos en el conjunto de validación.

Pérdidas Triplet Temporales (Figura 10.81 y Figura 10.82): Estas gráficas miden la capacidad del modelo para diferenciar la secuencia original de sus versiones invertida y permutada. En ambos casos muestran una alta inestabilidad en las pérdidas de entrenamiento. Sin embargo, la pérdida de validación para la variante invertida se mantiene estable en 10.0, sin mostrar ningún cambio destacable, mientras que para la variante permutada muestra una tendencia descendente.

- Evolución del Espacio Latente (PCA y UMAP).

Épocas Iniciales (5-25): En las primeras etapas del entrenamiento, los puntos están mezclados sin mostrar algúna estructura clara. Posteriormente, hacia la Época 25, los tres grupos de colores, es decir, las etiquetas, comienzan a separarse en regiones distintas, especialmente en la visualización UMAP.

Épocas Intermedias (45-65): A medida que avanza el entrenamiento, la separación entre los clústeres semánticos se vuelve mucho más evidente y la distancia entre ellos aumenta. Además, se comienza a ver una sub-estructura temporal, sin embargo, aún no son claras las distancias que las variaciones tienen entre sí.

Épocas Finales (85-100): En las gráficas finales, se observa que hay tres clústeres de colores distintos y muy bien separados. El espacio latente ha diferenciado exitosamente las tres señas diferentes. También se puede observar una estructura temporal dentro de cada clúster, donde los videos originales, representados por un círculo, y desplazados, representados por la cruz, están agrupados juntos, pero con la misma distancia o mayor de las variantes invertidas, representadas por una equis. Por otro lado, las secuencias permutadas, representadas por un cuadrado, están separadas de la original en la mayoría de los casos. También se puede observar que hay varios casos donde la secuencia invertida está más cerca de la original que la desplazada.

- Análisis de los Resultados Numéricos Finales.

Resumen del Entrenamiento (Tabla 8.7): Esta tabla presenta las métricas finales en la época 100. La Pérdida Total de Validación, con un valor de 19.4987, es por poco mayor a la de Entrenamiento, con un valor de 18.6857, lo cual es un comportamiento que indica un leve sobreajuste. Es válido anotar que, la Pérdida de Reconstrucción en Validación, con un valor de 0.0452, es incluso menor que la de entrenamiento con 0.0468, mostrando una buena generalización del autoencoder. Por otro lado, la Pérdida Triplet Inter-Glosario de Validación, siendo 0.0014, es decir, un valor casi nulo, respalda que hubo una separación casi perfecta. Sin embargo, esto contrasta con el bajo ratio semántico, esto finalmente valida que el modelo logra una separación triplet perfecta, pero lo hace a costa de aumentar la distancia Intra-Clase, es decir, el denominador del ratio, probablemente al separar las variantes temporales dentro del mismo clúster semántico. Finalmente, la Tasa de Aprendizaje en 0.0000 confirma la activación del callback ReduceLROnPlateau.

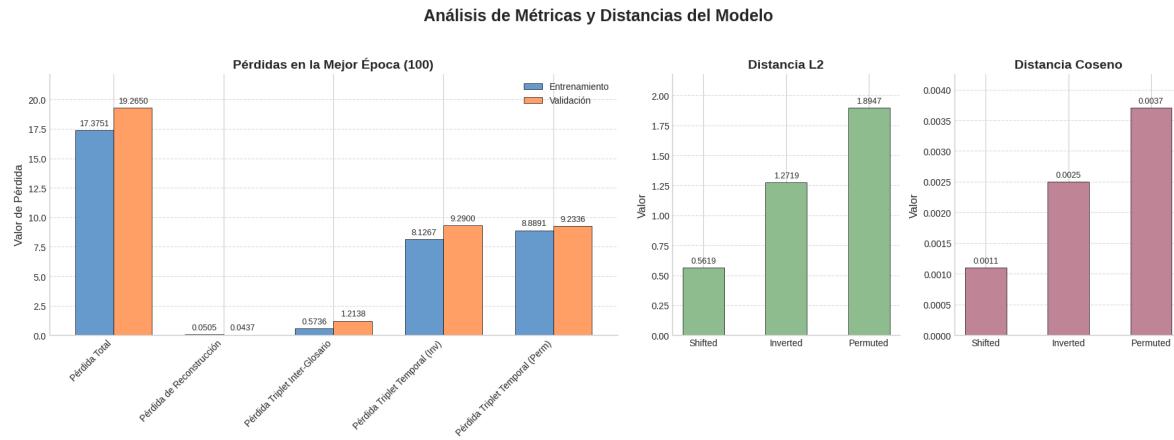
Evaluación de Sensibilidad Temporal (Tabla 8.8): Esta tabla muestra el resultado final del aprendizaje temporal. Donde los valores de Distancia L2 son 0.5210 para la Shifted, 0.5224 para la Inverted y 0.8721 para la Permuted. Estos valores confirman que las distancias para las secuencias Shifted e Inverted son casi iguales. Por lo tanto, la jerarquía de distancias deseada, Shifted < Inverted < Permuted, no se cumple. Queriendo decir que el modelo solo aprendió la relación Shifted ≈ Inverted > Permuted a pesar de que el modelo aprendió a penalizar fuertemente el desorden aleatorio con Permuted, falló en diferenciar la gravedad entre una alteración leve y una severa, tratándolas como igualmente diferentes del original. Los valores de la Distancia Coseno refuerzan este análisis, mostrando una falta de distinción entre la inversión y el desplazamiento.

2. Análisis de las gráficas y tablas por experimento

2.3. Experimentos con el dataset SLOVO

Experimento con 2 etiquetas

Figura 8.5: Estas gráficas muestran los valores de las tablas del resumen de entrenamiento y de la evaluación de la sensibilidad temporal.



Cuadro 8.9: Resumen del entrenamiento

Métrica	Valor en la Mejor Época (100)
Pérdida Total (Entrenamiento)	17.3751
Pérdida de Reconstrucción (Entrenamiento)	0.0505
Pérdida Triplet Inter-Glosario (Entrenamiento)	0.5736
Pérdida Triplet Temporal (Inv) (Entrenamiento)	8.1267
Pérdida Triplet Temporal (Perm) (Entrenamiento)	8.8891
Pérdida Total (Validación)	19.2650
Pérdida de Reconstrucción (Validación)	0.0437
Pérdida Triplet Inter-Glosario (Validación)	1.2138
Pérdida Triplet Temporal (Inv) (Validación)	9.2900
Pérdida Triplet Temporal (Perm) (Validación)	9.2336
Tasa de Aprendizaje (lr)	0.0000

Cuadro 8.10: Evaluación de Sensibilidad Temporal

Métrica de Distancia	Shifted	Inverted	Permuted
Distancia L2	0.5619	1.2719	1.8947
Distancia Coseno	0.0011	0.0025	0.0037

Interpretación de todas las gráficas y tablas

- Separación Semántica y Sensibilidad Temporal.

Evolución del Ratio de Separación Semántica vs. Baseline (Figura 10.97 y Figura 10.100): Estas gráficas miden la capacidad del modelo para separar señas de diferentes clases, con el valor Inter-clase en relación con la compactación de señas de la misma clase, es decir, el valor Intra-clase. Se puede ver que muestra un comportamiento inestable, iniciando por debajo del baseline no entrenado con un valor de aproximadamente 0.973, para luego tener un pico inicial que supera brevemente el baseline no entrenado, el cual tiene un valor de aproximadamente 1.00. Para luego caer y alcanzar un segundo pico más suave de aproximadamente 0.995 cerca de la época 40 y finaliza el entrenamiento con un ratio de aproximadamente 0.979, por debajo del baseline no entrenado. Si bien el modelo no logra superar consistentemente al baseline aleatorio, sí supera de gran manera y sostenida al baseline de PCA. Esto quiere decir que, aunque la separación semántica final no es la mejor, el aprendizaje no lineal del modelo captura relaciones más valiosas que una simple reducción de dimensionalidad lineal.

Evolución de la Sensibilidad Temporal vs. Baseline (Figura 10.98): Esta gráfica evalúa cómo el modelo diferencia las secuencias originales de sus variantes temporales. Los baselines no entrenados muestran distancias bajas con valores de 0.31 para la Shifted, de 0.48 para la Inverted Inverted y 0.40 para la Permuted. Como se puede ver, el modelo entrenado, tras un descenso inicial, aumenta en gran medida la distancia para todas las variantes, superando en gran medida a los baselines. Es importante anotar que la distancia Shifted se estabiliza alrededor de 0.74, mientras que las distancias Inverted y Permuted crecen mucho más y se comportan de una manera muy parecida, finalizando ambas en valores altos y parecidos, alrededor de 1.07 y 1.08 respectivamente. Esto quiere decir que el modelo demuestra haber aprendido exitosamente la sensibilidad temporal, penalizando fuertemente las alteraciones, además, cumple la jerarquía deseada que indica que las alteraciones leves, es decir, la secuencia Shifted, son más cercanas al original que las Inverted, Permuted. Sin embargo, a diferencia de otros experimentos, el modelo parece tratar las secuencias invertidas y permutadas como casi igualmente diferentes del original.

Evolución de la Separación Semántica en el Espacio Latente (Figura 10.101): Este gráfico muestra los componentes del ratio semántico, señalando el comportamiento de la distancia Promedio Inter-Clase, es decir, entre clases diferentes, y la distancia Promedio Intra-Clase, dentro de la misma clase. El resultado muestra que ambas líneas, Intra-Clase y Inter-Clase, son casi iguales durante las 100 épocas, pues ambas siguen el mismo patrón de un descenso inicial, un ascenso marcado y una estabilización final alrededor de un valor de 4.5. Esto explica por qué el ratio semántico en la Figura 7.147 se mantiene cercano a 1.0, el modelo no

2. Análisis de las gráficas y tablas por experimento

está logrando que la distancia entre clases sea mayor que la distancia dentro de las clases con una gran diferencia. Esto quiere decir que el espacio latente se expande y contrae en su totalidad, pero la separación semántica relativa no mejora.

- Comparación del Modelo y Análisis de Pérdidas.

Comparación de Evolución de Pérdida Total en Validación (Figura 10.99): Esta gráfica se encarga de justificar la complejidad de la arquitectura del modelo comparándolo con un modelo más simple. Se puede ver que el modelo simple desciende rápidamente, pero se estanca en una pérdida de validación de aproximadamente 20.6, activando el EarlyStopping alrededor de la época 82. Mientras que el modelo principal, aunque tiene un pico inicial, continúa un descenso sostenido hasta finalizar en una pérdida más baja, de aproximadamente un valor de 19.3, justificando el uso de la arquitectura más compleja.

Pérdida Total (Figura 10.102): La gráfica de pérdida total muestra una curva de validación suave y con una tendencia descendente constante, lo que indica un proceso de aprendizaje estable y una buena generalización. A diferencia de la pérdida de entrenamiento, que es mucho más inestable, con bastantes picos, aunque mantiene la misma tendencia general a bajar.

Pérdida de Reconstrucción (Figura 10.103): Esta gráfica mide la capacidad del autoencoder para reconstruir el video original. Ambas curvas, de entrenamiento y validación, muestran un descenso constante, donde se aprecia que la pérdida de validación es menor que la pérdida de entrenamiento durante todas las épocas, finalizando en un valor de aproximadamente 0.044. Este resultado sugiere una muy buena capacidad de generalización del componente autoencoder.

Pérdida Triplet Semántica (Figura 10.104): Esta gráfica se encarga de mostrar la métrica que evalúa la separación entre las dos señas. Se puede evidenciar que la pérdida de entrenamiento es inestable, pero tiene un comportamiento a disminuir. Sin embargo, la pérdida de validación presenta una bajada brusca que luego presenta una leve subida hasta la época 20 aproximadamente, para luego descender y se estabiliza en un valor de aproximadamente 1.2. Esto indica que, si bien el modelo aprende a separar las señas en el conjunto de entrenamiento, tiene dificultades para generalizar esta separación semántica al conjunto de validación, lo que se alinea con el bajo ratio semántico observado.

Pérdidas Triplet Temporales (Figura 10.105 y Figura 10.106): Estas gráficas miden la capacidad de diferenciar la secuencia original de sus versiones invertida y permutada. Se puede ver que ambas presentan un comportamiento parecido, donde la pérdida de entrenamiento es muy inestable y oscila en un amplio rango. Por otro lado, la pérdida de validación en ambos casos desciende ligeramente en las primeras épocas y luego se mantiene notablemente estable, casi pareciendo plana, alrededor de un valor de 9.3 para la Invertida y de 9.2 para la Permutada.

Esto quiere decir que el modelo encuentra rápidamente una solución que es suficientemente buena para la separación temporal y no la logra mejorar durante el resto del entrenamiento.

- Evolución del Espacio Latente (PCA y UMAP).

Épocas Iniciales (5-25): En la época 5, las visualizaciones de PCA y UMAP muestran un espacio latente completamente desorganizado. Los puntos de ambas clases y todas las variantes temporales están mezcladas sin una estructura aparente. Luego, hacia la época 25, se comienzan a formar ciertos grupos semánticos, especialmente en la visualización UMAP, aunque todavía con mucha dispersión.

Épocas Intermedias (45-65): A mitad del entrenamiento, la separación semántica se vuelve mucho más definida, donde las visualizaciones, tanto en PCA como en UMAP, muestran regiones distintas para las clases. Sin embargo, dentro de estos grupos semánticos, la estructura temporal permanece mezclada y desordenada.

Épocas Finales (85-100): Al final del entrenamiento, el espacio latente muestra una clara separación semántica. Donde las clases ocupan regiones distintas del espacio. Sin embargo, el objetivo secundario de organizar el espacio temporalmente dentro de cada clúster no se logra. Teniendo las diferentes variantes mezcladas. Esto refuerza que el modelo está aplicando la misma transformación a todas las variantes, logrando separar las clases, pero fallando en organizar las sub-estructuras temporales.

- Análisis de los Resultados Numéricos Finales.

Resumen del Entrenamiento (Tabla 8.9): La tabla de métricas finales confirma las observaciones pasadas, donde la Pérdida Total de Validación, con un valor de 19.2650, es superior a la de Entrenamiento, la cual tiene un valor de 17.3751, lo que indica un comportamiento normal. Por otro lado, la Pérdida de Reconstrucción de Validación es inferior a la de Entrenamiento, destacando la excelente generalización del autoencoder. A su vez, la Pérdida Triplet Inter-Glosario de Validación es más del doble que la de Entrenamiento, lo que señala la dificultad en la separación semántica. Además, las pérdidas temporales de validación, con valores de 9.2900 para la invertida y 9.2336 para la permutada, son altas y muy similares. Por último, la Tasa de Aprendizaje final en 0.0000 indica que el callback ReduceLROnPlateau se activó, reduciendo la tasa al detectar un estancamiento.

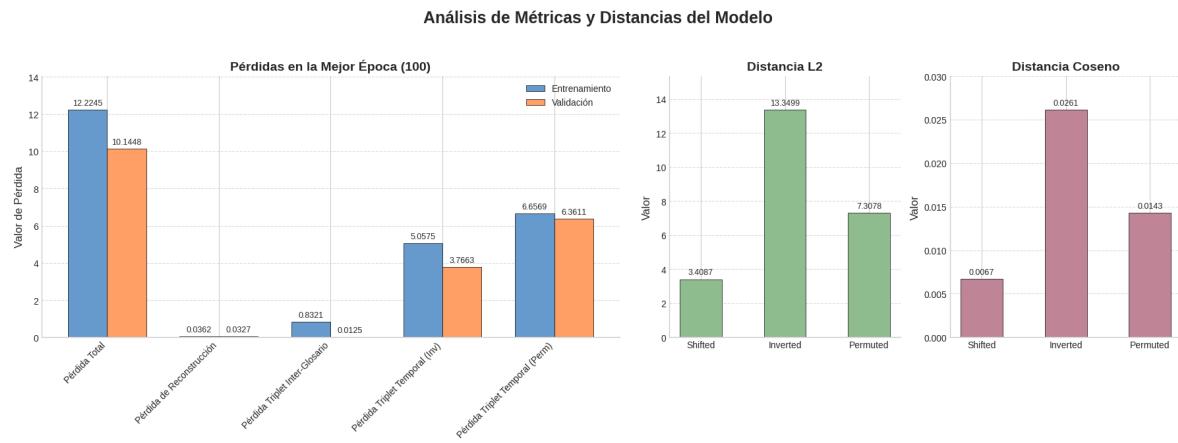
Evaluación de Sensibilidad Temporal (Tabla 8.10): Esta tabla presenta la evaluación final de la sensibilidad temporal. Donde, a pesar de las trayectorias de la separación de las secuencias, los valores numéricos finales demuestran que el modelo aprendió la jerarquía temporal esperada. Con valores que cumplen la relación Shifted Inverted Permuted, con valores de 0.5619 1.2719 1.8947. Y con una distancia Coseno que también cumple la misma jerarquía, con valores de

2. Análisis de las gráficas y tablas por experimento

0.0011 0.0025 0.0037. Esto demuestra finalmente que el modelo ha aprendido con éxito a estructurar el espacio latente de manera que las permutaciones aleatorias están más lejos del original, seguidas por las inversiones, mientras que los desplazamientos leves se mantienen más cerca.

Experimento con 3 etiquetas

Figura 8.6: Estas gráficas muestran los valores de las tablas del resumen de entrenamiento y de la evaluación de la sensibilidad temporal.



Cuadro 8.11: Resumen del entrenamiento

Métrica	Valor en la Mejor Época (100)
Pérdida Total (Entrenamiento)	12.2245
Pérdida de Reconstrucción (Entrenamiento)	0.0362
Pérdida Triplet Inter-Glosario (Entrenamiento)	0.8321
Pérdida Triplet Temporal (Inv) (Entrenamiento)	5.0575
Pérdida Triplet Temporal (Perm) (Entrenamiento)	6.6569
Pérdida Total (Validación)	10.1448
Pérdida de Reconstrucción (Validación)	0.0327
Pérdida Triplet Inter-Glosario (Validación)	0.0125
Pérdida Triplet Temporal (Inv) (Validación)	3.7663
Pérdida Triplet Temporal (Perm) (Validación)	6.3611
Tasa de Aprendizaje (lr)	0.0000

Interpretación de todas las gráficas y tablas

Cuadro 8.12: Evaluación de Sensibilidad Temporal

Métrica de Distancia	Shifted	Inverted	Permuted
Distancia L2	3.4087	13.3499	7.3078
Distancia Coseno	0.0067	0.0261	0.0143

- Separación Semántica y Sensibilidad Temporal.

Evolución del Ratio de Separación Semántica vs. Baseline (Figura 10.121 y Figura 10.124): Estas gráficas se encargan de medir el ratio entre la distancia Inter-Clase, es decir, entre señas diferentes, y la distancia Intra-Clase, en otras palabras, dentro de la misma señal. Como se puede observar, el baseline del modelo no entrenado tiene un valor de 0.97 y el modelo entrenado nunca supera este baseline. El modelo original, tras un pico inicial cercano a 0.966, desciende y permanece por debajo del baseline durante el resto del entrenamiento de manera inestable. Esto indica que el entrenamiento no logró mejorar de la mejor manera posible la separación semántica en comparación con la inicialización aleatoria. Sin embargo, el modelo sí supera de manera muy grande al baseline de PCA, el cual tiene un valor de 0.52, lo que justifica el uso de un modelo de aprendizaje no lineal sobre un método lineal simple.

Evolución de la Sensibilidad Temporal vs. Baseline (Figura 10.122): Esta gráfica evalúa cómo el modelo diferencia las secuencias originales de sus variantes temporales. Se puede ver que los baselines no entrenados tienen distancias bajas con valores de 0.39 para la Shifted, 0.50 para la Inverted y 0.42 para la Permuted. A diferencia de las distancias del modelo entrenado, que aumentan de forma constante y significativa, superando a todos los baselines. Sin embargo, el modelo aprendió a ser sensible a las alteraciones, pero considera la versión invertida como una disrupción temporal mayor que la versión permutada, lo cual contradice la hipótesis de la metodología, teniendo la siguiente relación, Shifted > Permuted > Inverted.

Evolución de la Separación Semántica en el Espacio Latente (Figura 10.125): Esta gráfica es importante para poder entender el bajo ratio semántico, porque muestra que la Distancia Promedio Inter-Clase es menor a la Distancia Promedio Intra-Clase y siguen la misma trayectoria durante las 100 épocas. Esto significa que el modelo está logrando lo contrario al objetivo de maximizar la distancia entre las clases diferentes y poder minimizar la distancia dentro de la misma clase. Este comportamiento quiere decir que se están agrupando las señas de la misma clase de forma menos compacta que las señas de clases diferentes.

- Comparación del Modelo y Análisis de Pérdidas.

Comparación de Evolución de Pérdida Total en Validación (Figura 10.123):

SERGIO ARBOLEDA

2. Análisis de las gráficas y tablas por experimento

Esta gráfica se encarga de justificar la complejidad del modelo. Donde se puede observar que el Modelo Principal muestra una disminución sostenida en cuanto a su valor de pérdida, finalizando el entrenamiento en un valor cercano a 16.8. Mientras que el Modelo Simple se estanca rápidamente en un valor por encima de 19.0, activando el EarlyStopping en la época 40, significa que la arquitectura más compleja es mucho más buena para este problema.

Pérdida Total (Figura 10.126): Esta gráfica muestra cómo avanza la pérdida de validación. Muestra una curva suave y descendente, indicando un proceso de aprendizaje estable y una buena generalización, finalizando en un valor inferior a 10.

Pérdida de Reconstrucción (Figura 10.127): Esta gráfica muestra cómo evoluciona la pérdida del autoencoder, el cual está encargado de la fidelidad visual luego de la reconstrucción, el cual muestra un buen rendimiento. Porque se puede observar que la pérdida de validación es menor que la pérdida de entrenamiento durante todo el entrenamiento.

Pérdida Triplet Semántica (Figura 10.128): Esta gráfica muestra cómo se comporta la pérdida de validación para la separación semántica, es decir, la Inter-Glosario. Se puede ver que la pérdida de validación para la separación semántica Inter-Glosario siempre permanece menor a la de entrenamiento y cae a 0.0 alrededor de la época 70 y no vuelve a subir durante todo el entrenamiento. Esto indicaría una separación semántica perfecta, lo cual no puede ser interpretado de esta manera por el bajo ratio semántico y la posible solapación de puntos debido a que la distancia Intra-Clase es mayor a la Inter-Clase.

Pérdidas Triplet Temporales (Figura 10.129 y Figura 10.130): Estas gráficas miden la capacidad de diferenciar la secuencia original de sus versiones invertida y permutada. Se puede observar que para ambas pérdidas muestran una tendencia a bajar de manera suave y constante, descendiendo desde un valor de aproximadamente 10.0 a uno de aproximadamente 5 para la invertida y 4.5 para la permutada. Sin embargo, la diferencia reside en el valor de la pérdida de entrenamiento, el cual es menor en la invertida por aproximadamente una unidad, lo que puede evidenciar un pequeño sobreajuste. Mientras que para la permutada el valor de entrenamiento es mayor que el de validación, lo que indica una buena generalización.

- Evolución del Espacio Latente (PCA y UMAP).

Épocas Iniciales (5-25): Al comienzo del entrenamiento, las visualizaciones de PCA y UMAP muestran un espacio latente completamente desorganizado. Los puntos de ambas clases y todas las variantes temporales están mezcladas sin una estructura aparente. Luego, hacia la época 25, se comienzan a formar ciertos grupos semánticos, especialmente en la visualización UMAP, aunque todavía con mucha dispersión.

Épocas Intermedias (45-65): A mitad del entrenamiento, la separación semántica no se puede apreciar de buena manera. Sin embargo, dentro de estos grupos semánticos, la estructura temporal comienza a mostrar signos de un comportamiento no deseado, teniendo las secuencias Permuted muy cerca de la original y la invertida más separada.

Épocas Finales (85-100): Finalizando el entrenamiento, el espacio latente no logra una separación semántica clara y definida. En las gráficas finales, los puntos de diferentes colores permanecen muy superpuestos y distribuidos por el espacio, sin formar clústeres definidos, debido a su mayor distancia dentro de cada cluster.

- Análisis de los Resultados Numéricos Finales.

Resumen del Entrenamiento (Tabla 8.11): En esta tabla se puede ver que la pérdida total es buena en validación y menor que la de entrenamiento, a simple vista se podría pensar que el modelo generaliza bien. Sin embargo, la pérdida total, al ser una suma ponderada de varias de las otras pérdidas, puede ocultar un mal rendimiento en un componente si otros componentes tienen un rendimiento excepcionalmente bueno. Como se puede ver en este caso, las pérdidas temporales y de reconstrucción, que tienen valores numéricos más altos, influyen en la suma, ocultando el problema con la separación semántica. Por otro lado, la pérdida Inter-glosario con un valor tan cercano a cero indica una separación semántica casi perfecta, lo cual va en contra de las gráficas del ratio semántico y la separación en el espacio latente. De esta misma manera, los valores de las secuencias invertidas y permutadas demuestran que el modelo aprendió a detectar alteraciones en el orden temporal y a penalizarlas aumentando su distancia en el espacio latente, logrando de esta manera minimizar exitosamente las pérdidas correspondientes. Sin embargo, esta sensibilidad es superficial y extraña, priorizando la inversión sobre la permutación, lo que indica que no aprendió la estructura temporal subyacente del movimiento, sino que aplicó un atajo en todo el aprendizaje.

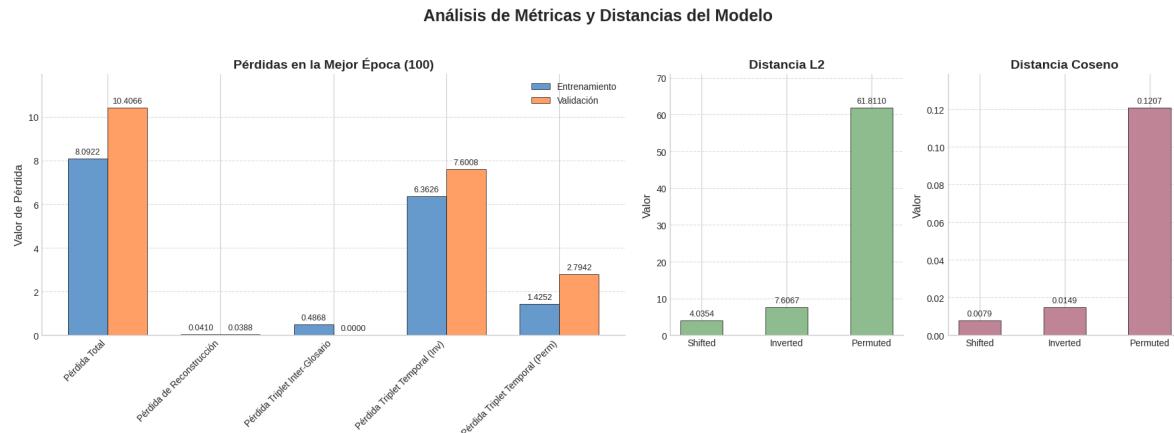
Evaluación de Sensibilidad Temporal (Tabla 8.12): En esta tabla se puede ver que los valores de distancia L2 son altos, lo que es bueno, pues confirma que el modelo mapea las variaciones lejos de las originales, demostrando una alta sensibilidad temporal. Sin embargo, el orden es el incorrecto, pues el modelo considera la Inverted como una alteración mucho mayor, con un valor de 13.349, que una Permuted con un valor de 7.307. Por otro lado, los valores de la longitud Coseno, al estar cerca de cero, indican que los vectores de las secuencias apuntan en direcciones parecidas. Cuando se tienen en cuenta ambas distancias, se puede ver que el modelo no está reorganizando los puntos de forma compleja, sino que simplemente los está empujando radialmente lejos del origen, haciendo que su distancia euclídea aumente bastante sin cambiar mucho su ángulo relativo con un atajo, lo que impidió la correcta agrupación semántica.

2. Análisis de las gráficas y tablas por experimento

2.4. Experimentos con los tres datasets

Experimento con 2 etiquetas

Figura 8.7: Estas gráficas muestran los valores de las tablas del resumen de entrenamiento y de la evaluación de la sensibilidad temporal.



Cuadro 8.13: Resumen del entrenamiento

Métrica	Valor en la Mejor Época (100)
Pérdida Total (Entrenamiento)	8.0922
Pérdida de Reconstrucción (Entrenamiento)	0.0410
Pérdida Triplet Inter-Glosario (Entrenamiento)	0.4868
Pérdida Triplet Temporal (Inv) (Entrenamiento)	6.3626
Pérdida Triplet Temporal (Perm) (Entrenamiento)	1.4252
Pérdida Total (Validación)	10.4066
Pérdida de Reconstrucción (Validación)	0.0388
Pérdida Triplet Inter-Glosario (Validación)	0.0000
Pérdida Triplet Temporal (Inv) (Validación)	7.6008
Pérdida Triplet Temporal (Perm) (Validación)	2.7942
Tasa de Aprendizaje (lr)	0.0000

Cuadro 8.14: Evaluación de Sensibilidad Temporal

Métrica de Distancia	Shifted	Inverted	Permuted
Distancia L2	4.0354	7.6067	61.8110
Distancia Coseno	0.0079	0.0149	0.1207

Interpretación de todas las gráficas y tablas

- Separación Semántica y Sensibilidad Temporal.

Evolución del Ratio de Separación Semántica vs. Baseline (Figura 10.145 y Figura 10.148): Estas gráficas se encargan de medir el ratio entre la distancia Inter-Clase, es decir, entre señas diferentes, y la distancia Intra-Clase, en otras palabras, dentro de la misma señal. Como se puede observar, el baseline del modelo no entrenado tiene un valor de 0.95 y el modelo entrenado muestra un comportamiento estable, superando consistentemente el baseline después de las primeras épocas. Sin embargo, el ratio se mantiene muy cercano a 1.0 durante todo el entrenamiento, finalizando en un valor de aproximadamente 0.99. Esto indica que, aunque el modelo mejora ligeramente con respecto a la inicialización aleatoria, no logra que la distancia Inter-Clase supere significativamente a la Intra-Clase. Sin embargo, el modelo sí supera de manera muy grande al baseline de PCA, el cual tiene un valor de 0.25, lo que justifica el uso de un modelo de aprendizaje no lineal sobre un método lineal simple.

Evolución de la Sensibilidad Temporal vs. Baseline (Figura 10.146): Esta gráfica evalúa cómo el modelo diferencia las secuencias originales de sus variantes temporales. Se puede ver que los baselines no entrenados tienen distancias bajas con valores de 0.52 para la Shifted, 0.61 para la Inverted y 0.48 para la Permuted. A diferencia de las distancias del modelo entrenado, que aumentan de forma constante y significativa, superando a todos los baselines. Esto quiere decir que el modelo aprende con éxito la jerarquía de disruptión temporal, teniendo a la distancia a la versión Shifted con un valor de aproximadamente 1.9, siendo la más baja, seguida por la Inverted con aproximadamente 2.7 y la Permuted con aproximadamente 4.8, siendo la más alta. Esto demuestra que el modelo es sensible al orden temporal y cumple con la metodología de evaluación Shifted Inverted Permuted.

Evolución de la Separación Semántica en el Espacio Latente (Figura 10.149): Esta gráfica es importante para poder entender el bajo ratio semántico, porque muestra que la Distancia Promedio Intra-Clase es, un poco más grande que la Distancia Promedio Inter-Clase durante casi todo el entrenamiento. Ambas distancias siguen la misma trayectoria, finalizando en valores muy cercanos, pero esto significa que el modelo no está logrando el objetivo de maximizar la distancia entre las clases diferentes y poder minimizar la distancia dentro de la misma clase. Este comportamiento ocurre por una alta varianza, donde el modelo considera que una señal de brother del dataset ISL es tan diferente de una señal de brother del dataset WSL como lo es de una señal de cold.

- Comparación del Modelo y Análisis de Pérdidas.

Comparación de Evolución de Pérdida Total en Validación (Figura 10.147):

SERGIO ARBOLEDA

2. Análisis de las gráficas y tablas por experimento

Esta gráfica se encarga de justificar la complejidad del modelo. Donde se puede observar que el Modelo Principal muestra una disminución drástica y sostenida en cuanto a su valor de pérdida, finalizando el entrenamiento en un valor cercano a 13.0. Mientras que el Modelo Simple se estanca rápidamente en un valor por encima de 19.0, significando que la arquitectura más compleja es mucho más buena para este problema.

Pérdida Total (Figura 10.150): Esta gráfica muestra cómo avanza la pérdida de validación. La cual presenta una curva suave con una tendencia descendente constante, finalizando en un valor de aproximadamente 13.0. Sin embargo, esta pérdida es notablemente más alta que la de entrenamiento, a pesar de la alta inestabilidad de la pérdida de la misma, esta finaliza en un valor más bajo. Lo que sugiere un sobreajuste del modelo a los datos de entrenamiento.

Pérdida de Reconstrucción (Figura 10.151): Esta gráfica muestra cómo evoluciona la pérdida del autoencoder, el cual está encargado de la fidelidad visual luego de la reconstrucción, el cual muestra un buen rendimiento. Puesto que la pérdida de validación desciende bruscamente para luego descender de manera constante tiendo valores bajos muy parecidos a los del entrenamiento, terminando ambas en un valor de aproximadamente 0.04.

Pérdida Triplet Semántica (Figura 10.152): Esta gráfica muestra cómo se comporta la pérdida de validación para la separación semántica, es decir, la Inter-Glosario. La cual se puede ver que cae a un valor cercano a 0.4 alrededor de la época 20 para estabilizarse en este punto. Esto aparentemente quiere decir que hay una separación semántica casi perfecta. Sin embargo, este resultado cambia con las demás gráficas, demostrando que resulta en cero no porque haya agrupado las clases, minimizando la distancia Intra-Clase, sino porque la distancia Inter-Clase e Intra-Clase son tan similares que satisfacen matemáticamente la condición de la pérdida.

Pérdidas Triplet Temporales (Figura 10.153 y Figura 10.154): Estas gráficas miden la capacidad de diferenciar la secuencia original de sus versiones invertida y permutada. Se puede observar que para ambas pérdidas, sus líneas de validación muestran una tendencia a bajar. El problema reside cuando se comparan con sus rendimientos de entrenamiento, donde ambas pérdidas de entrenamiento son más bajas que las de validación, con la permutada teniendo una diferencia entre estas dos perdidas mucho más grande. Esto indicaría que, aunque las pérdidas están descendiendo, presentan un sobreajuste.

- Evolución del Espacio Latente (PCA y UMAP).

Épocas Iniciales (5-25): En las primeras etapas, el espacio latente está desorganizado semánticamente, donde las etiquetas y las variaciones están completamente mezcladas. Como se puede apreciar en las diferentes gráficas por idioma, el modelo intenta hacer algunas separaciones tiendiendo a agrupar por idioma en algunas

ocasiones, y los puntos que representan a cada dataset ya forman clústeres que, aunque no estén perfectamente definidos, superan la agrupación de las secuencias y sus variaciones.

Épocas Intermedias (45-65): A medida que avanza el entrenamiento, no se empieza a observar una agrupación en las etiquetas, los clústeres siguen estando muy dispersos y solapados. Lo mismo no se puede aplicar para el caso de los datasets, los cuales siguen mostrando una separación por idioma como al inicio del entrenamiento, separando al dataset WSL por completo en una esquina.

Épocas Finales (85-100): Las visualizaciones finales muestran que no hay una separación semántica, debido a que los puntos de las etiquetas no forman clústeres distintos, sino que estos están en gran medida superpuestos y distribuidos por todo el espacio, mezclados dentro de las agrupaciones por idioma. Esto confirma lo que se vio en las gráficas de ratio semántico y de distancias. Sin embargo, a pesar de esto, se puede observar una subestructura temporal que, aunque no es muy consistente, muestra que para las originales, la desplazada está más cerca, mientras que las invertidas, están un poco más lejos, y permutadas están extremadamente más lejos.

- Análisis de los Resultados Numéricos Finales.

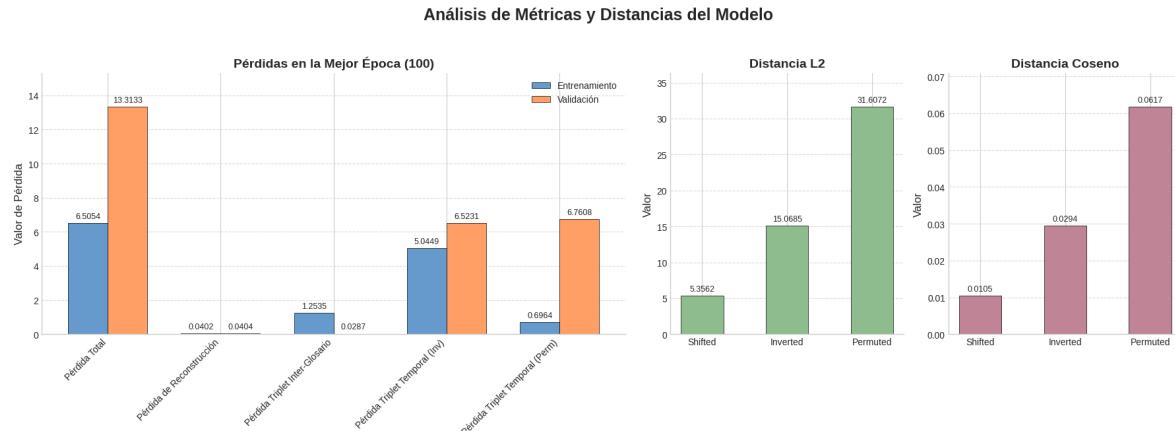
Resumen del Entrenamiento (Tabla 8.13): La tabla muestra que para el valor de la Pérdida Total de Validación se tuvo un valor de 10.4066, y para la Pérdida Total de Entrenamiento, uno de 8.0922. El hecho de que la pérdida de entrenamiento sea más baja que la de validación sugiere un sobreajuste del modelo a los datos de entrenamiento. Por otro lado, la Pérdida Triplet Inter-Glosario de Validación tuvo un valor de 0.0000. Este valor es muy bajo, llegando a cero, pero es engañoso. Este valor podría indicar un éxito en un escenario diferente, pero en este caso señala que el modelo no aprendió a agrupar las secuencias, falló en entender el concepto abstracto de cada seña, tratando todas las secuencias como diferentes, resultando en una distancia Intra-Clase alta e incluso mayor que la Inter-Clase.

Evaluación de Sensibilidad Temporal (Tabla 8.14): Esta tabla señala que la tarea de aprendizaje temporal fue exitosa, esto porque la distancia L2 cumple con la relación Shifted Inverted Permuted con valores de 4.0354 para la Shifted, 7.6067 para la Inverted y 61.8110 para la Permuted. De igual manera, la distancia Coseno también lo cumple con valores de 0.0079 para la Shifted, 0.0149 para la Inverted y 0.0617 para la Permuted. Este comportamiento confirma que, a pesar del fallo en la agrupación semántica por etiqueta, el aprendizaje de la estructura temporal fue exitoso.

2. Análisis de las gráficas y tablas por experimento

Experimento con 3 etiquetas

Figura 8.8: Estas gráficas muestran los valores de las tablas del resumen de entrenamiento y de la evaluación de la sensibilidad temporal.



Cuadro 8.15: Resumen del entrenamiento

Métrica	Valor en la Mejor Época (100)
Pérdida Total (Entrenamiento)	6.5054
Pérdida de Reconstrucción (Entrenamiento)	0.0402
Pérdida Triplet Inter-Glosario (Entrenamiento)	1.2535
Pérdida Triplet Temporal (Inv) (Entrenamiento)	5.0449
Pérdida Triplet Temporal (Perm) (Entrenamiento)	0.6964
Pérdida Total (Validación)	13.3133
Pérdida de Reconstrucción (Validación)	0.0404
Pérdida Triplet Inter-Glosario (Validación)	0.0287
Pérdida Triplet Temporal (Inv) (Validación)	6.5231
Pérdida Triplet Temporal (Perm) (Validación)	6.7608
Tasa de Aprendizaje (lr)	0.0000

Cuadro 8.16: Evaluación de Sensibilidad Temporal

Métrica de Distancia	Shifted	Inverted	Permuted
Distancia L2	5.3562	15.0685	31.6072
Distancia Coseno	0.0105	0.0294	0.0617

Interpretación de todas las gráficas y tablas

- Separación Semántica y Sensibilidad Temporal.

Evolución del Ratio de Separación Semántica vs. Baseline (Figura 10.183 y Figura 10.186): Estas gráficas se encargan de medir el ratio entre la distancia Inter-Clase, es decir, entre señas diferentes, y la distancia Intra-Clase, en otras palabras, dentro de la misma señal. Como se puede observar, el baseline del modelo no entrenado tiene un valor de 1.04 y el modelo entrenado muestra un comportamiento inestable, el cual supera el baseline en las primeras épocas, para luego caer por debajo de él durante unas épocas para volver a superarlo de manera inestable y volver a caer por debajo del mismo. La mayor parte del entrenamiento, aproximadamente desde la época 40 a la 95, estuvo por debajo del baseline. Esto indica que el entrenamiento no logró mejorar de manera consistente la separación semántica en comparación con la inicialización aleatoria. Sin embargo, el modelo sí supera de manera muy grande al baseline de PCA, el cual tiene un valor de 0.52, lo que justifica el uso de un modelo de aprendizaje no lineal sobre un método lineal simple.

Evolución de la Sensibilidad Temporal vs. Baseline (Figura 10.184): Esta gráfica evalúa cómo el modelo diferencia las secuencias originales de sus variantes temporales. Se puede ver que los baselines no entrenados tienen distancias bajas con valores de 0.50 para la Shifted, 0.57 para la Inverted y 0.42 para la Permuted. A diferencia de las distancias del modelo entrenado, que aumentan de forma constante y significativa, superando a todos los baselines. Esto quiere decir que el modelo aprende con éxito la jerarquía de disruptión temporal, teniendo a la distancia a la versión Shifted con un valor de aproximadamente 2.05, siendo la más baja, seguida por la Inverted con aproximadamente 3.4 y la Permuted con aproximadamente 4.1, siendo la más alta. Esto demostraría que el modelo es sensible al orden temporal y cumple con la metodología de evaluación Shifted Inverted Permuted. Sin embargo, teniendo en cuenta las otras gráficas, este resultado es engañoso.

Evolución de la Separación Semántica en el Espacio Latente (Figura 10.187): Esta gráfica es importante para poder entender el bajo ratio semántico, porque muestra que la Distancia Promedio Inter-Clase y la Distancia Promedio Intra-Clase son casi iguales y siguen la misma trayectoria durante las 100 épocas, donde la Inter-Clase es mayor por una pequeña diferencia. Esto significa que el modelo no está logrando el objetivo de maximizar la distancia entre las clases diferentes y poder minimizar la distancia dentro de la misma clase. Este comportamiento ocurre por una alta varianza, donde el modelo considera que una señal de brother del dataset ISL es tan diferente de una señal de brother del dataset WSL como lo es de una señal de cold.

- Comparación del Modelo y Análisis de Pérdidas.

2. Análisis de las gráficas y tablas por experimento

Comparación de Evolución de Pérdida Total en Validación (Figura 10.185): Esta gráfica se encarga de justificar la complejidad del modelo. Donde se puede observar que el Modelo Principal muestra una disminución drástica y sostenida en cuanto a su valor de pérdida, finalizando el entrenamiento en un valor cercano a 13. Mientras que el Modelo Simple se estanca rápidamente en un valor por encima de 19.0, significando que la arquitectura más compleja es mucho más buena para este problema.

Pérdida Total (Figura 10.188): Esta gráfica muestra cómo avanza la pérdida de validación. La cual presenta una curva suave con una tendencia descendente constante, después de un pequeño salto que tiene en las primeras épocas, finalizando en un valor de aproximadamente 13. Sin embargo, esta pérdida es mucho más alta que la de entrenamiento, a pesar de la alta inestabilidad de la pérdida de la misma, esta finaliza en un valor de aproximadamente 6.5, lo que indica un sobreajuste del modelo a los datos de entrenamiento.

Pérdida de Reconstrucción (Figura 10.189): Esta gráfica muestra cómo evoluciona la pérdida del autoencoder, el cual está encargado de la fidelidad visual luego de la reconstrucción, el cual muestra un buen rendimiento. Puesto que la pérdida de validación desciende brusca y constantemente con valores bajos, pero muy parecidos a los del entrenamiento, terminando en un valor de aproximadamente 0.04.

Pérdida Triplet Semántica (Figura 10.190): Esta gráfica muestra cómo se comporta la pérdida de validación para la separación semántica, es decir, la Inter-Glosario. La cual se puede ver que inicia muy bajo para tener un pico muy alto para luego caer a un valor cercano a 0.0 alrededor de la época 15 para estabilizarse en este punto. Esto quiere decir que hay una separación semántica casi perfecta. Sin embargo, este resultado cambia con las demás graficas, demostrando que resulta en cero no porque haya agrupado las clases, minimizando la distancia Intra-Clase, sino porque separó todo de todo por un margen muy pequeño.

Pérdidas Triplet Temporales (Figura 10.191 y Figura 10.192): Estas gráficas miden la capacidad de diferenciar la secuencia original de sus versiones invertida y permutada. Se puede observar que para ambas pérdidas muestran una tendencia a bajar, sin embargo, es la invertida la que muestra un descenso más suave y pronunciado, a diferencia de la permutada, que presenta un cambio menos pronunciado, aunque constante. El problema reside cuando se comparan con sus rendimientos de entrenamiento, donde la permutada tiene una diferencia muy grande, teniendo un valor mucho más bajo de entrenamiento, de la misma manera, la invertida tiene un mayor valor en su validación, pero en esta ocasión por una diferencia pequeña. Esto indicaría que, aunque las pérdidas están descendiendo, presentan un sobreajuste.

- Evolución del Espacio Latente (PCA y UMAP).

Épocas Iniciales (5-25): En las primeras etapas, el espacio latente está desorganizado, donde las etiquetas y las variaciones están completamente mezcladas, mostrando una fuerte separación basada en el idioma. Como se puede apreciar en las diferentes gráficas, los puntos que representan a cada dataset forman clústeres que, aunque no estén bien definidos, superan la agrupación de las secuencias y sus variaciones.

Épocas Intermedias (45-65): A medida que avanza el entrenamiento, se empieza a observar una ligera sensación de agrupación en las etiquetas, pero los clústeres siguen estando muy dispersos y solapados. Lo mismo se puede aplicar para el caso de los datasets.

Épocas Finales (85-100): Las visualizaciones finales muestran que no hay una separación semántica, esto debido a que los puntos de las etiquetas no forman clústeres distintos, sino que estos están en gran medida superpuestos y distribuidos por todo el espacio. Esto confirma lo que se vio en las gráficas de ratio semántico y de distancias. Sin embargo, a pesar de esto, se puede observar una subestructura temporal que, aunque no es muy consistente, muestra que para las originales, la desplazada está más cerca, mientras que las invertidas y permutadas están más lejos.

- Análisis de los Resultados Numéricos Finales.

Resumen del Entrenamiento (Tabla 8.15): La tabla muestra que para el valor de la Pérdida Total de Validación se tuvo un valor de 13.3133, y para la Pérdida Total de Entrenamiento, uno de 6.5054. El hecho de que la pérdida de entrenamiento sea significativamente más baja que la de validación sugiere un sobreajuste del modelo a los datos de entrenamiento. Por otro lado, la Pérdida Triplet Inter-Glosario de Validación tuvo un valor de 0.0287, este valor es muy bajo, llegando casi a cero. Este valor podría indicar un éxito en un escenario diferente, pero en este caso señala que el modelo no aprendió a agrupar las secuencias, falló en entender el concepto abstracto de cada seña, tratando todas las secuencias como diferentes, resultando en una distancia Intra-Clase alta.

Evaluación de Sensibilidad Temporal (Tabla 8.16): Esta tabla señala que la tarea de aprendizaje temporal fue exitosa, esto porque la distancia L2 cumple con la relación con valores de 5.3562 para la Shifted, 15.0685 para la Inverted y 31.6072 para la Permuted. De igual manera, la distancia Coseno también lo cumple con valores de 0.0105 para la Shifted, 0.0294 para la Inverted y 0.0617 para la Permuted. Este comportamiento confirma que, a pesar del fallo en la agrupación semántica por etiqueta, el aprendizaje de la estructura temporal fue exitoso.

3. Comparación entre experimentos

3. Comparación entre experimentos

3.1. Experimentos con WSL

Al comparar los experimentos de 2 y 3 etiquetas del conjunto de datos WSL, se puede ver que ambos demuestran la capacidad de la arquitectura para aprender representaciones latentes buenas, pero es el experimento con 3 etiquetas el que revela una mayor robustez. En ambos escenarios, el modelo logró una separación semántica perfecta en el conjunto de validación, alcanzando una pérdida triplet inter-glosario de 0.0000. Esto indica que la arquitectura, que combina un autoencoder Conv3D con una Bi-GRU, es altamente eficaz para diferenciar las señas, y esta capacidad no se vio comprometida al aumentar la complejidad de 2 a 3 clases.

Una diferencia que se puede encontrar es la que surge en las métricas de sensibilidad temporal. Si bien ambos modelos aprendieron con éxito la jerarquía temporal deseada, donde las secuencias desplazadas están más cerca de la original que las invertidas o permutadas, el experimento con 3 etiquetas produjo un espacio latente más expandido. Las distancias L2 para todas las variantes temporales son mayores en el experimento de 3 etiquetas en comparación con el de 2 etiquetas. Esto sugiere que forzar al modelo a discriminar entre más clases semánticas también lo hizo a crear diferencias temporales más grandes, ampliando de esta manera el espacio de características para poder acomodar la nueva información sin perder la estructura temporal.

Lo anterior indica que el modelo de 3 etiquetas no solo pudo manejar la complejidad adicional, sino que pareció mejorar con esta misma. Logró una pérdida total de validación final más baja que el modelo de 2 etiquetas. Sugiriendo que entrenar con una mayor variedad semántica puede, ayudar al modelo a generalizar mejor, llevándolo a una convergencia más eficiente en el objetivo de aprendizaje multitarea.

En cuanto al tratamiento de las secuencias, el modelo prioriza claramente la estructura temporal correcta. Esto se puede ver en los resultados numéricos de las tablas que muestran que el modelo aprende a ser invariante a pequeños cambios temporales, al tener la distancia al Positive Shifted mínima, lo cual es muy importante para reconocer señas realizadas a diferentes velocidades. Sin embargo, también penaliza severamente las interrupciones temporales. Porque la distancia L2 es mucho más grande para las variantes Permuted en comparación con las Inverted. Finalmente, esto demuestra que el modelo, impulsado por la capa Bi-GRU, ha aprendido que el orden de los fotogramas es fundamental para la identidad de una señal, mucho más que la simple dirección de reproducción.

3.2. Experimentos con ISL

Al analizar los experimentos del conjunto de datos ISL, se observa una notable diferencia en el rendimiento y comportamiento del modelo al escalar de 2 a 3 etiquetas, especialmente en la forma en que trata la estructura temporal de las secuencias, también se puede observar que hay un leve sobreajuste en ambos experimentos.

En el experimento con 2 etiquetas, si bien no se alcanzó una separación semántica perfecta, sí se pudo lograr un aprendizaje temporal exitoso. Se puede sustentar lo anterior en que los resultados numéricos muestran que el modelo aprendió la jerarquía de distancias deseada tanto para la métrica L2 como para la Coseno. Esto indica que el modelo diferenció correctamente las alteraciones temporales, identificando los desplazamientos leves como los más similares al original y las permutaciones como las más diferentes.

En comparación, el experimento con 3 etiquetas presentó un resultado complejo y contradictorio. Aunque el modelo logró una separación semántica casi perfecta con una pérdida triplet Inter-Glosario de 0.0014, falló de manera crítica en la tarea de sensibilidad temporal. Los valores finales de distancia L2 para las secuencias Shifted e Inverted fueron casi iguales. Esto demuestra que el modelo, si bien aprendió a penalizar el desorden aleatorio, no pudo diferenciar entre una alteración leve y una severa, tratándolas como igualmente diferentes del video original.

Este comportamiento sugiere que, para el dataset ISL, el modelo enfrenta un conflicto entre los objetivos de aprendizaje. Al aumentar la complejidad semántica a 3 etiquetas, el modelo priorizó agresivamente la separación de las clases, logrando una pérdida inter-glosario muy baja, sacrificando la correcta comprensión temporal. El modelo logró su objetivo semántico, pero a costa de perder la capacidad de diferenciar la direccionalidad y el orden secuencial correcto, tratando las secuencias invertidas y las levemente desplazadas como iguales.

3.3. Experimentos con SLOVO

En cuanto a los experimentos con el dataset SLOVO, se puede ver que los resultados son negativos en ambos escenarios. A diferencia de otros conjuntos, el modelo mostró una clara incapacidad para dominar simultáneamente las tareas de separación semántica y sensibilidad temporal, exhibiendo un comportamiento diferente en cada experimento.

El experimento con 2 etiquetas logró un éxito en la sensibilidad temporal, pero fracasó en la separación semántica. Aunque el modelo tuvo dificultades para generalizar la separación de las señas y un sobreajuste, como lo demuestra una alta pérdida Triplet Inter-Glosario de validación y un ratio semántico que no logró superar al baseline no entrenado, sí logró aprender la jerarquía temporal. Los valores numéricos finales cumplieron la relación deseada en las distancias L2 y Coseno, demostrando que el modelo pudo identificar correctamente las permutaciones como la alteración más.

Por el contrario, el experimento con 3 etiquetas falló en gran medida en la sensibilidad temporal. El modelo aprendió una jerarquía temporal incorrecta, donde el modelo

3. Comparación entre experimentos

consideró la inversión como una alteración mucho más drástica que la permutación, un resultado que contradice la hipótesis de la metodología. Por lo tanto, sugiere que el modelo no aprendió la estructura subyacente del movimiento, sino que encontró un atajo en el aprendizaje. Además, el éxito semántico es contradictorio, porque si bien la pérdida Inter-Glosario de validación cayó casi a cero, indicando una separación triplet perfecta, esto no se reflejó en una verdadera organización del espacio latente. Donde el ratio semántico general se mantuvo por debajo del baseline, y las visualizaciones finales mostraron que los clústeres permanecían superpuestos. Finalmente, esto indica que el modelo minimizó la pérdida sin lograr el objetivo de agrupación, empujando los vectores lejos radialmente en lugar de organizarlos de forma estructural.

3.4. Experimentos con los tres datasets

Al comparar los experimentos de 2 y 3 etiquetas, se puede ver que aunque el modelo demuestra una notable capacidad para comprender la estructura temporal de las secuencias de señas, falla al tener que hacer la separación semántica. Este problema se pone peor con el aumento de la complejidad. Se puede observar que el experimento con 3 etiquetas muestra un sobreajuste mayor que el de 2 etiquetas, mientras que el modelo de 2 etiquetas tiene una diferencia de aproximadamente 2.3 puntos entre las pérdidas, el modelo de 3 etiquetas muestra una mucho más amplia de casi 6.8 puntos. Esto muestra que, aunque el modelo puede ajustarse más a los datos de entrenamiento con más etiquetas con menor pérdida de entrenamiento, pierde en gran medida la capacidad de generalización.

Ambos experimentos fallan en su objetivo principal de agrupar las señas por su significado. Porque el modelo de 2 etiquetas logra un ratio semántico apenas superior del baseline aleatorio, mientras que el rendimiento del modelo de 3 etiquetas es inestable y cae por debajo del baseline aleatorio durante gran parte del entrenamiento. En ambos casos, la pérdida semántica Inter-Glosario en validación es casi cero, sin embargo, este valor es engañoso, pues no indica una agrupación exitosa, sino un fallo en el que las distancias dentro de una misma clase son tan grandes como las distancias entre clases diferentes. El modelo no aprende el concepto abstracto de la señal, sino que está influenciado por variaciones entre datasets.

A pesar de este gran fallo semántico, el éxito de los experimentos radica en cómo el modelo trata la estructura temporal de las secuencias. Siendo este un objetivo clave del diseño de la arquitectura, que utiliza una combinación de Conv3D para el espacio y Bi-GRU para el tiempo, entrenada con pérdidas triplet temporales para aprender exitosamente la jerarquía de disruptión temporal. Las métricas de evaluación de sensibilidad temporal, muestran que en el experimento de 2 etiquetas, las distancias L2 siguen la regla con valores de 4.0 7.6 61.8, y la separación en el de 3 etiquetas es mejor, con valores de 5.3 15.0 31.6, tratando las secuencias no como un conjunto de fotogramas, sino como una dinámica ordenada.

Capítulo 4: Conclusiones y trabajos futuros

1. Conclusiones

Los experimentos realizados en esta investigación demuestran que la técnica propuesta para la creación de un espacio latente estructurado es viable y aporta valor significativo en el contexto del procesamiento de lenguaje de señas. Esto debido a que uno de los hallazgos más importantes fue la capacidad del modelo para aprender y codificar el orden temporal de las secuencias de video. Esto se evidenció en casi todos los experimentos, a excepción del experimento con ISL y SLOVO usando tres etiquetas, donde se cumplió la relación esperada de las distancias Coseno y L2. En estos mismos casos, el modelo diferenció exitosamente entre las secuencias originales y sus variantes, manteniendo la distancia de la variante desplazada como la más baja y la permutada como la más alta. Este comportamiento indica que el modelo no procesa los videos como fotogramas solos, sino que puede comprender la dinámica temporal, lo cual es muy importante para ver la importancia de esta nueva técnica. Además, en todos los casos, el modelo propuesto logró superar el baseline de PCA y mejorar de gran manera la pérdida del modelo simple, justificando su mayor complejidad en la arquitectura.

Así mismo, el análisis comparativo de los datasets reveló el impacto crítico que tienen las características de los datos en los resultados del modelo. Porque los experimentos ejecutados con el dataset WSL obtuvieron mejores resultados, a pesar de que este conjunto posee un menor número de videos en comparación con ISL y SLOVO. La razón de este comportamiento radica en la calidad y uniformidad de WSL, el cual proviene de fuentes educativas y fue estrictamente filtrado para palabras individuales. Por el contrario, los datasets ISL y SLOVO, si bien más grandes en volumen, presentan bastante variabilidad con múltiples intérpretes, condiciones de grabación improvisadas y fondos diversos. Esta heterogeneidad provocó que el modelo, en lugar de aprender las características semánticas de las señas, comenzara a sobreajustarse. Esto mismo ayuda a mostrar que el sobreajuste no siempre se debe a la falta de datos, en este caso, más datos con ruido y alta variabilidad impidieron una convergencia exitosa hacia la separación semántica.

De igual manera, se pudo evidenciar la importancia de la creación de más sets de datos en el contexto del lenguaje de señas, y más aún los que hacen énfasis en palabras

2. Trabajos futuros y recomendaciones

y no simples deletreos, ya que la mayoría de los datasets encontrados son pequeños y no cuentan con una gran cantidad de videos por palabra o directamente no existen, lo cual limita el desarrollo de modelos más robustos y complejos.

Por otro lado, este proyecto de investigación también busca animar a los estudiantes de la Universidad Sergio Arboleda, para demostrarles que no se necesita de una solución definitiva, los mejores recursos computacionales o invertir mucho dinero, para hacer aportes valiosos que puedan contribuir con la construcción de un mundo mejor. Si no más bien, de una investigación rigurosa, honesta y fundamental, que por medio de la excelencia académica que caracteriza a los Sergistas, se puedan hacer aportes más significativos, sentando las piedras angulares como un puente hacia un futuro donde grandes soluciones, iniciando con este tipo de aportes, permitan que todos tengan las mismas oportunidades, así se tenga una discapacidad auditiva.

2. Trabajos futuros y recomendaciones

El poder demostrar un aprendizaje temporal mediante las métricas de distancia y la creación exitosa de un espacio latente estructurado sienta una base sólida para futuras investigaciones. La metodología que se propone en esta investigación, la cual combina una arquitectura espacial, con Conv3D, temporal, con Bi-GRU, y una función de pérdida triplet diseñada para la dinámica de video, ha demostrado ser un punto de partida valioso. Como abstracción positiva para la investigación científica, este enfoque puede extenderse a otros dominios donde el orden secuencial de la información es semánticamente crítico, más allá del lenguaje de señas.

Para poder mejorar teóricamente los resultados del modelo actual, se proponen varias líneas de acción. En primera instancia, se debe realizar algo para evitar variabilidad y mala calidad de los datos, pues aunque aumentar la cantidad de datos es una estrategia común, los hallazgos muestran que priorizar la calidad de los mismos podría ser más efectivo. Un conjunto de datos con condiciones controladas minimizaría la necesidad de un preprocesamiento complejo y reduciría el ruido, facilitando al modelo la extracción de características semánticas relevantes.

Por otro lado, si se quiere construir un modelo robusto capaz de generalizar en entornos variables, es buena idea mantener la diversidad de los datos, pero la estrategia de entrenamiento debe cambiar. Sería necesario un entrenamiento más prolongado, aumentando significativamente las épocas, dimensiones latentes, entre otros parámetros, para permitir que el modelo aprenda a separar de mejor manera las secuencias. Asimismo, se podrían explorar modificaciones en la arquitectura, como la incorporación de mecanismos de atención, para ayudar al modelo a ponderar la importancia de las características espaciales y temporales y a ignorar el ruido. También se puede pensar

en ignorar la pérdida de reconstrucción, puesto que no interesa en gran medida esta característica, haciendo que el modelo se concentre puramente en la separación de las secuencias. Sin embargo, sí se recomienda que al menos las cámaras con las que se toman los videos sean las mismas, pues al preprocesar y redimensionar los datos, las secuencias presentarán deformaciones que pueden confundir al modelo.

Finalmente, se recomienda altamente desarrollar esta técnica con datasets de mayor calidad, que sean más grandes, limpios y estén enfocados en palabras. Esto permitiría explorar verdaderamente los límites de la técnica y su capacidad para escalar a vocabularios más complejos de otros lenguajes.

Anexos

1. Gráficas de los Experimentos Realizados

1.1. Experimentos del dataset de WLSL

Con 2 etiquetas

Figura 10.1: Esta gráfica compara el ratio Semántico del modelo comparado con un «baseline» o punto de referencia. Siendo este el modelo sin entrenar. El eje Y representa el ratio semántico, que mide la capacidad del modelo para diferenciar entre palabras diferentes, mientras que el eje X representa las épocas de entrenamiento.

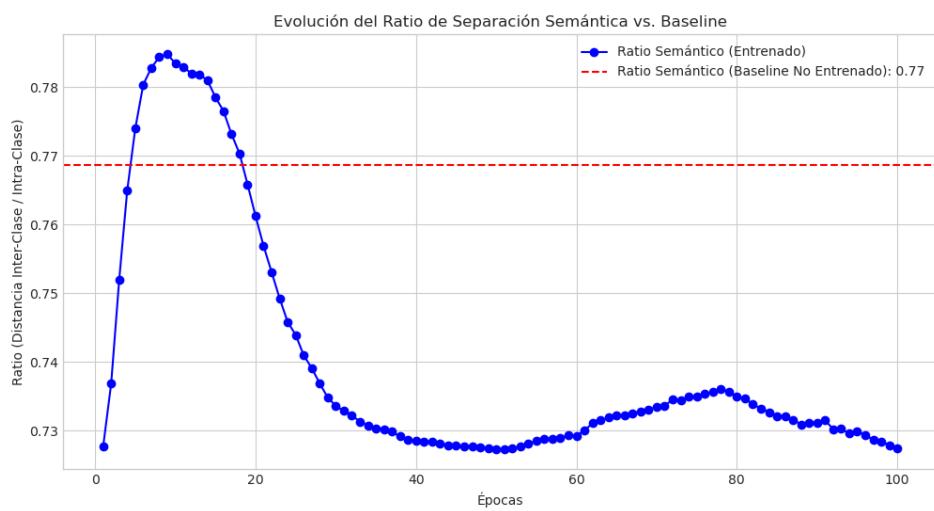


Figura 10.2: Esta gráfica evalúa la capacidad del modelo para entender el orden temporal de las secuencias de video comparadas con sus respectivos «baselines» del modelo no entrenado. Muestra la distancia euclíadiana promedio entre la secuencia original y sus versiones alteradas (shifted, inverted, permuted).

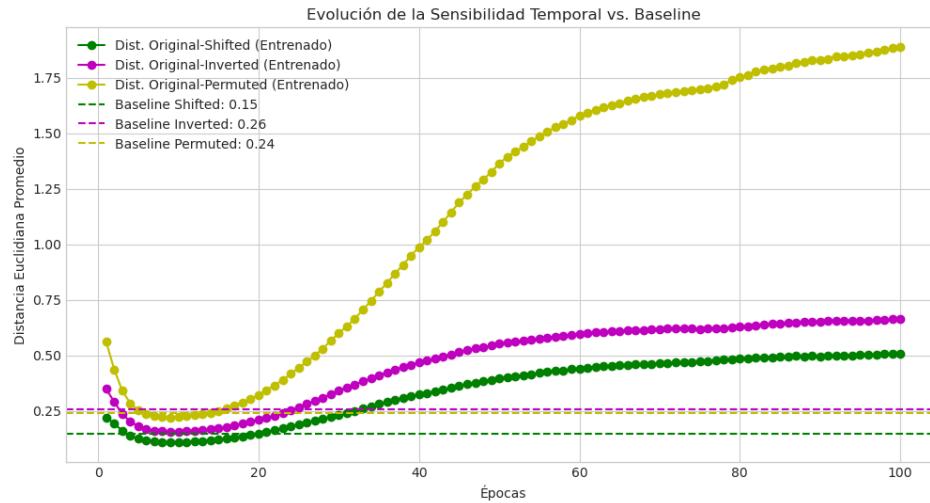
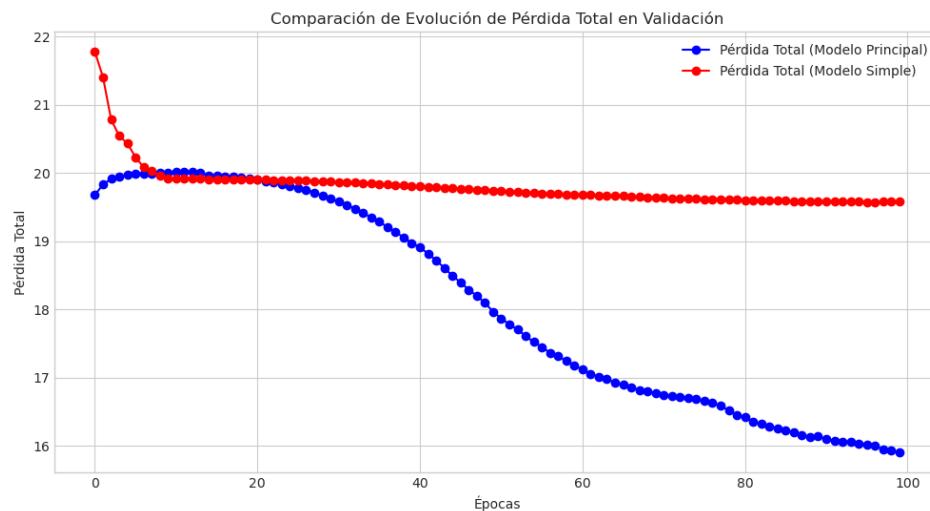


Figura 10.3: Esta gráfica compara el rendimiento del modelo principal con un modelo más simple en términos de la pérdida total de validación. El eje Y representa el valor de la pérdida, una métrica que indica cuán bien el modelo está aprendiendo, donde los valores más bajos son mejores, y el eje X representa las épocas.



1. Gráficas de los Experimentos Realizados

Figura 10.4: Esta gráfica compara el ratio semántico del modelo comparado con dos «baselines» o puntos de referencia. Siendo estos el modelo sin entrenar y la representación de PCA. El eje Y representa el ratio semántico, que mide la capacidad del modelo para diferenciar entre palabras diferentes, mientras que el eje X representa las épocas de entrenamiento.

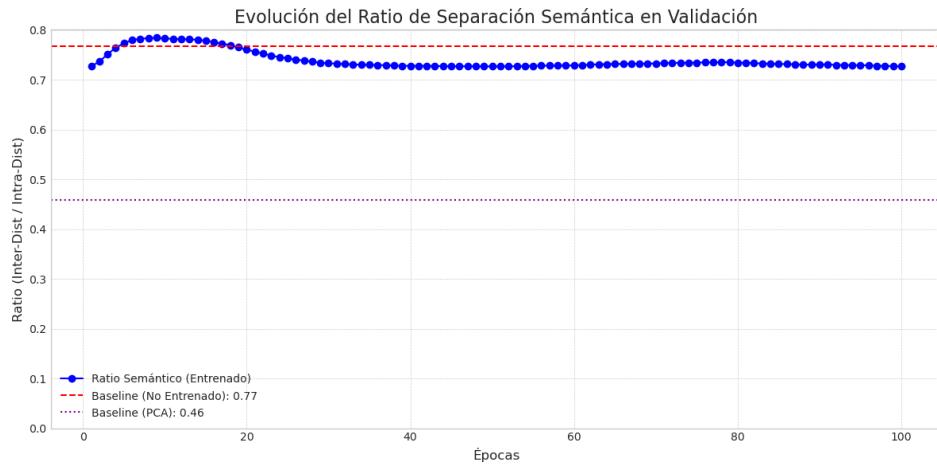


Figura 10.5: Este gráfico mide directamente la calidad de la separación semántica en el espacio latente para palabras con la misma y diferente clase. El eje Y representa la distancia euclídea promedio y el eje X son las épocas.

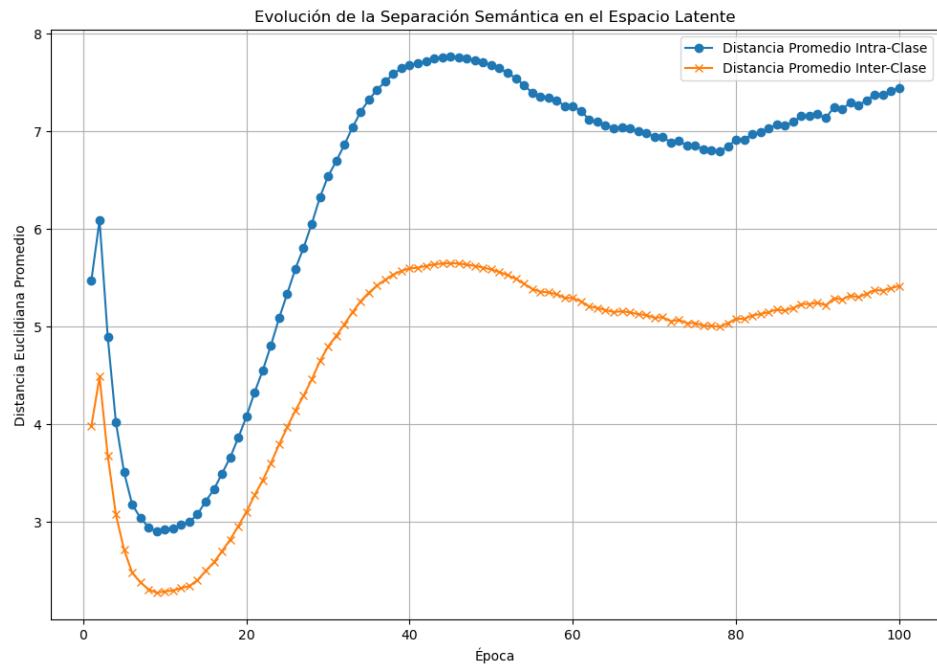
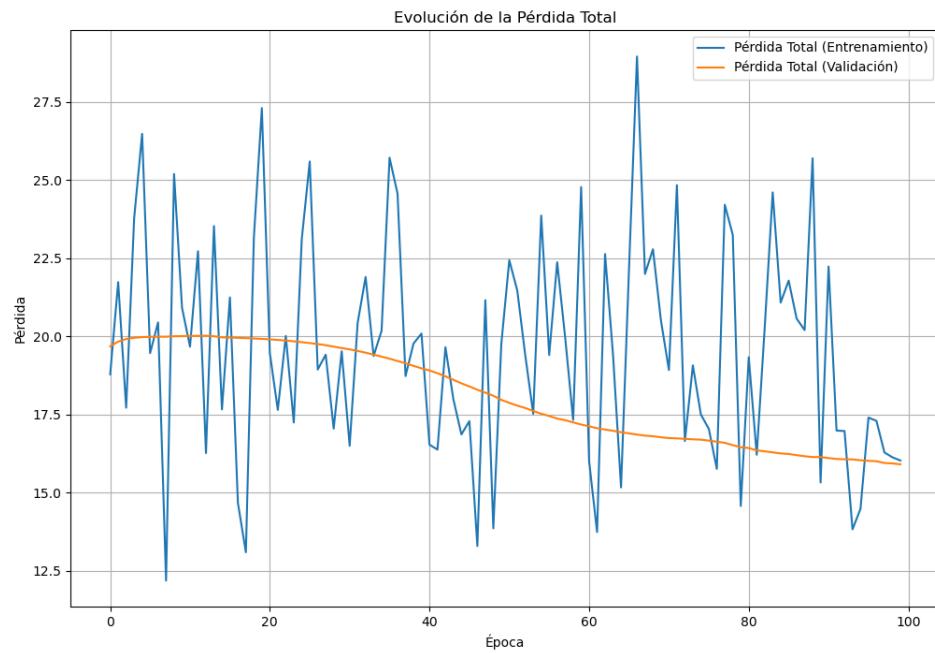


Figura 10.6: Este gráfico muestra la evolución de la «Pérdida Total» a lo largo de 100 épocas de entrenamiento. El eje Y representa el valor de la pérdida, una métrica que indica cuán bien el modelo está aprendiendo donde valores más bajos son mejores. El eje X representa las épocas, es decir, cada ciclo completo de entrenamiento sobre el conjunto de datos.



1. Gráficas de los Experimentos Realizados

Figura 10.7: Esta gráfica ilustra la «Pérdida de Reconstrucción», que mide qué tan bien el autoencoder del modelo puede reconstruir la entrada original después de haberla comprimido en un espacio latente. Al igual que en la gráfica anterior, el eje Y es el valor de la pérdida y el eje X son las épocas.

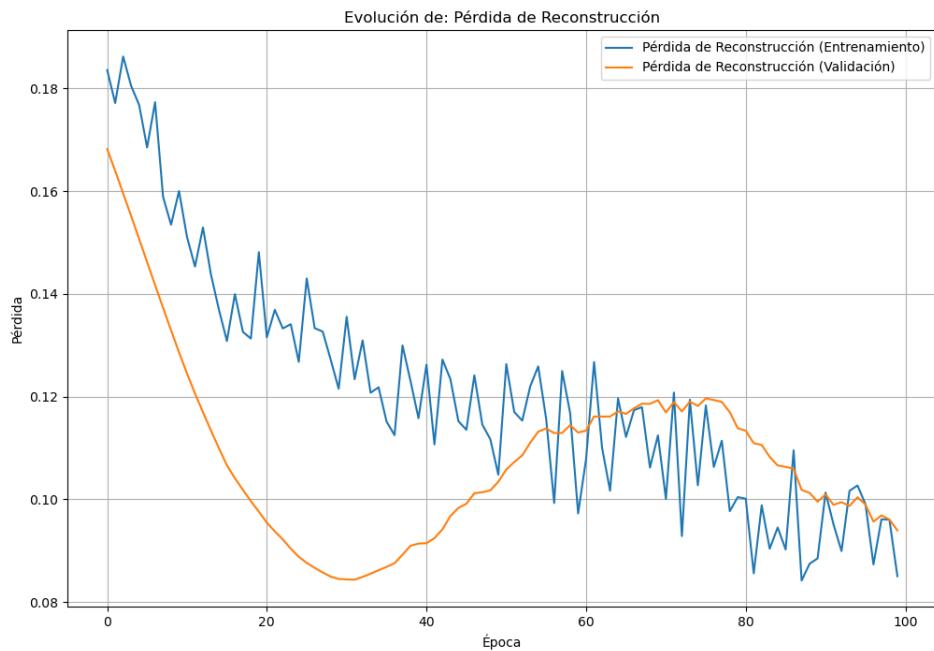
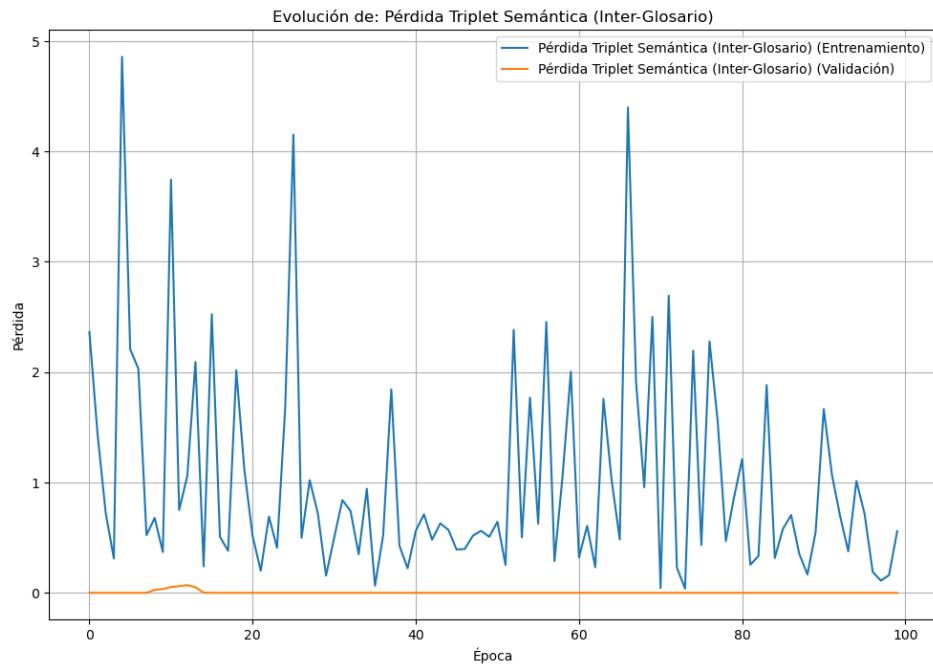


Figura 10.8: Este gráfico muestra la «Pérdida Triplet Semántica», una métrica clave que evalúa si el modelo puede diferenciar entre distintos glosarios, en este caso, las señas «brother» y «cold». El objetivo es que las representaciones de un mismo glosario estén más cerca entre sí que las de glosarios diferentes. Igualmente, el eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.



1. Gráficas de los Experimentos Realizados

Figura 10.9: Esta visualización se enfoca en la sensibilidad temporal del modelo, puesto que, mide la diferencia entre la secuencia original de un video y su versión invertida. Esto quiere decir que el modelo debe aprender que una secuencia invertida es significativamente diferente de la original. El eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.

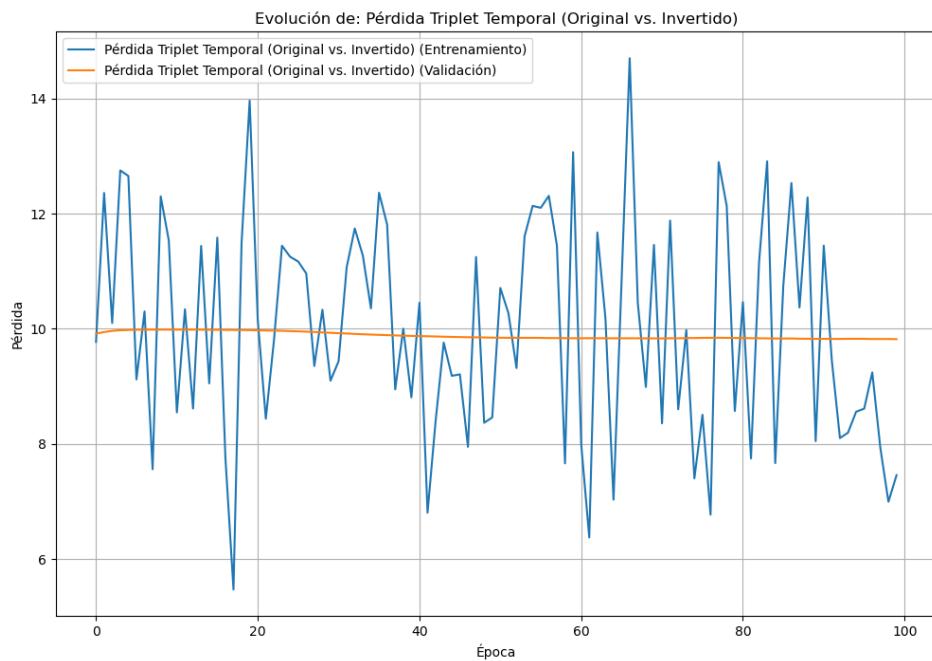
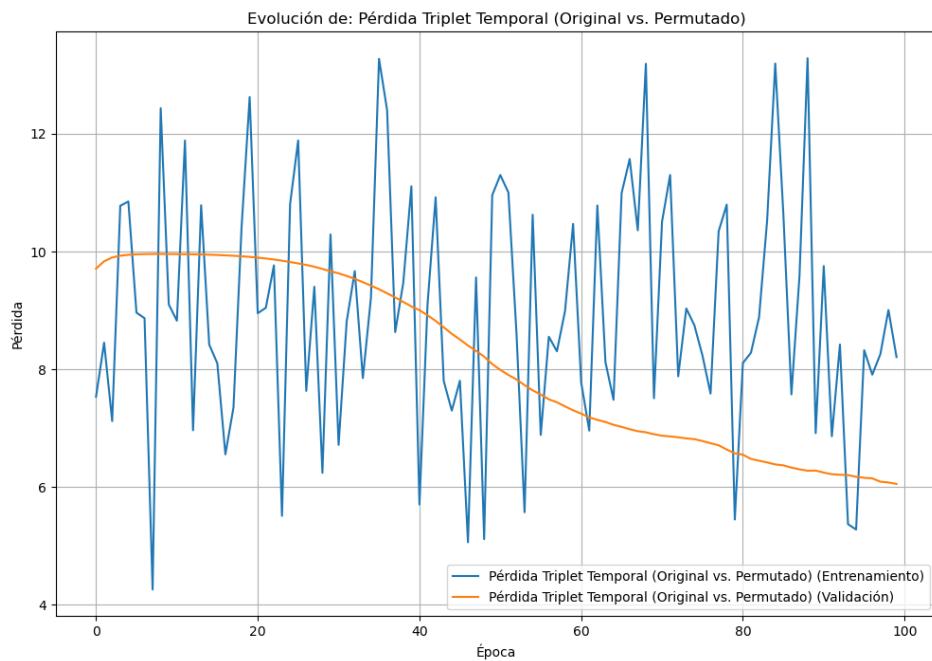


Figura 10.10: Similar a la gráfica anterior, esta también evalúa la sensibilidad temporal, pero en este caso, compara la secuencia original con una versión donde los fotogramas han sido desordenados aleatoriamente. Donde el objetivo es que el modelo reconozca que una secuencia permutada es muy diferente de la original. El eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.



1. Gráficas de los Experimentos Realizados

Figura 10.11: Esta grafica muestra el espacio latente en la mejor epoca (100) utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

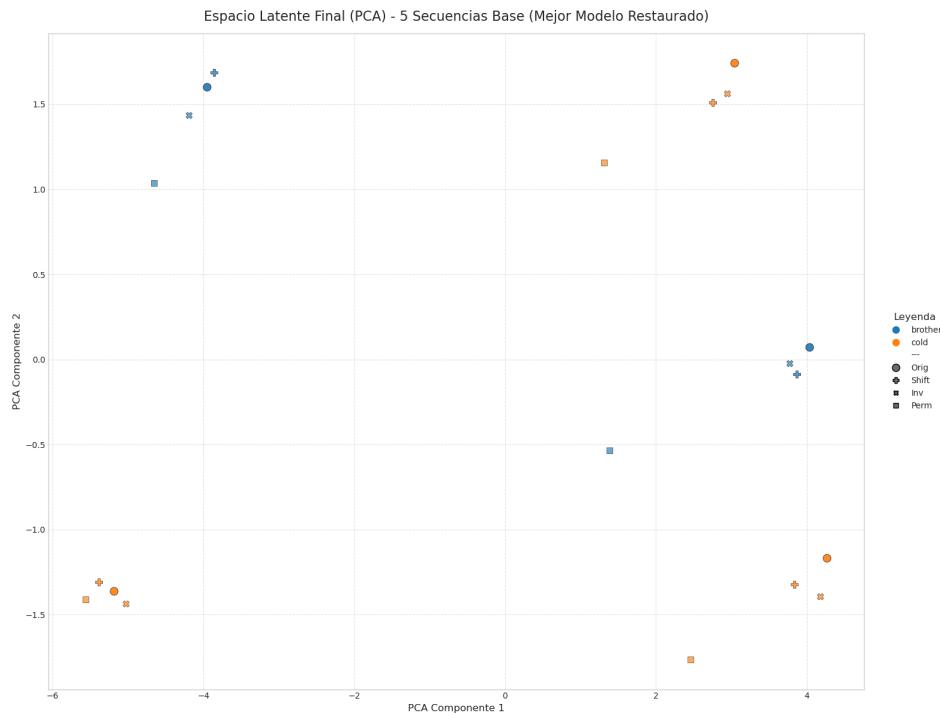
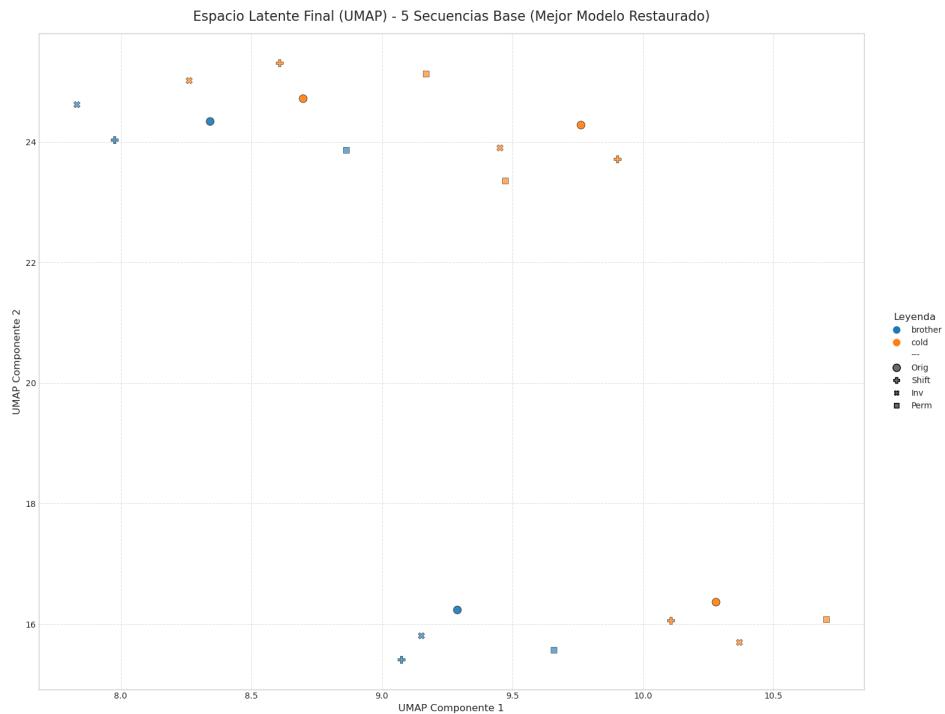


Figura 10.12: Esta grafica muestra el espacio latente en la mejor epoca (100) utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.13: Esta grafica muestra el espacio latente en la epoca 5 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

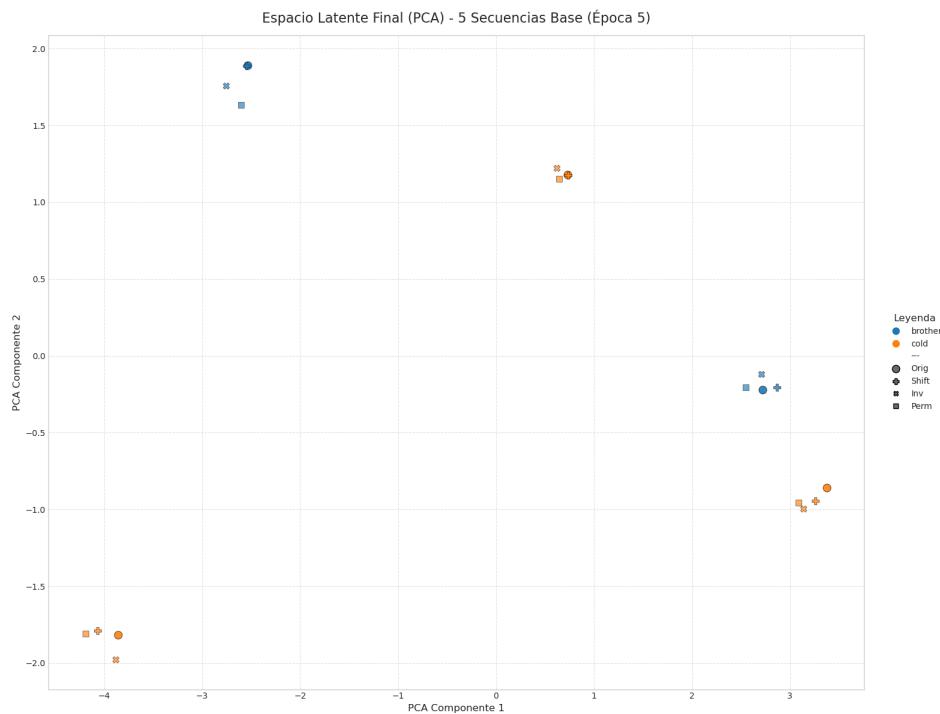
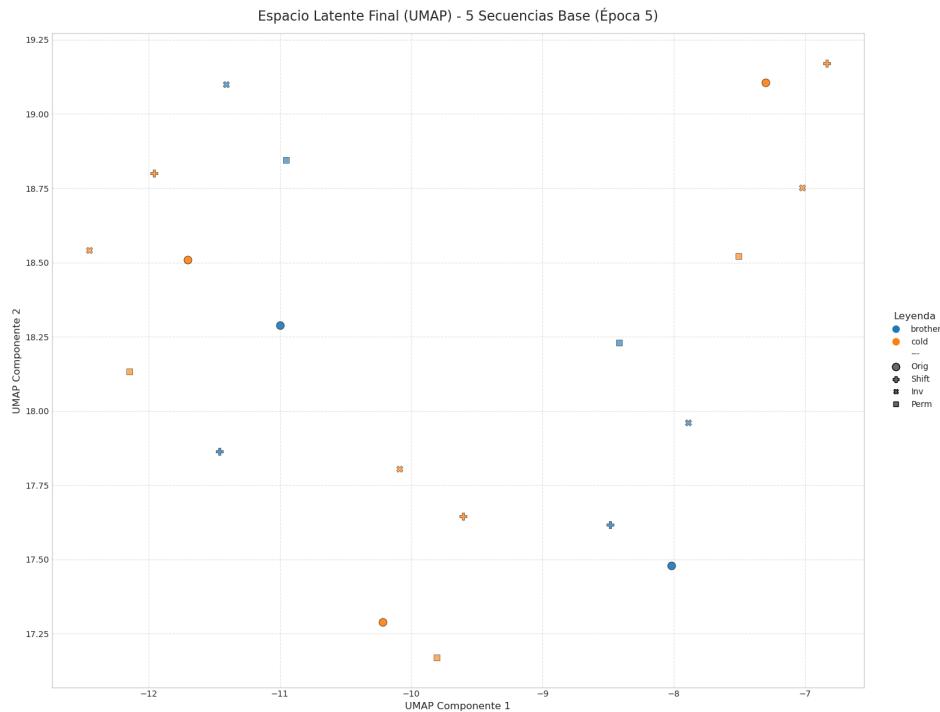


Figura 10.14: Esta grafica muestra el espacio latente en la epoca 5 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.15: Esta grafica muestra el espacio latente en la epoca 15 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

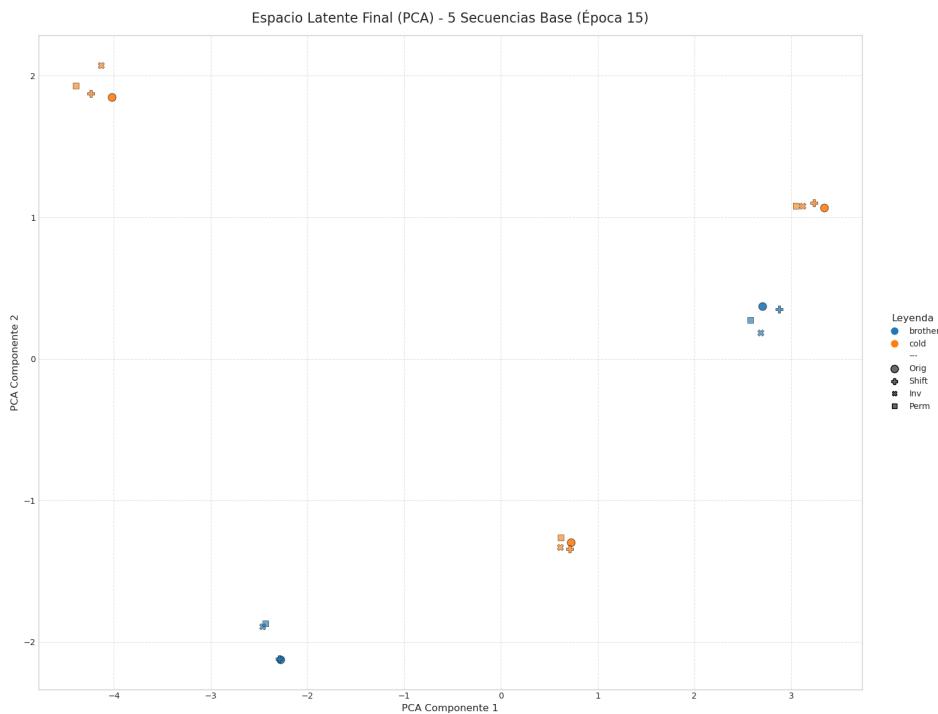
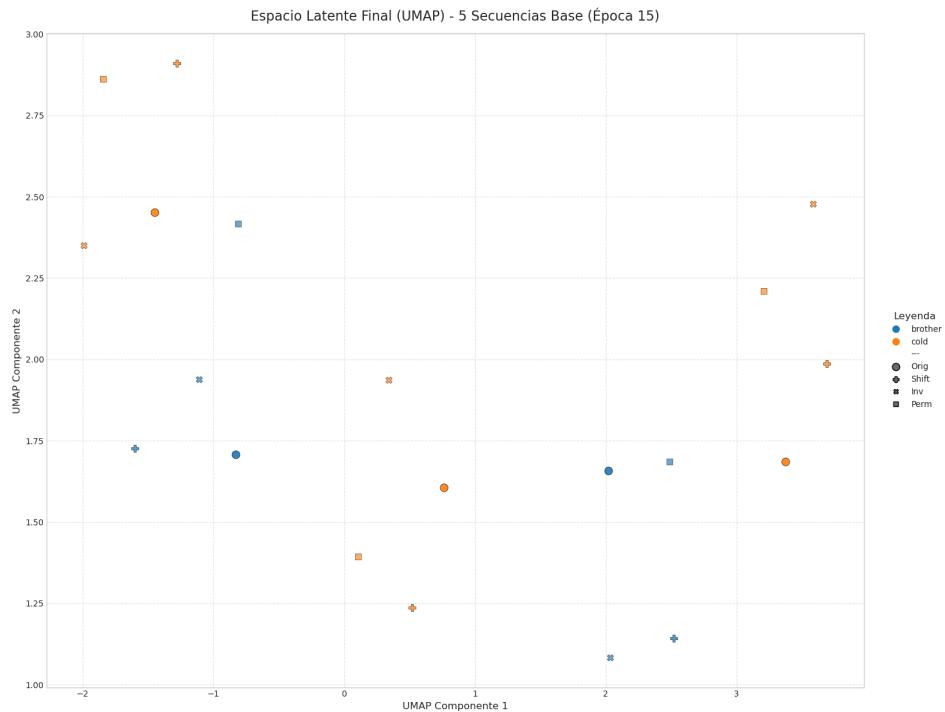


Figura 10.16: Esta grafica muestra el espacio latente en la epoca 15 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.17: Esta grafica muestra el espacio latente en la epoca 25 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

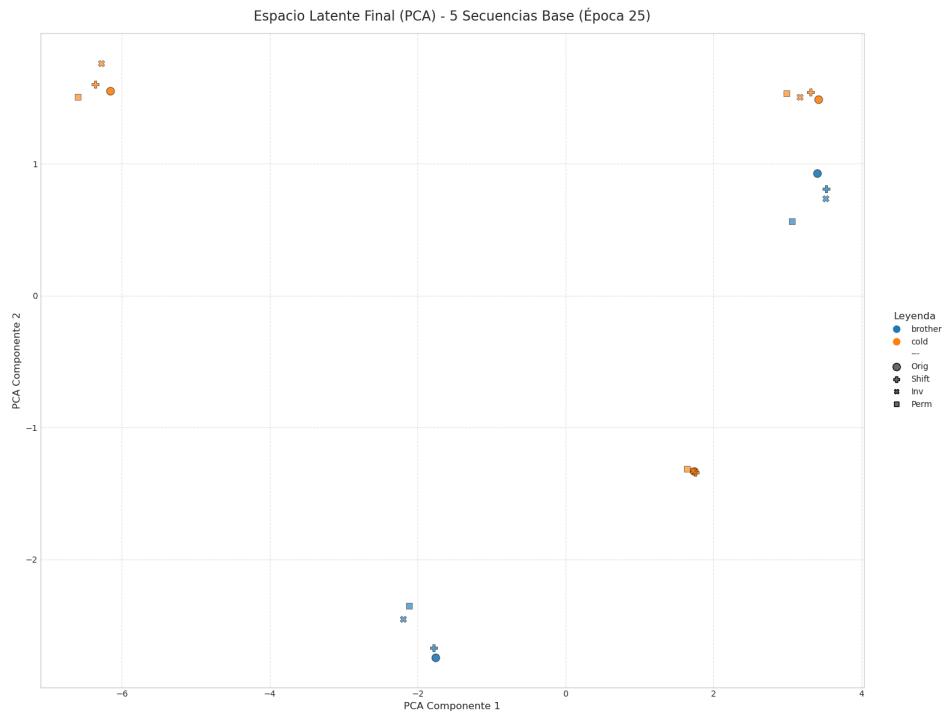
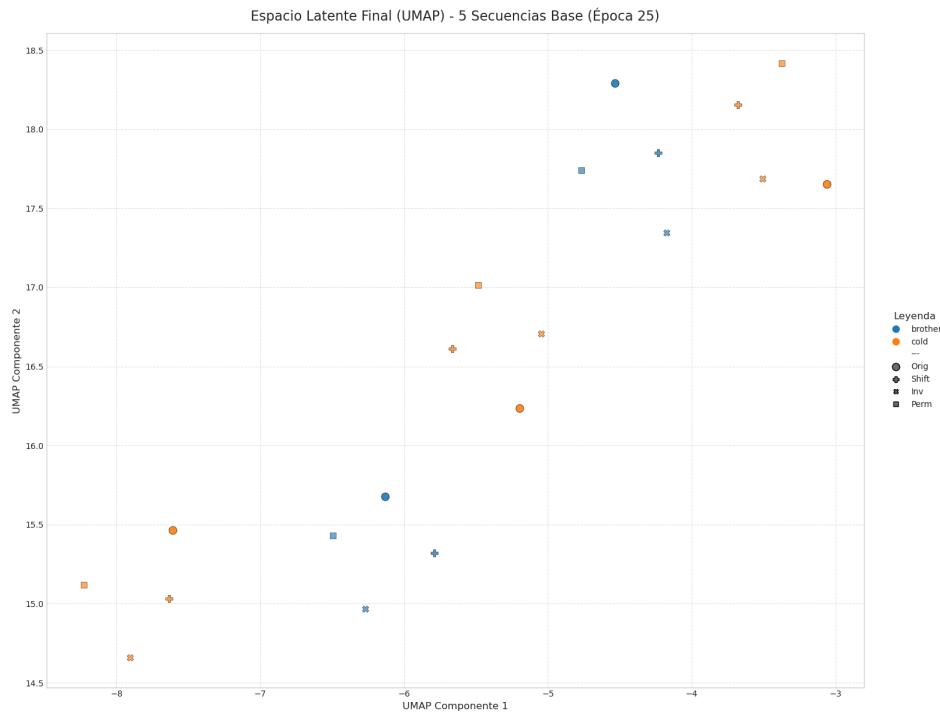


Figura 10.18: Esta grafica muestra el espacio latente en la epoca 25 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.19: Esta grafica muestra el espacio latente en la epoca 45 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

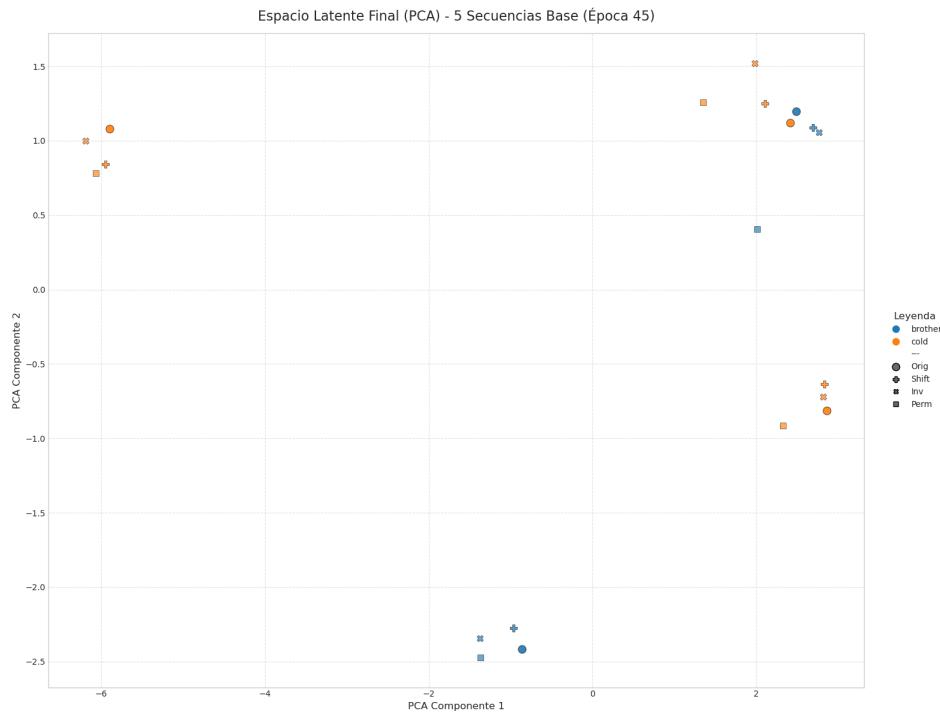
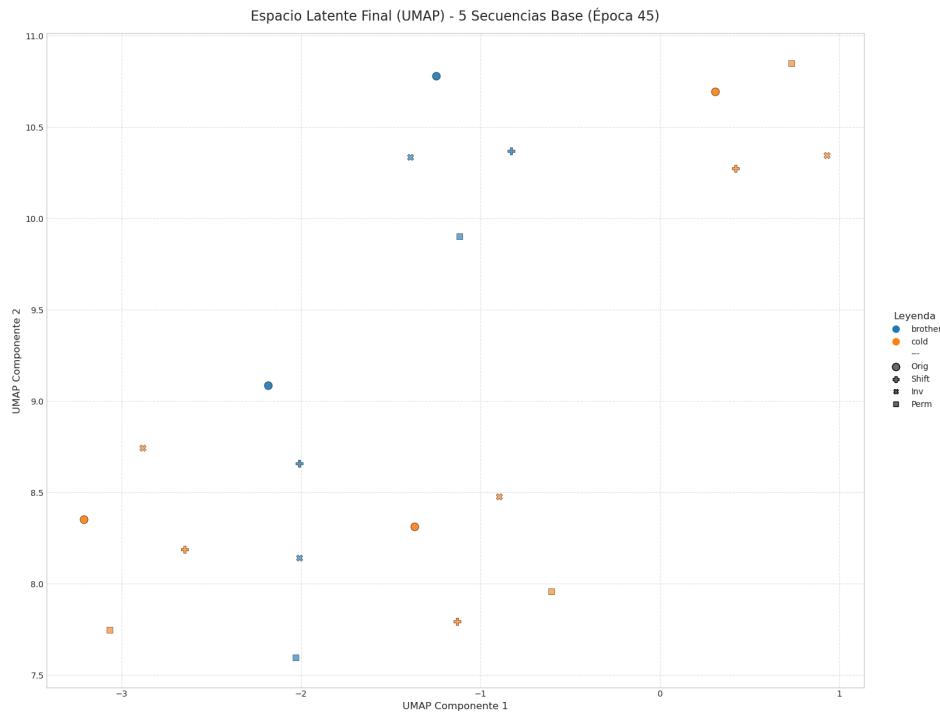


Figura 10.20: Esta grafica muestra el espacio latente en la epoca 45 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.21: Esta grafica muestra el espacio latente en la epoca 65 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

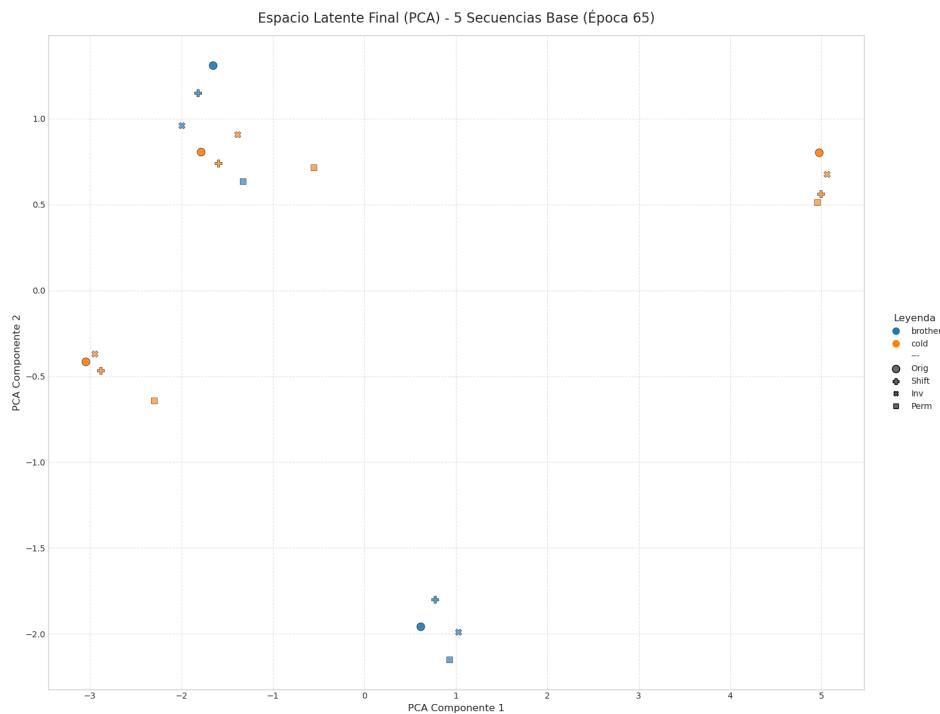
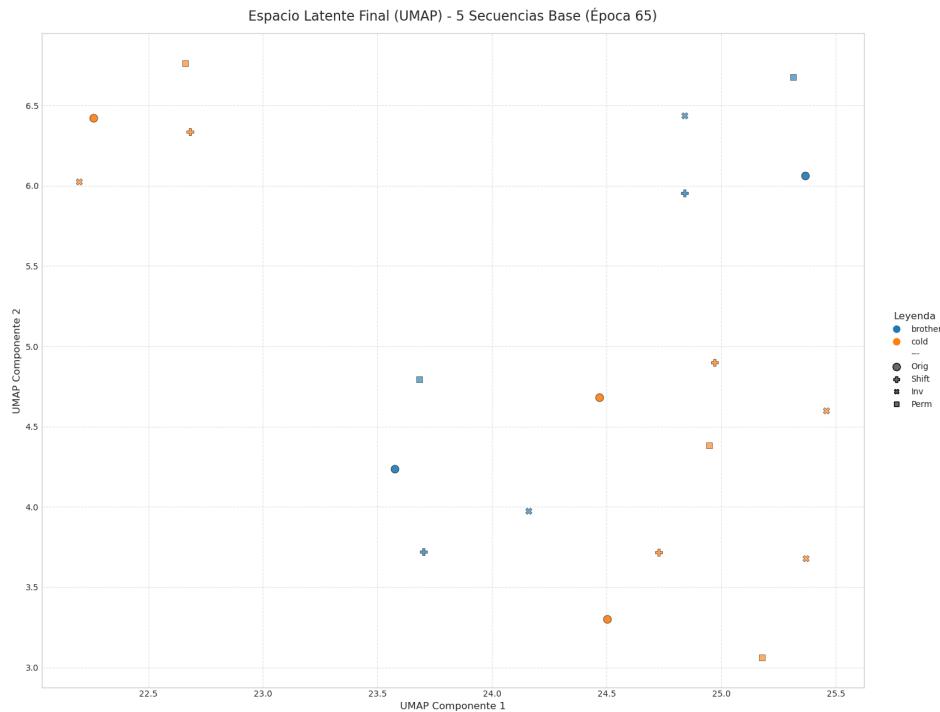


Figura 10.22: Esta grafica muestra el espacio latente en la epoca 65 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.23: Esta grafica muestra el espacio latente en la epoca 85 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

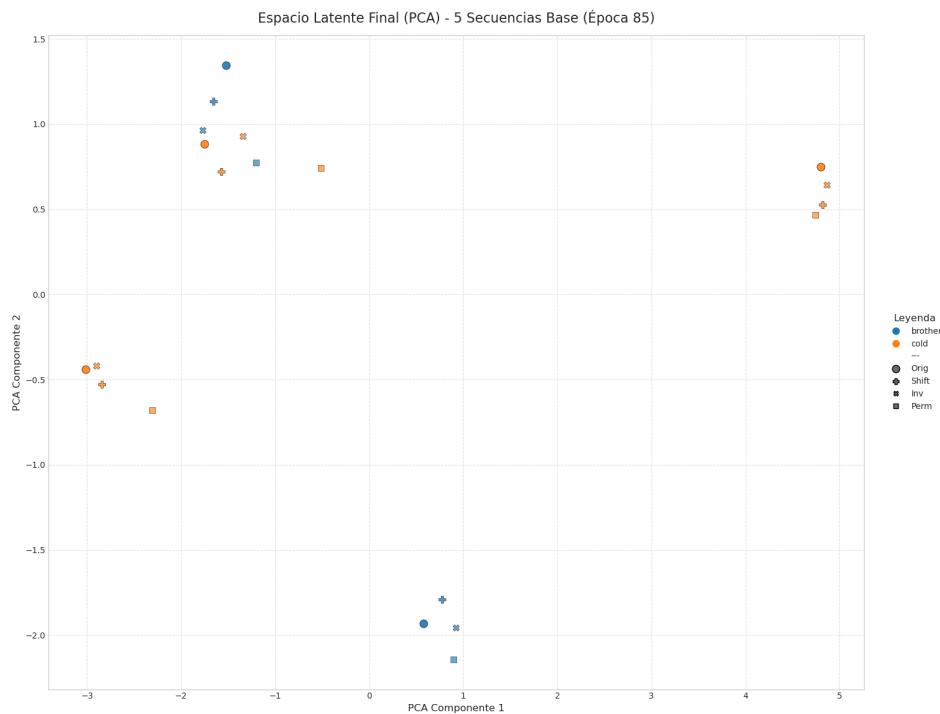
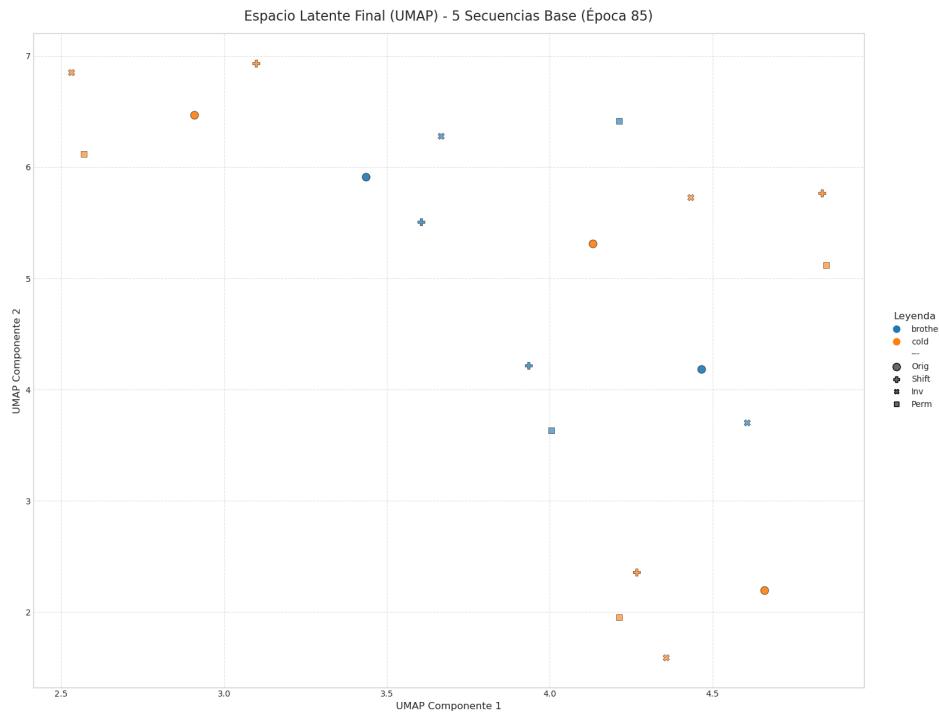


Figura 10.24: Esta grafica muestra el espacio latente en la epoca 85 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Con 3 etiquetas

Figura 10.25: Esta gráfica compara el ratio Semántico del modelo comparado con un «baseline» o punto de referencia. Siendo este el modelo sin entrenar. El eje Y representa el ratio semántico, que mide la capacidad del modelo para diferenciar entre palabras diferentes, mientras que el eje X representa las épocas de entrenamiento.

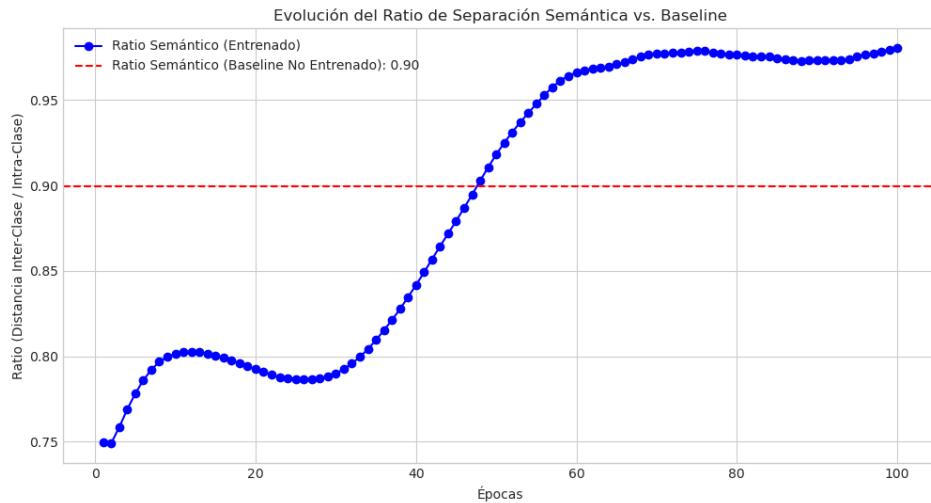


Figura 10.26: Esta gráfica evalúa la capacidad del modelo para entender el orden temporal de las secuencias de video comparadas con sus respectivos «baselines» del modelo no entrenado. Muestra la distancia euclídea promedio entre la secuencia original y sus versiones alteradas (shifted, inverted, permuted).

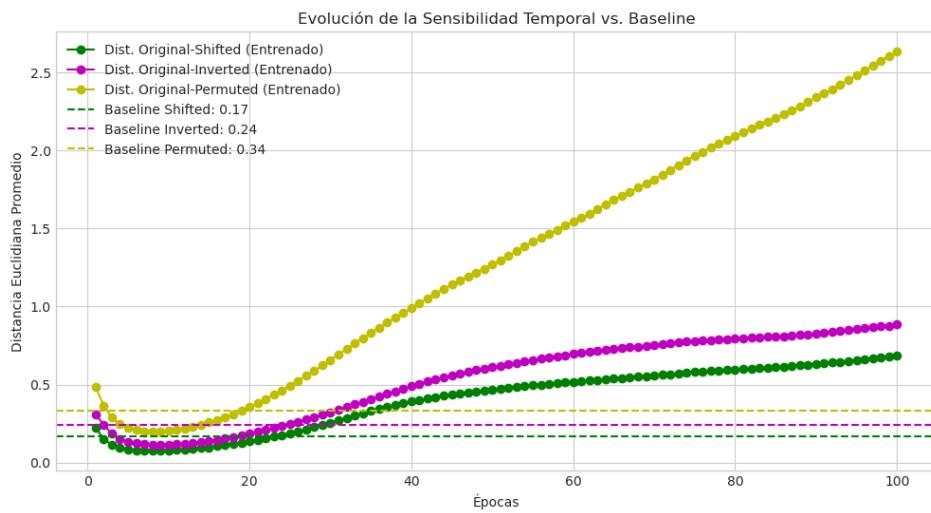


Figura 10.27: Esta gráfica compara el rendimiento del modelo principal con un modelo más simple en términos de la pérdida total de validación. El eje Y representa el valor de la pérdida, una métrica que indica cuán bien el modelo está aprendiendo, donde los valores más bajos son mejores, y el eje X representa las épocas.

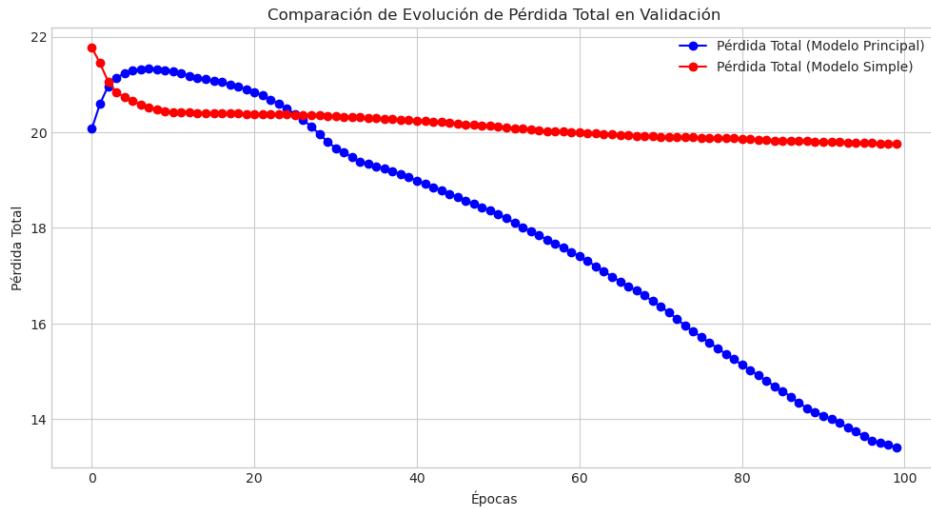
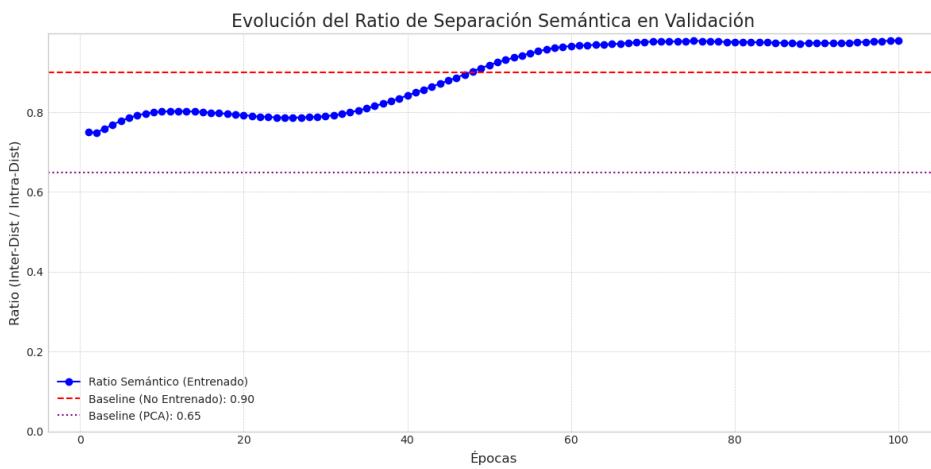
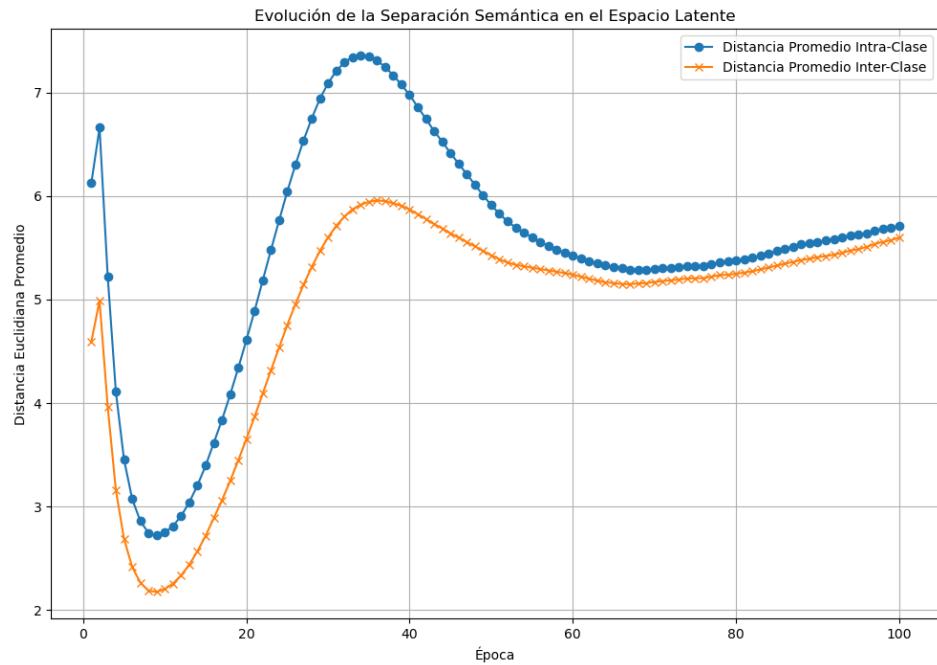


Figura 10.28: Esta gráfica compara el ratio semántico del modelo comparado con dos «baselines» o puntos de referencia. Siendo estos el modelo sin entrenar y la representación de PCA. El eje Y representa el ratio semántico, que mide la capacidad del modelo para diferenciar entre palabras diferentes, mientras que el eje X representa las épocas de entrenamiento.



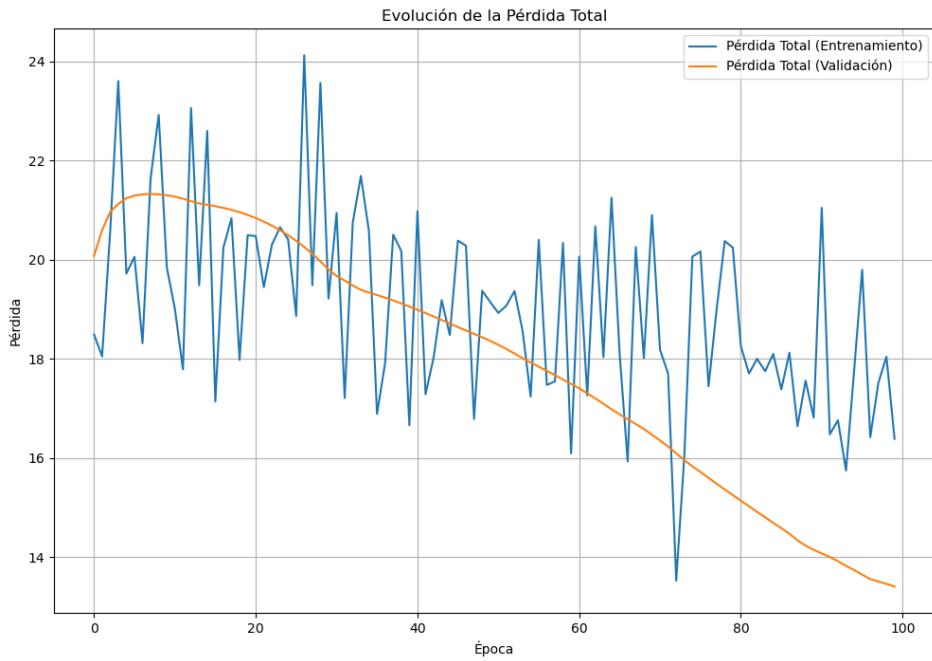
1. Gráficas de los Experimentos Realizados

Figura 10.29: Este gráfico mide directamente la calidad de la separación semántica en el espacio latente para palabras con la misma y diferente clase. El eje Y representa la distancia euclídea promedio y el eje X son las épocas.



UNIVERSIDAD
SERGIO ARBOLEDA

Figura 10.30: Este gráfico muestra la evolución de la «Pérdida Total» a lo largo de 100 épocas de entrenamiento. El eje Y representa el valor de la pérdida, una métrica que indica cuán bien el modelo está aprendiendo donde valores más bajos son mejores. El eje X representa las épocas, es decir, cada ciclo completo de entrenamiento sobre el conjunto de datos.



1. Gráficas de los Experimentos Realizados

Figura 10.31: Esta gráfica ilustra la «Pérdida de Reconstrucción», que mide qué tan bien el autoencoder del modelo puede reconstruir la entrada original después de haberla comprimido en un espacio latente. Al igual que en la gráfica anterior, el eje Y es el valor de la pérdida y el eje X son las épocas.

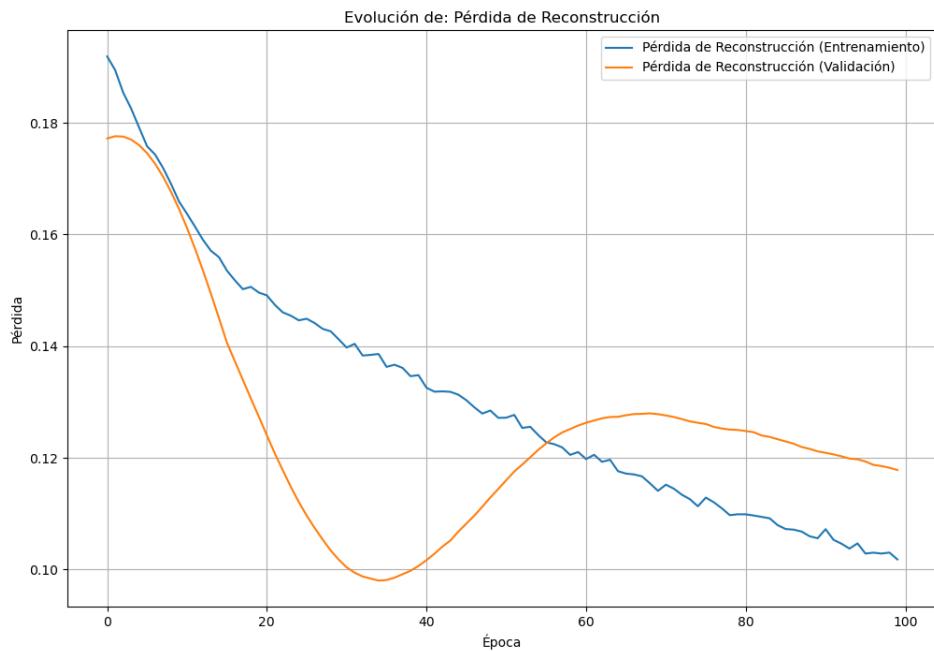
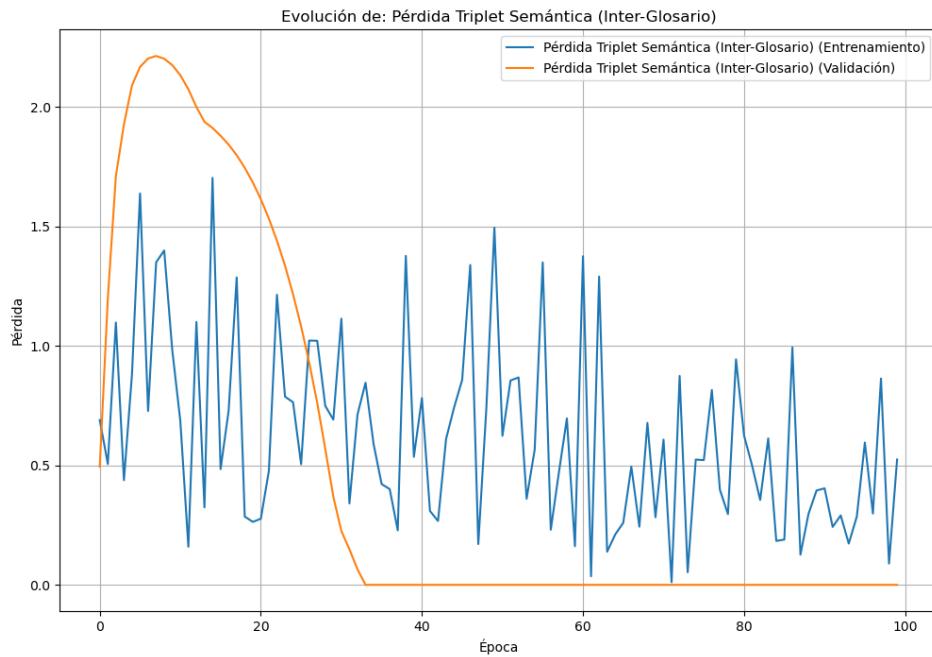


Figura 10.32: Este gráfico muestra la «Pérdida Triplet Semántica», una métrica clave que evalúa si el modelo puede diferenciar entre distintos glosarios, en este caso, las señas «brother», «cold» y «man». El objetivo es que las representaciones de un mismo glosario estén más cerca entre sí que las de glosarios diferentes. Igualmente, el eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.



1. Gráficas de los Experimentos Realizados

Figura 10.33: Esta visualización se enfoca en la sensibilidad temporal del modelo, puesto que, mide la diferencia entre la secuencia original de un video y su versión invertida. Esto quiere decir que el modelo debe aprender que una secuencia invertida es significativamente diferente de la original. El eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.

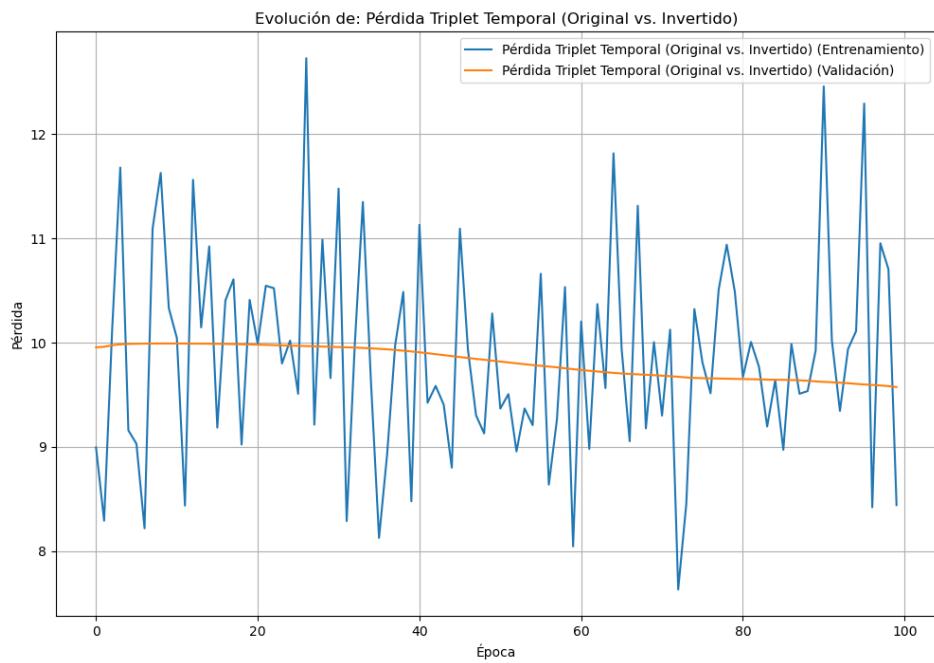
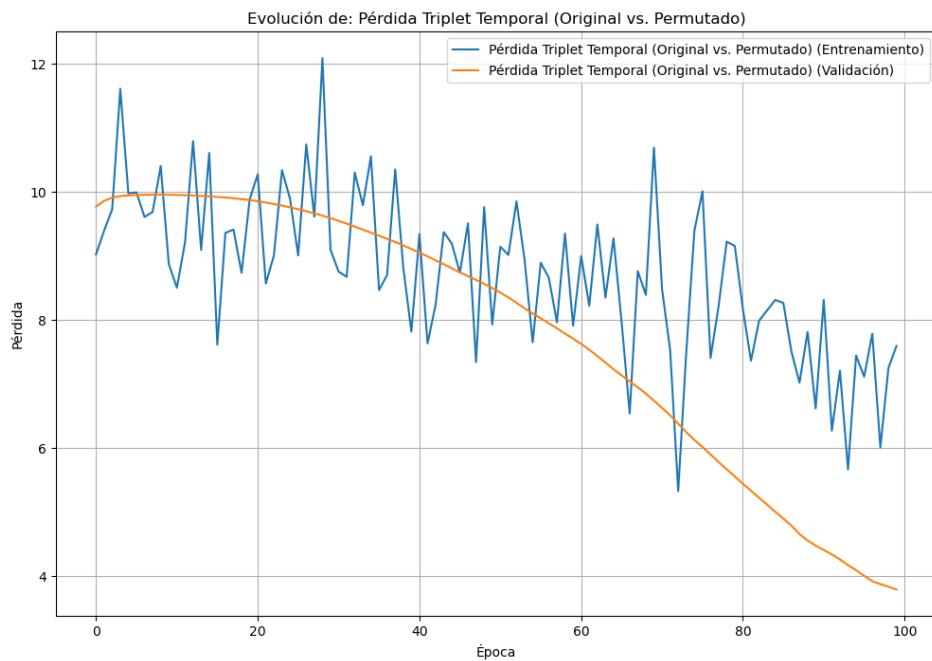


Figura 10.34: Similar a la gráfica anterior, esta también evalúa la sensibilidad temporal, pero en este caso, compara la secuencia original con una versión donde los fotogramas han sido desordenados aleatoriamente. Donde el objetivo es que el modelo reconozca que una secuencia permutada es muy diferente de la original. El eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.



1. Gráficas de los Experimentos Realizados

Figura 10.35: Esta grafica muestra el espacio latente en la mejor epoca (100) utilizando pca, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis invertida.

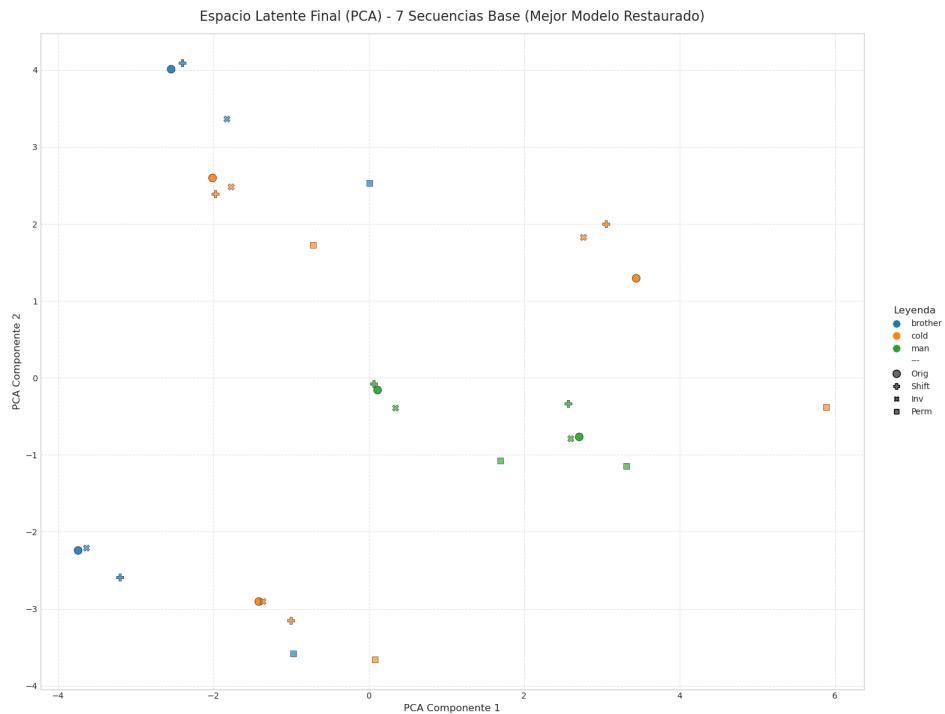
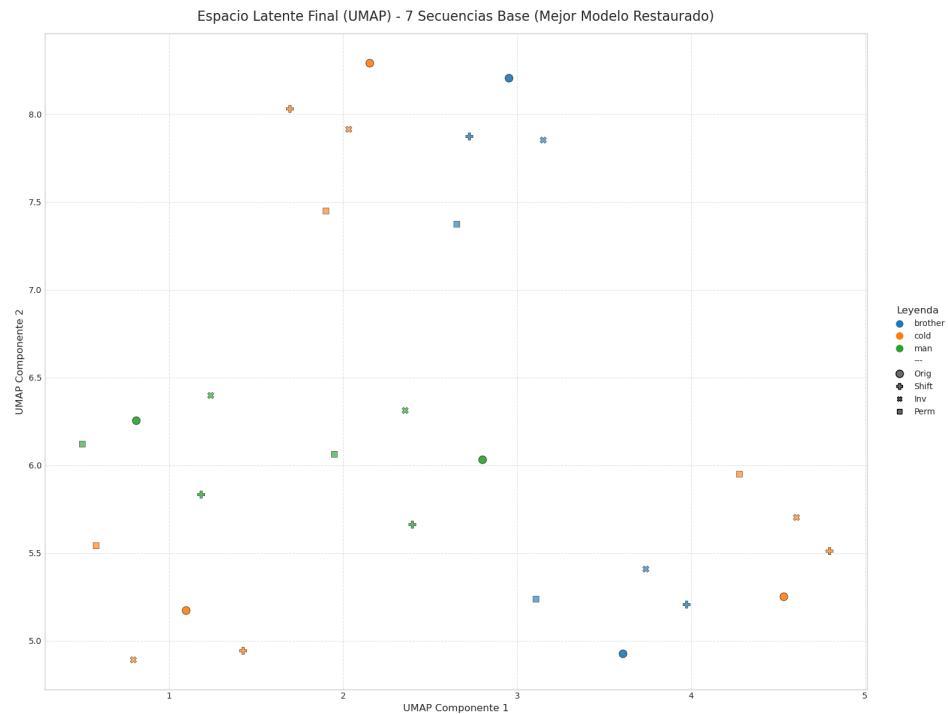


Figura 10.36: Esta grafica muestra el espacio latente en la mejor epoca (100) utilizando umap, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.37: Esta grafica muestra el espacio latente en la epoca 5 utilizando pca, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

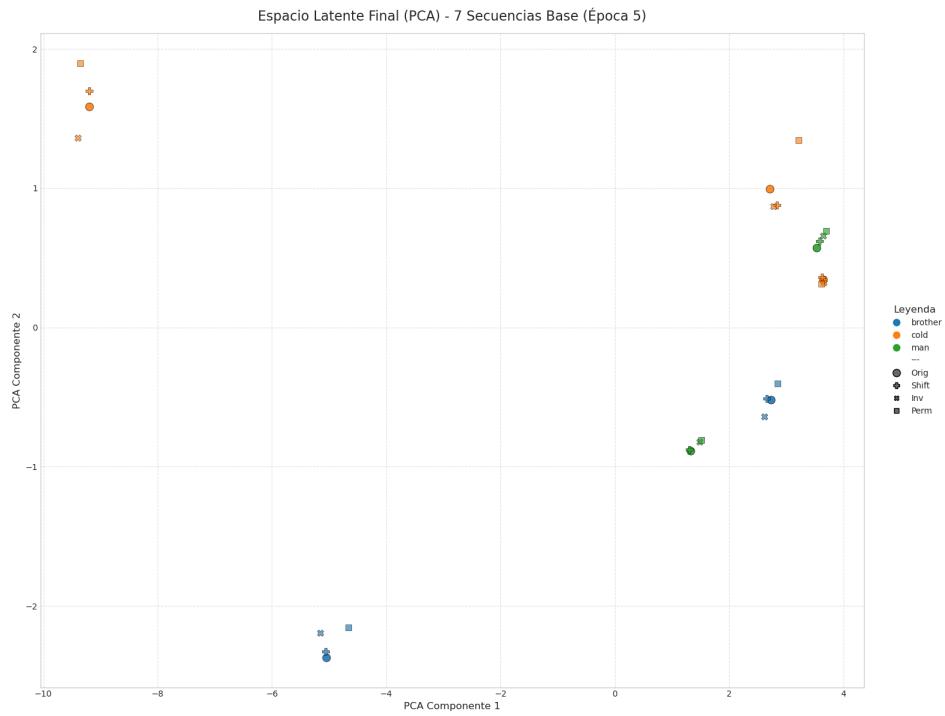
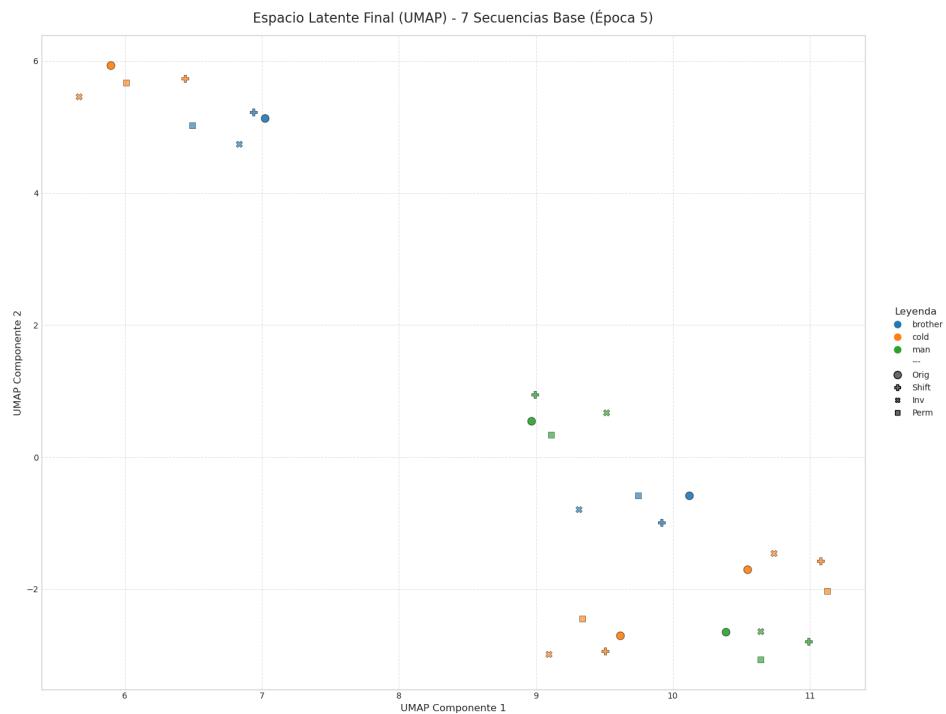


Figura 10.38: Esta grafica muestra el espacio latente en la epoca 5 utilizando umap, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.39: Esta grafica muestra el espacio latente en la epoca 15 utilizando pca, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

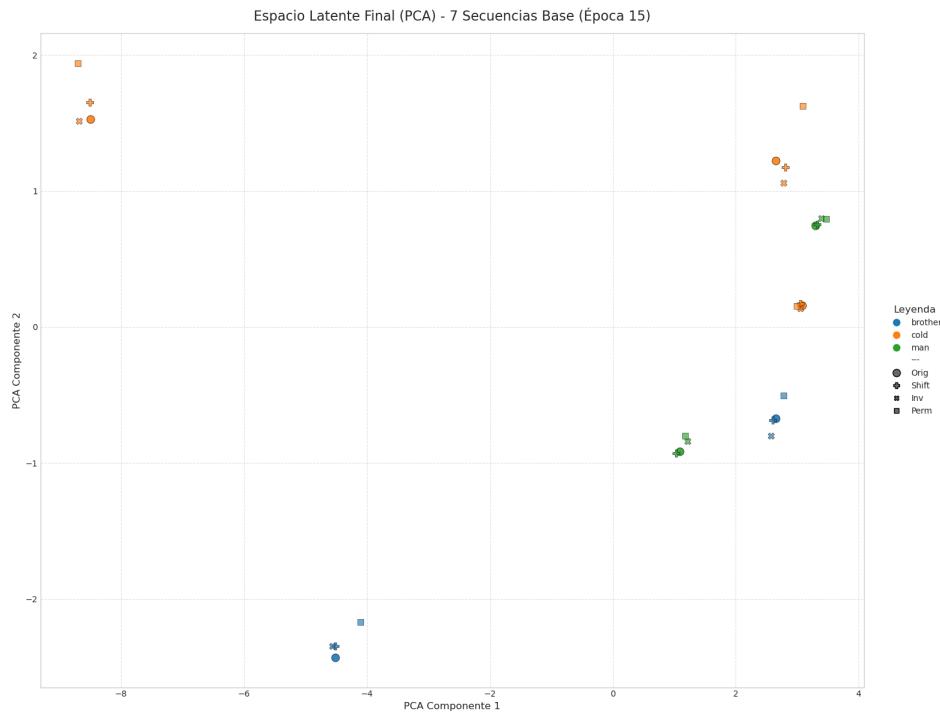
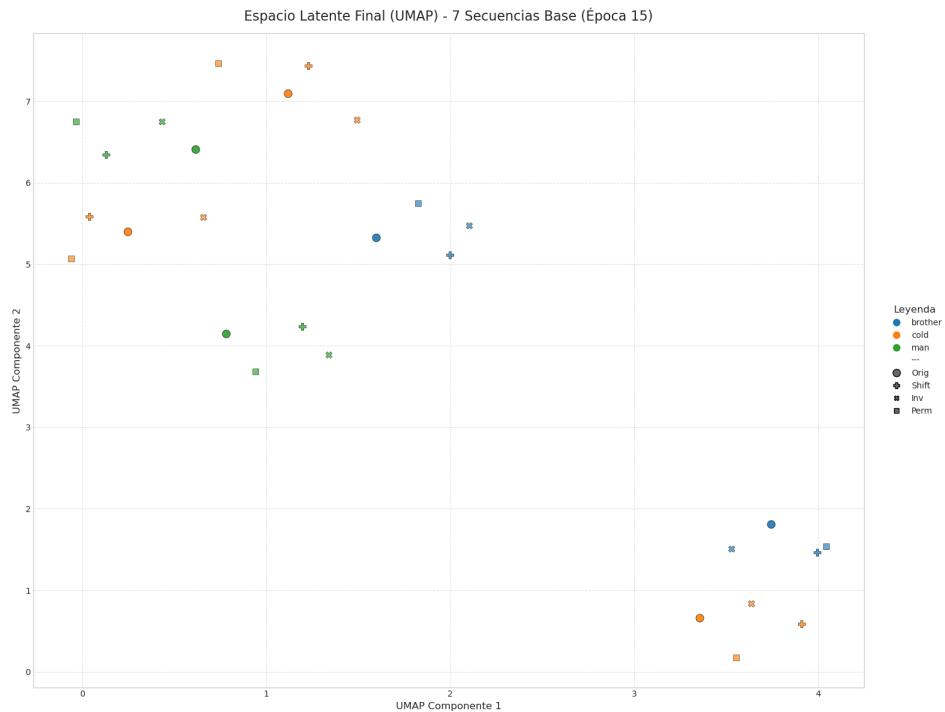


Figura 10.40: Esta grafica muestra el espacio latente en la epoca 15 utilizando umap, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.41: Esta grafica muestra el espacio latente en la epoca 25 utilizando pca, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

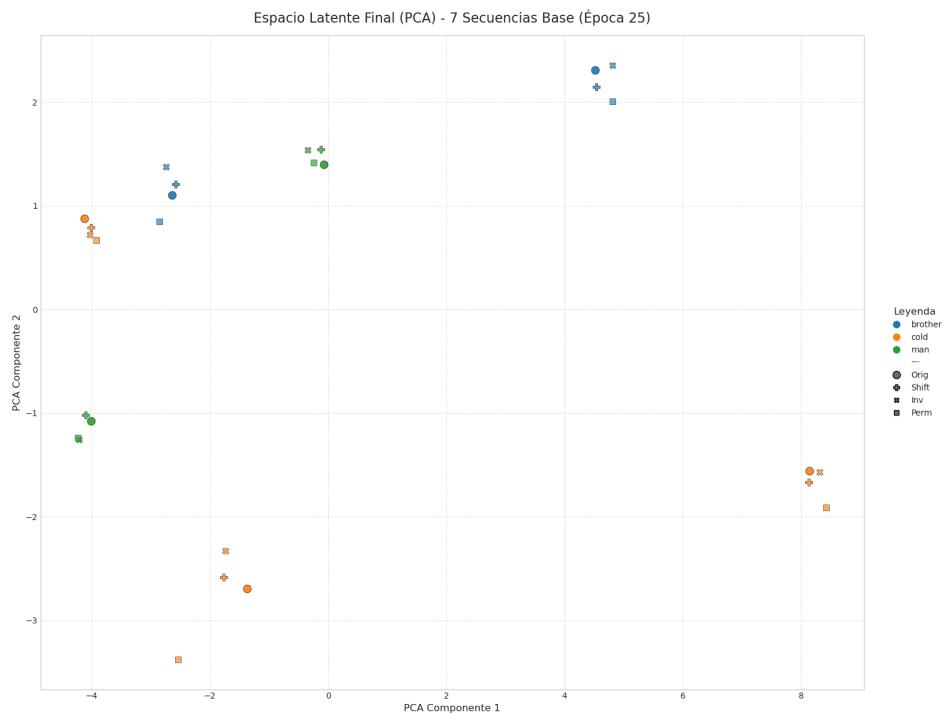
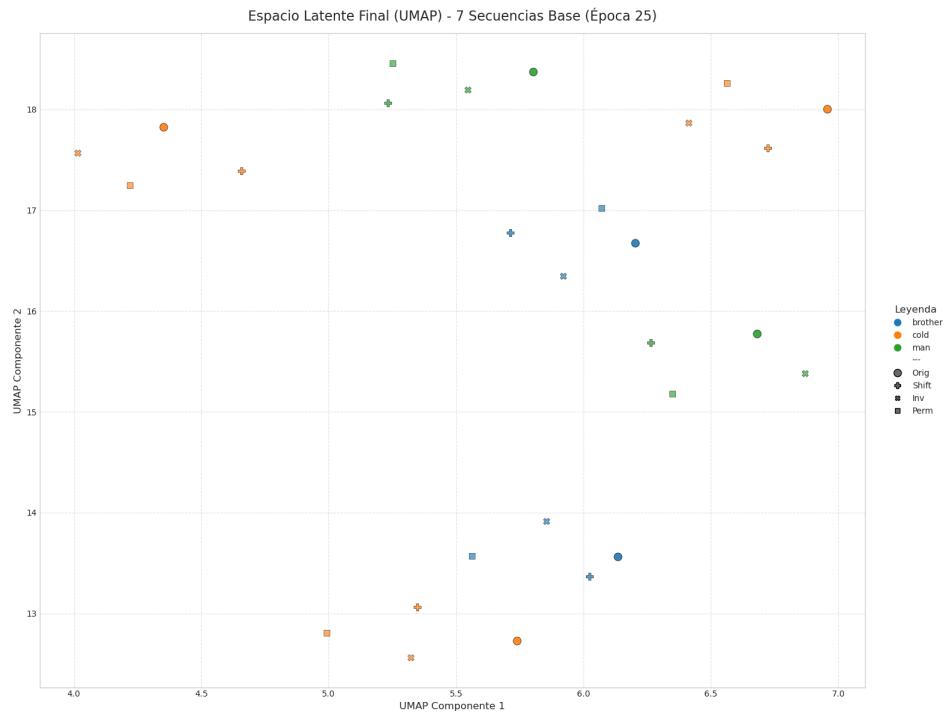


Figura 10.42: Esta grafica muestra el espacio latente en la epoca 25 utilizando umap, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.43: Esta grafica muestra el espacio latente en la epoca 45 utilizando pca, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

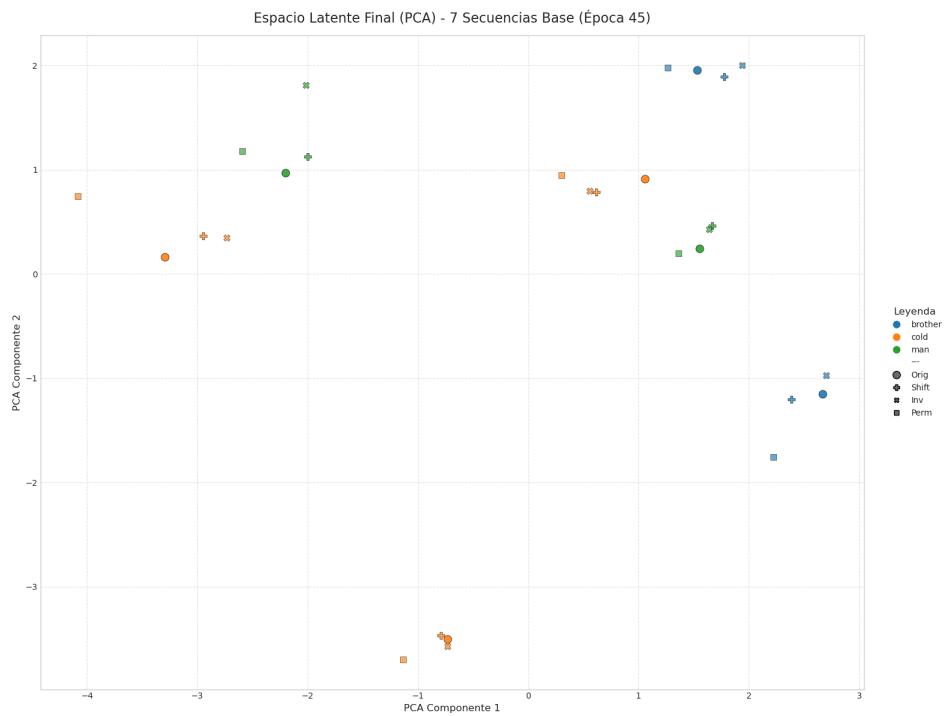
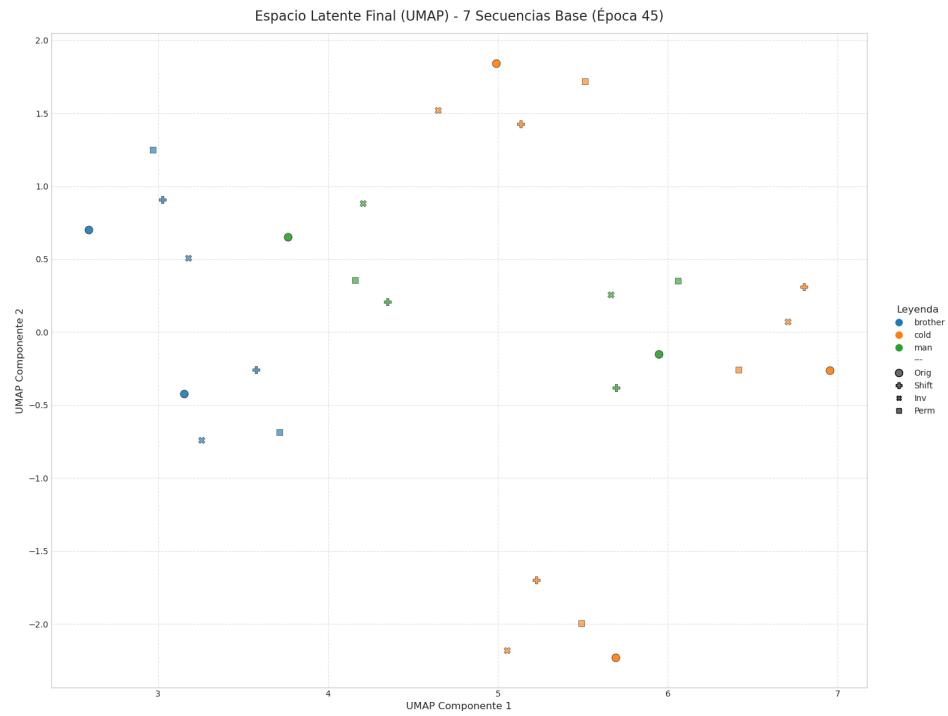


Figura 10.44: Esta grafica muestra el espacio latente en la epoca 45 utilizando umap, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.45: Esta grafica muestra el espacio latente en la epoca 65 utilizando pca, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

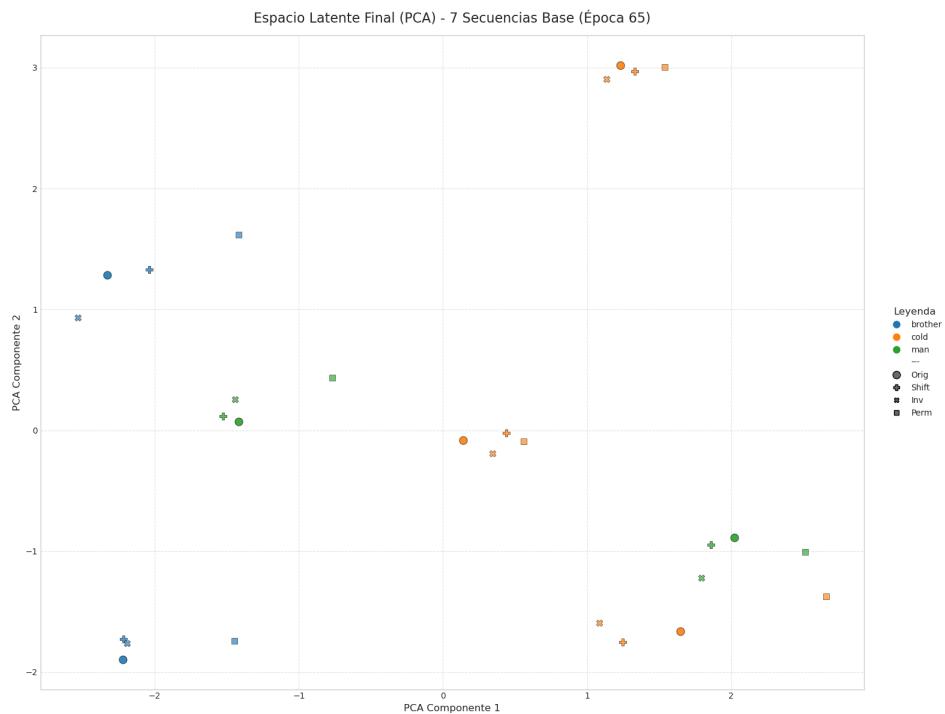
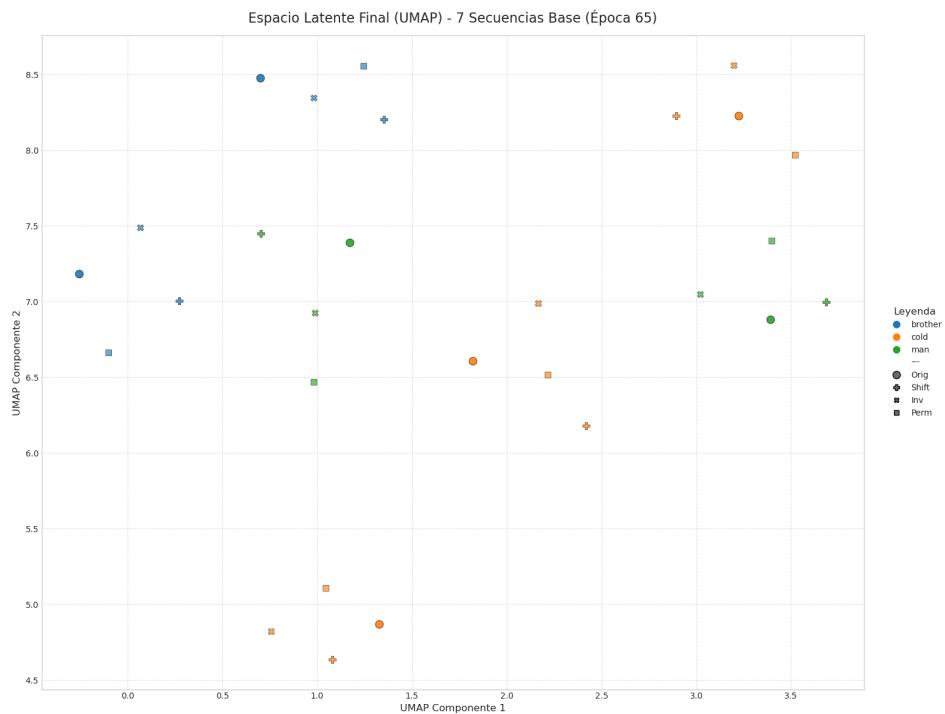


Figura 10.46: Esta grafica muestra el espacio latente en la epoca 65 utilizando umap, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.47: Esta grafica muestra el espacio latente en la epoca 85 utilizando pca, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

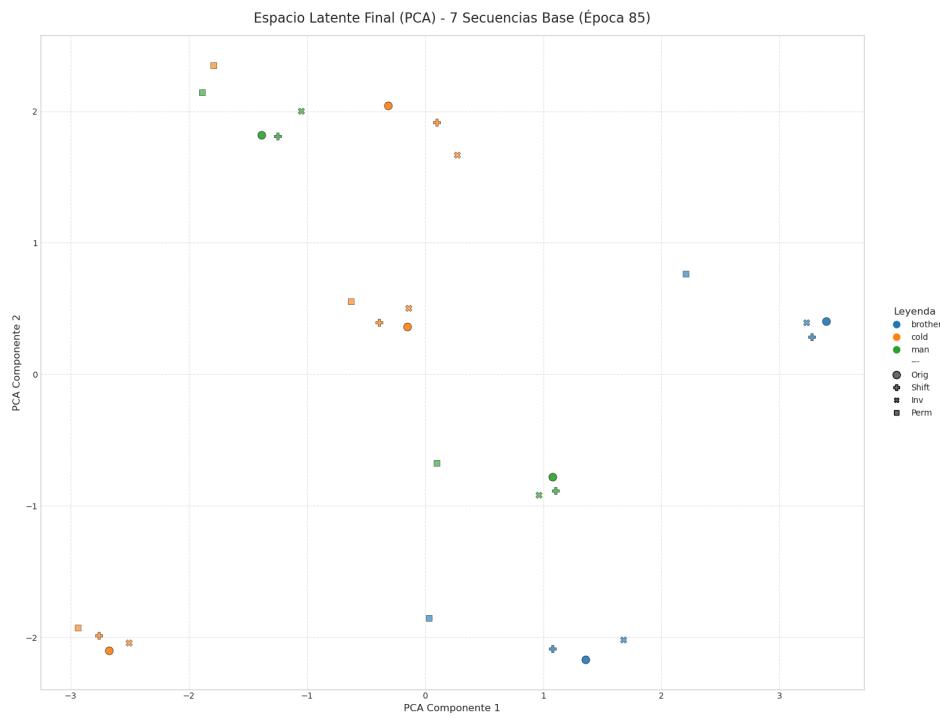
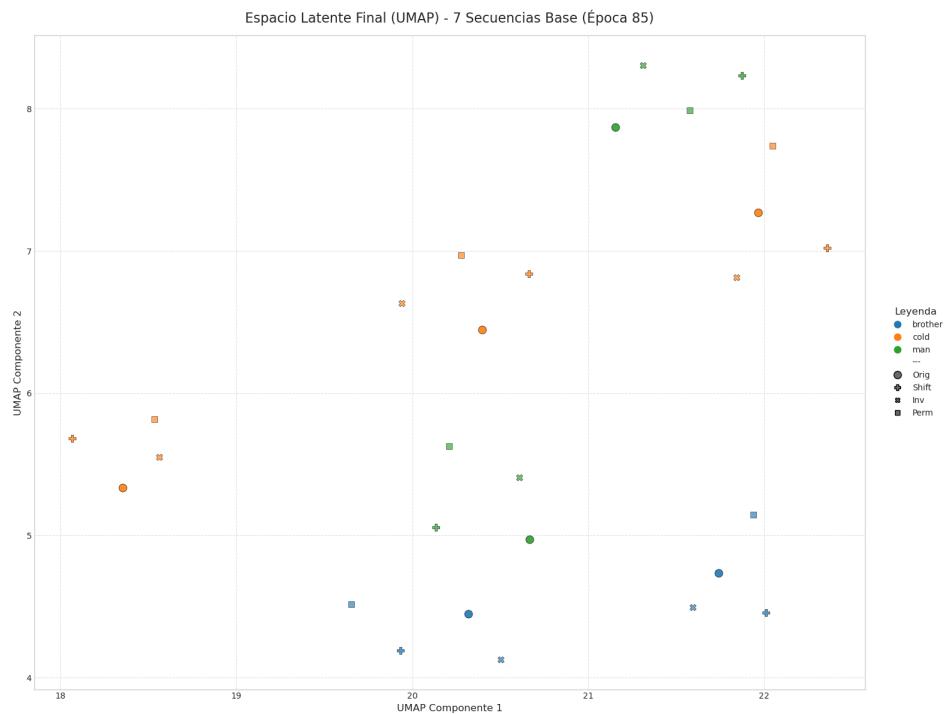


Figura 10.48: Esta grafica muestra el espacio latente en la epoca 85 utilizando umap, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

1.2. Con el dataset de ISL

Con 2 etiquetas

Figura 10.49: Esta gráfica compara el ratio Semántico del modelo comparado con un «baseline» o punto de referencia. Siendo este el modelo sin entrenar. El eje Y representa el ratio semántico, que mide la capacidad del modelo para diferenciar entre palabras diferentes, mientras que el eje X representa las épocas de entrenamiento.

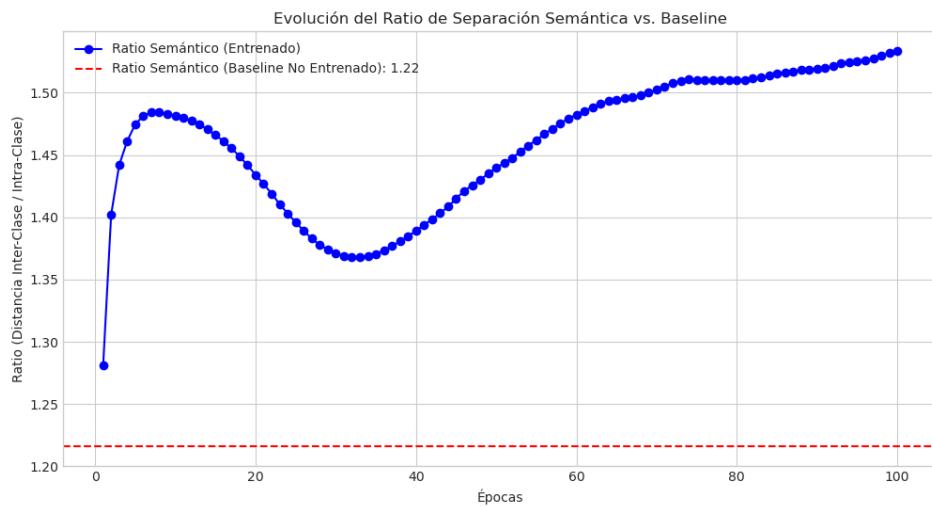


Figura 10.50: Esta gráfica evalúa la capacidad del modelo para entender el orden temporal de las secuencias de video comparadas con sus respectivos «baselines» del modelo no entrenado. Muestra la distancia euclídea promedio entre la secuencia original y sus versiones alteradas (shifted, inverted, permuted).

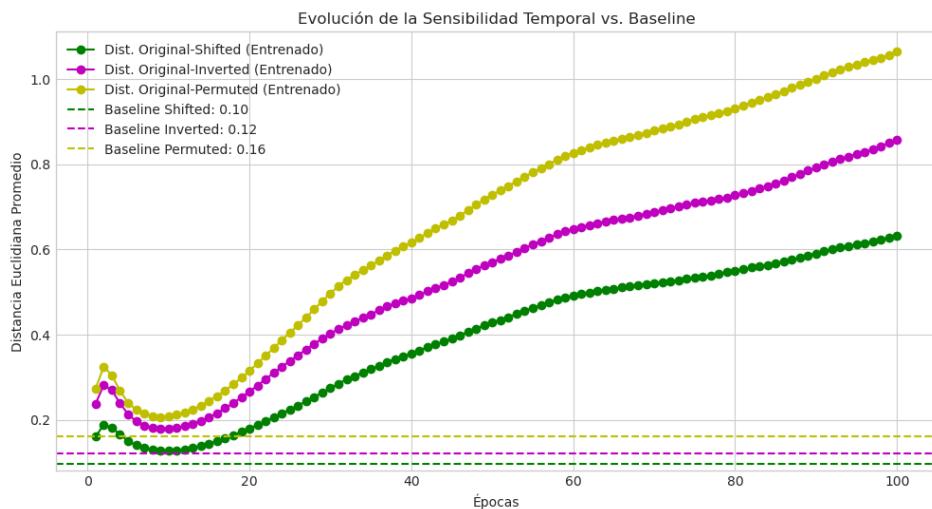


Figura 10.51: Esta gráfica compara el rendimiento del modelo principal con un modelo más simple en términos de la pérdida total de validación. El eje Y representa el valor de la pérdida, una métrica que indica cuán bien el modelo está aprendiendo, donde los valores más bajos son mejores, y el eje X representa las épocas.

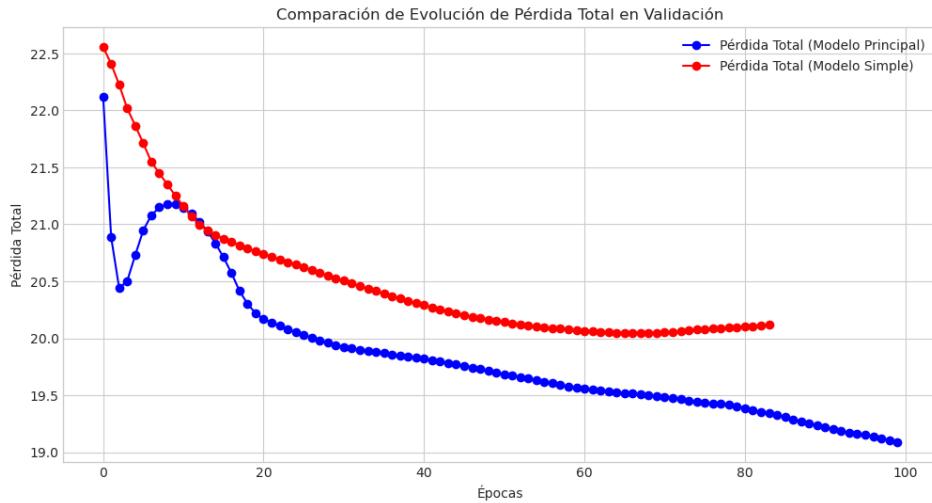
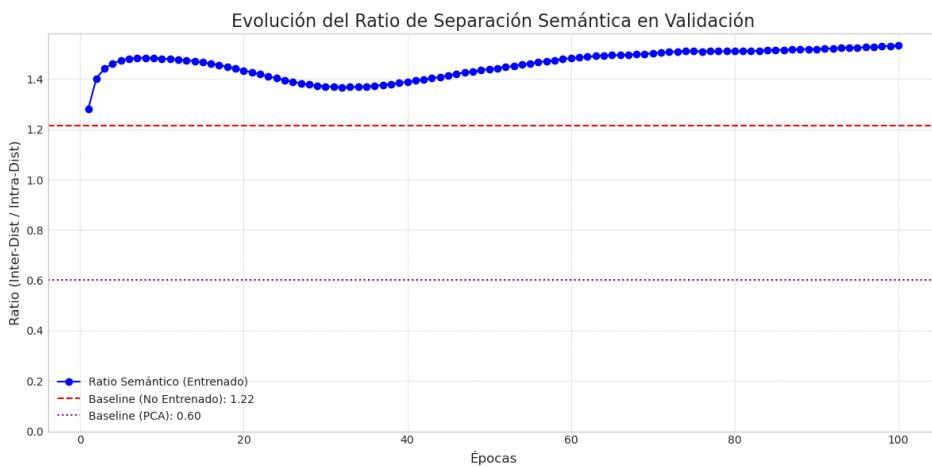


Figura 10.52: Esta gráfica compara el ratio semántico del modelo comparado con dos «baselines» o puntos de referencia. Siendo estos el modelo sin entrenar y la representación de PCA. El eje Y representa el ratio semántico, que mide la capacidad del modelo para diferenciar entre palabras diferentes, mientras que el eje X representa las épocas de entrenamiento.



1. Gráficas de los Experimentos Realizados

Figura 10.53: Este gráfico mide directamente la calidad de la separación semántica en el espacio latente para palabras con la misma y diferente clase. El eje Y representa la distancia euclíadiana promedio y el eje X son las épocas.

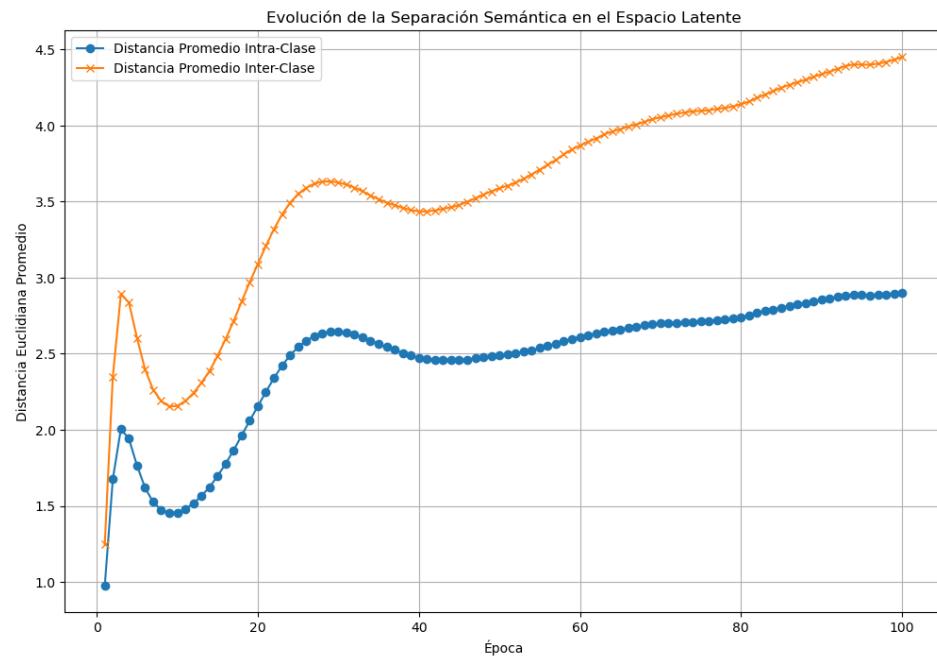
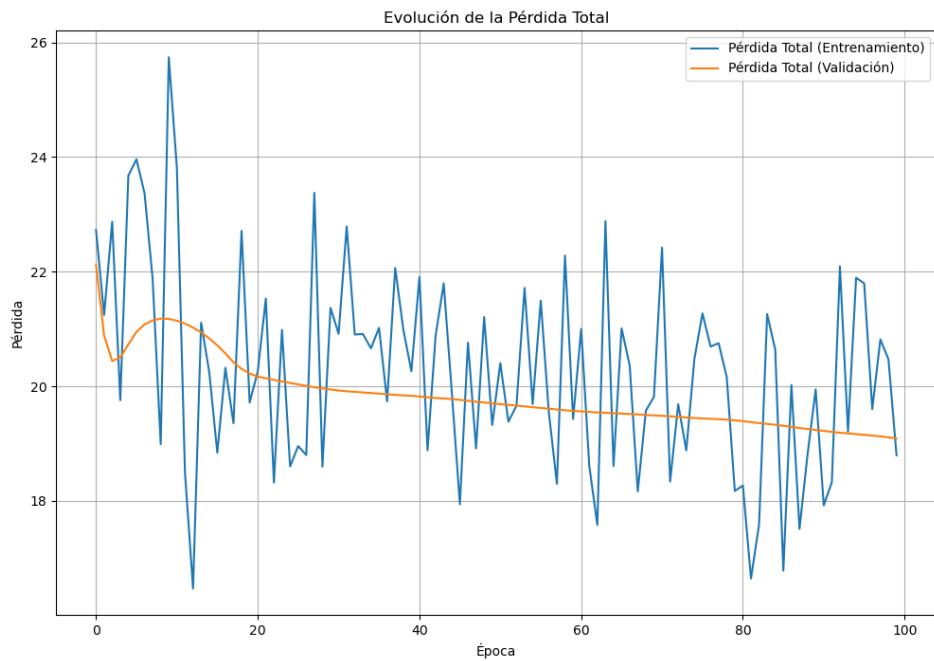


Figura 10.54: Este gráfico muestra la evolución de la «Pérdida Total» a lo largo de 100 épocas de entrenamiento. El eje Y representa el valor de la pérdida, una métrica que indica cuán bien el modelo está aprendiendo donde valores más bajos son mejores. El eje X representa las épocas, es decir, cada ciclo completo de entrenamiento sobre el conjunto de datos.



1. Gráficas de los Experimentos Realizados

Figura 10.55: Esta gráfica ilustra la «Pérdida de Reconstrucción», que mide qué tan bien el autoencoder del modelo puede reconstruir la entrada original después de haberla comprimido en un espacio latente. Al igual que en la gráfica anterior, el eje Y es el valor de la pérdida y el eje X son las épocas.

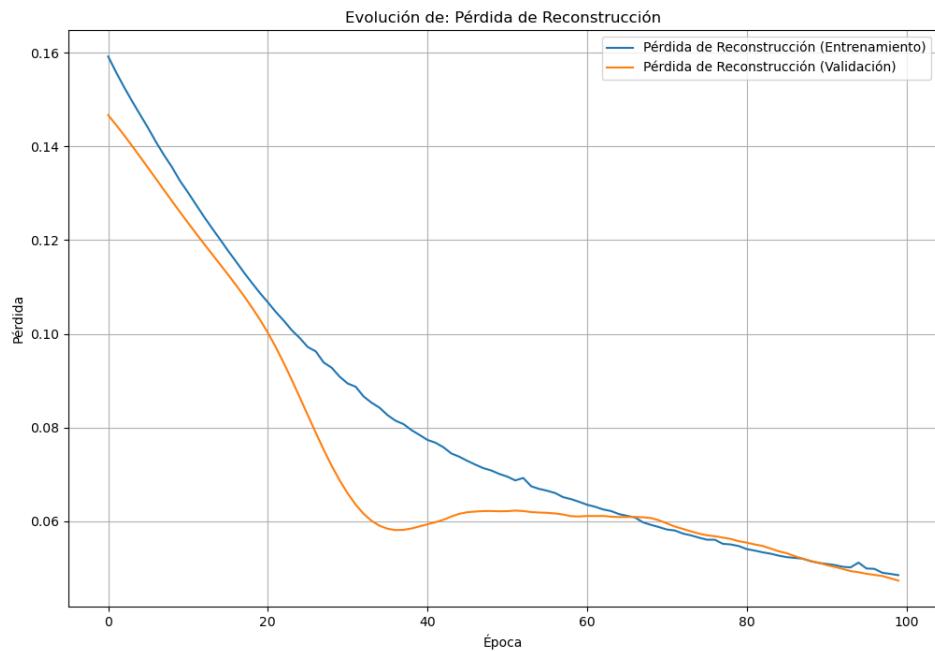
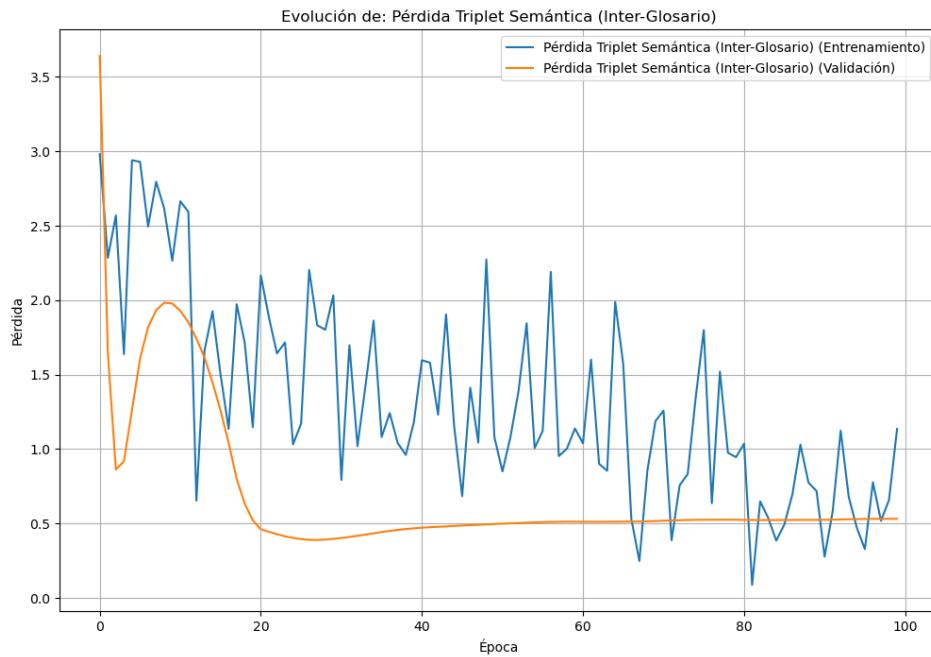


Figura 10.56: Este gráfico muestra la «Pérdida Triplet Semántica», una métrica clave que evalúa si el modelo puede diferenciar entre distintos glosarios, en este caso, las señas «brother» y «cold». El objetivo es que las representaciones de un mismo glosario estén más cerca entre sí que las de glosarios diferentes. Igualmente, el eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.



1. Gráficas de los Experimentos Realizados

Figura 10.57: Esta visualización se enfoca en la sensibilidad temporal del modelo, puesto que, mide la diferencia entre la secuencia original de un video y su versión invertida. Esto quiere decir que el modelo debe aprender que una secuencia invertida es significativamente diferente de la original. El eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.

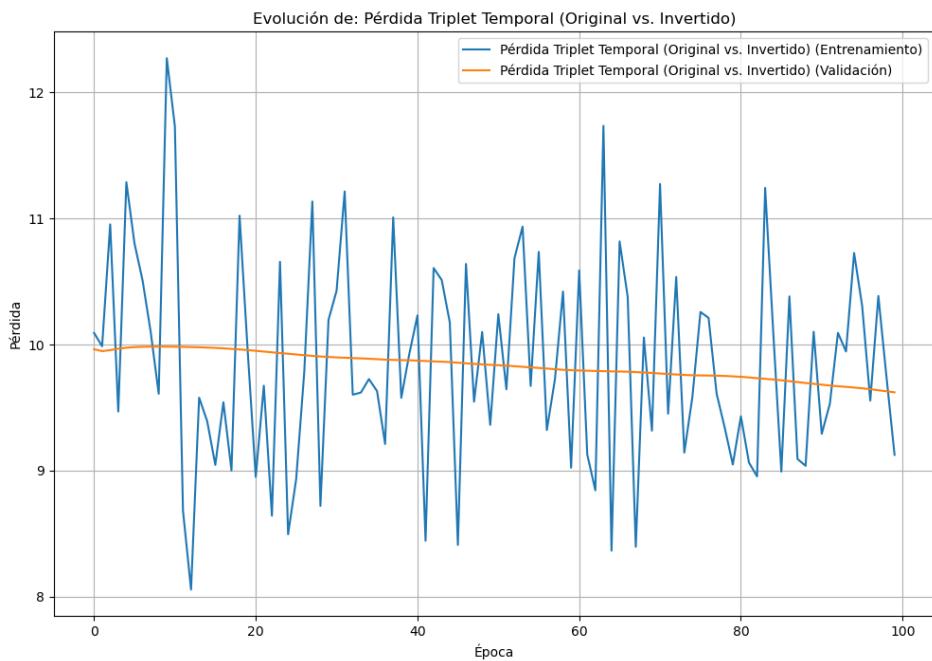
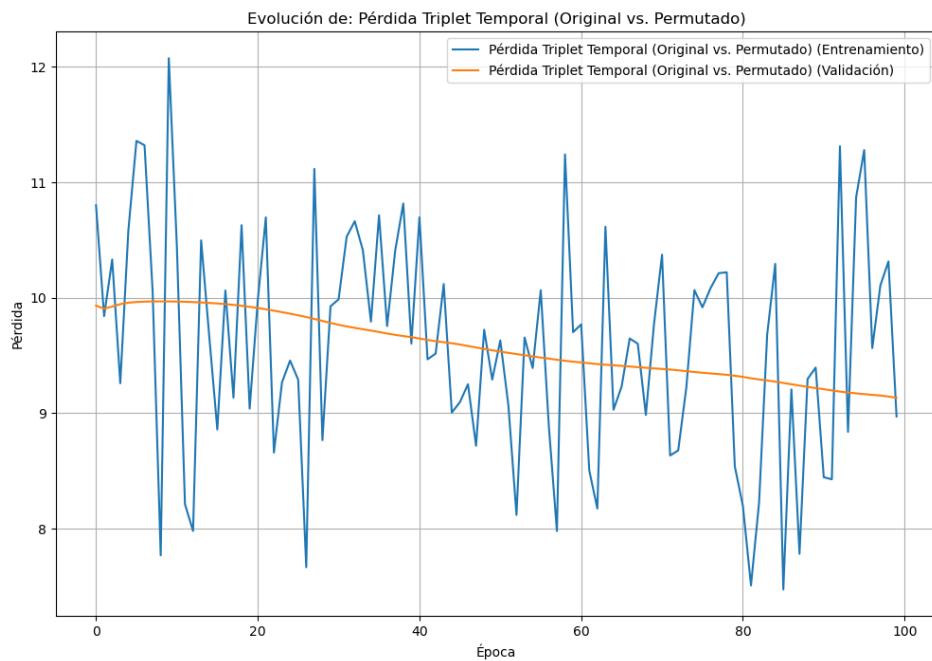


Figura 10.58: Similar a la gráfica anterior, esta también evalúa la sensibilidad temporal, pero en este caso, compara la secuencia original con una versión donde los fotogramas han sido desordenados aleatoriamente. Donde el objetivo es que el modelo reconozca que una secuencia permutada es muy diferente de la original. El eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.



1. Gráficas de los Experimentos Realizados

Figura 10.59: Esta grafica muestra el espacio latente en la mejor epoca (100) utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

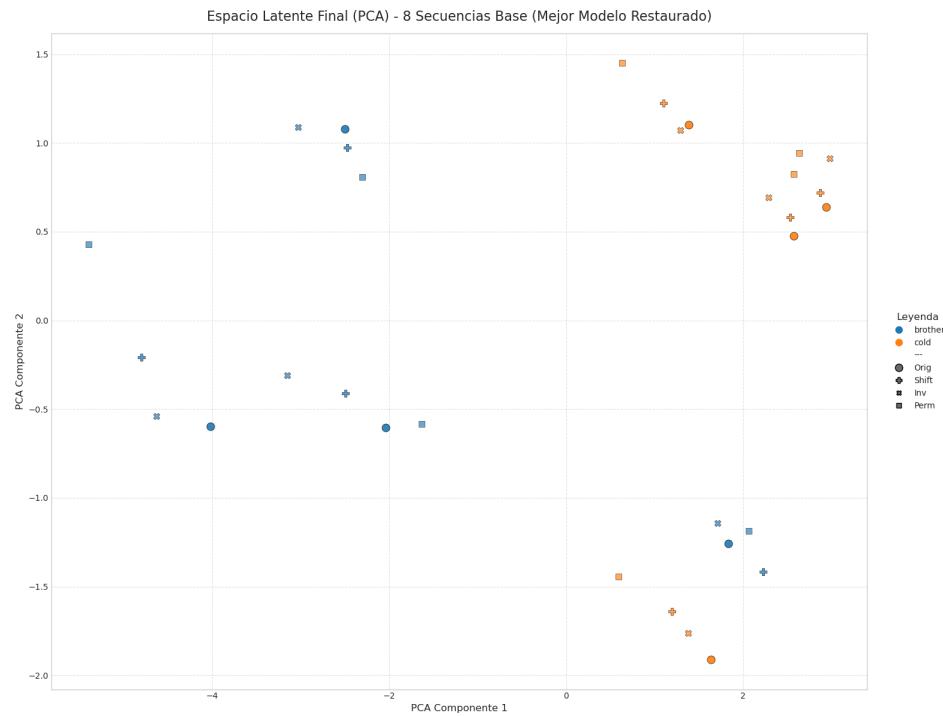
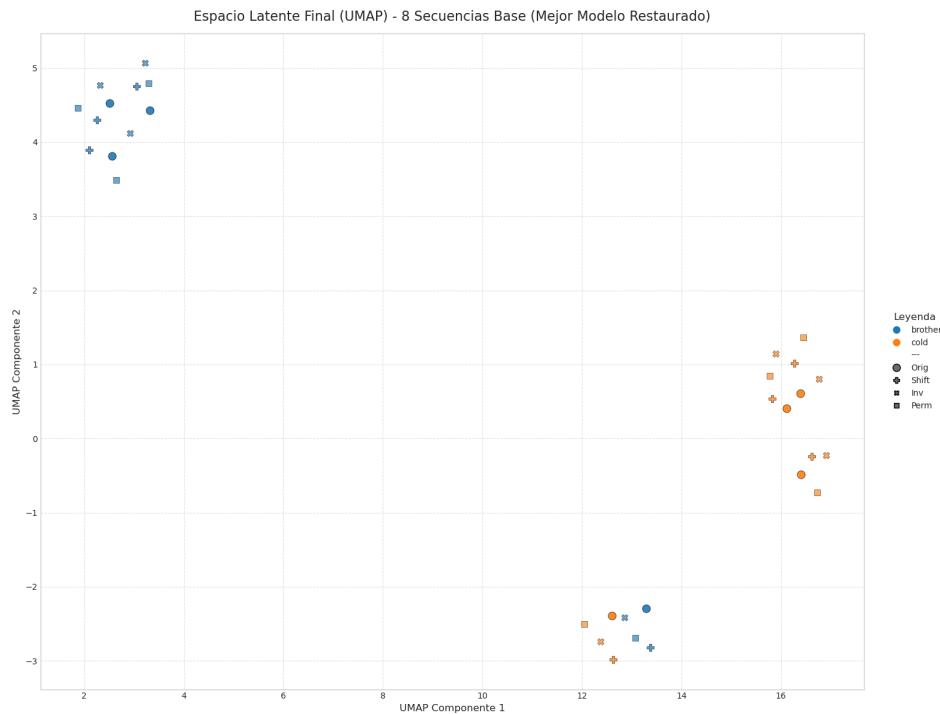


Figura 10.60: Esta grafica muestra el espacio latente en la mejor epoca (100) utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.61: Esta grafica muestra el espacio latente en la epoca 5 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

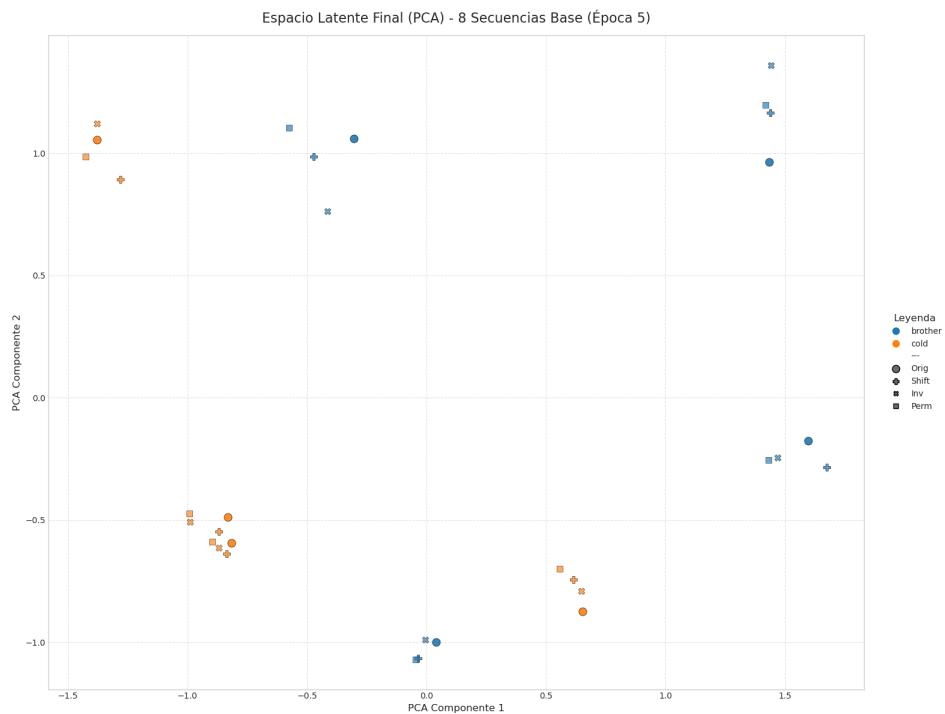
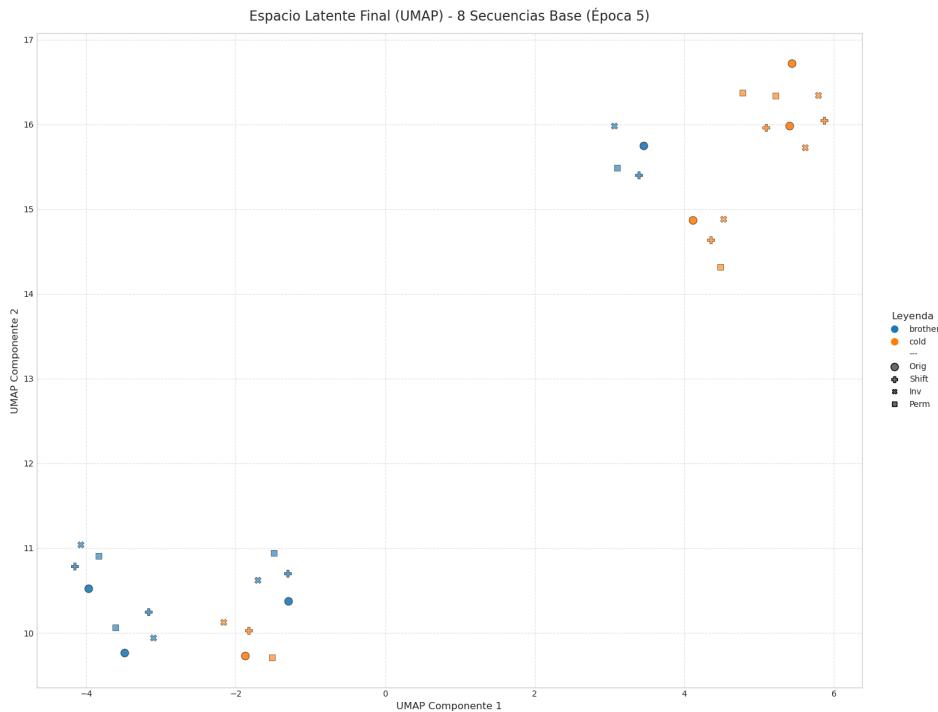


Figura 10.62: Esta grafica muestra el espacio latente en la epoca 5 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.63: Esta grafica muestra el espacio latente en la epoca 15 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

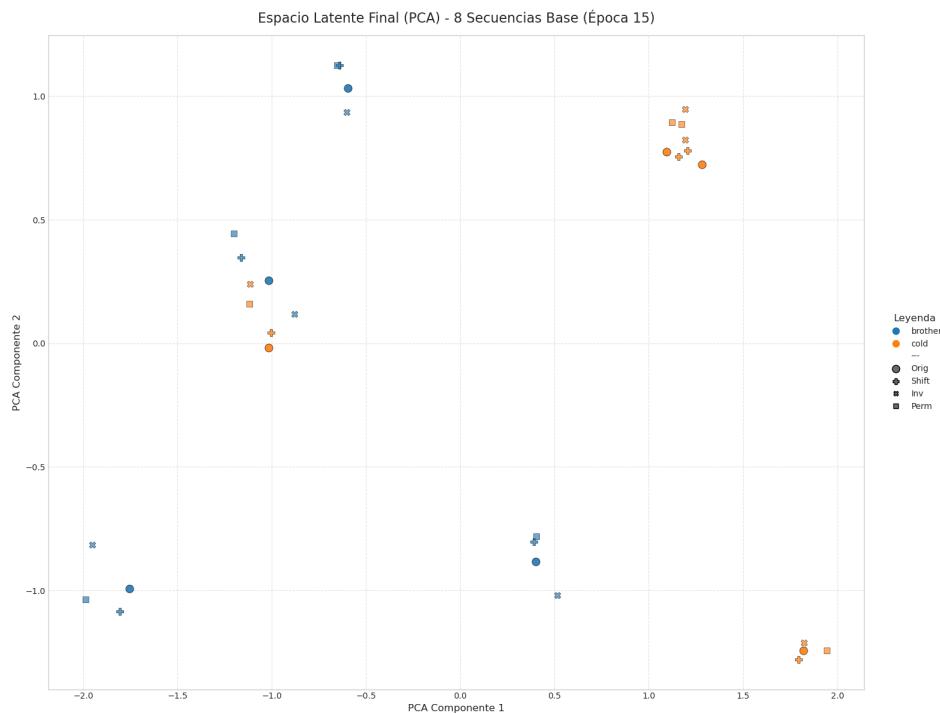
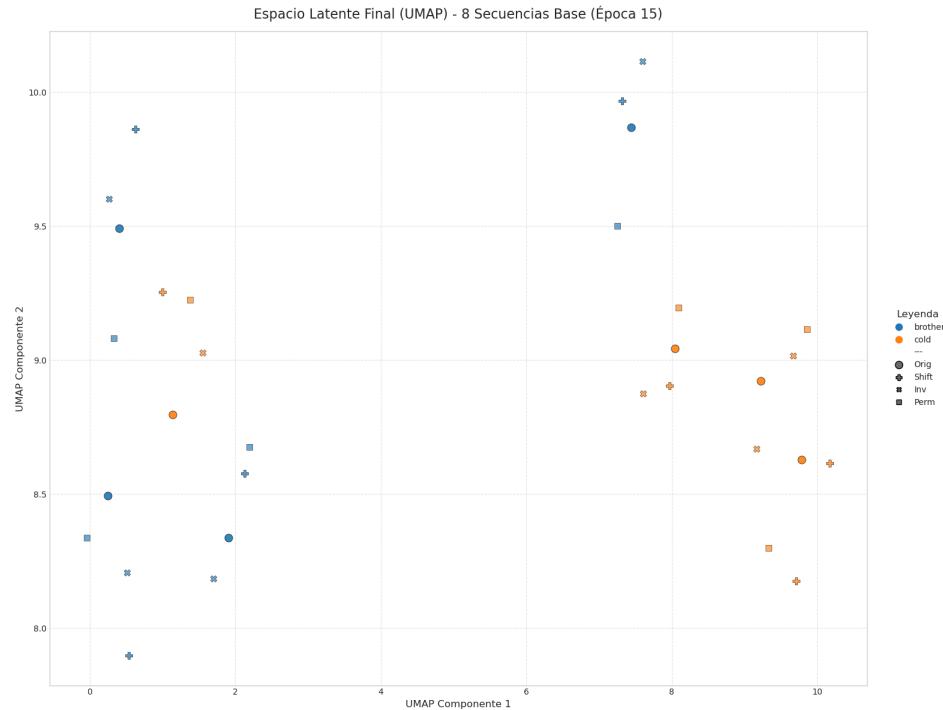


Figura 10.64: Esta grafica muestra el espacio latente en la epoca 15 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.65: Esta grafica muestra el espacio latente en la epoca 25 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

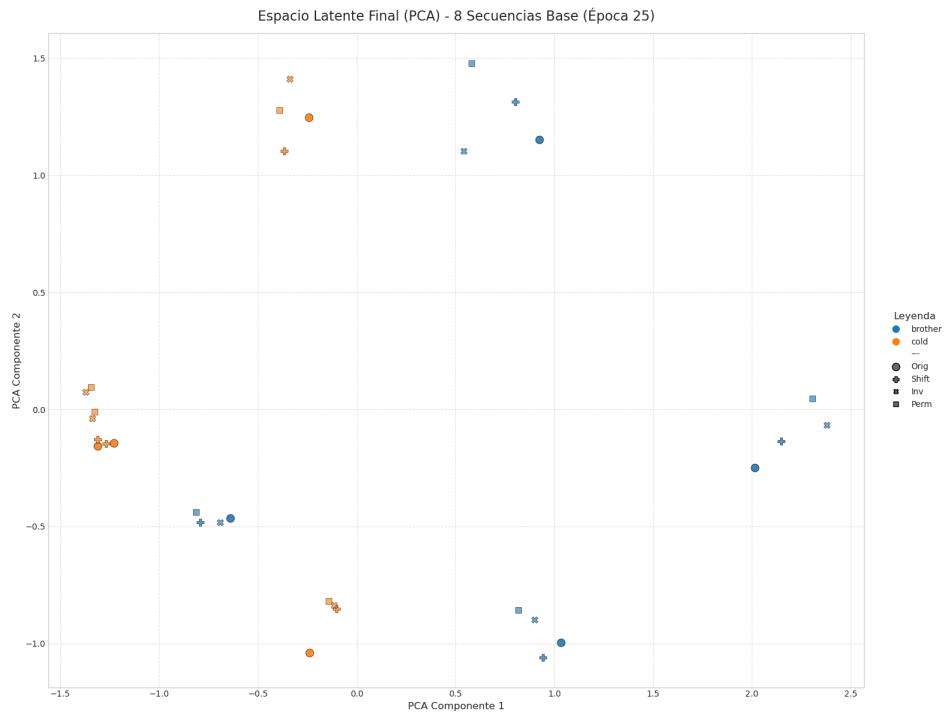
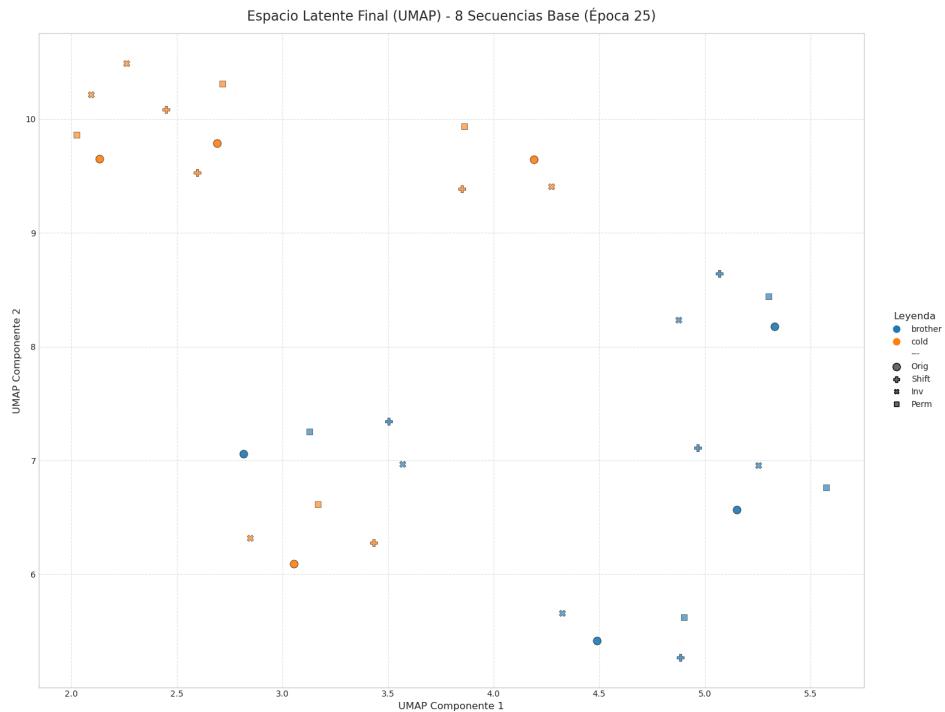


Figura 10.66: Esta grafica muestra el espacio latente en la epoca 25 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.67: Esta grafica muestra el espacio latente en la epoca 45 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

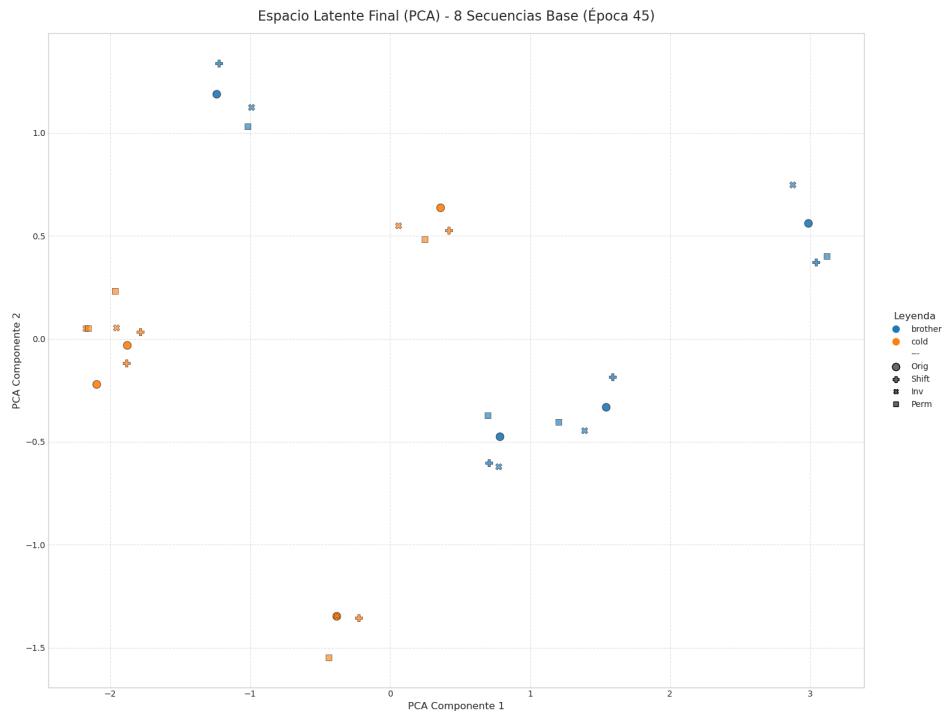
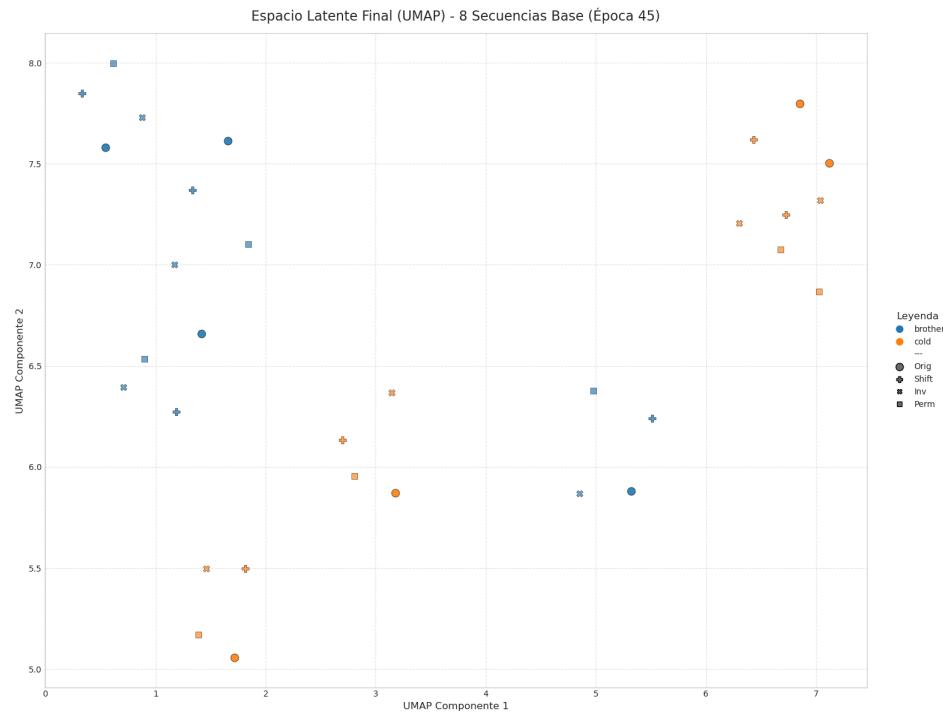


Figura 10.68: Esta grafica muestra el espacio latente en la epoca 45 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.69: Esta grafica muestra el espacio latente en la epoca 65 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

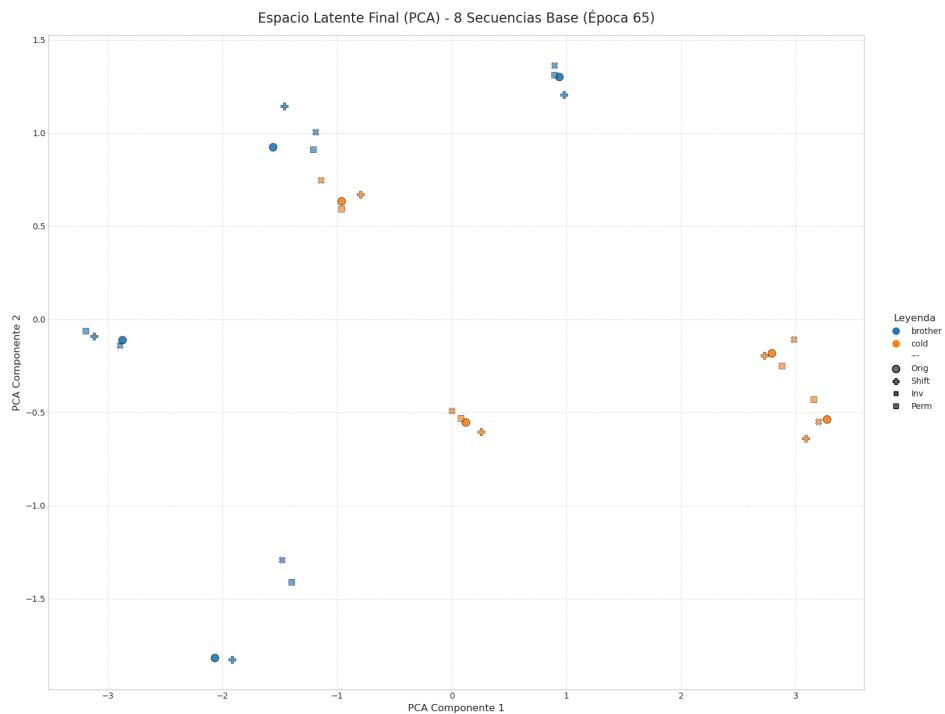
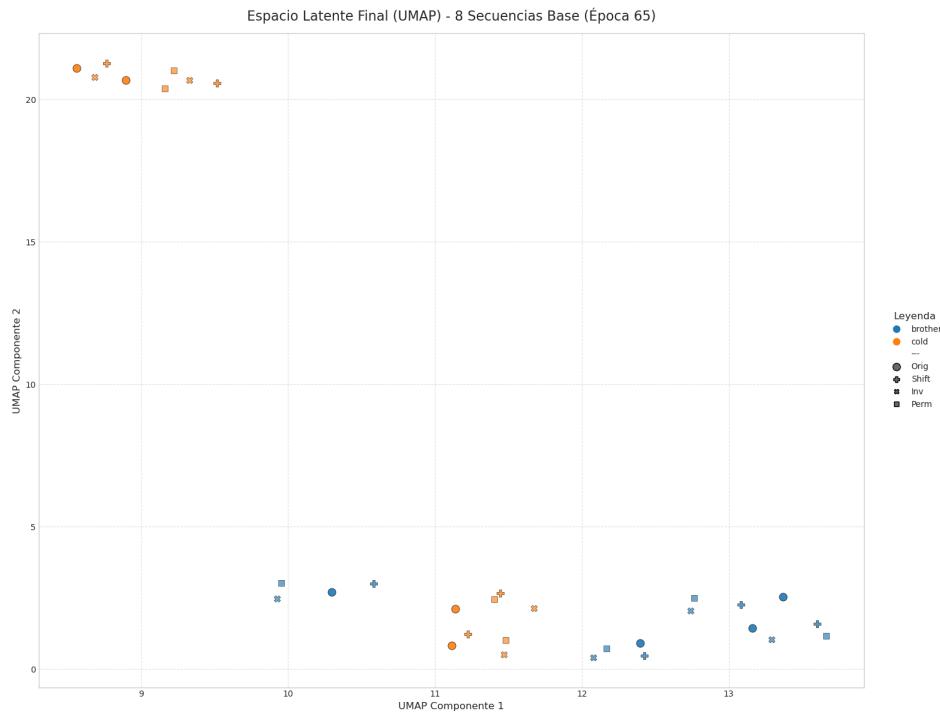


Figura 10.70: Esta grafica muestra el espacio latente en la epoca 65 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.71: Esta grafica muestra el espacio latente en la epoca 85 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

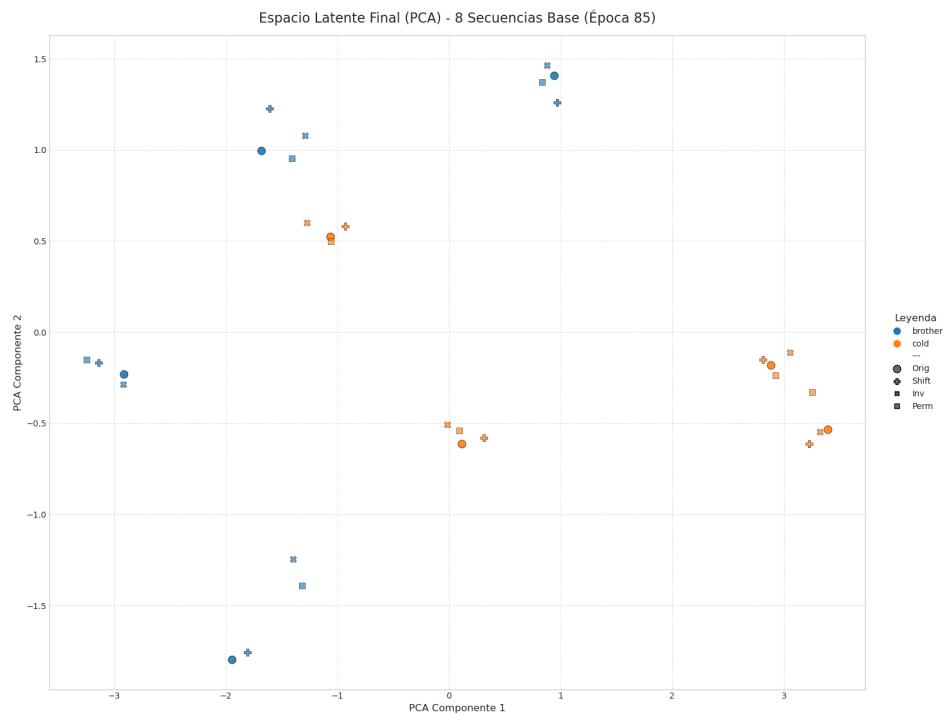
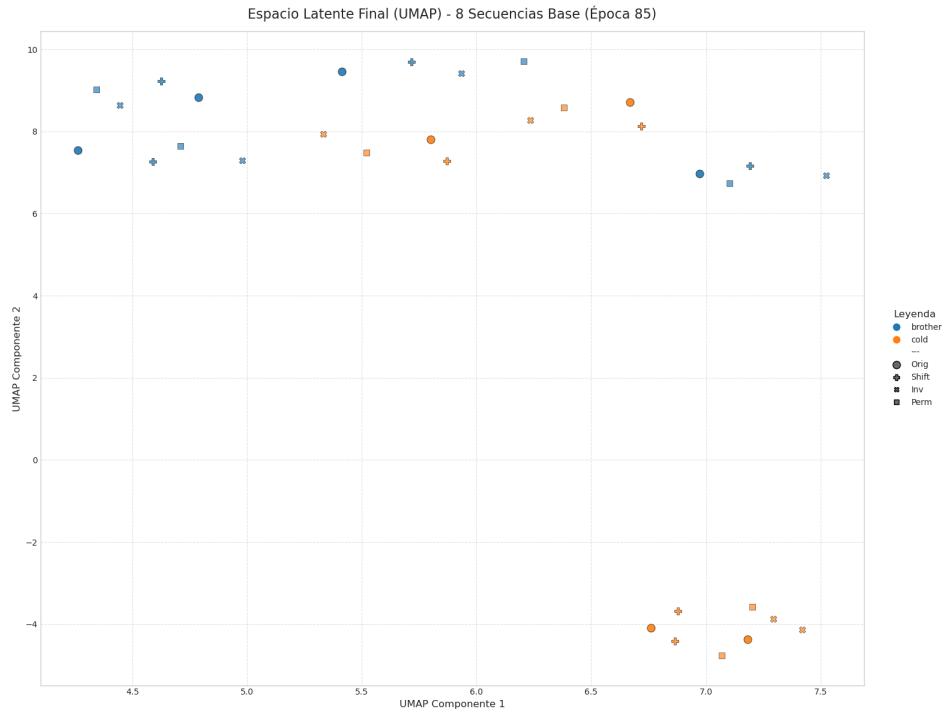


Figura 10.72: Esta grafica muestra el espacio latente en la epoca 85 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Con 3 etiquetas

Figura 10.73: Esta gráfica compara el ratio Semántico del modelo comparado con un «baseline» o punto de referencia. Siendo este el modelo sin entrenar. El eje Y representa el ratio semántico, que mide la capacidad del modelo para diferenciar entre palabras diferentes, mientras que el eje X representa las épocas de entrenamiento.

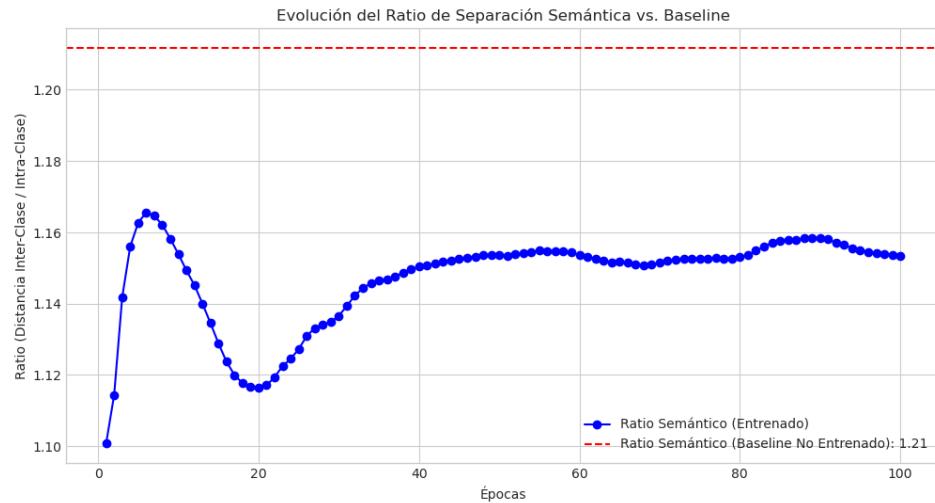


Figura 10.74: Esta gráfica evalúa la capacidad del modelo para entender el orden temporal de las secuencias de video comparadas con sus respectivos «baselines» del modelo no entrenado. Muestra la distancia euclídea promedio entre la secuencia original y sus versiones alteradas (shifted, inverted, permuted).

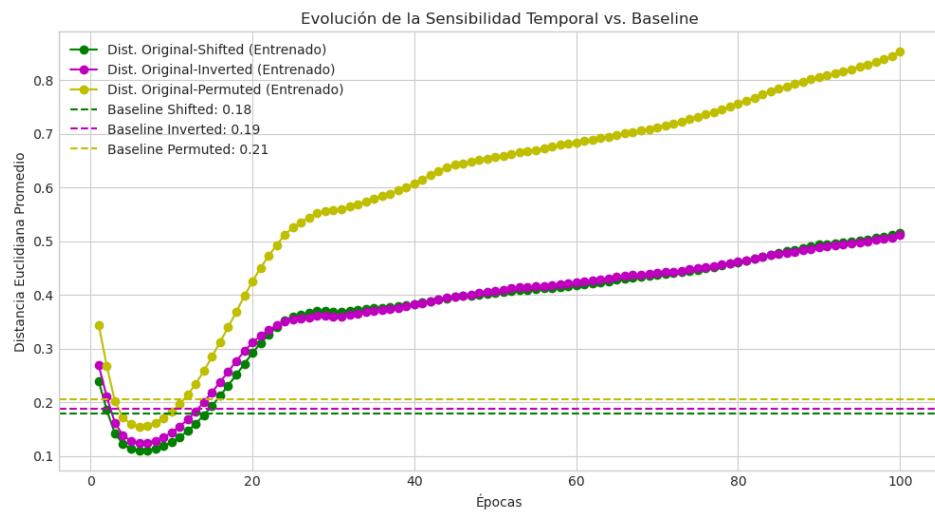


Figura 10.75: Esta gráfica compara el rendimiento del modelo principal con un modelo más simple en términos de la pérdida total de validación. El eje Y representa el valor de la pérdida, una métrica que indica cuán bien el modelo está aprendiendo, donde los valores más bajos son mejores, y el eje X representa las épocas.

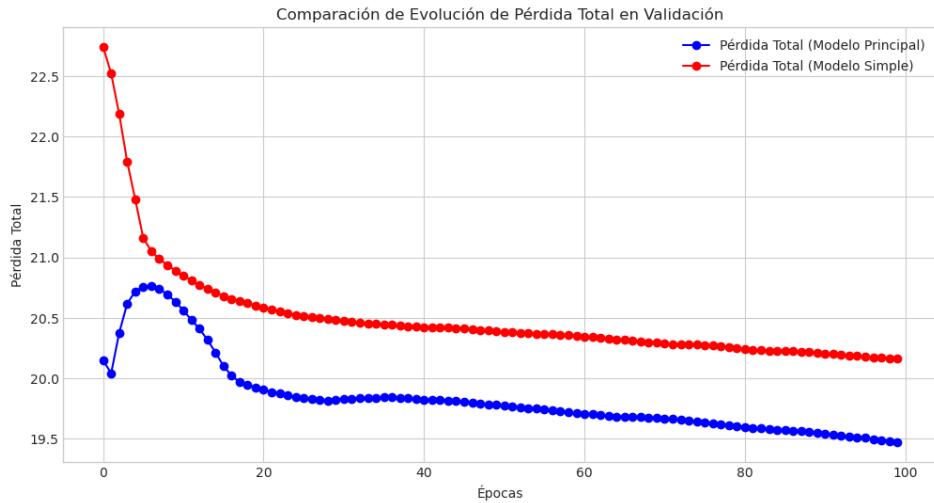
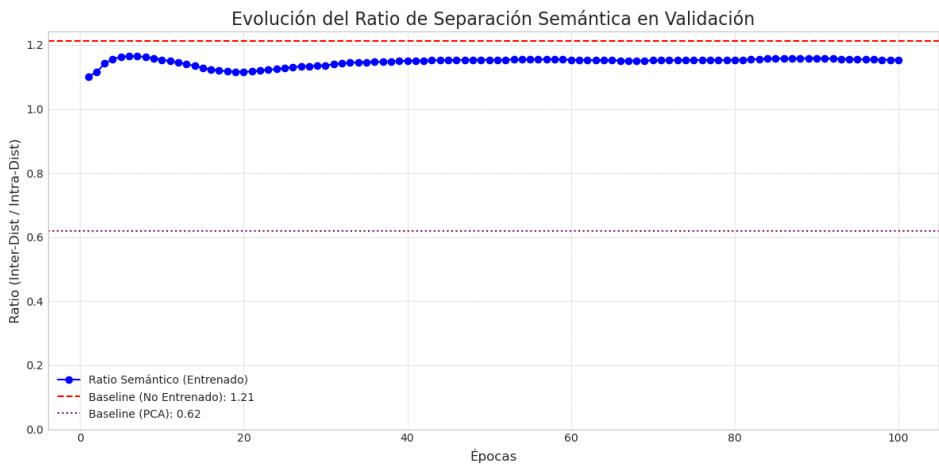


Figura 10.76: Esta gráfica compara el ratio semántico del modelo comparado con dos «baselines» o puntos de referencia. Siendo estos el modelo sin entrenar y la representación de PCA. El eje Y representa el ratio semántico, que mide la capacidad del modelo para diferenciar entre palabras diferentes, mientras que el eje X representa las épocas de entrenamiento.



1. Gráficas de los Experimentos Realizados

Figura 10.77: Este gráfico mide directamente la calidad de la separación semántica en el espacio latente para palabras con la misma y diferente clase. El eje Y representa la distancia euclíadiana promedio y el eje X son las épocas.

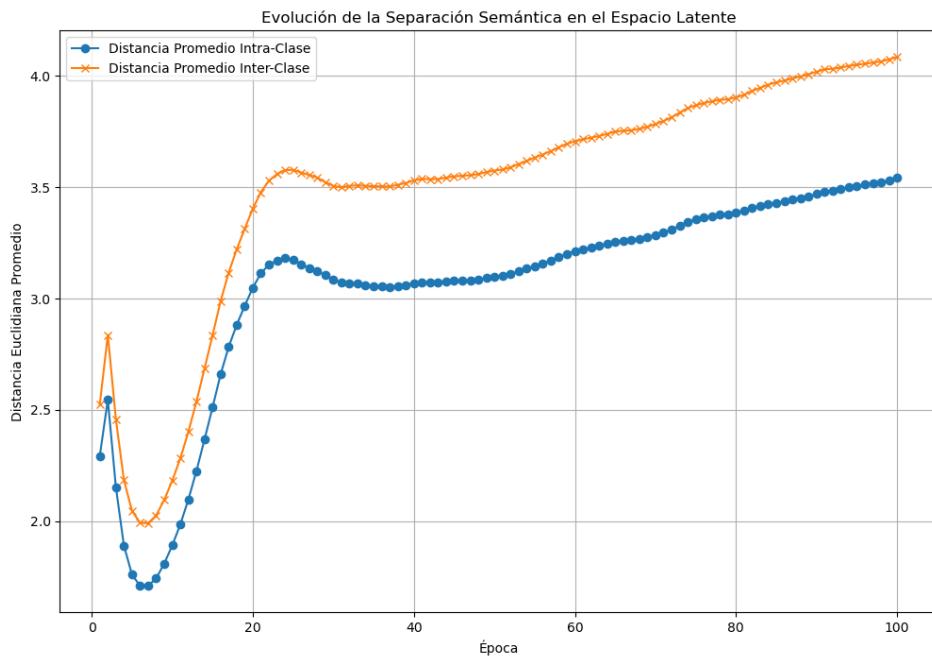
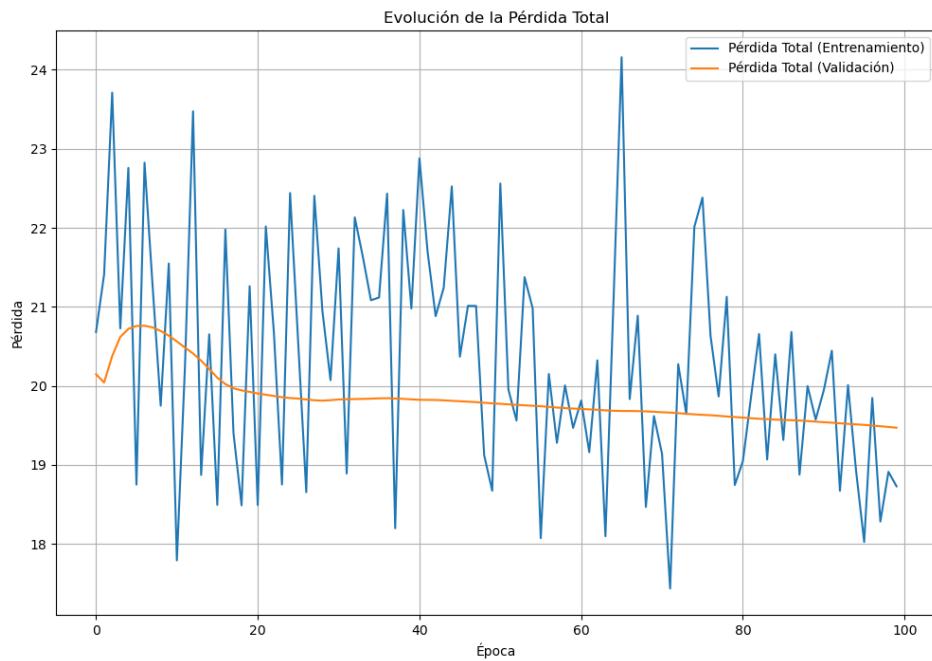


Figura 10.78: Este gráfico muestra la evolución de la «Pérdida Total» a lo largo de 100 épocas de entrenamiento. El eje Y representa el valor de la pérdida, una métrica que indica cuán bien el modelo está aprendiendo donde valores más bajos son mejores. El eje X representa las épocas, es decir, cada ciclo completo de entrenamiento sobre el conjunto de datos.



1. Gráficas de los Experimentos Realizados

Figura 10.79: Esta gráfica ilustra la «Pérdida de Reconstrucción», que mide qué tan bien el autoencoder del modelo puede reconstruir la entrada original después de haberla comprimido en un espacio latente. Al igual que en la gráfica anterior, el eje Y es el valor de la pérdida y el eje X son las épocas.

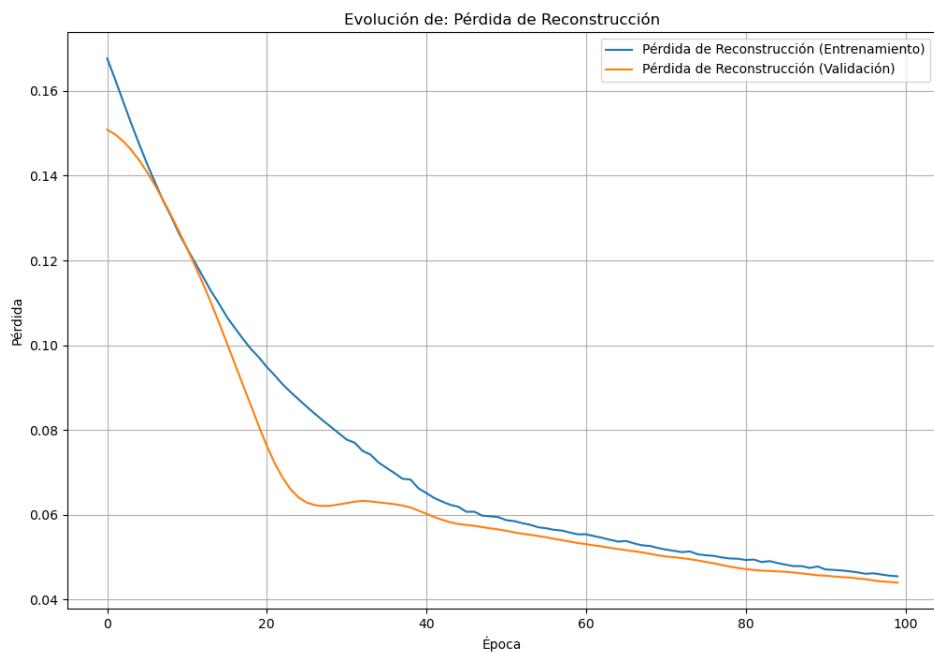
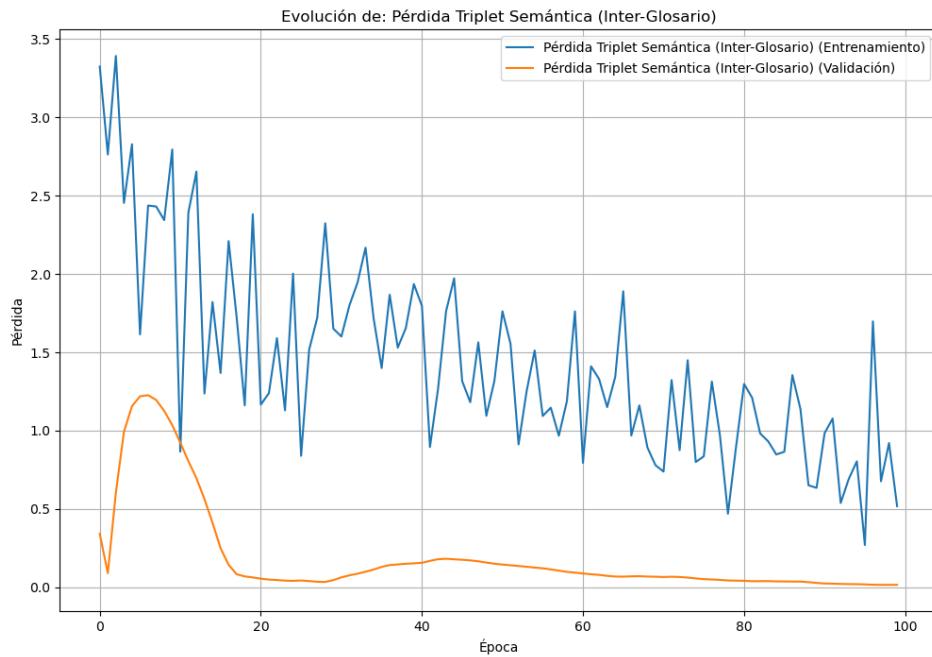


Figura 10.80: Este gráfico muestra la «Pérdida Triplet Semántica», una métrica clave que evalúa si el modelo puede diferenciar entre distintos glosarios, en este caso, las señas «brother», «cold» y «man». El objetivo es que las representaciones de un mismo glosario estén más cerca entre sí que las de glosarios diferentes. Igualmente, el eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.



1. Gráficas de los Experimentos Realizados

Figura 10.81: Esta visualización se enfoca en la sensibilidad temporal del modelo, puesto que, mide la diferencia entre la secuencia original de un video y su versión invertida. Esto quiere decir que el modelo debe aprender que una secuencia invertida es significativamente diferente de la original. El eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.

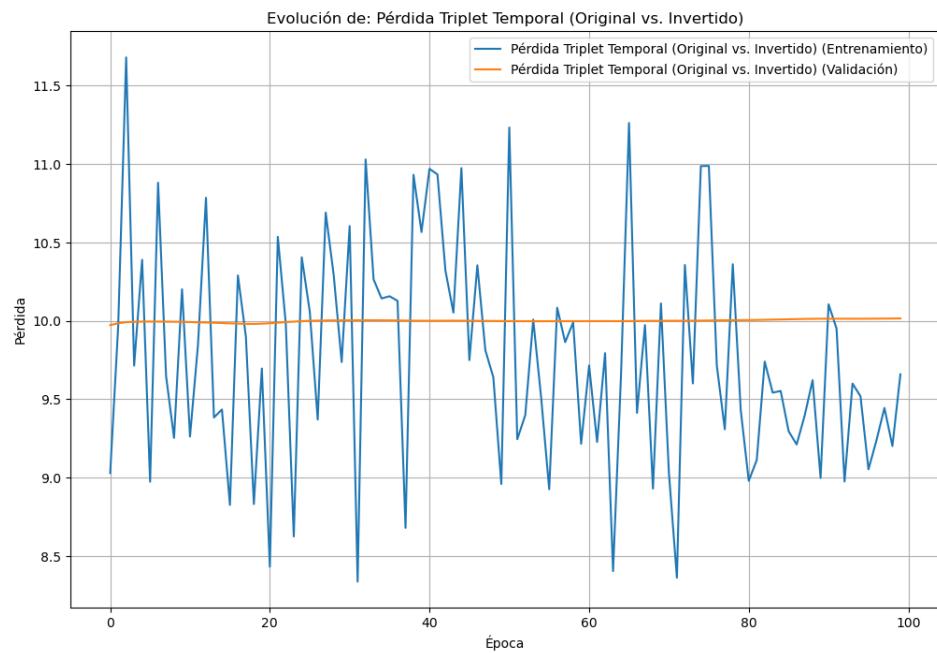
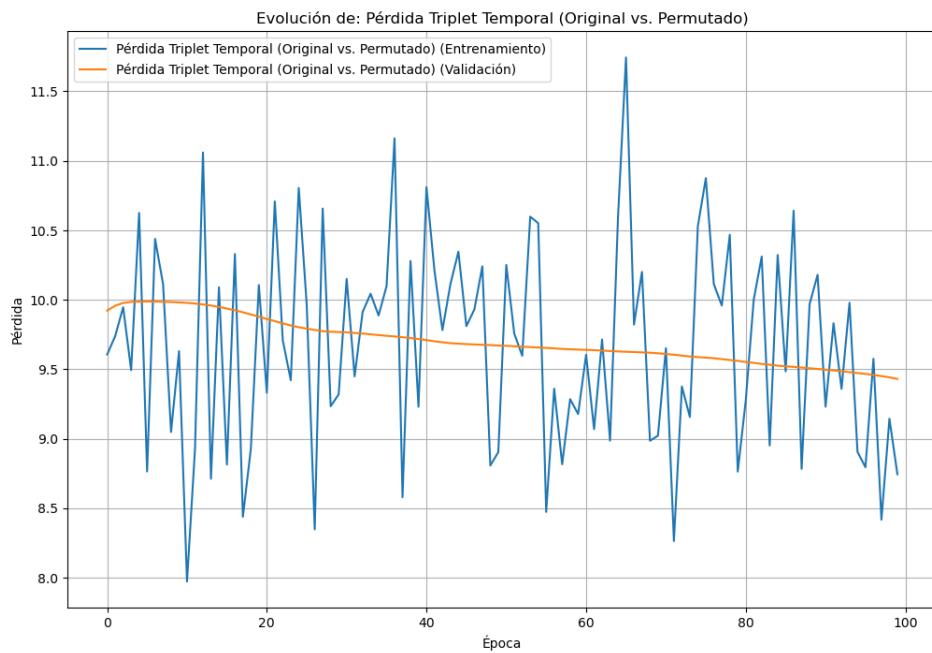


Figura 10.82: Similar a la gráfica anterior, esta también evalúa la sensibilidad temporal, pero en este caso, compara la secuencia original con una versión donde los fotogramas han sido desordenados aleatoriamente. Donde el objetivo es que el modelo reconozca que una secuencia permutada es muy diferente de la original. El eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.



1. Gráficas de los Experimentos Realizados

Figura 10.83: Esta grafica muestra el espacio latente en la mejor epoca (100) utilizando pca, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis invertida.

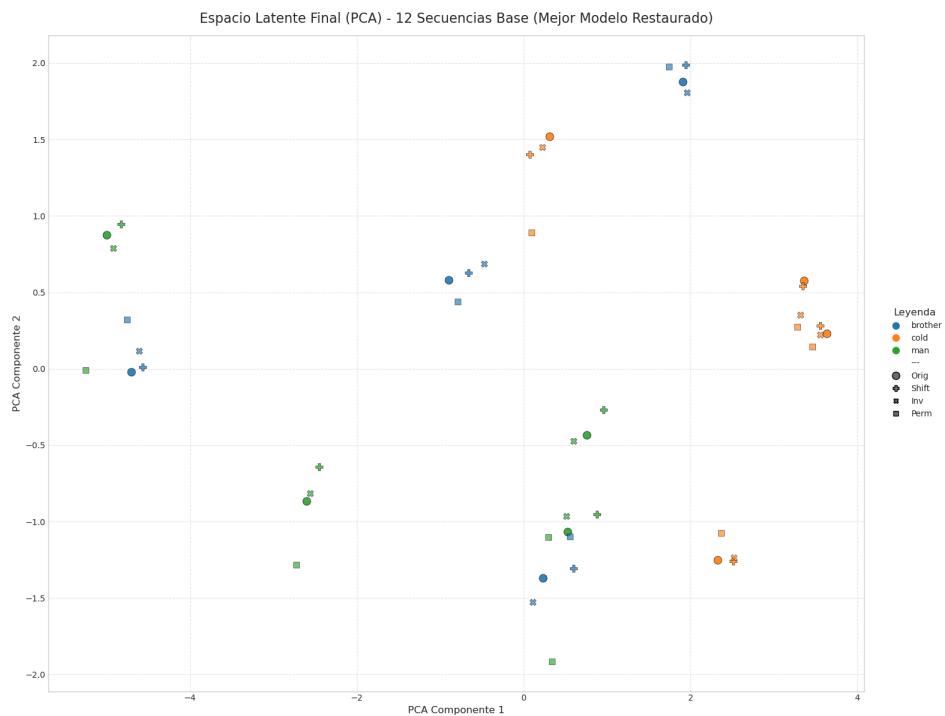
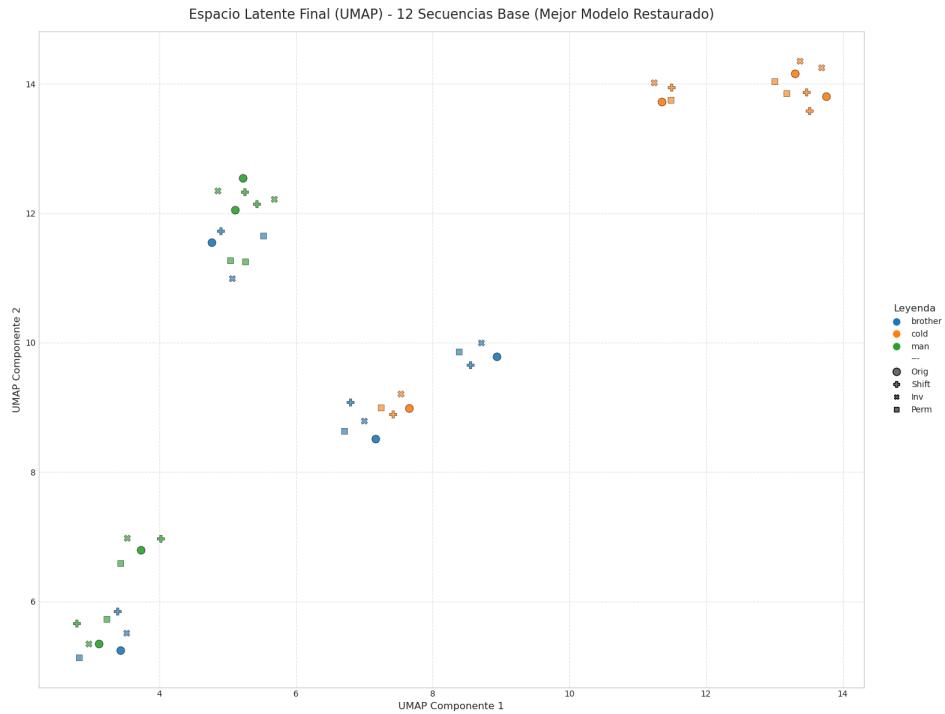


Figura 10.84: Esta grafica muestra el espacio latente en la mejor epoca (100) utilizando umap, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.85: Esta grafica muestra el espacio latente en la epoca 5 utilizando pca, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

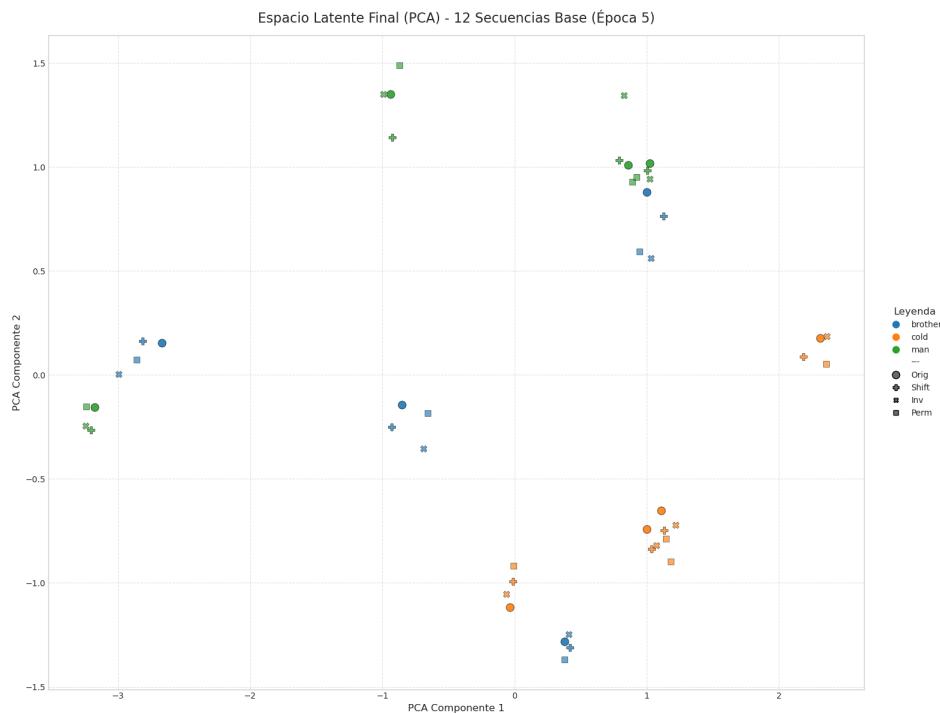
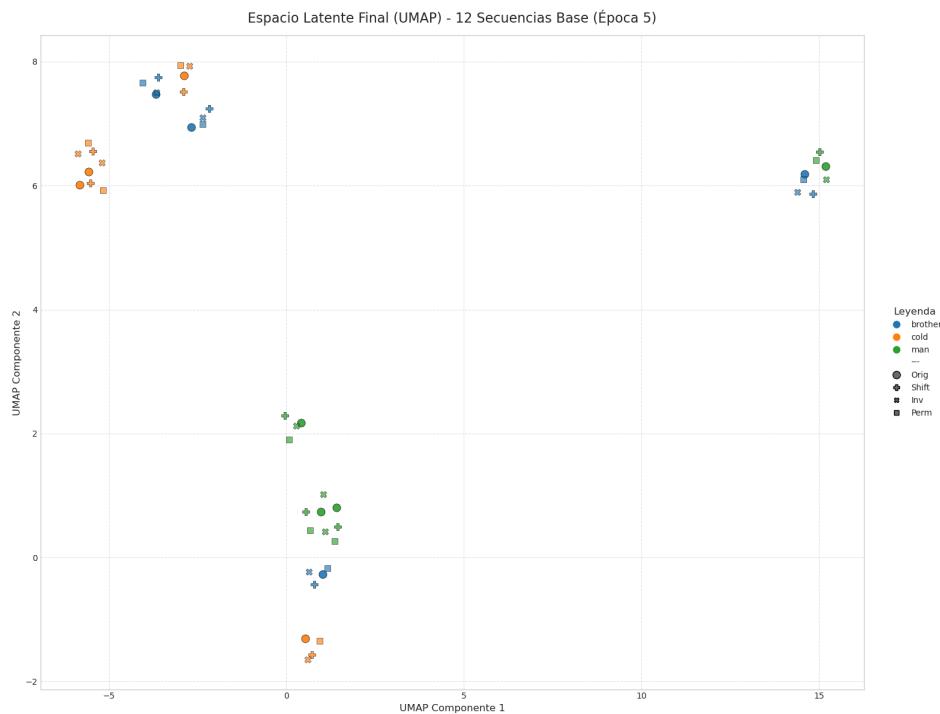


Figura 10.86: Esta grafica muestra el espacio latente en la epoca 5 utilizando umap, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.87: Esta grafica muestra el espacio latente en la epoca 15 utilizando pca, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

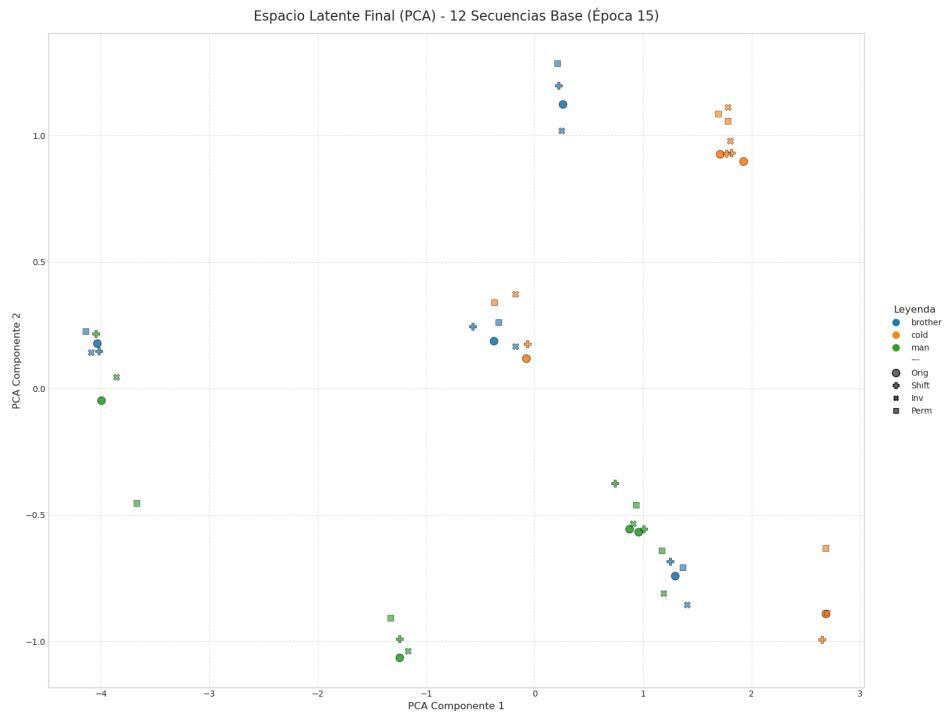
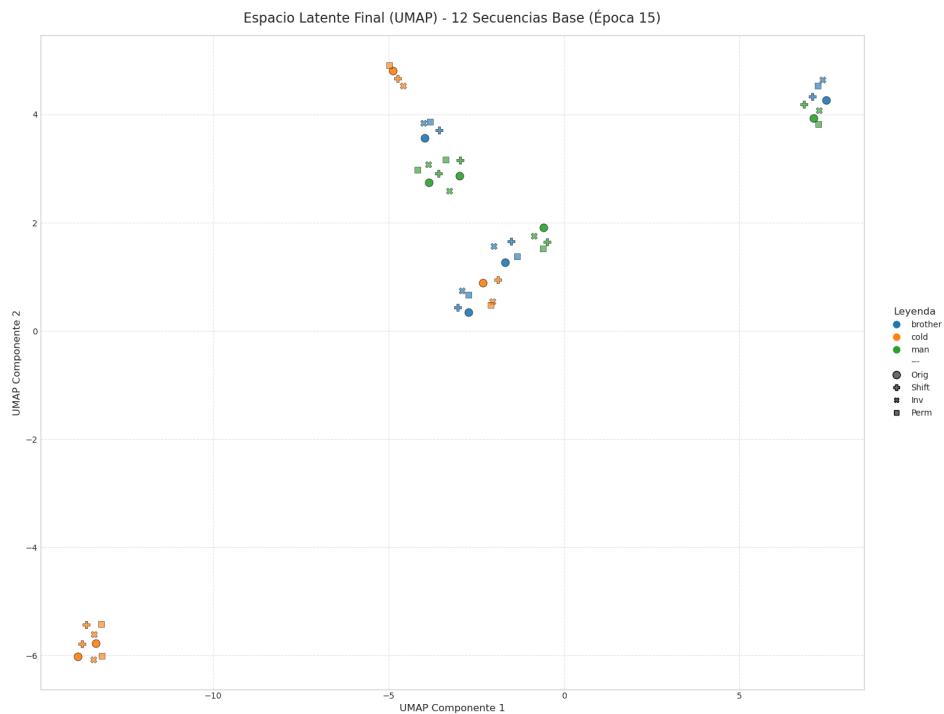


Figura 10.88: Esta grafica muestra el espacio latente en la epoca 15 utilizando umap, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.89: Esta grafica muestra el espacio latente en la epoca 25 utilizando pca, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

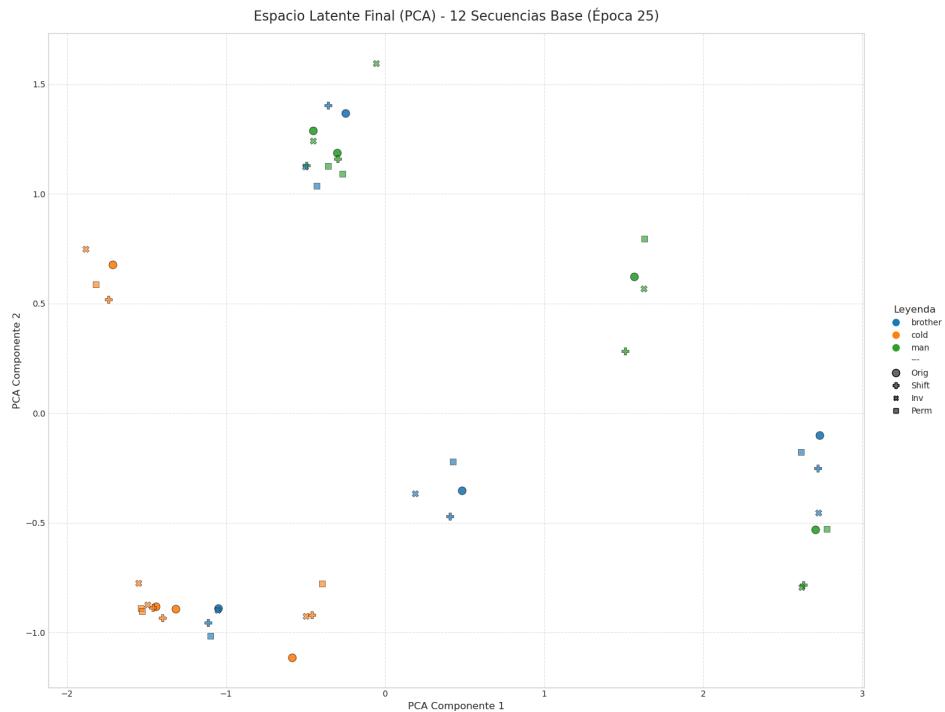
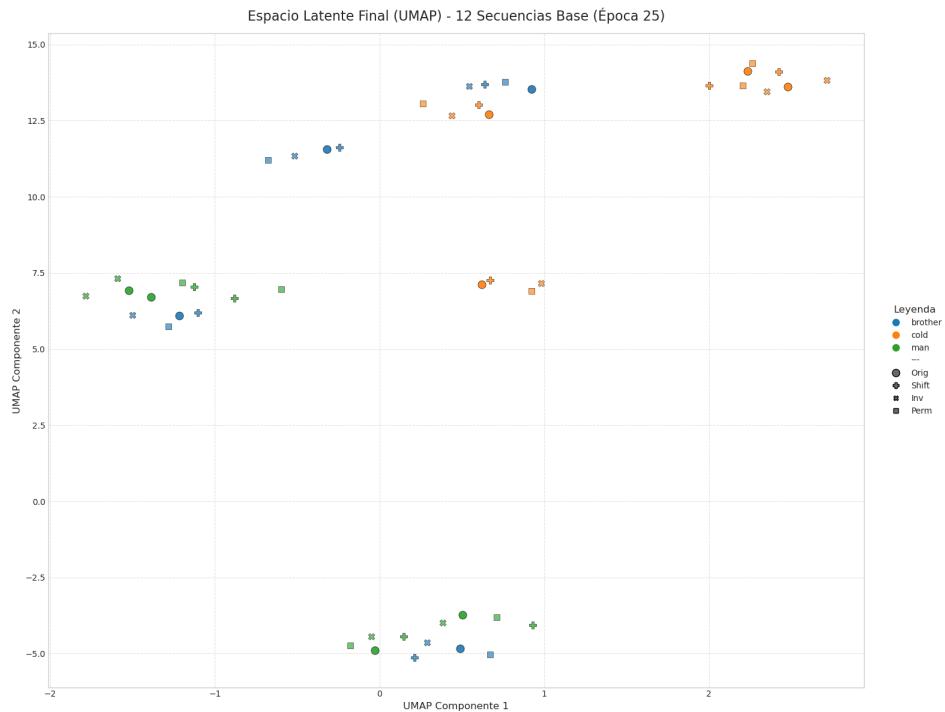


Figura 10.90: Esta grafica muestra el espacio latente en la epoca 25 utilizando umap, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.91: Esta grafica muestra el espacio latente en la epoca 45 utilizando pca, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

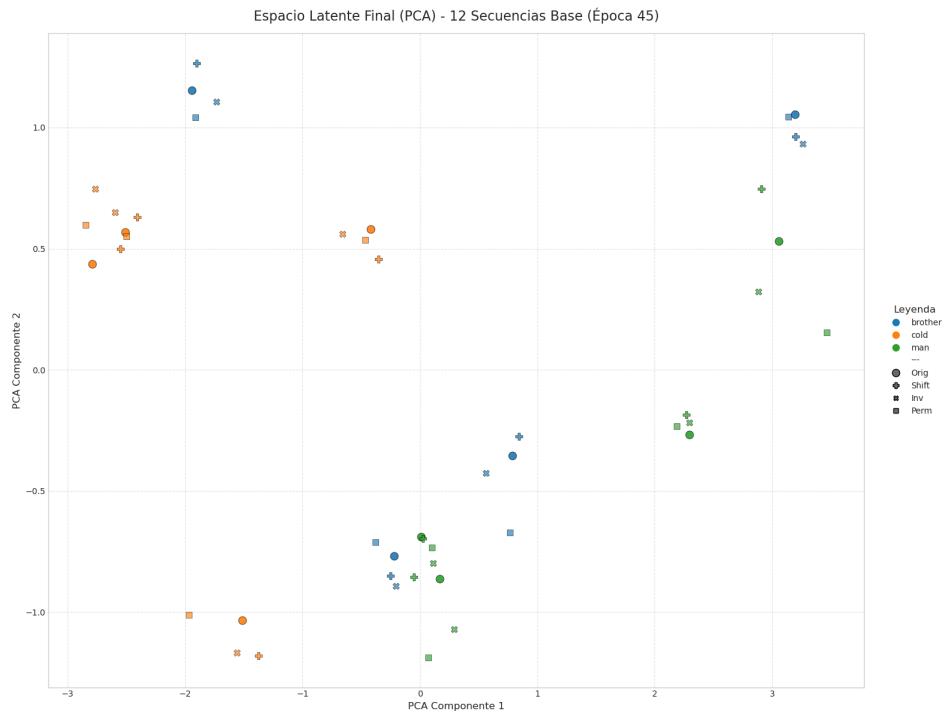
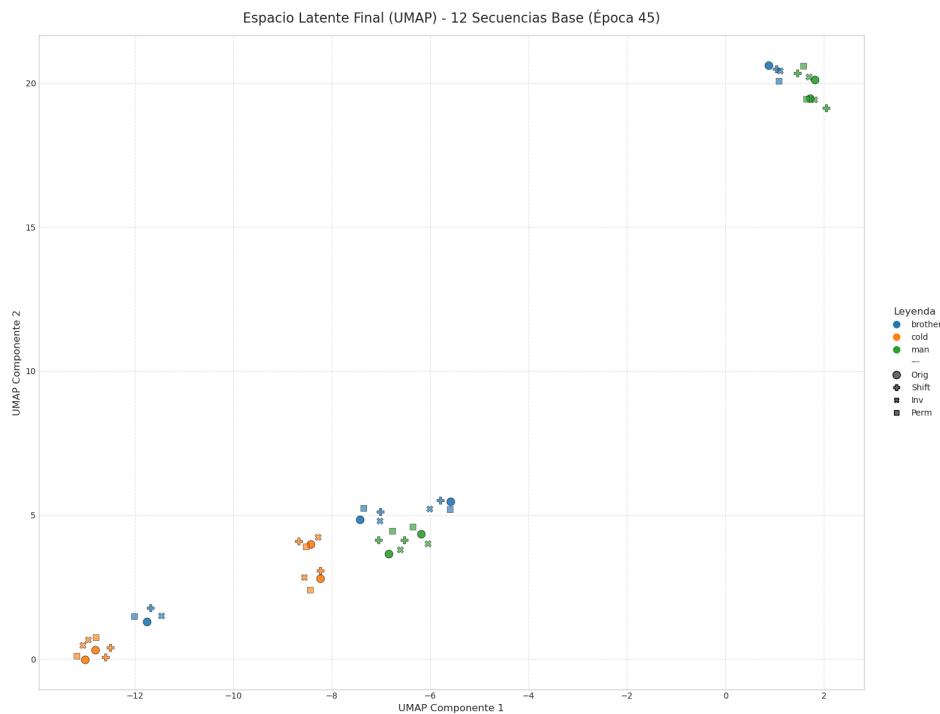


Figura 10.92: Esta grafica muestra el espacio latente en la epoca 45 utilizando umap, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.93: Esta grafica muestra el espacio latente en la epoca 65 utilizando pca, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

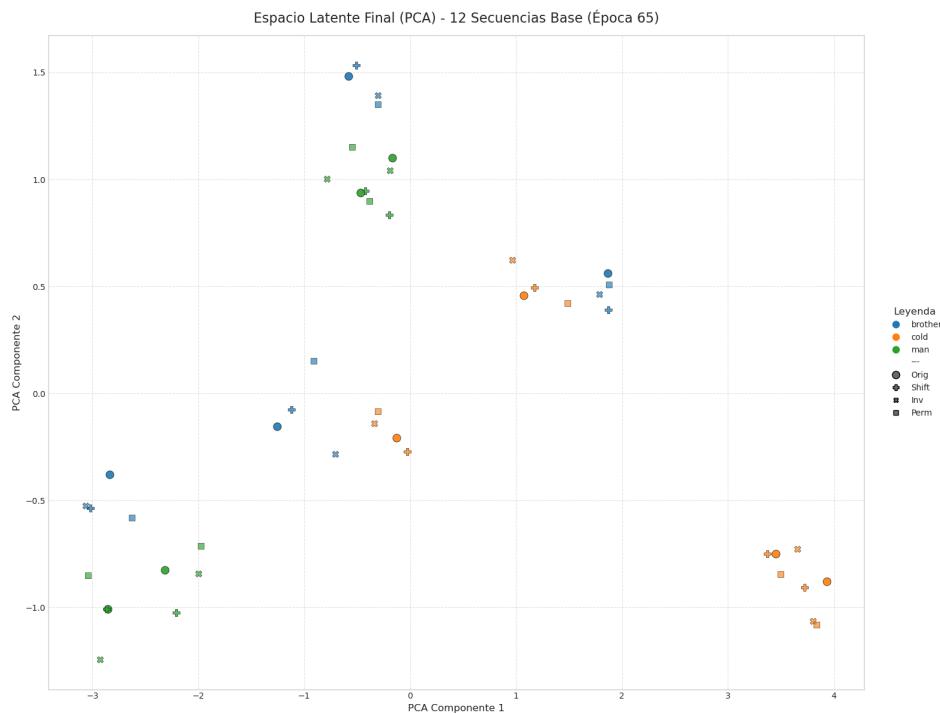
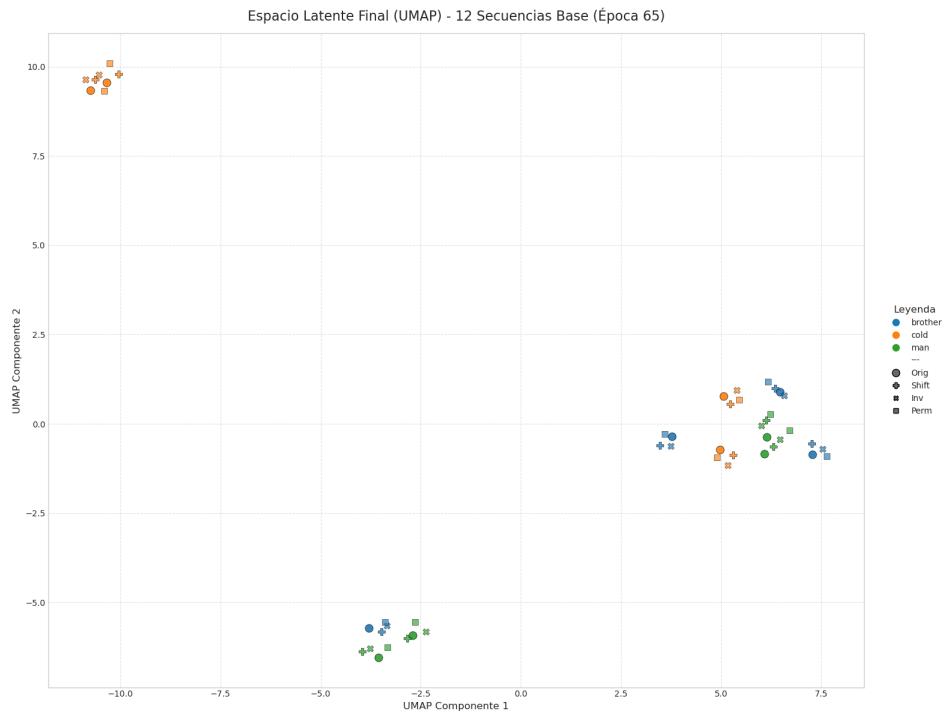


Figura 10.94: Esta grafica muestra el espacio latente en la epoca 65 utilizando umap, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.95: Esta grafica muestra el espacio latente en la epoca 85 utilizando pca, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

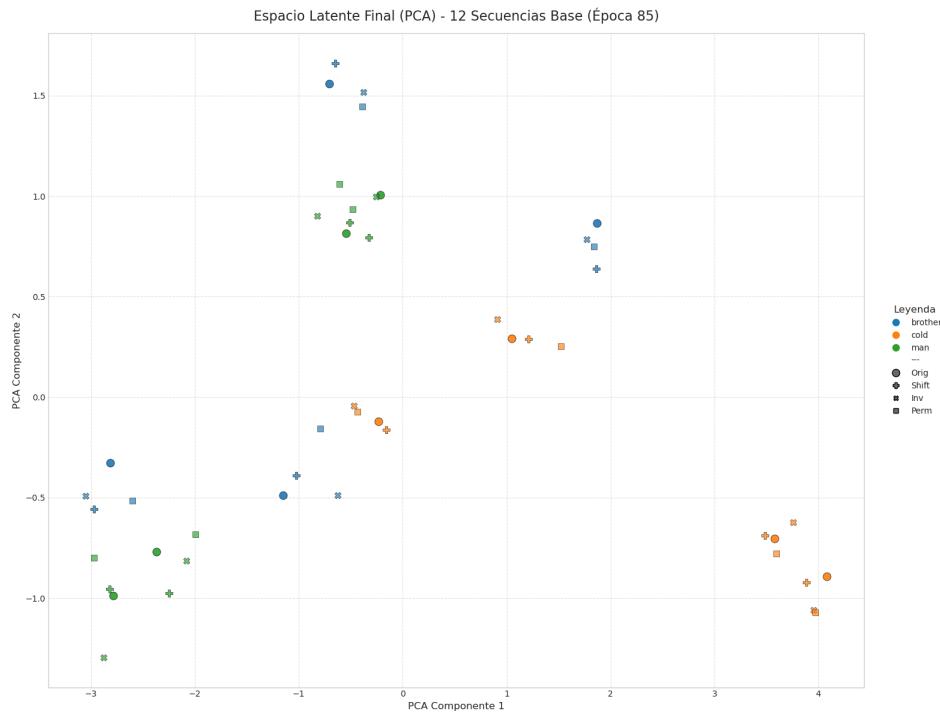
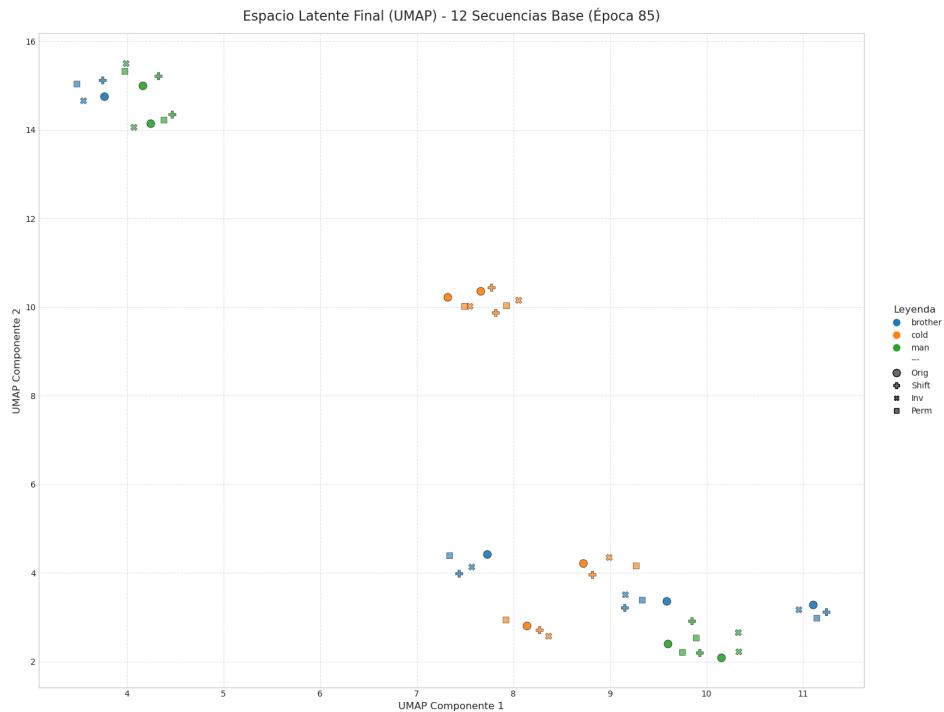


Figura 10.96: Esta grafica muestra el espacio latente en la epoca 85 utilizando umap, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

1.3. Con el dataset de SLOVO

Con 2 etiquetas

Figura 10.97: Esta gráfica compara el ratio Semántico del modelo comparado con un «baseline» o punto de referencia. Siendo este el modelo sin entrenar. El eje Y representa el ratio semántico, que mide la capacidad del modelo para diferenciar entre palabras diferentes, mientras que el eje X representa las épocas de entrenamiento.

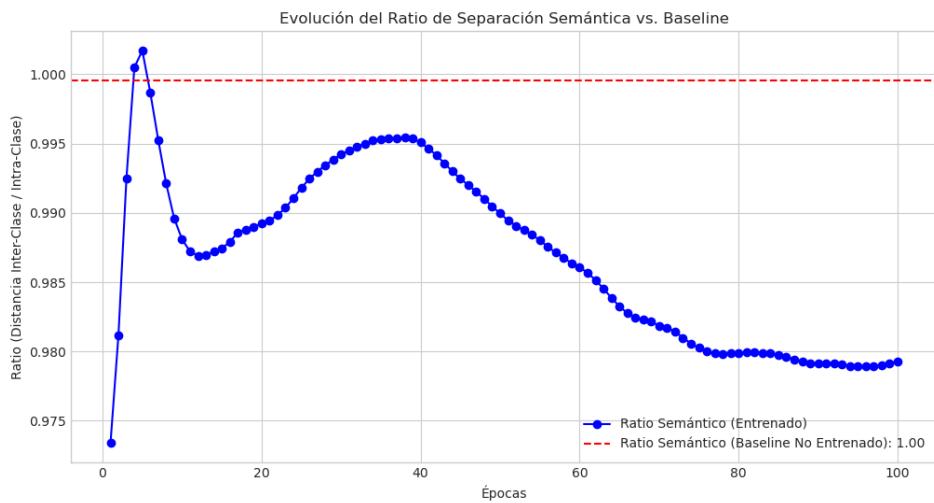


Figura 10.98: Esta gráfica evalúa la capacidad del modelo para entender el orden temporal de las secuencias de video comparadas con sus respectivos «baselines» del modelo no entrenado. Muestra la distancia euclídea promedio entre la secuencia original y sus versiones alteradas (shifted, inverted, permuted).

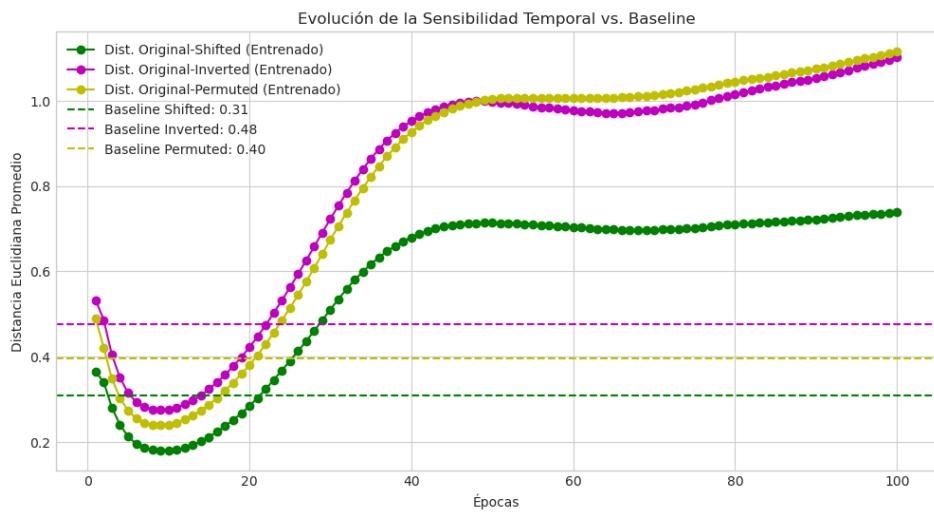


Figura 10.99: Esta gráfica compara el rendimiento del modelo principal con un modelo más simple en términos de la pérdida total de validación. El eje Y representa el valor de la pérdida, una métrica que indica cuán bien el modelo está aprendiendo, donde los valores más bajos son mejores, y el eje X representa las épocas.

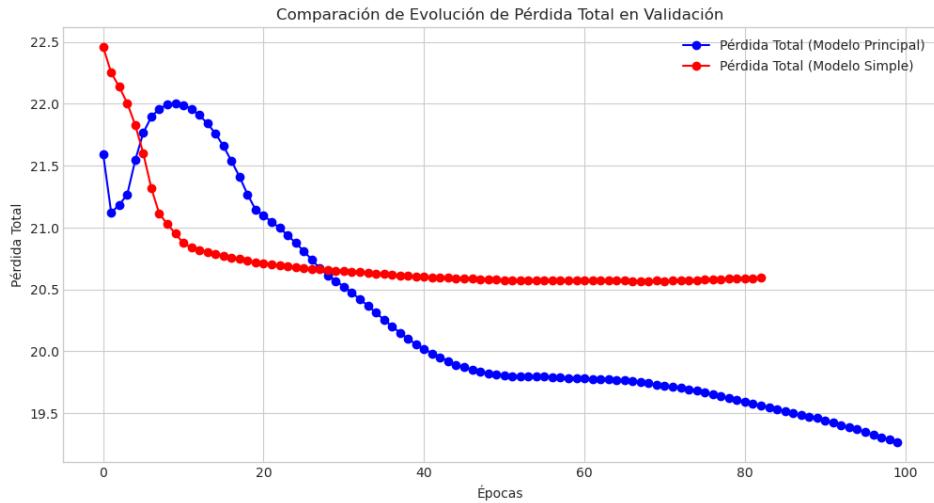
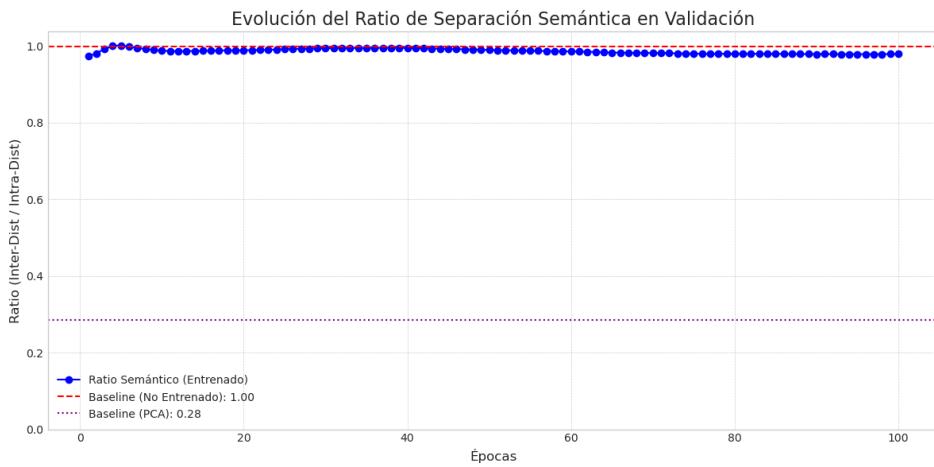


Figura 10.100: Esta gráfica compara el ratio semántico del modelo comparado con dos «baselines» o puntos de referencia. Siendo estos el modelo sin entrenar y la representación de PCA. El eje Y representa el ratio semántico, que mide la capacidad del modelo para diferenciar entre palabras diferentes, mientras que el eje X representa las épocas de entrenamiento.



1. Gráficas de los Experimentos Realizados

Figura 10.101: Este gráfico mide directamente la calidad de la separación semántica en el espacio latente para palabras con la misma y diferente clase. El eje Y representa la distancia euclíadiana promedio y el eje X son las épocas.

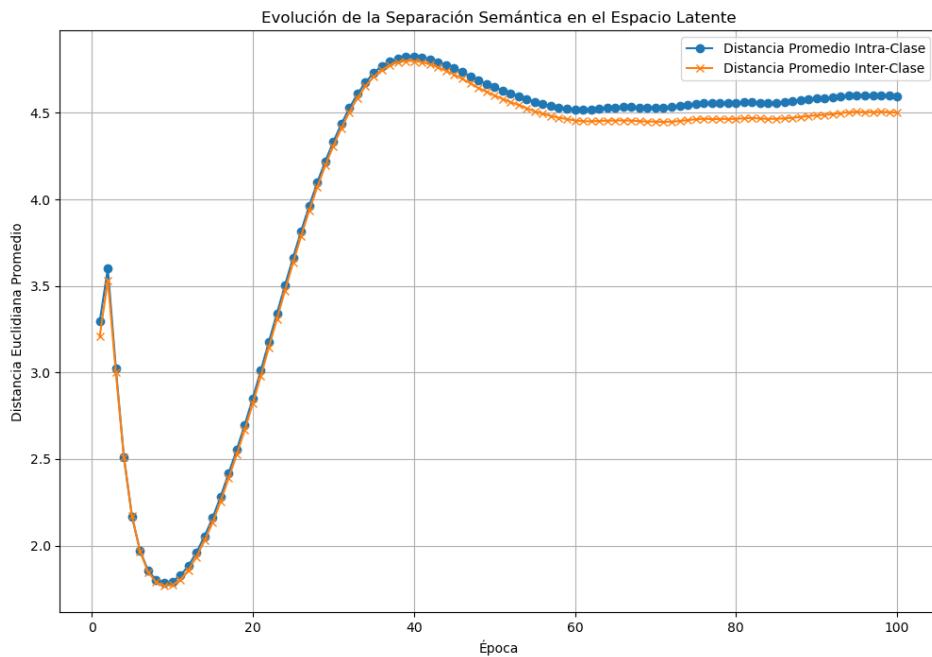
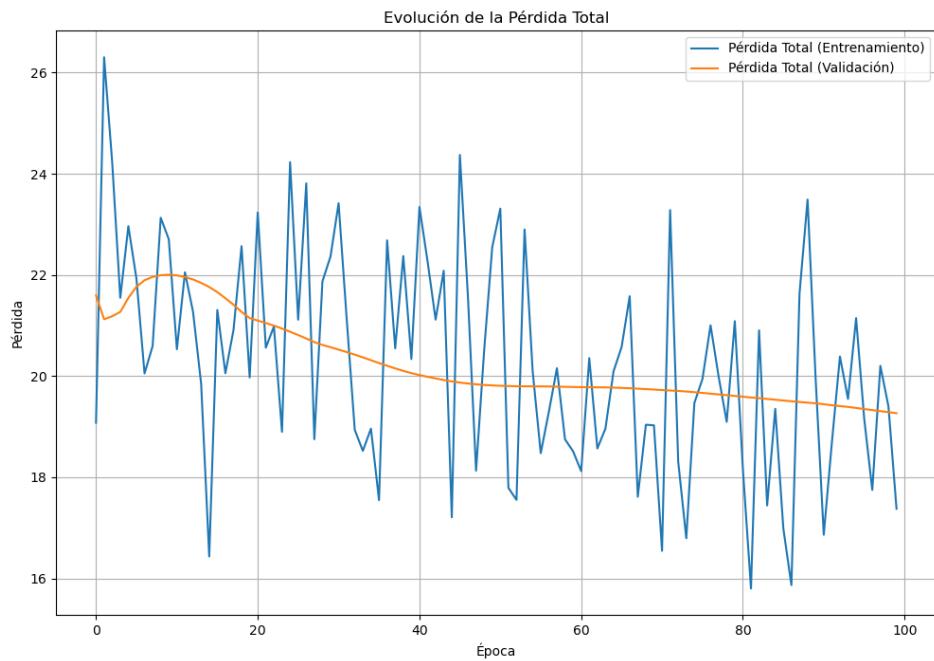


Figura 10.102: Este gráfico muestra la evolución de la «Pérdida Total» a lo largo de 100 épocas de entrenamiento. El eje Y representa el valor de la pérdida, una métrica que indica cuán bien el modelo está aprendiendo donde valores más bajos son mejores. El eje X representa las épocas, es decir, cada ciclo completo de entrenamiento sobre el conjunto de datos.



1. Gráficas de los Experimentos Realizados

Figura 10.103: Esta gráfica ilustra la «Pérdida de Reconstrucción», que mide qué tan bien el autoencoder del modelo puede reconstruir la entrada original después de haberla comprimido en un espacio latente. Al igual que en la gráfica anterior, el eje Y es el valor de la pérdida y el eje X son las épocas.

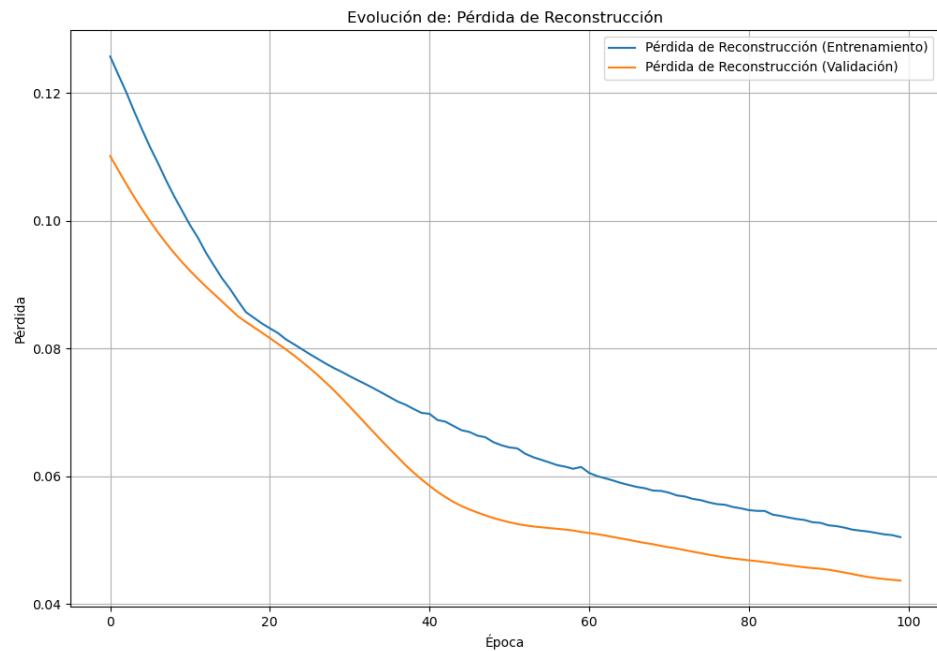
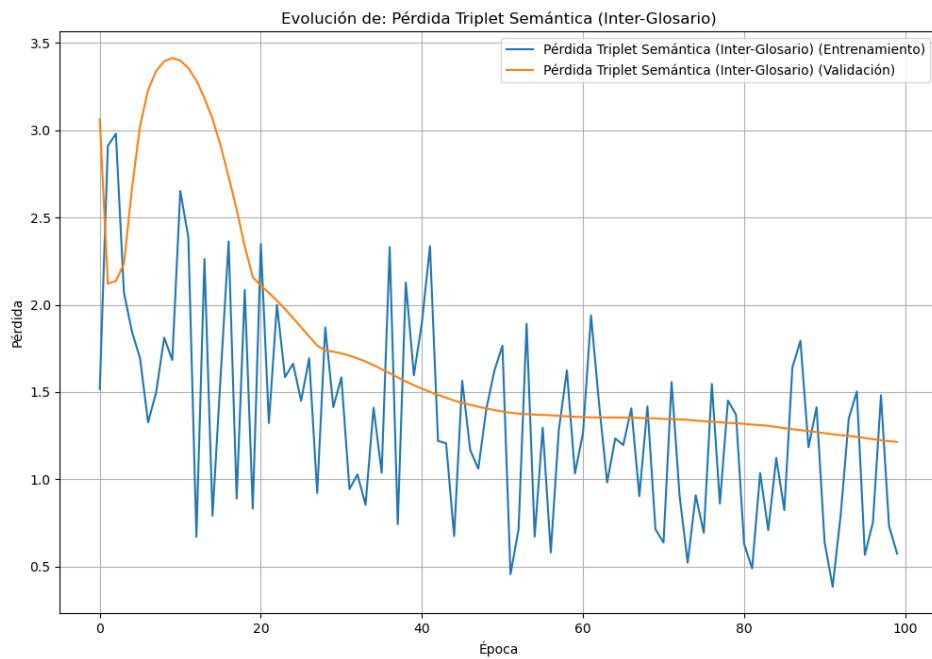


Figura 10.104: Este gráfico muestra la «Pérdida Triplet Semántica», una métrica clave que evalúa si el modelo puede diferenciar entre distintos glosarios, en este caso, las señas «brother» y «cold». El objetivo es que las representaciones de un mismo glosario estén más cerca entre sí que las de glosarios diferentes. Igualmente, el eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.



1. Gráficas de los Experimentos Realizados

Figura 10.105: Esta visualización se enfoca en la sensibilidad temporal del modelo, puesto que, mide la diferencia entre la secuencia original de un video y su versión invertida. Esto quiere decir que el modelo debe aprender que una secuencia invertida es significativamente diferente de la original. El eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.

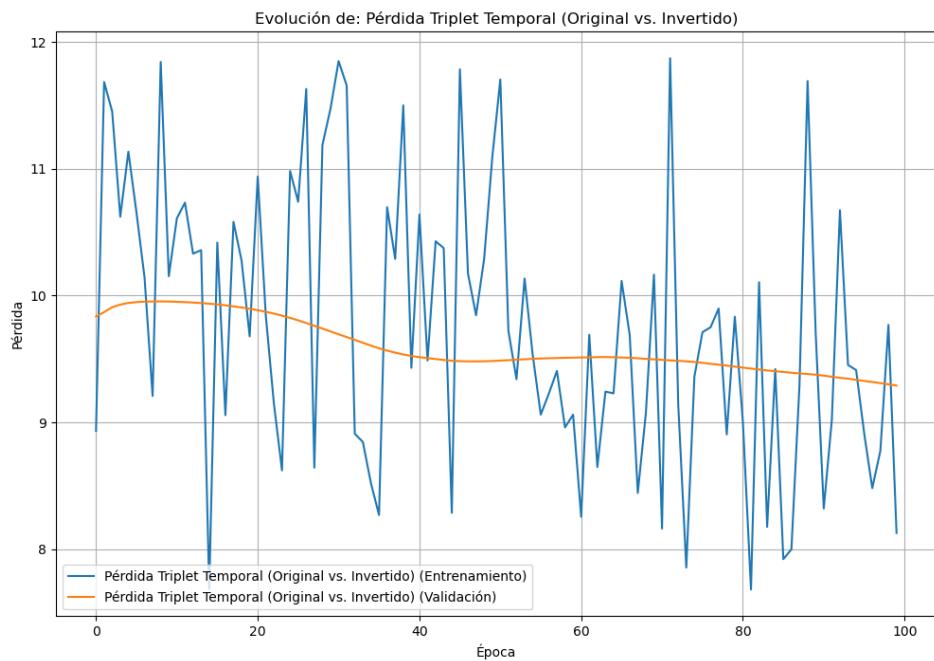
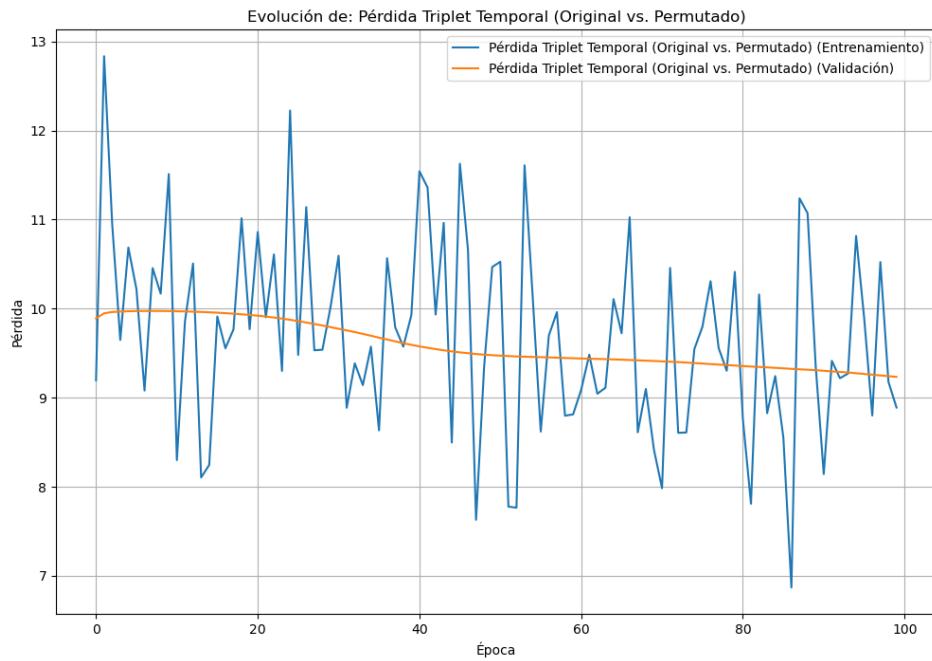


Figura 10.106: Similar a la gráfica anterior, esta también evalúa la sensibilidad temporal, pero en este caso, compara la secuencia original con una versión donde los fotogramas han sido desordenados aleatoriamente. Donde el objetivo es que el modelo reconozca que una secuencia permutada es muy diferente de la original. El eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.



1. Gráficas de los Experimentos Realizados

Figura 10.107: Esta grafica muestra el espacio latente en la mejor epoca (100) utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

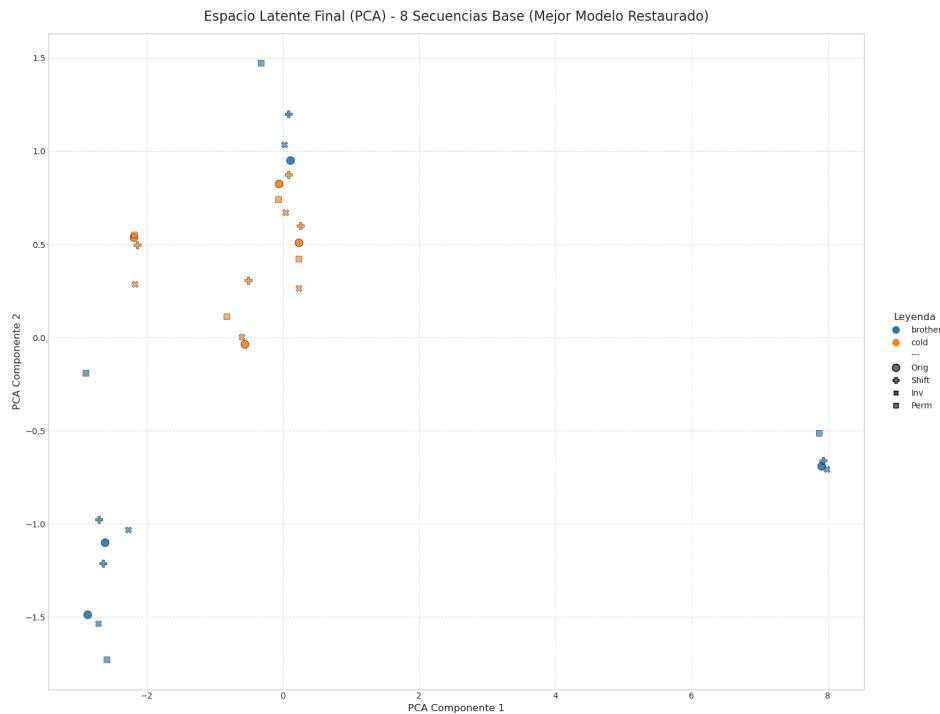
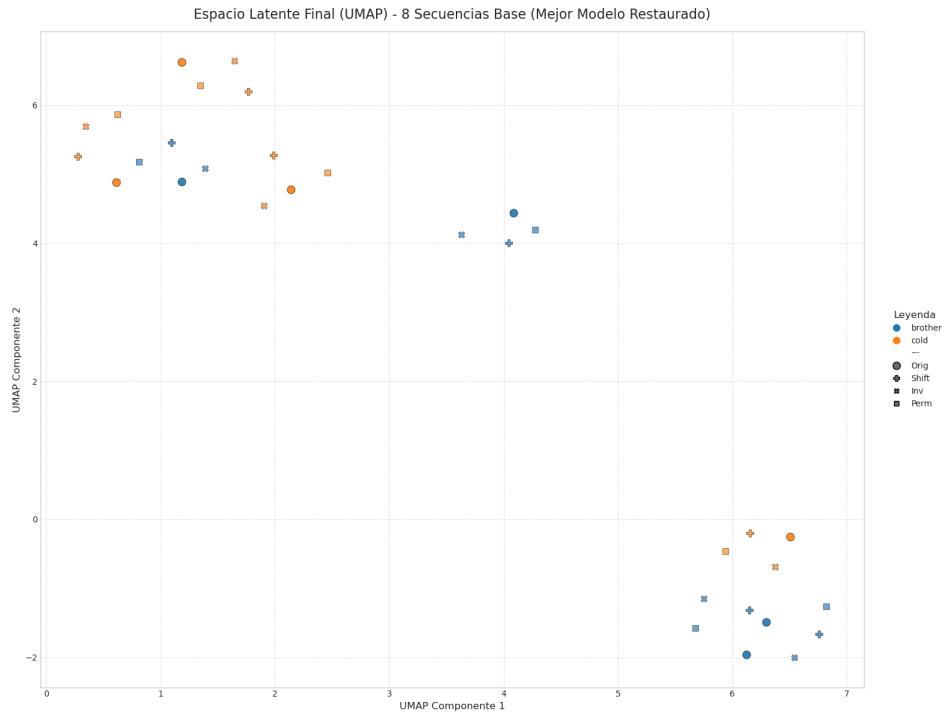


Figura 10.108: Esta grafica muestra el espacio latente en la mejor epoca (100) utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.109: Esta grafica muestra el espacio latente en la epoca 5 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

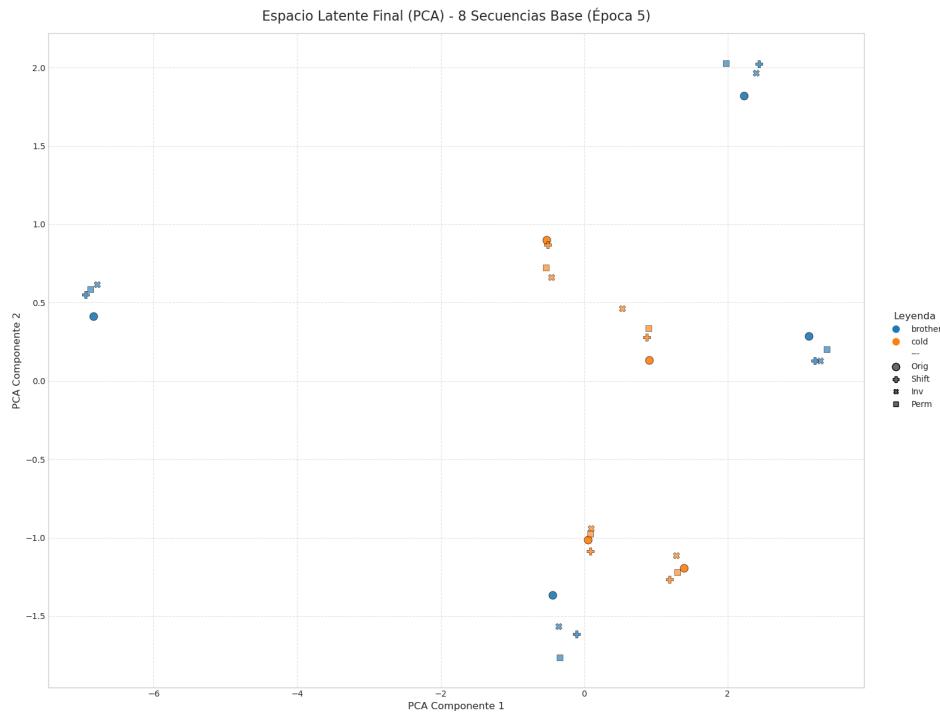
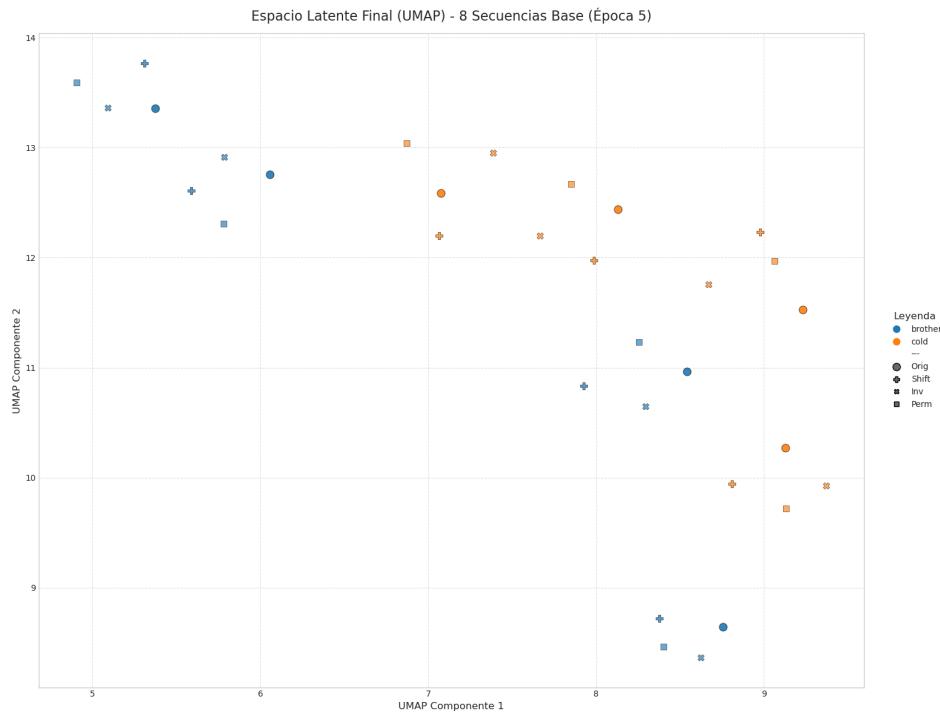


Figura 10.110: Esta grafica muestra el espacio latente en la epoca 5 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.111: Esta grafica muestra el espacio latente en la epoca 15 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

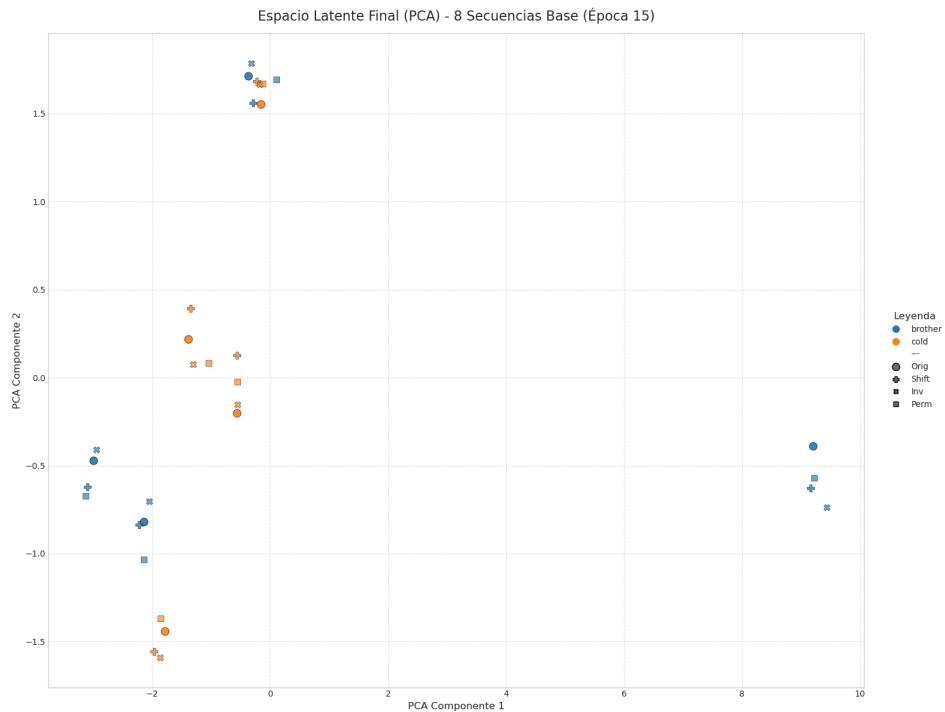
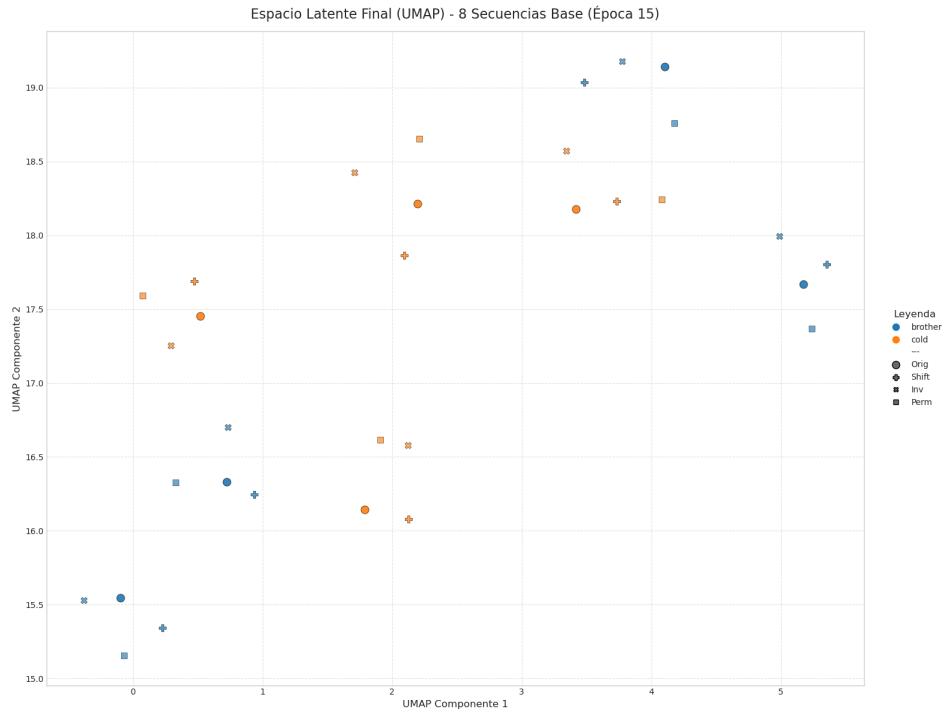


Figura 10.112: Esta grafica muestra el espacio latente en la epoca 15 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.113: Esta grafica muestra el espacio latente en la epoca 25 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

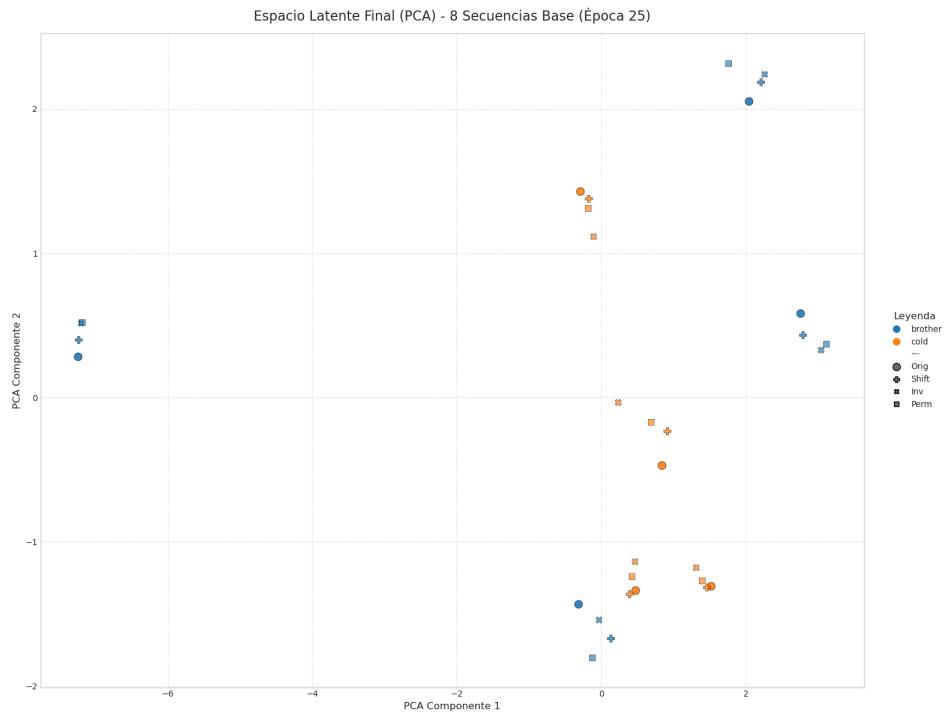
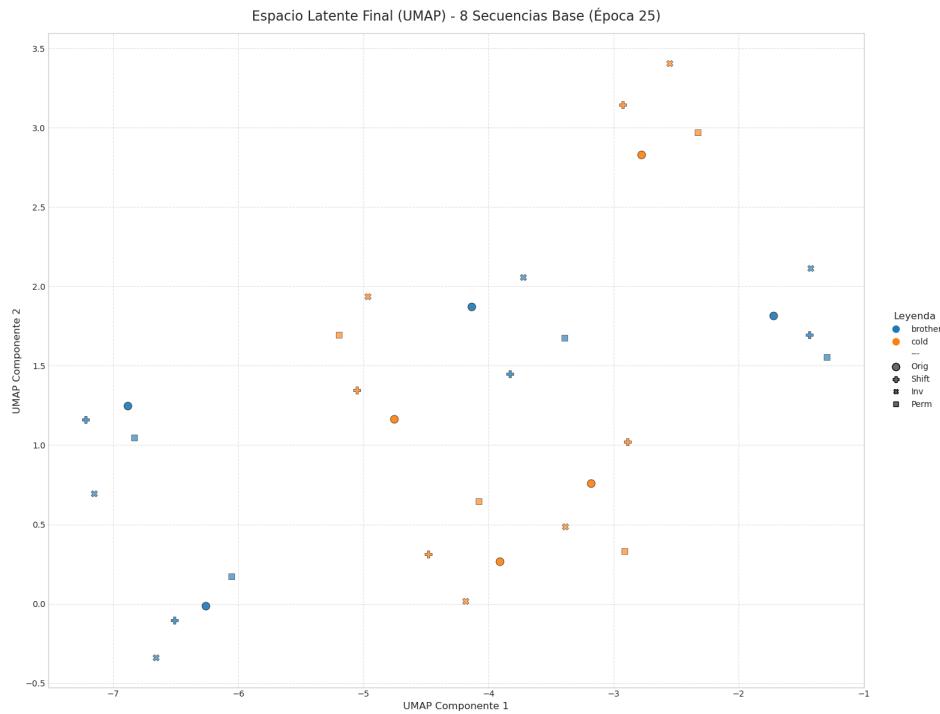


Figura 10.114: Esta grafica muestra el espacio latente en la epoca 25 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.115: Esta grafica muestra el espacio latente en la epoca 45 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

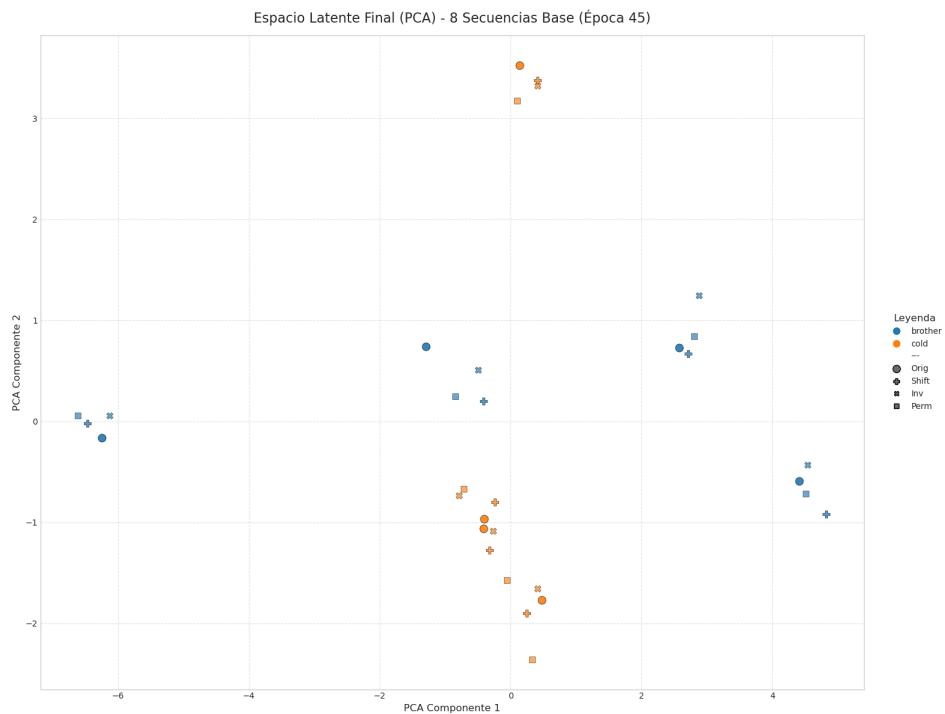
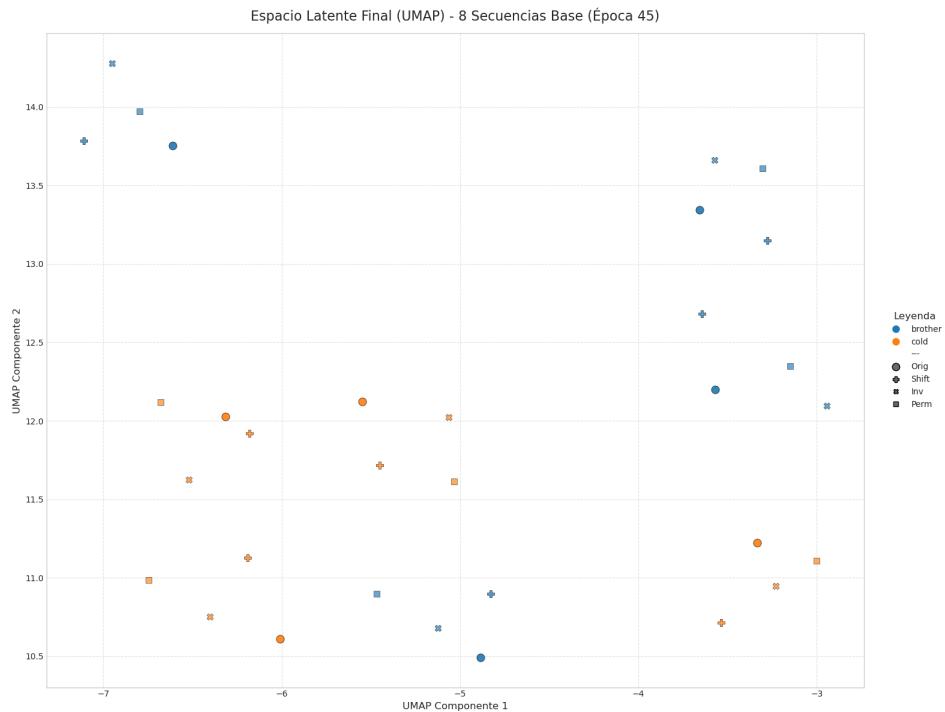


Figura 10.116: Esta grafica muestra el espacio latente en la epoca 45 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.117: Esta grafica muestra el espacio latente en la epoca 65 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

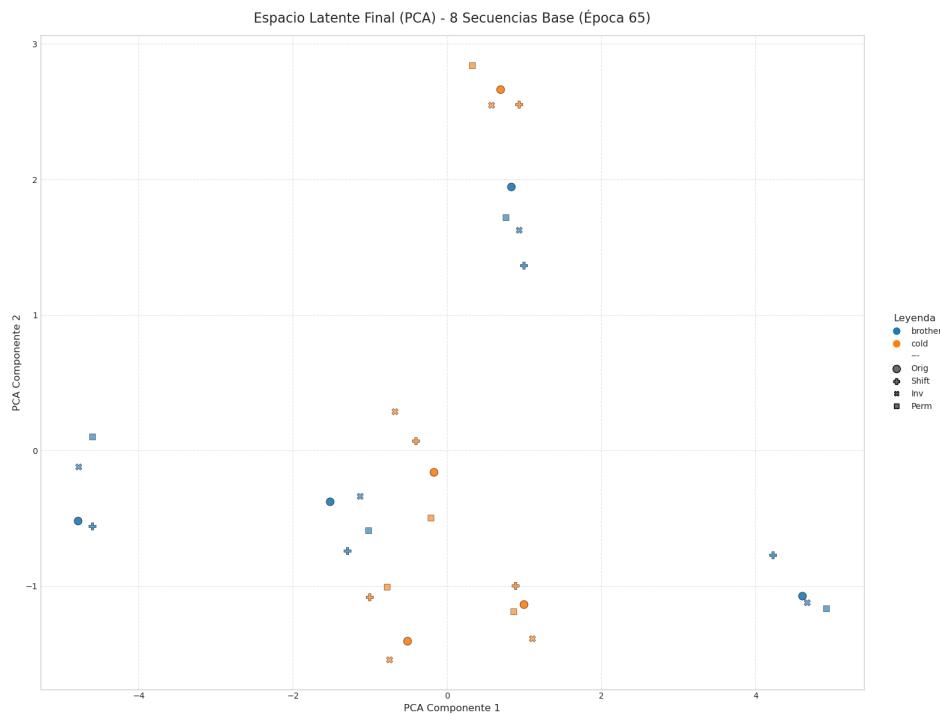
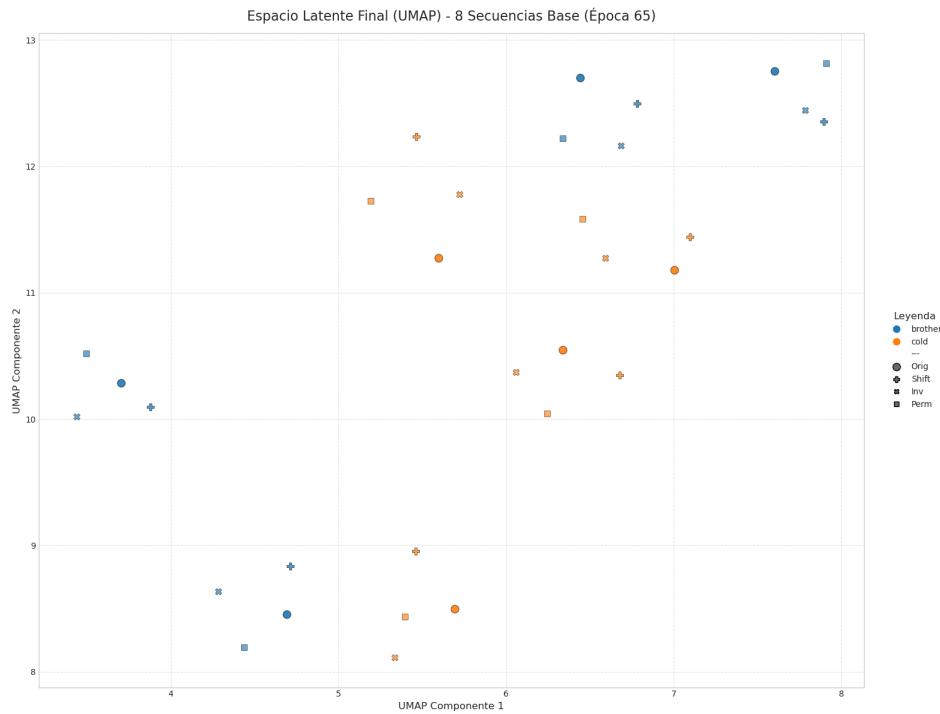


Figura 10.118: Esta grafica muestra el espacio latente en la epoca 65 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.119: Esta grafica muestra el espacio latente en la epoca 85 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

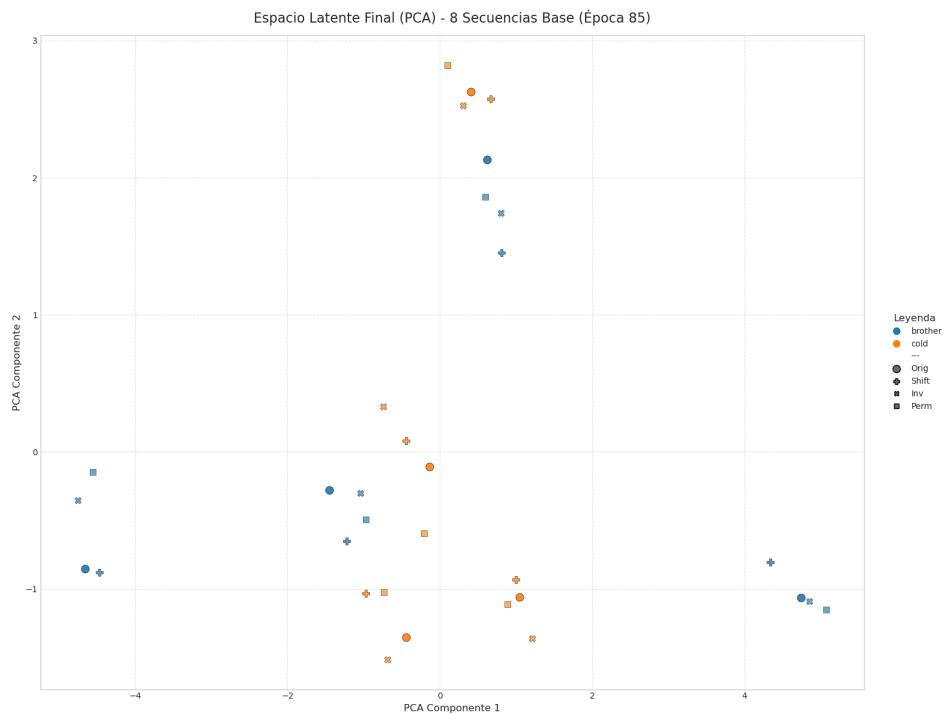
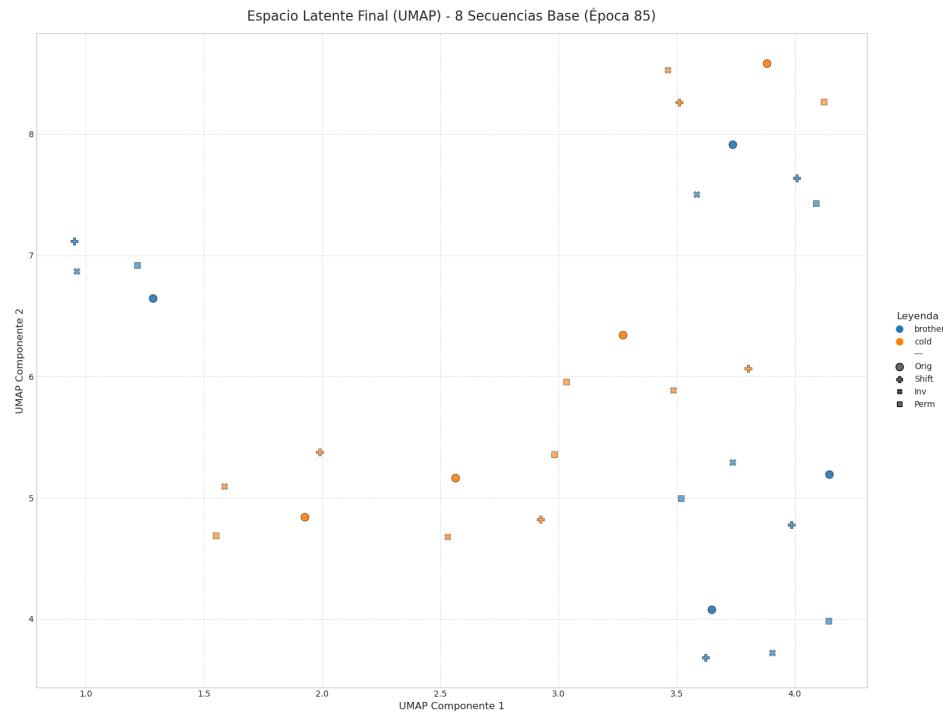


Figura 10.120: Esta grafica muestra el espacio latente en la epoca 85 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Con 3 etiquetas

Figura 10.121: Esta gráfica compara el ratio Semántico del modelo comparado con un «baseline» o punto de referencia. Siendo este el modelo sin entrenar. El eje Y representa el ratio semántico, que mide la capacidad del modelo para diferenciar entre palabras diferentes, mientras que el eje X representa las épocas de entrenamiento.

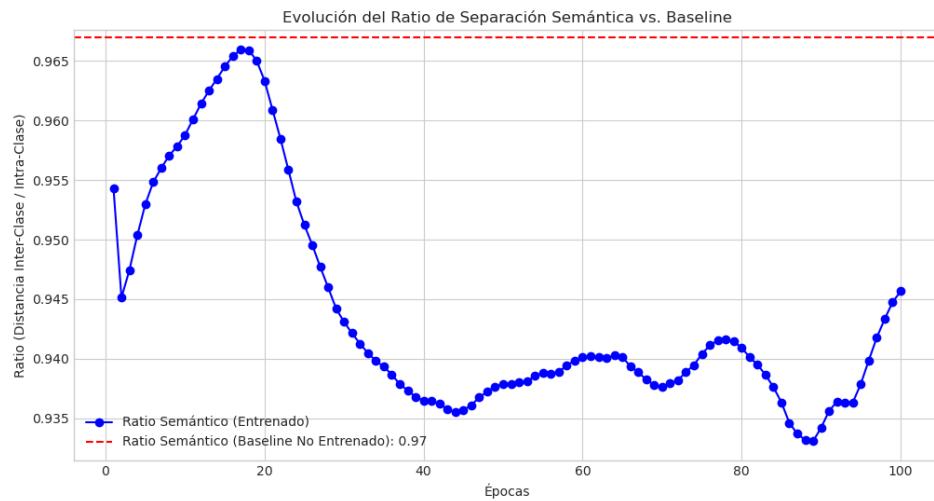


Figura 10.122: Esta gráfica evalúa la capacidad del modelo para entender el orden temporal de las secuencias de video comparadas con sus respectivos «baselines» del modelo no entrenado. Muestra la distancia euclíadiana promedio entre la secuencia original y sus versiones alteradas (shifted, inverted, permuted).

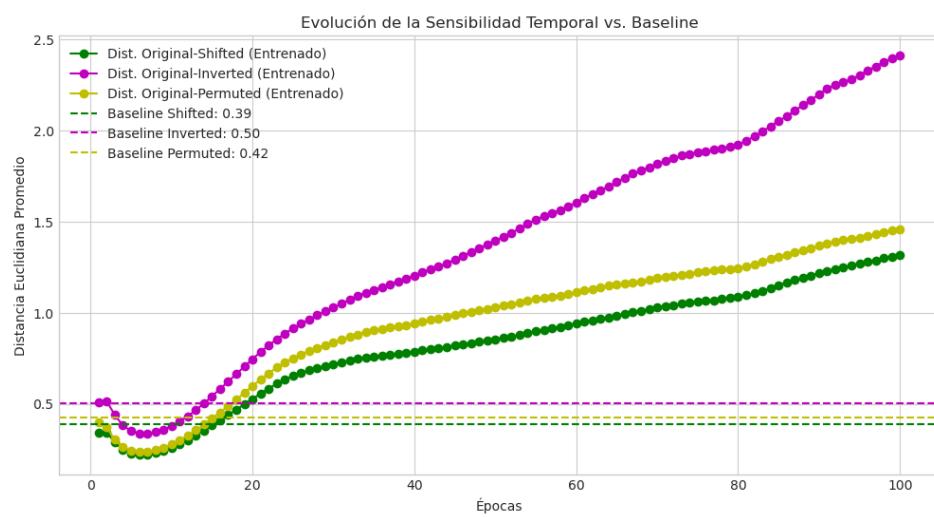


Figura 10.123: Esta gráfica compara el rendimiento del modelo principal con un modelo más simple en términos de la pérdida total de validación. El eje Y representa el valor de la pérdida, una métrica que indica cuán bien el modelo está aprendiendo, donde los valores más bajos son mejores, y el eje X representa las épocas.

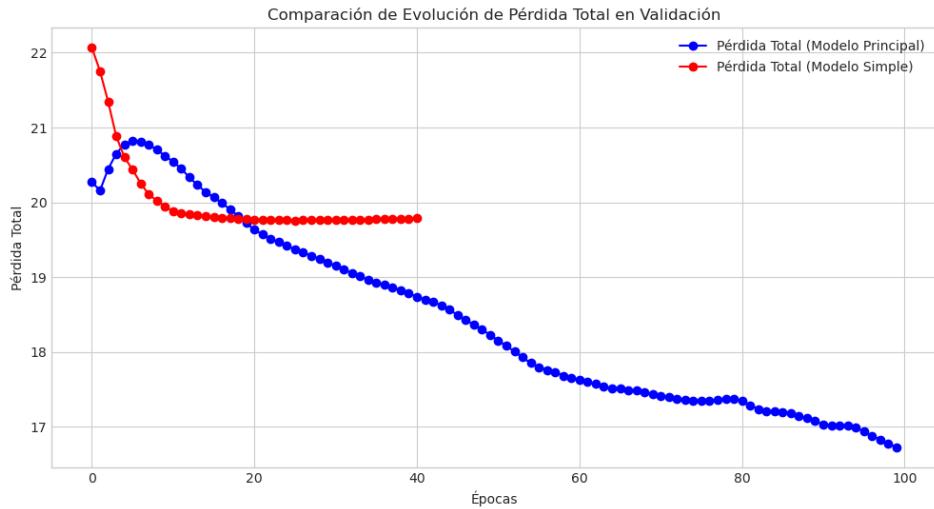
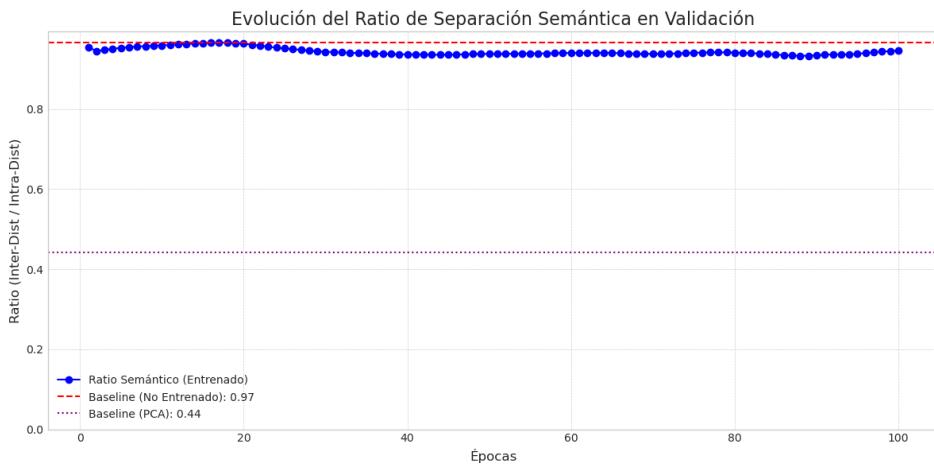
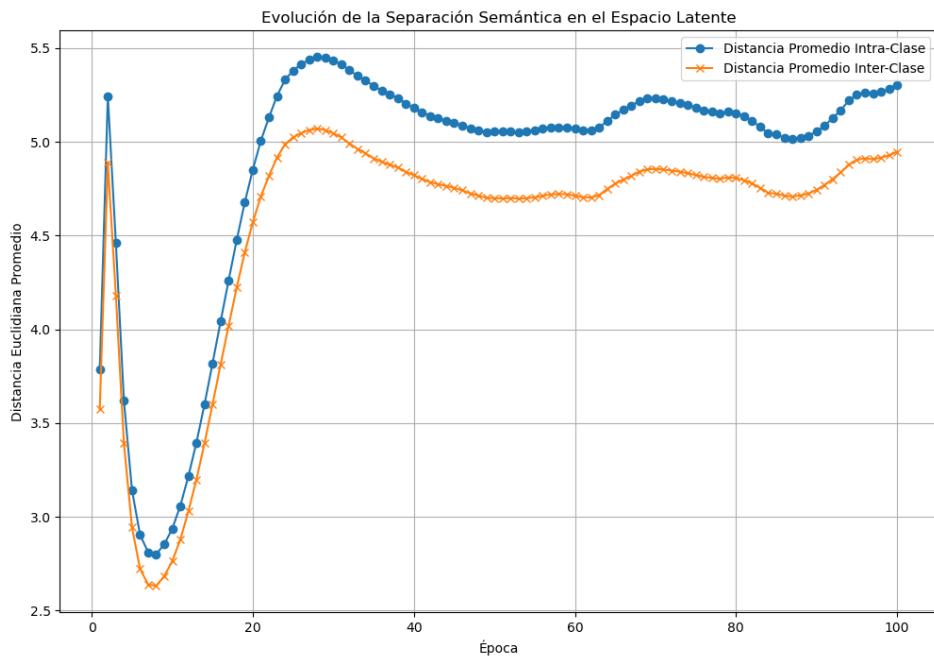


Figura 10.124: Esta gráfica compara el ratio semántico del modelo comparado con dos «baselines» o puntos de referencia. Siendo estos el modelo sin entrenar y la representación de PCA. El eje Y representa el ratio semántico, que mide la capacidad del modelo para diferenciar entre palabras diferentes, mientras que el eje X representa las épocas de entrenamiento.



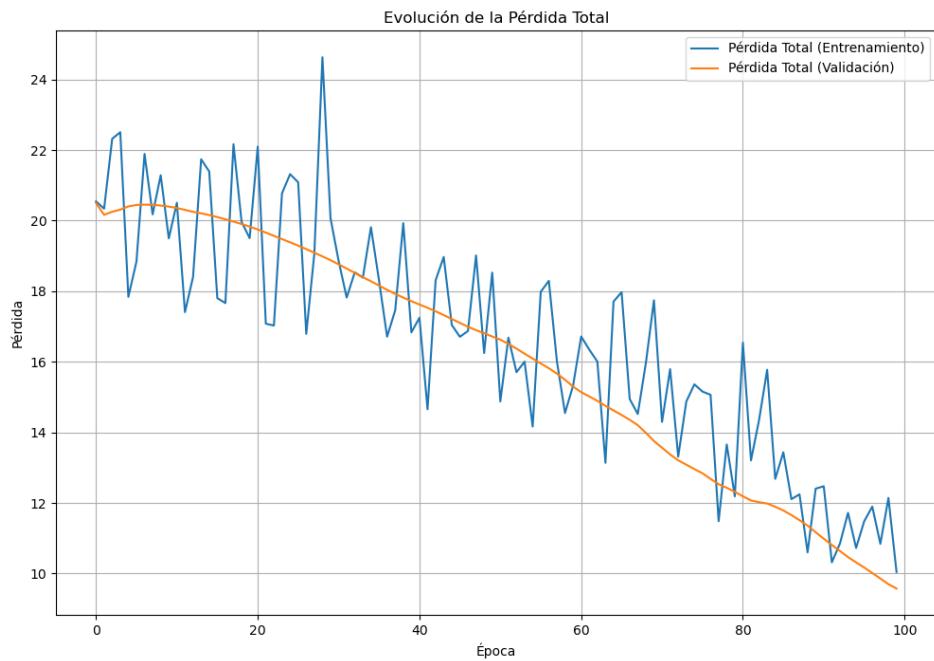
1. Gráficas de los Experimentos Realizados

Figura 10.125: Este gráfico mide directamente la calidad de la separación semántica en el espacio latente para palabras con la misma y diferente clase. El eje Y representa la distancia euclíadiana promedio y el eje X son las épocas.



UNIVERSIDAD
SERGIO ARBOLEDA

Figura 10.126: Este gráfico muestra la evolución de la «Pérdida Total» a lo largo de 100 épocas de entrenamiento. El eje Y representa el valor de la pérdida, una métrica que indica cuán bien el modelo está aprendiendo donde valores más bajos son mejores. El eje X representa las épocas, es decir, cada ciclo completo de entrenamiento sobre el conjunto de datos.



1. Gráficas de los Experimentos Realizados

Figura 10.127: Esta gráfica ilustra la «Pérdida de Reconstrucción», que mide qué tan bien el autoencoder del modelo puede reconstruir la entrada original después de haberla comprimido en un espacio latente. Al igual que en la gráfica anterior, el eje Y es el valor de la pérdida y el eje X son las épocas.

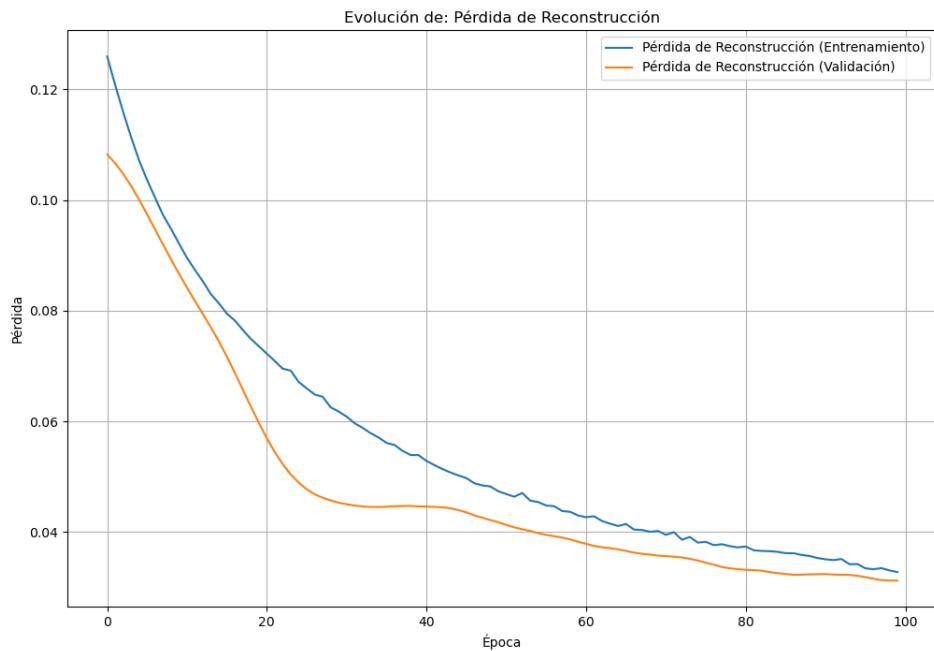
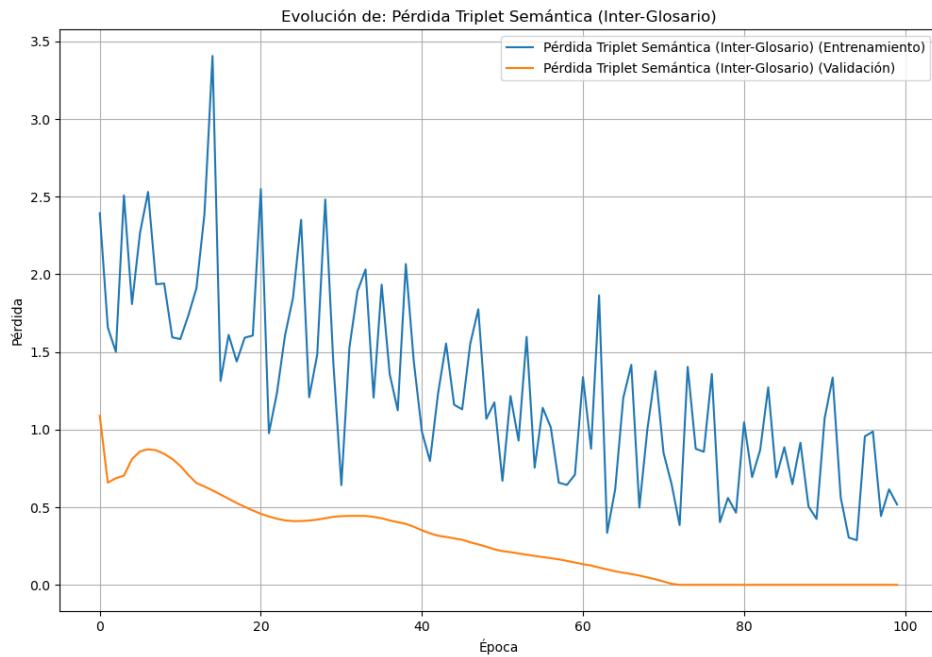


Figura 10.128: Este gráfico muestra la «Pérdida Triplet Semántica», una métrica clave que evalúa si el modelo puede diferenciar entre distintos glosarios, en este caso, las señas «brother», «cold» y «man». El objetivo es que las representaciones de un mismo glosario estén más cerca entre sí que las de glosarios diferentes. Igualmente, el eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.



1. Gráficas de los Experimentos Realizados

Figura 10.129: Esta visualización se enfoca en la sensibilidad temporal del modelo, puesto que, mide la diferencia entre la secuencia original de un video y su versión invertida. Esto quiere decir que el modelo debe aprender que una secuencia invertida es significativamente diferente de la original. El eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.

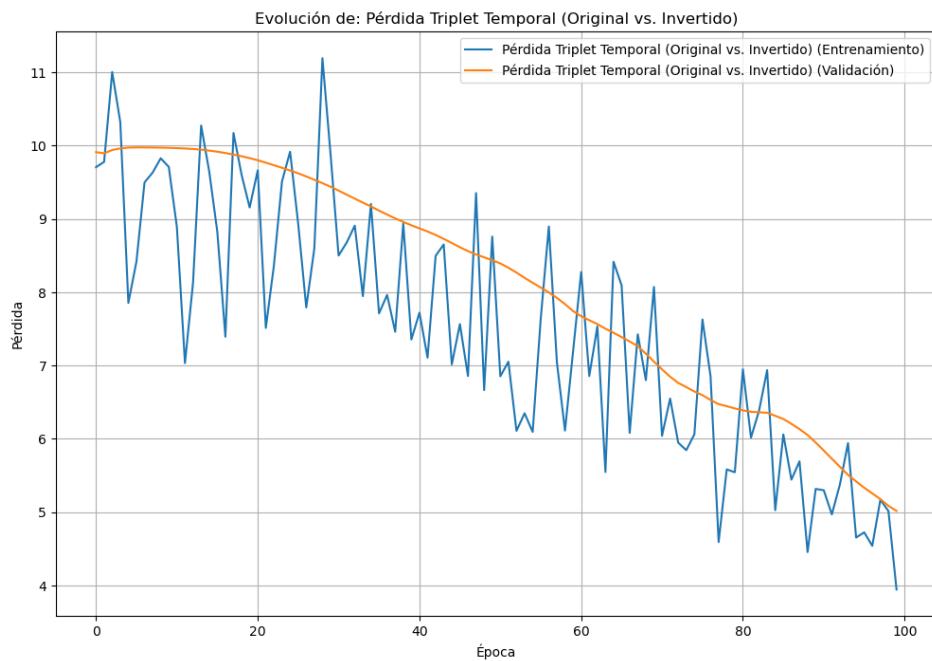
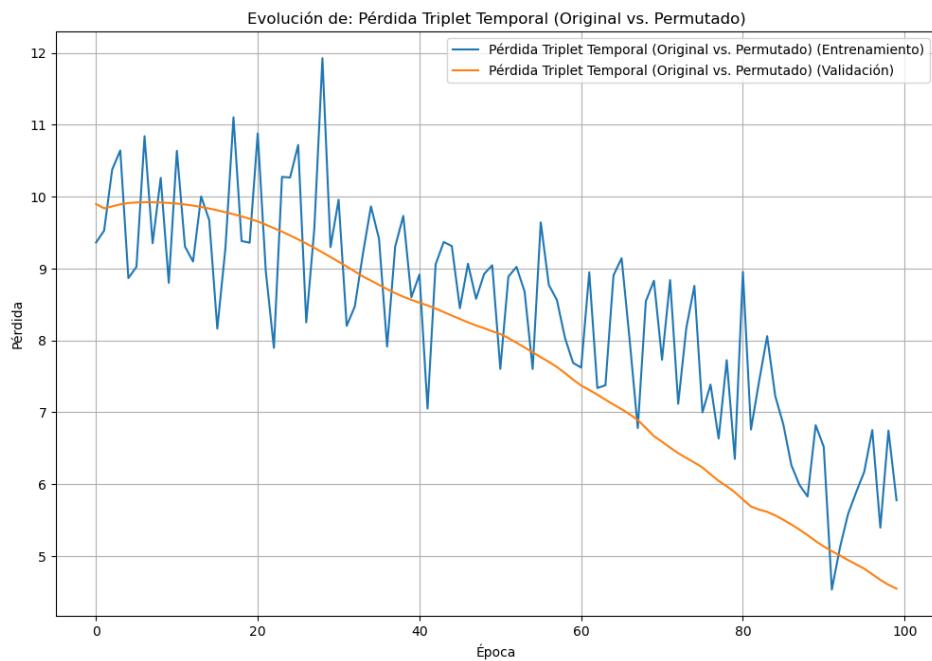


Figura 10.130: Similar a la gráfica anterior, esta también evalúa la sensibilidad temporal, pero en este caso, compara la secuencia original con una versión donde los fotogramas han sido desordenados aleatoriamente. Donde el objetivo es que el modelo reconozca que una secuencia permutada es muy diferente de la original. El eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.



1. Gráficas de los Experimentos Realizados

Figura 10.131: Esta grafica muestra el espacio latente en la mejor epoca (100) utilizando pca, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

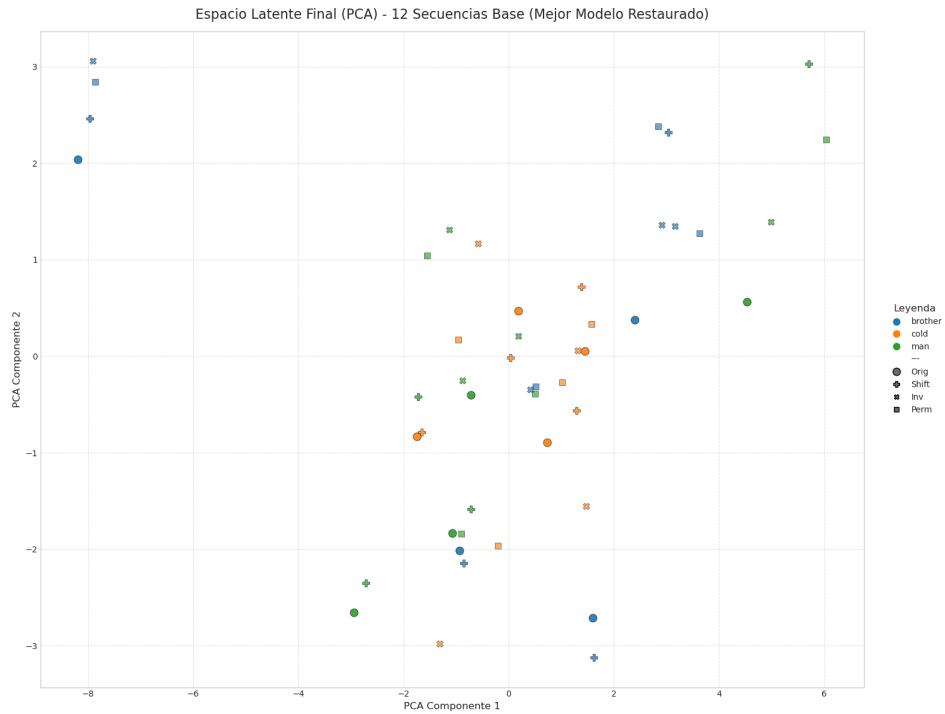
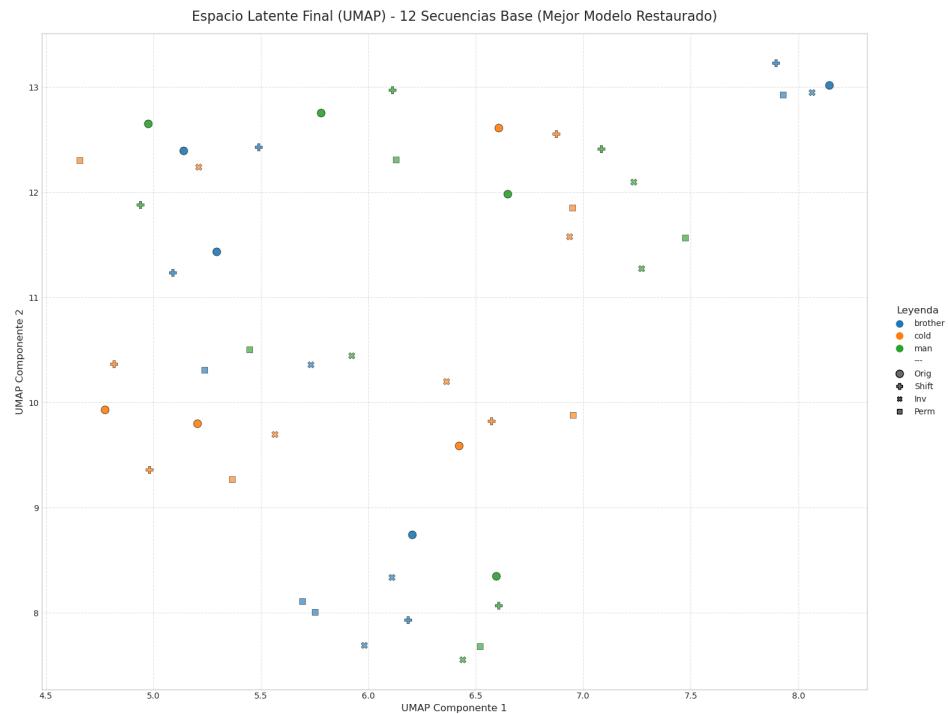


Figura 10.132: Esta grafica muestra el espacio latente en la mejor epoca (100) utilizando umap, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.133: Esta grafica muestra el espacio latente en la epoca 5 utilizando pca, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

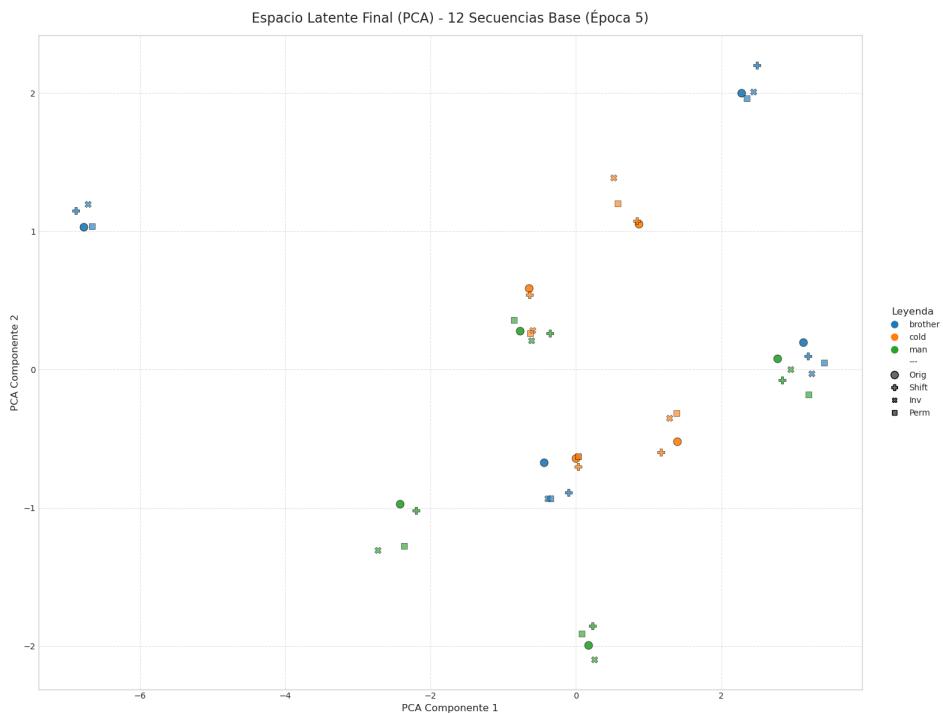
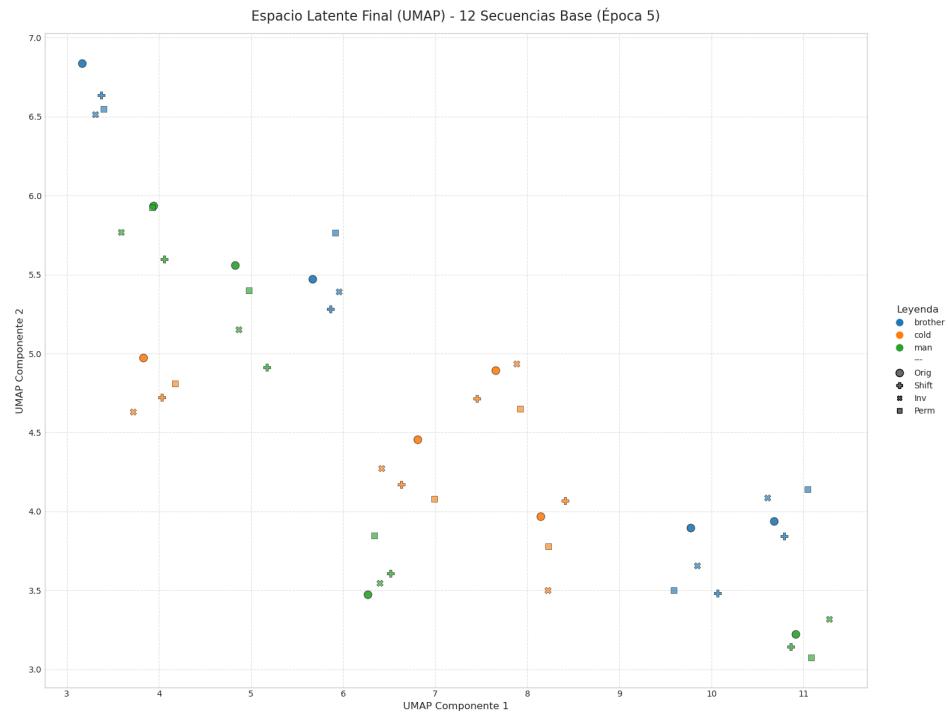


Figura 10.134: Esta grafica muestra el espacio latente en la epoca 5 utilizando umap, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.135: Esta grafica muestra el espacio latente en la epoca 15 utilizando pca, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

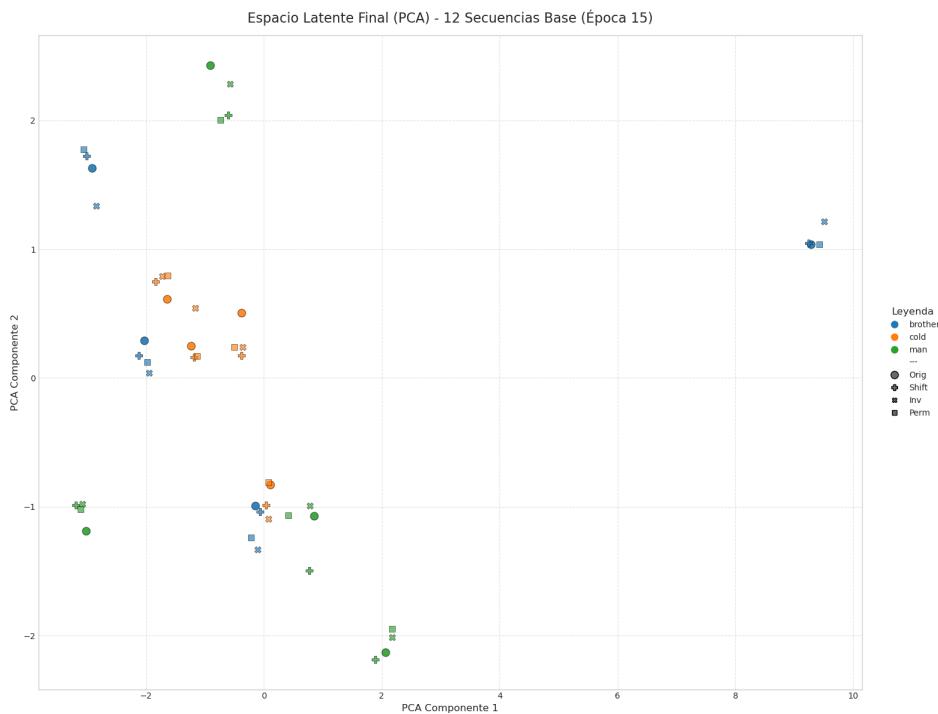
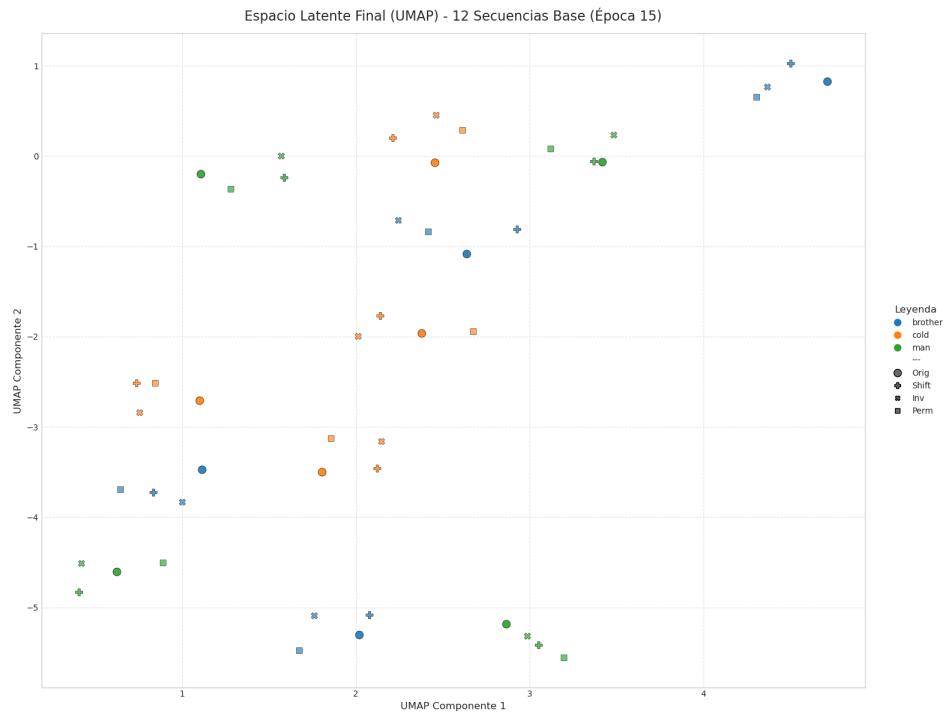


Figura 10.136: Esta grafica muestra el espacio latente en la epoca 15 utilizando umap, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.137: Esta grafica muestra el espacio latente en la epoca 25 utilizando pca, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

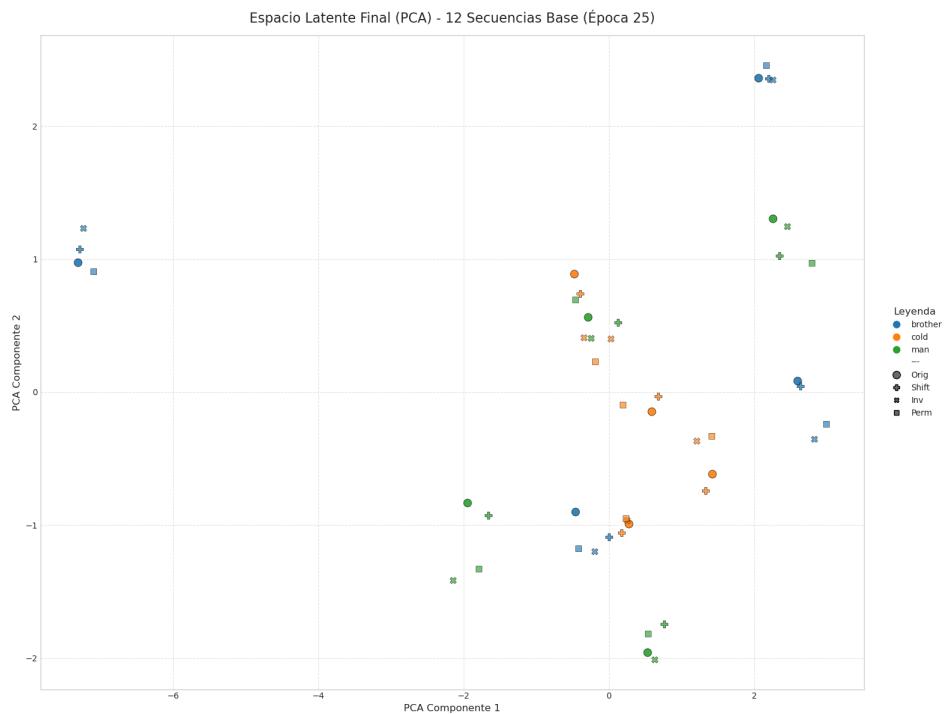
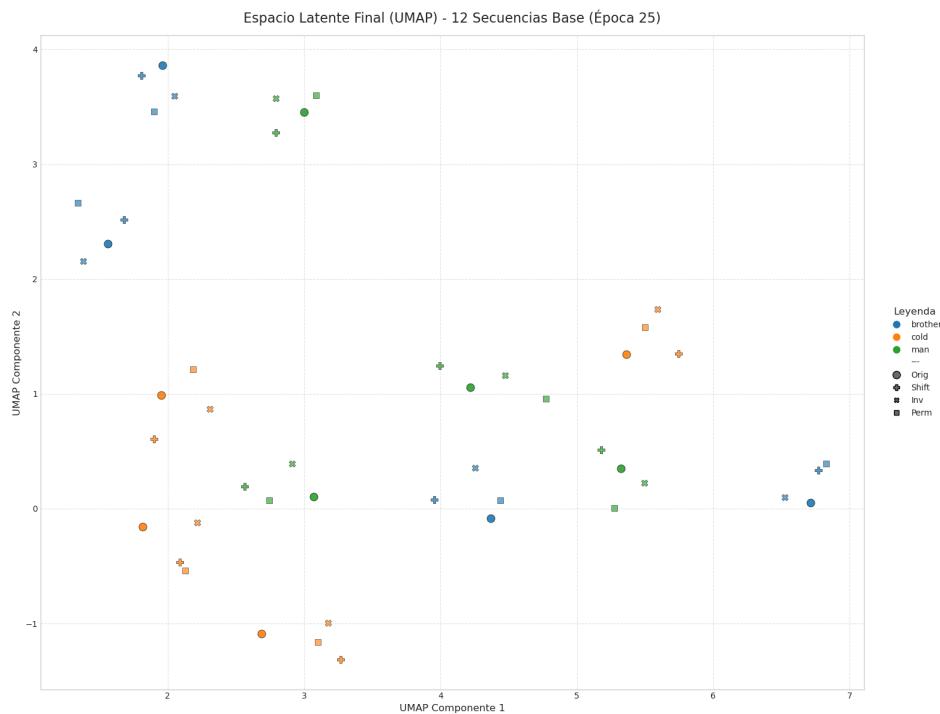


Figura 10.138: Esta grafica muestra el espacio latente en la epoca 25 utilizando umap, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.139: Esta grafica muestra el espacio latente en la epoca 45 utilizando pca, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

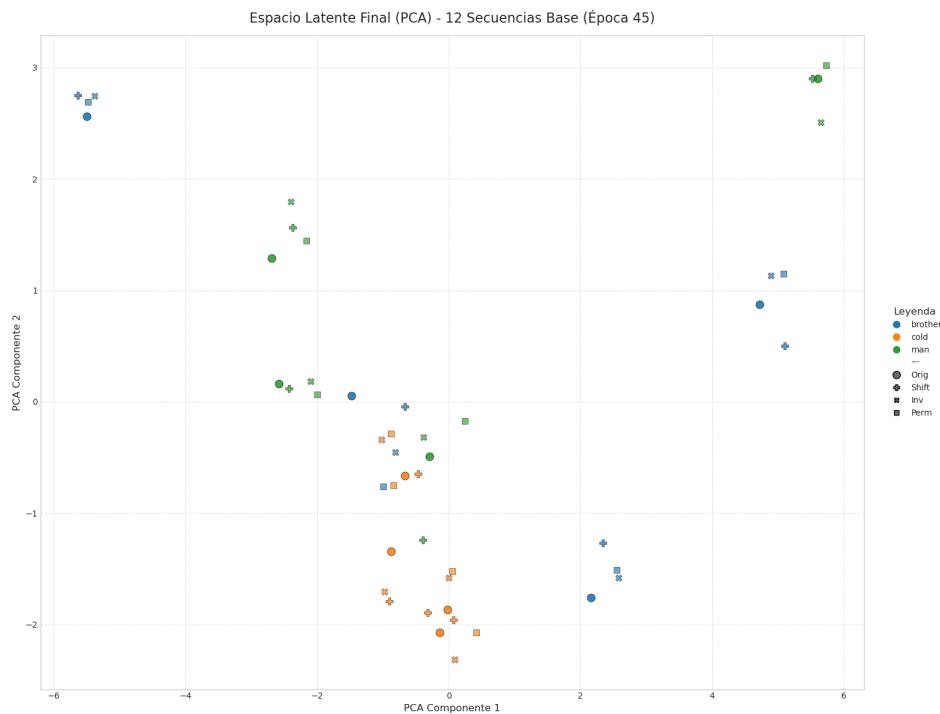
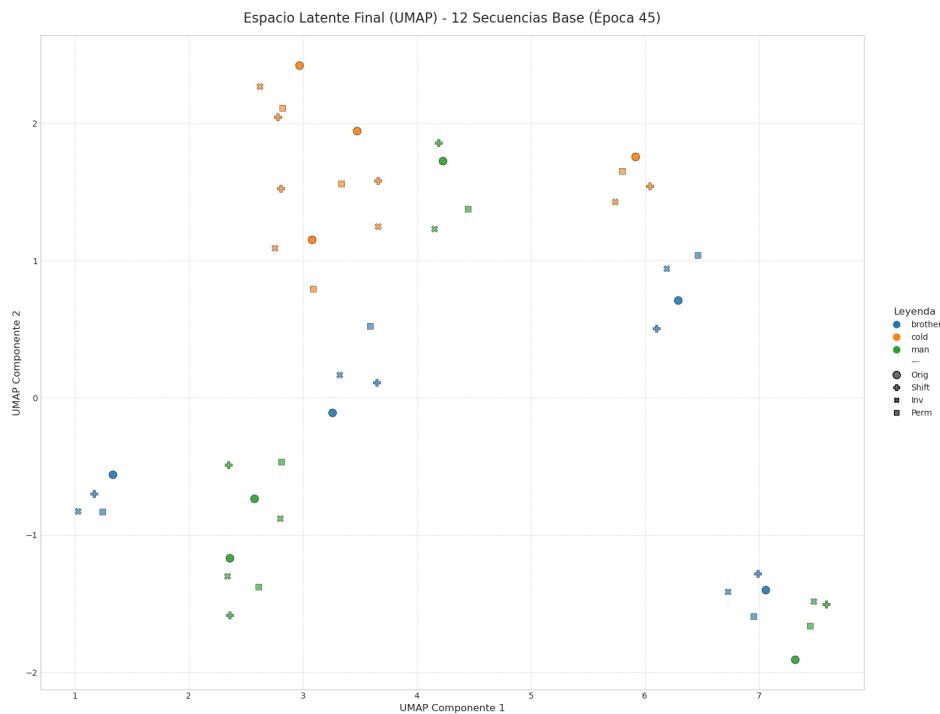


Figura 10.140: Esta grafica muestra el espacio latente en la epoca 45 utilizando umap, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.141: Esta grafica muestra el espacio latente en la epoca 65 utilizando pca, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

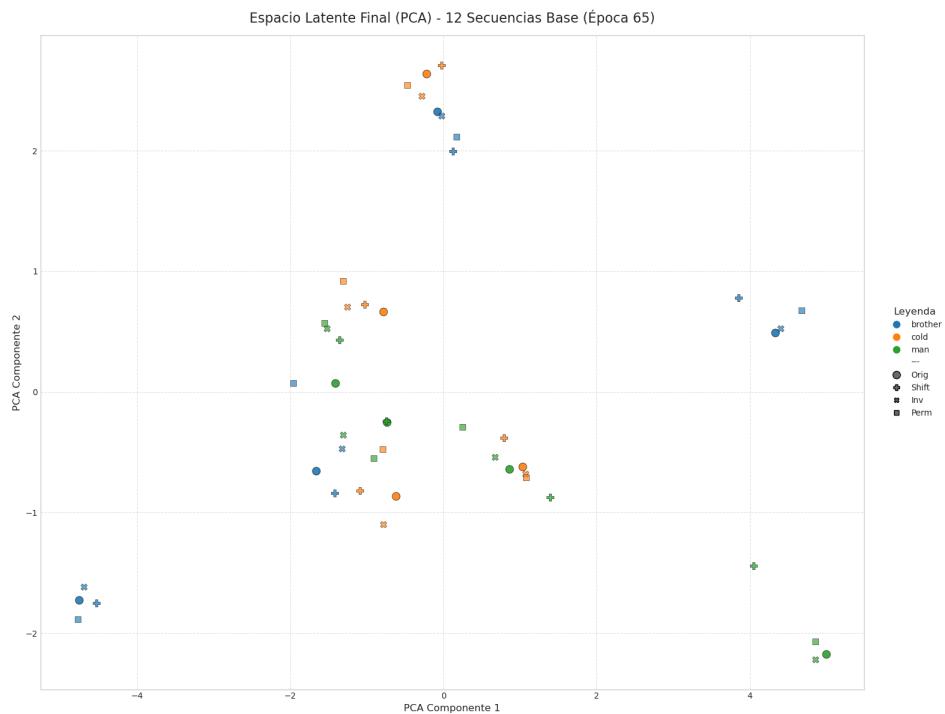
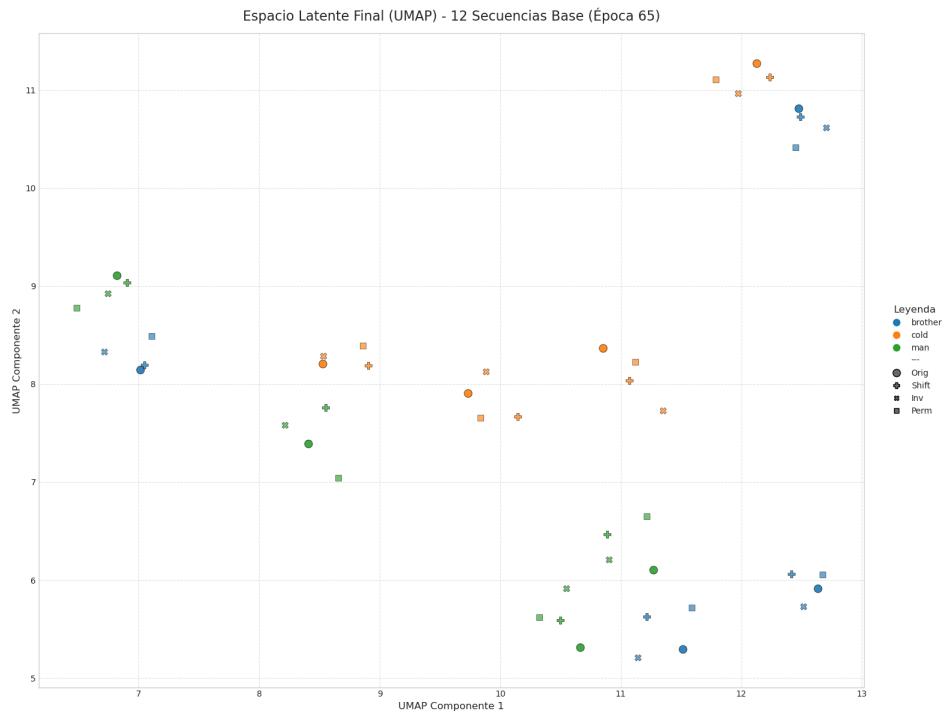


Figura 10.142: Esta grafica muestra el espacio latente en la epoca 65 utilizando umap, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.143: Esta grafica muestra el espacio latente en la epoca 85 utilizando pca, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

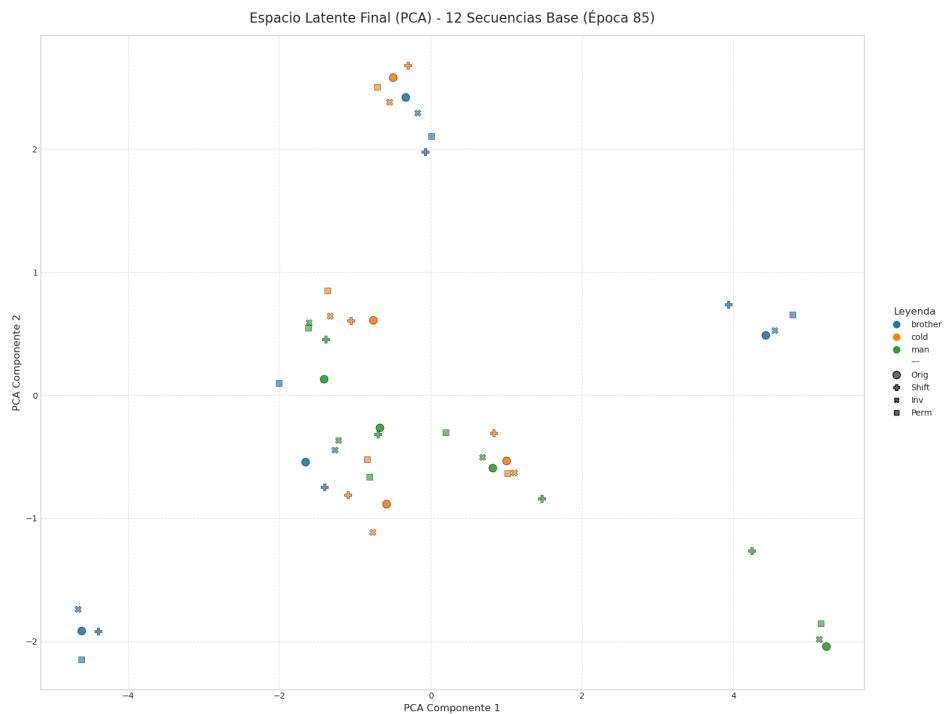
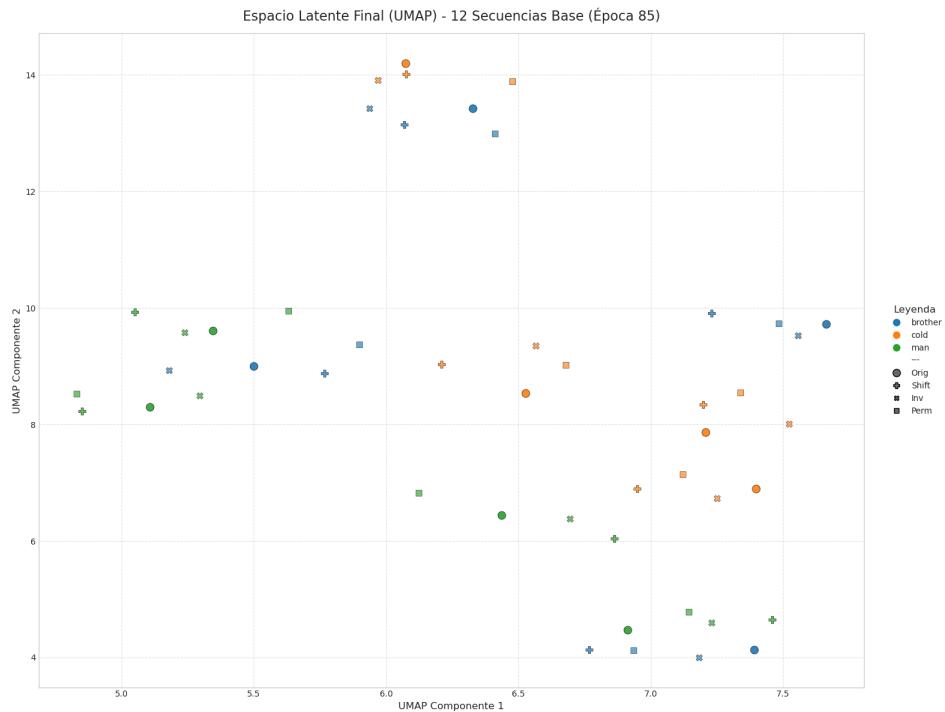


Figura 10.144: Esta grafica muestra el espacio latente en la epoca 85 utilizando umap, donde las señas «brother» estan con color azul, «cold» en naranja y «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

1.4. Con los tres conjuntos de datos

Con 2 etiquetas

Figura 10.145: Esta gráfica compara el ratio Semántico del modelo comparado con un «baseline» o punto de referencia. Siendo este el modelo sin entrenar. El eje Y representa el ratio semántico, que mide la capacidad del modelo para diferenciar entre palabras diferentes, mientras que el eje X representa las épocas de entrenamiento.

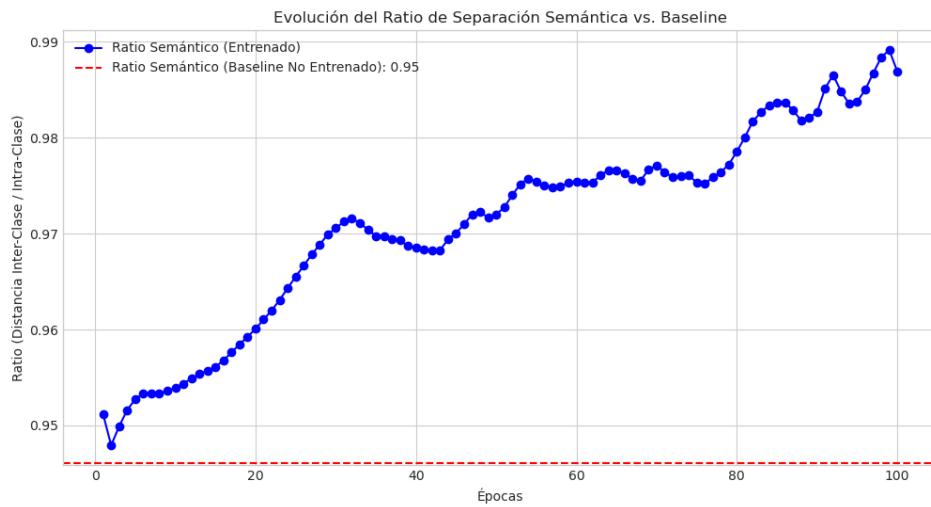


Figura 10.146: Esta gráfica evalúa la capacidad del modelo para entender el orden temporal de las secuencias de video comparadas con sus respectivos «baselines» del modelo no entrenado. Muestra la distancia euclidiana promedio entre la secuencia original y sus versiones alteradas (shifted, inverted, permuted).

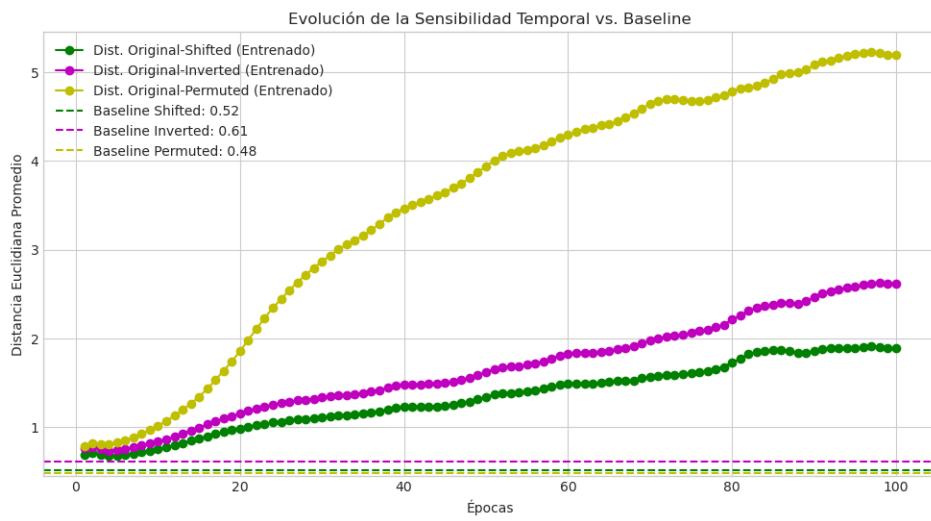


Figura 10.147: Esta gráfica compara el rendimiento del modelo principal con un modelo más simple en términos de la pérdida total de validación. El eje Y representa el valor de la pérdida, una métrica que indica cuán bien el modelo está aprendiendo, donde los valores más bajos son mejores, y el eje X representa las épocas.

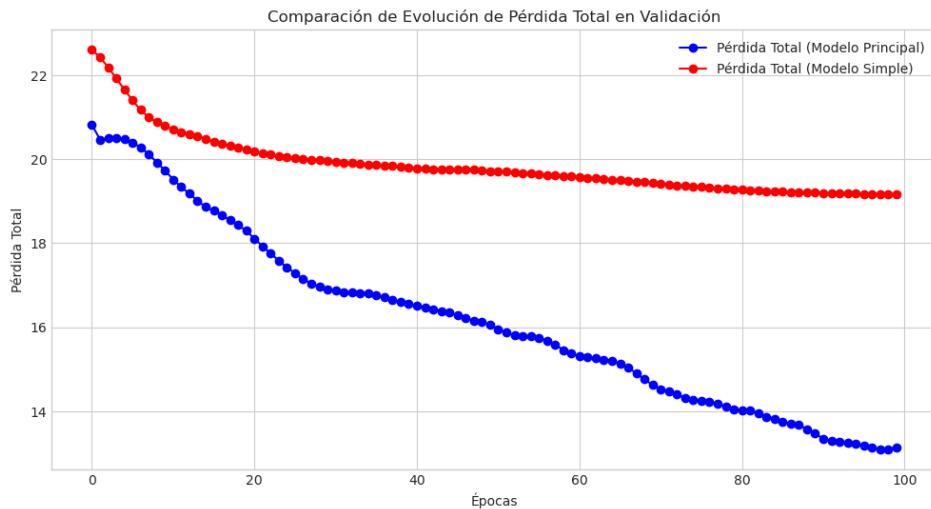
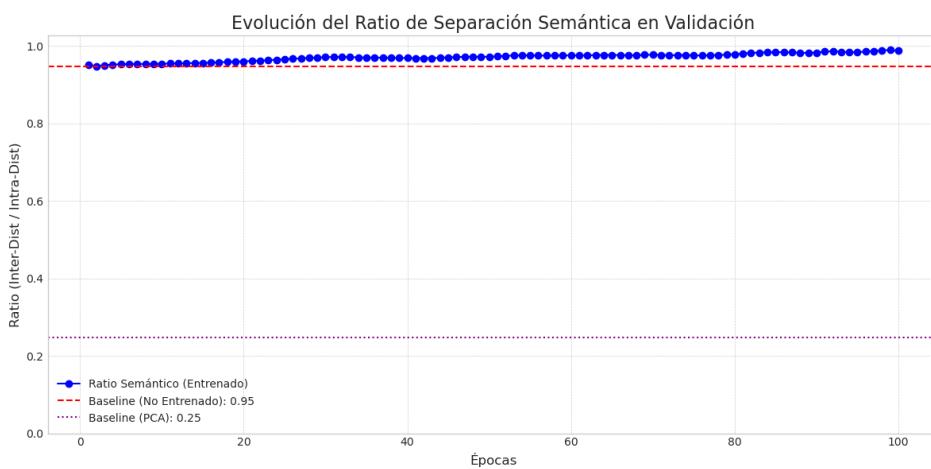


Figura 10.148: Esta gráfica compara el ratio semántico del modelo comparado con dos «baselines» o puntos de referencia. Siendo estos el modelo sin entrenar y la representación de PCA. El eje Y representa el ratio semántico, que mide la capacidad del modelo para diferenciar entre palabras diferentes, mientras que el eje X representa las épocas de entrenamiento.



1. Gráficas de los Experimentos Realizados

Figura 10.149: Este gráfico mide directamente la calidad de la separación semántica en el espacio latente para palabras con la misma y diferente clase. El eje Y representa la distancia euclíadiana promedio y el eje X son las épocas.

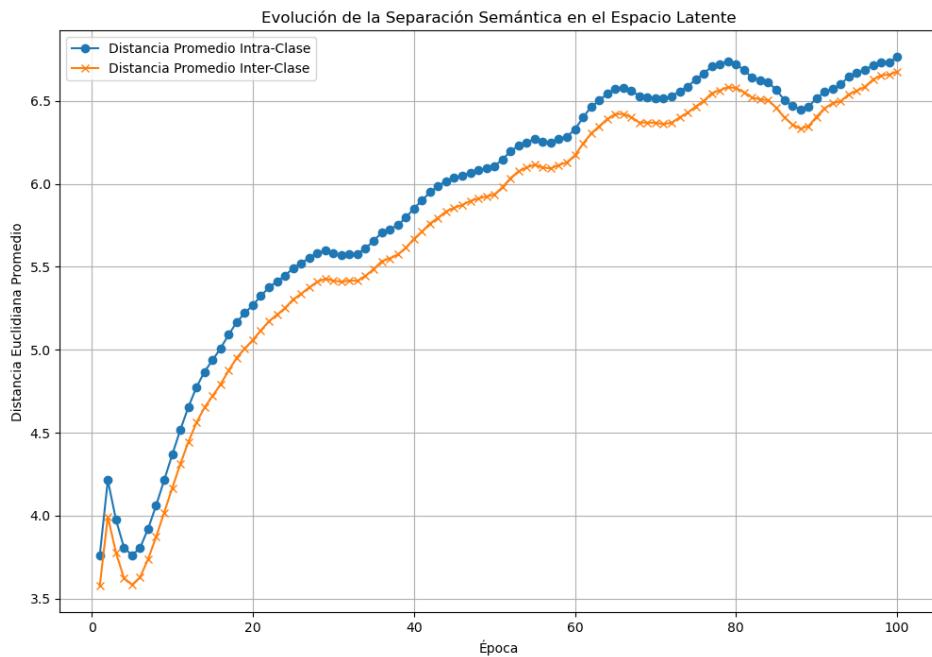
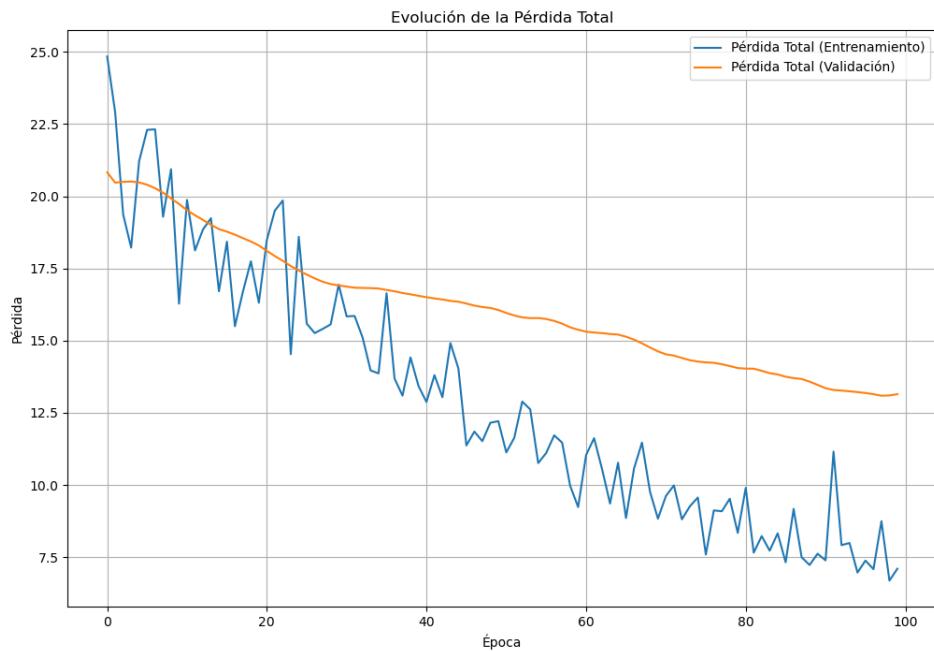


Figura 10.150: Este gráfico muestra la evolución de la «Pérdida Total» a lo largo de 100 épocas de entrenamiento. El eje Y representa el valor de la pérdida, una métrica que indica cuán bien el modelo está aprendiendo donde valores más bajos son mejores. El eje X representa las épocas, es decir, cada ciclo completo de entrenamiento sobre el conjunto de datos.



1. Gráficas de los Experimentos Realizados

Figura 10.151: Esta gráfica ilustra la «Pérdida de Reconstrucción», que mide qué tan bien el autoencoder del modelo puede reconstruir la entrada original después de haberla comprimido en un espacio latente. Al igual que en la gráfica anterior, el eje Y es el valor de la pérdida y el eje X son las épocas.

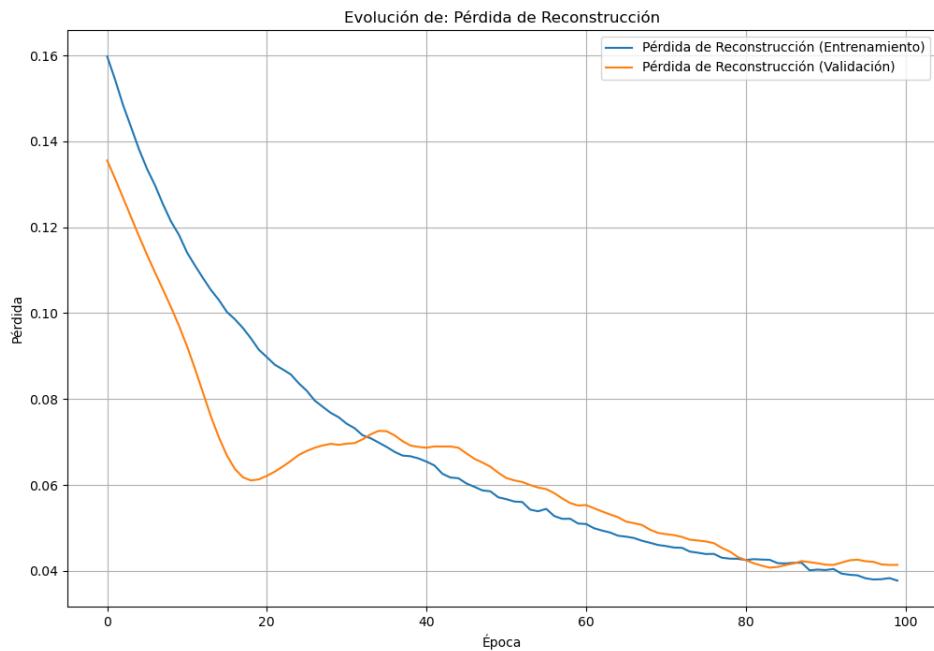
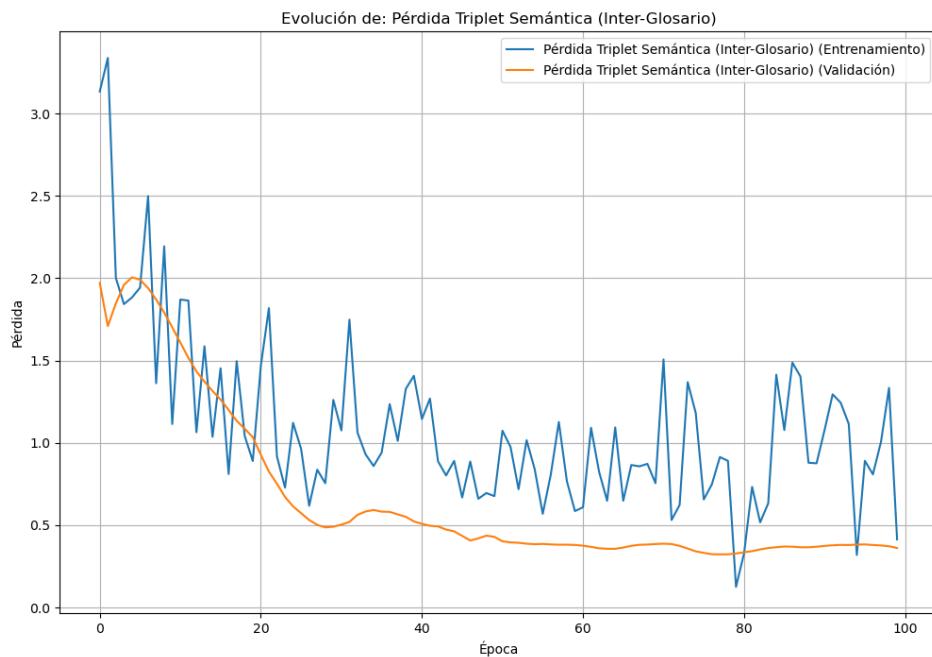


Figura 10.152: Este gráfico muestra la «Pérdida Triplet Semántica», una métrica clave que evalúa si el modelo puede diferenciar entre distintos glosarios, en este caso, las señas «brother» y «cold». El objetivo es que las representaciones de un mismo glosario estén más cerca entre sí que las de glosarios diferentes. Igualmente, el eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.



1. Gráficas de los Experimentos Realizados

Figura 10.153: Esta visualización se enfoca en la sensibilidad temporal del modelo, puesto que, mide la diferencia entre la secuencia original de un video y su versión invertida. Esto quiere decir que el modelo debe aprender que una secuencia invertida es significativamente diferente de la original. El eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.

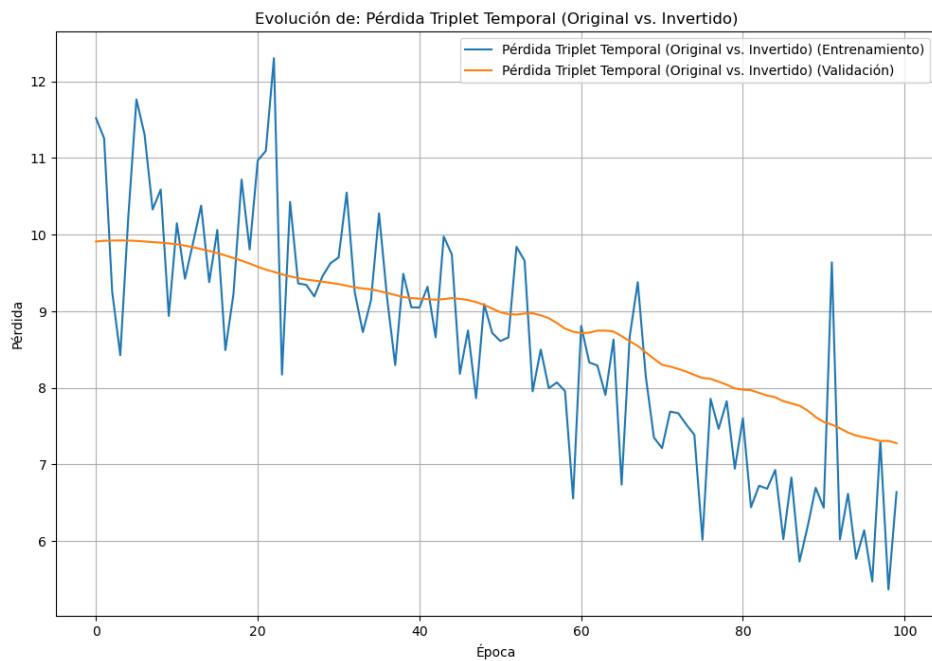
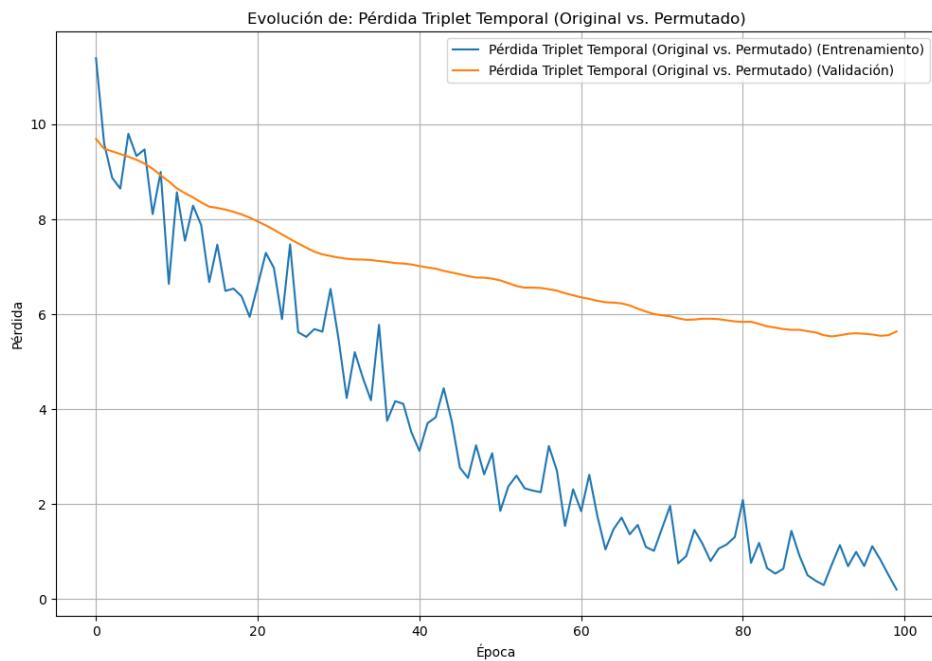


Figura 10.154: Similar a la gráfica anterior, esta también evalúa la sensibilidad temporal, pero en este caso, compara la secuencia original con una versión donde los fotogramas han sido desordenados aleatoriamente. Donde el objetivo es que el modelo reconozca que una secuencia permutada es muy diferente de la original. El eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.



1. Gráficas de los Experimentos Realizados

Figura 10.155: Esta grafica muestra el espacio latente en la mejor epoca (100) utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

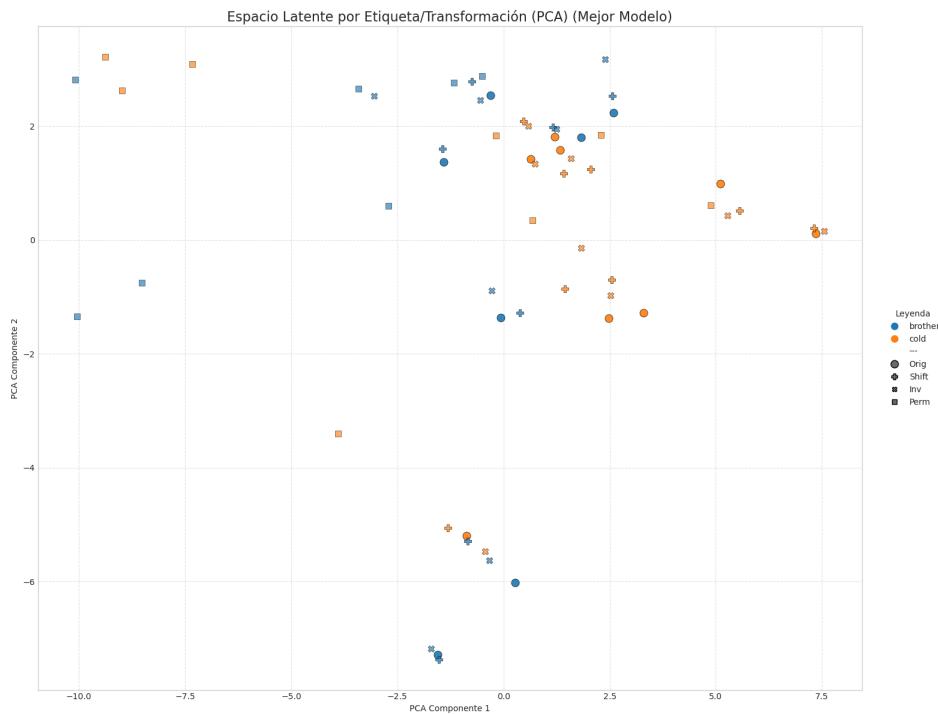
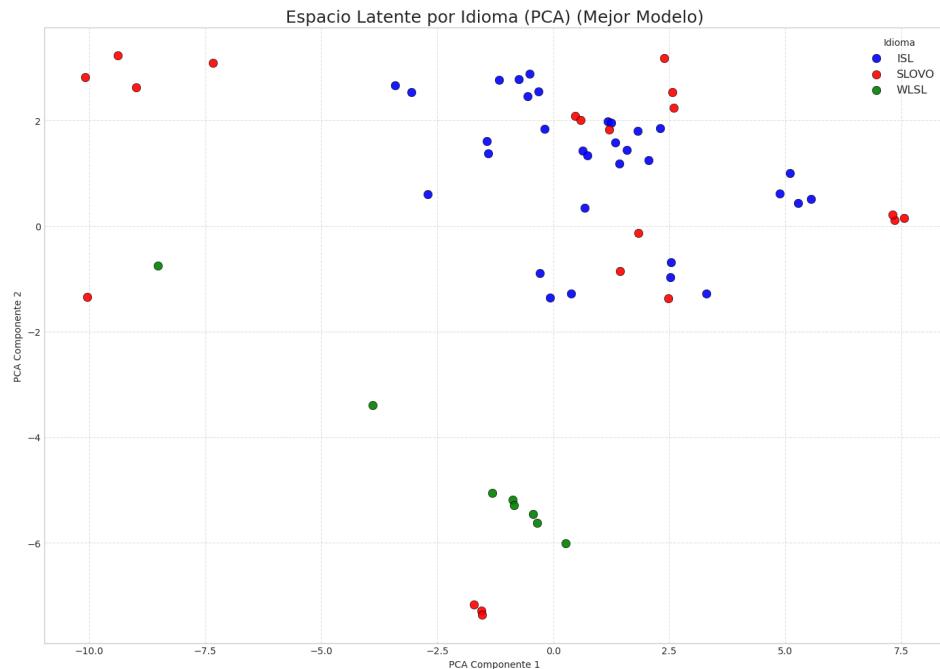


Figura 10.156: Esta grafica muestra el espacio latente en la mejor epoca (100) utilizando pca, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.



1. Gráficas de los Experimentos Realizados

Figura 10.157: Esta grafica muestra el espacio latente en la mejor epoca (100) utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

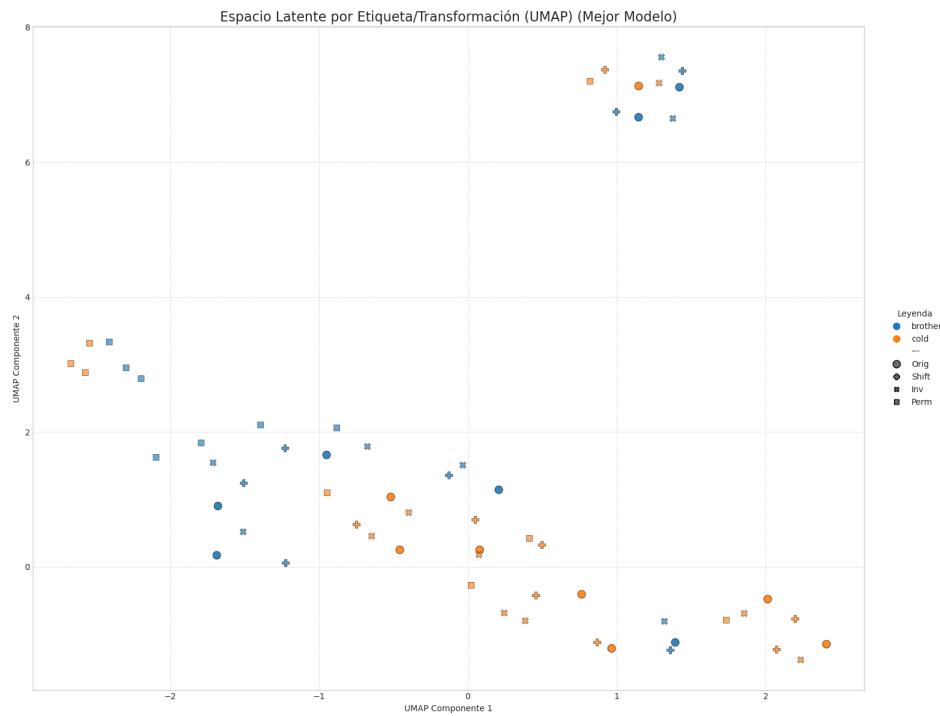
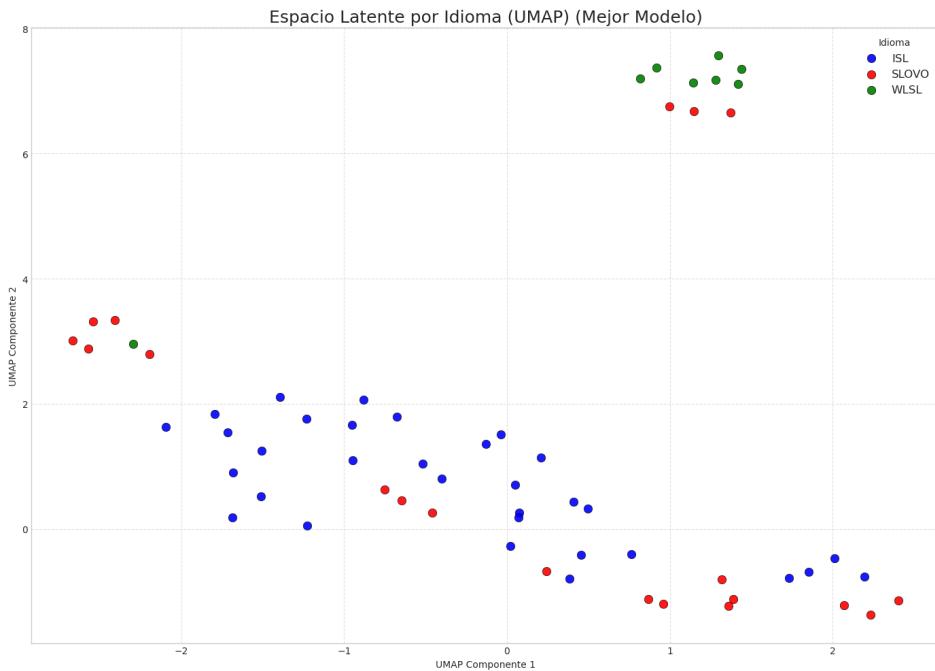


Figura 10.158: Esta grafica muestra el espacio latente en la mejor epoca (100) utilizando umap, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.



1. Gráficas de los Experimentos Realizados

Figura 10.159: Esta grafica muestra el espacio latente en la epoca 5 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

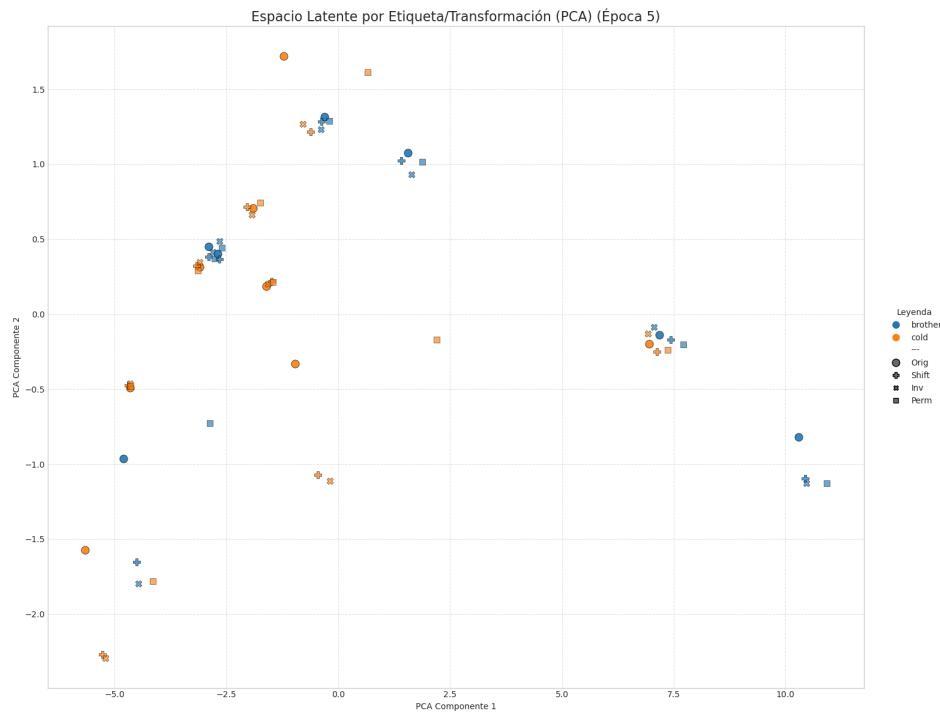
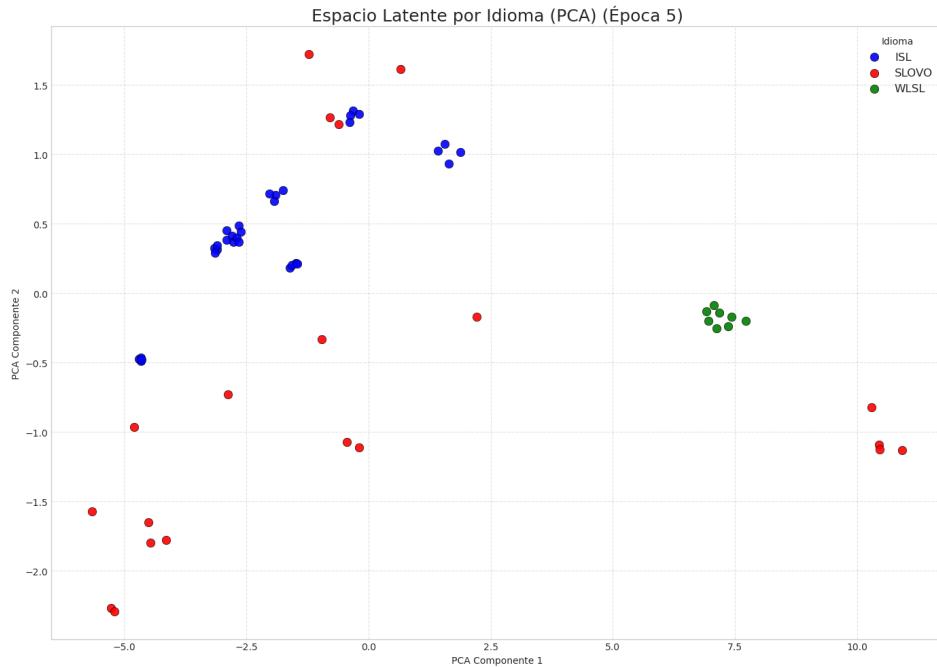


Figura 10.160: Esta grafica muestra el espacio latente en la epoca 5 utilizando pca, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.



1. Gráficas de los Experimentos Realizados

Figura 10.161: Esta grafica muestra el espacio latente en la epoca 5 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

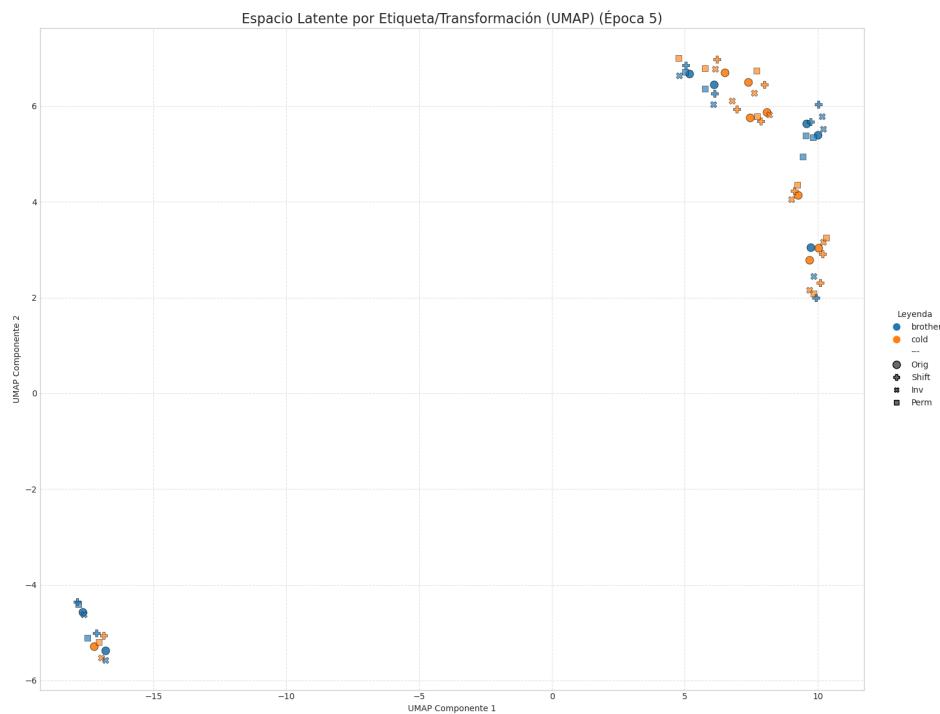
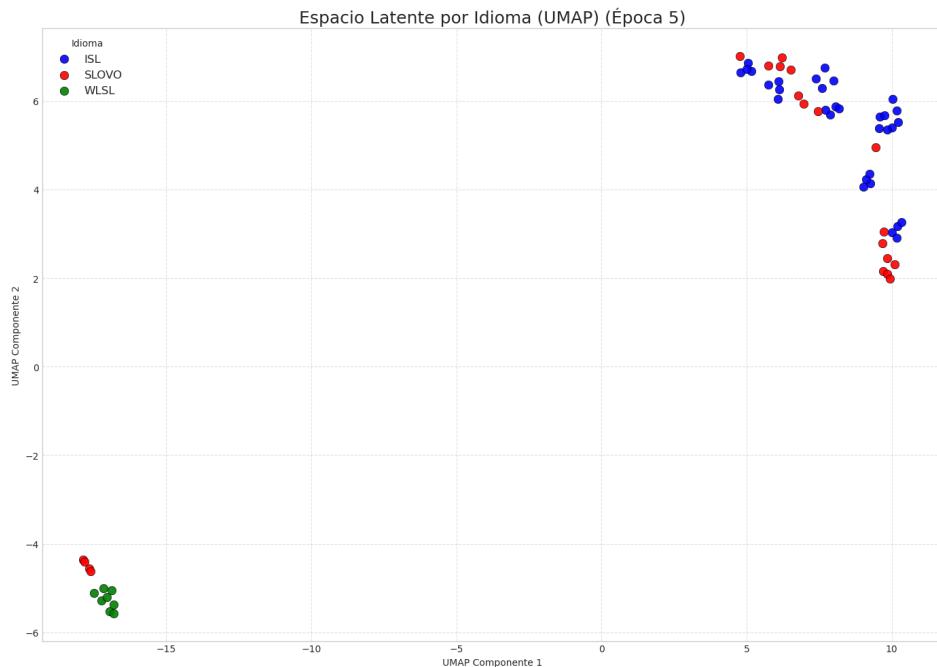


Figura 10.162: Esta grafica muestra el espacio latente en la epoca 5 utilizando umap, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.



1. Gráficas de los Experimentos Realizados

Figura 10.163: Esta grafica muestra el espacio latente en la epoca 15 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

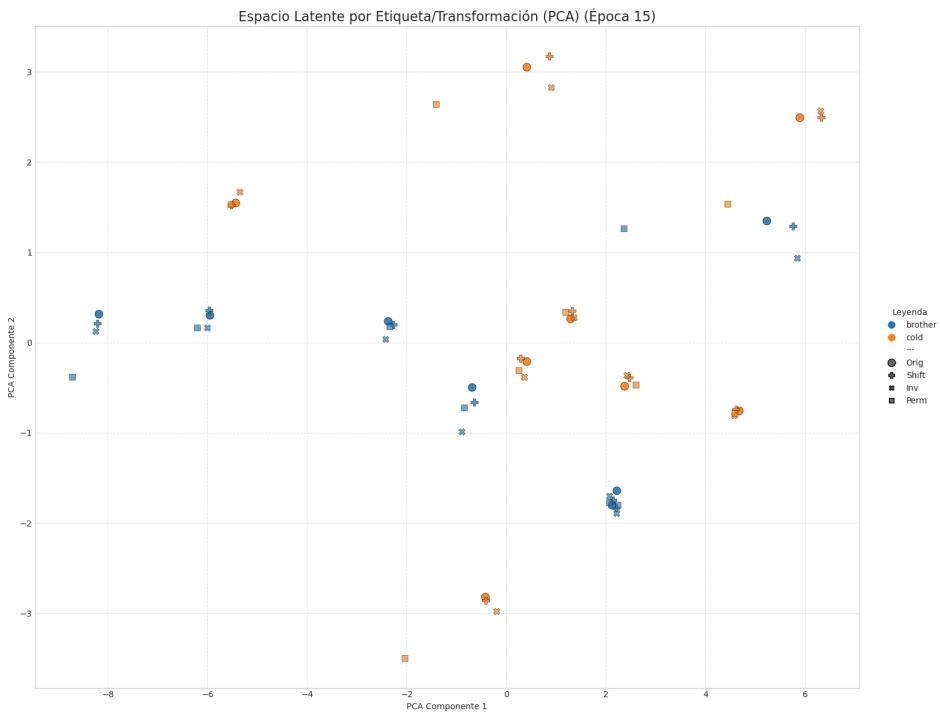
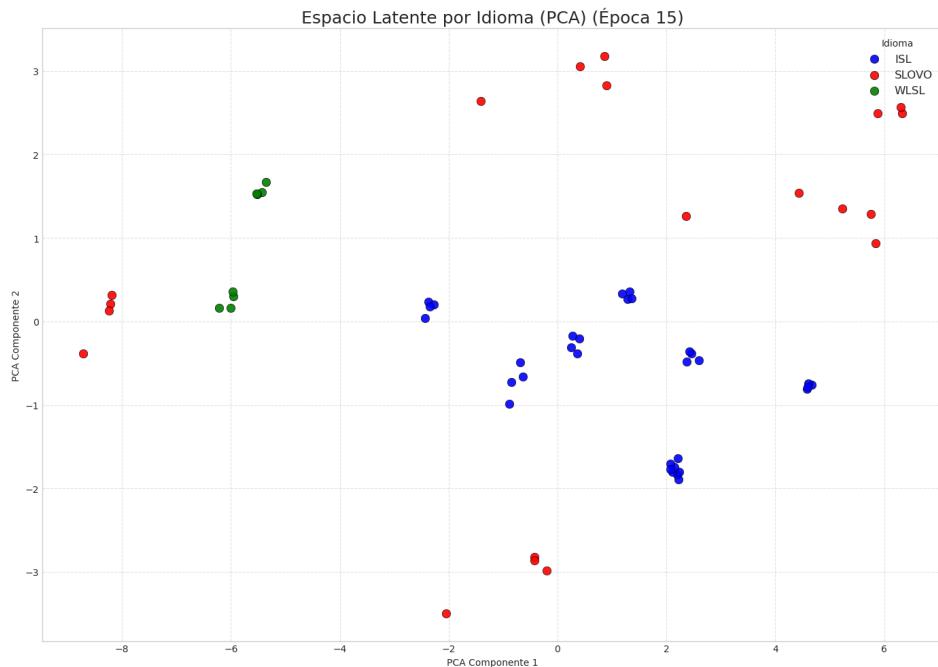


Figura 10.164: Esta grafica muestra el espacio latente en la epoca 15 utilizando pca, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.



1. Gráficas de los Experimentos Realizados

Figura 10.165: Esta grafica muestra el espacio latente en la epoca 15 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

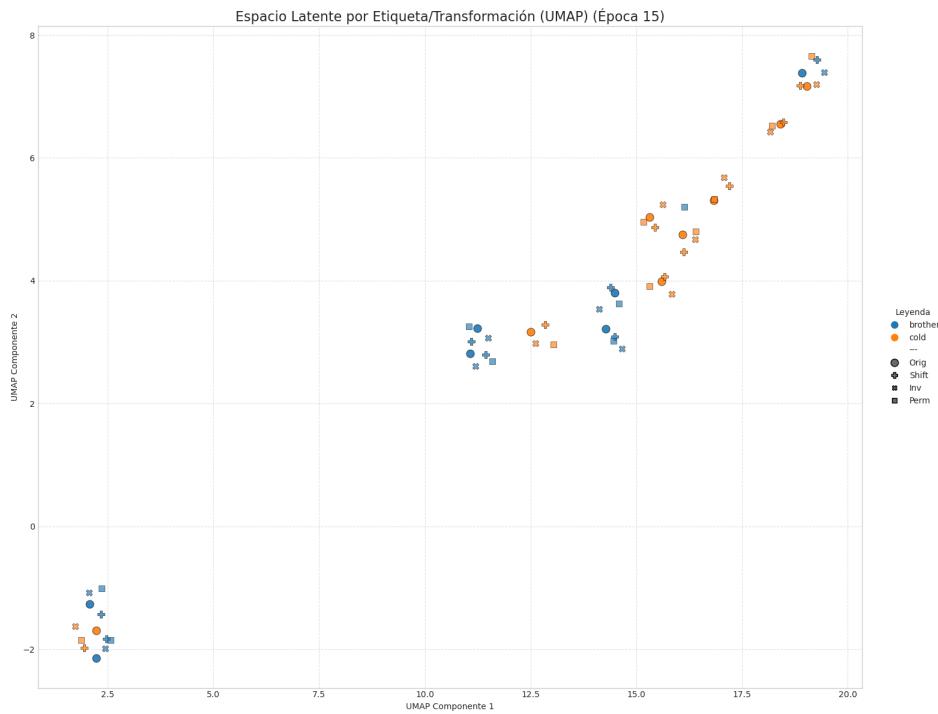
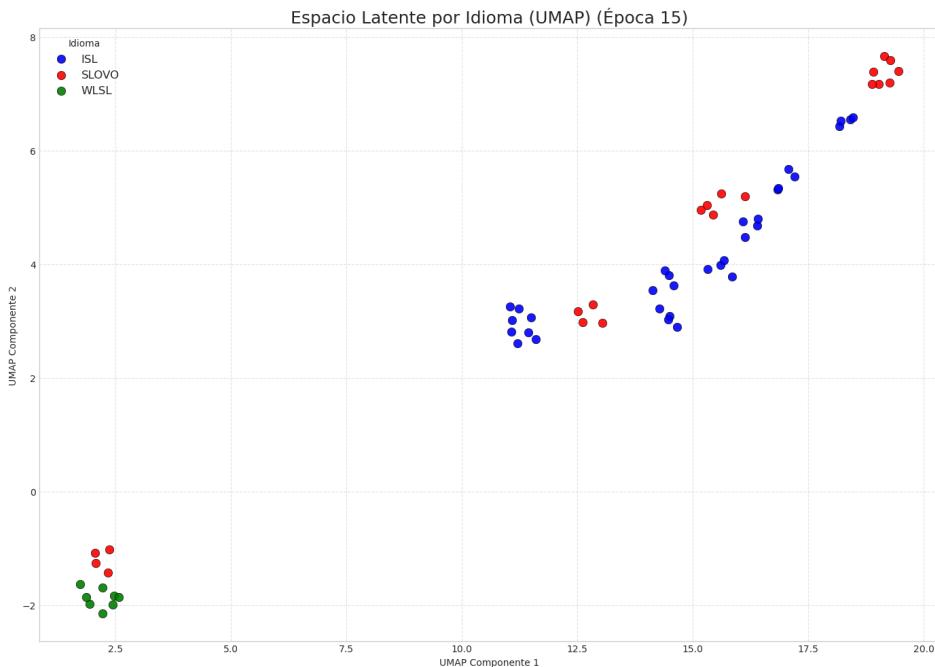


Figura 10.166: Esta grafica muestra el espacio latente en la epoca 15 utilizando umap, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.



1. Gráficas de los Experimentos Realizados

Figura 10.167: Esta grafica muestra el espacio latente en la epoca 25 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

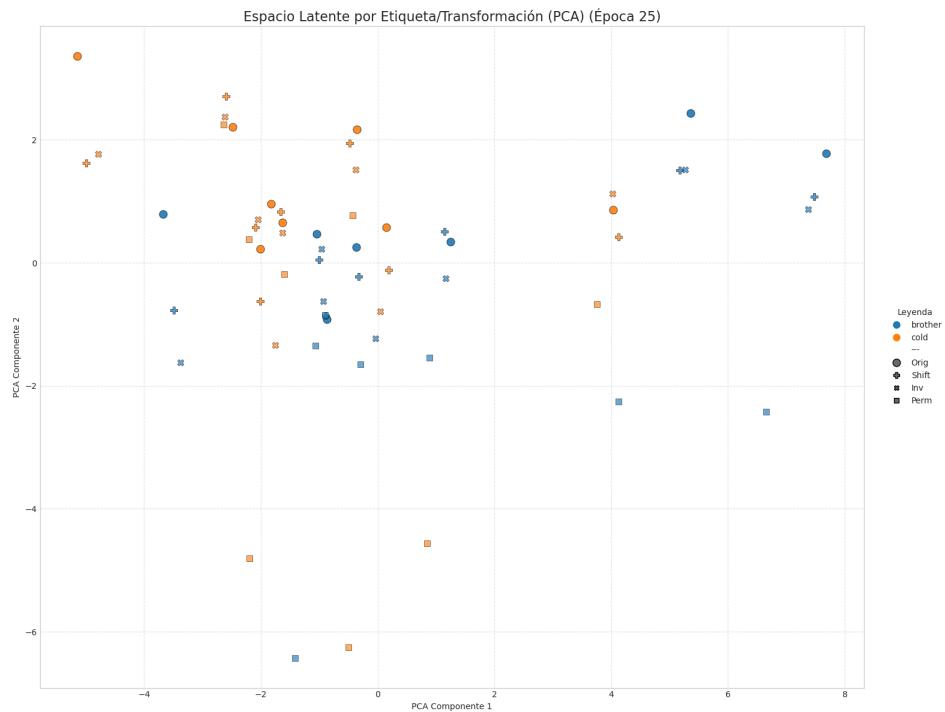
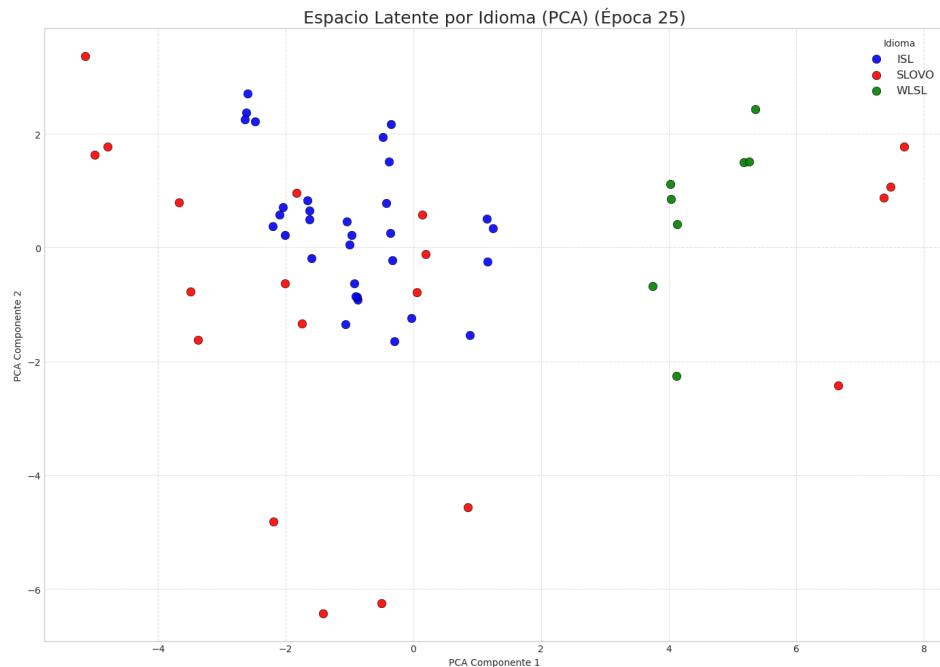


Figura 10.168: Esta grafica muestra el espacio latente en la epoca 25 utilizando pca, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.



1. Gráficas de los Experimentos Realizados

Figura 10.169: Esta grafica muestra el espacio latente en la epoca 25 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

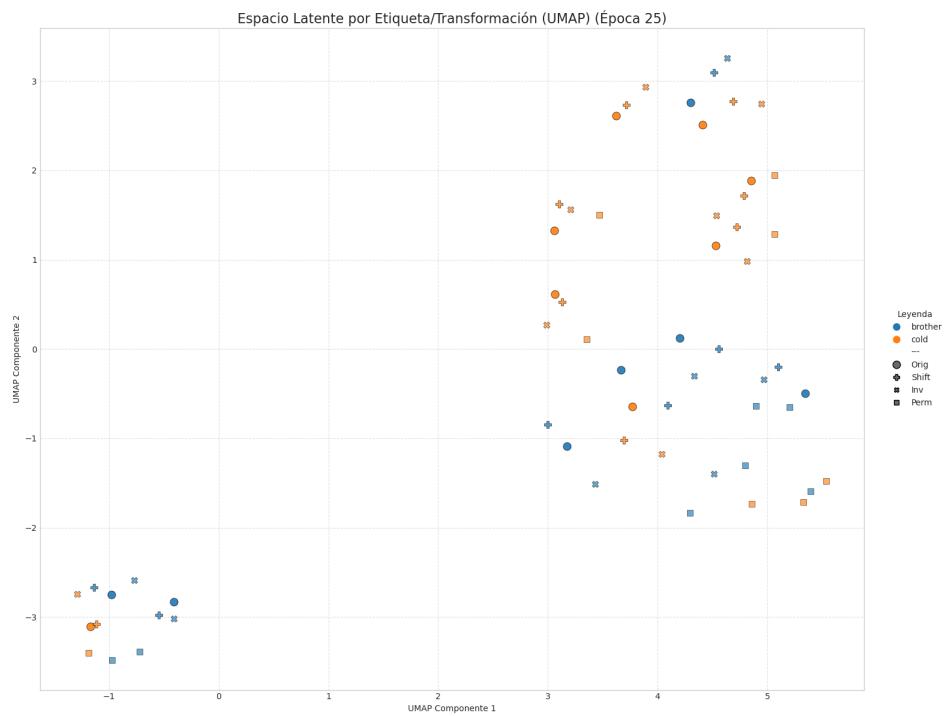
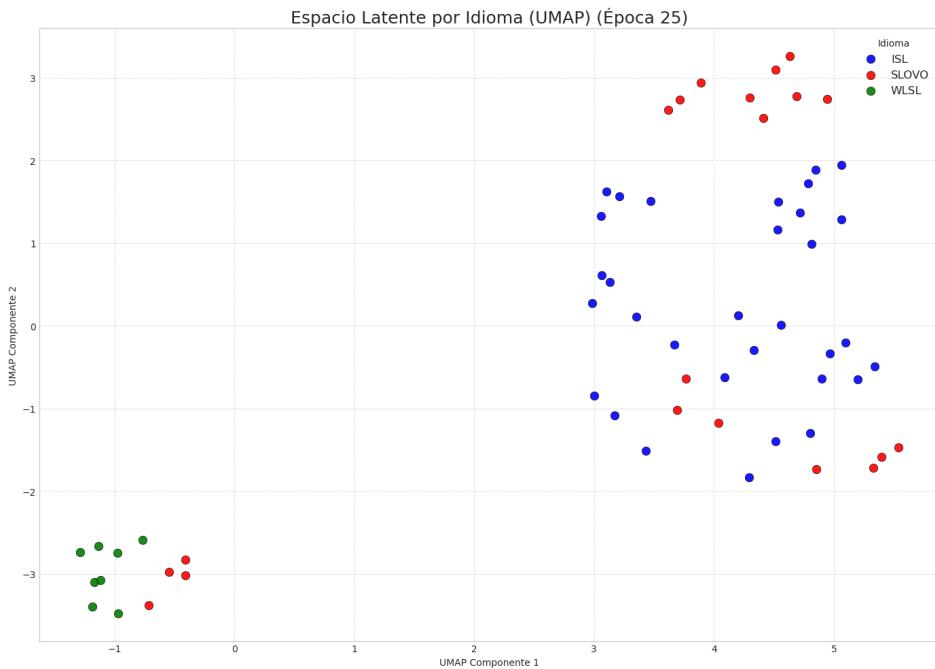


Figura 10.170: Esta grafica muestra el espacio latente en la epoca 25 utilizando umap, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.



1. Gráficas de los Experimentos Realizados

Figura 10.171: Esta grafica muestra el espacio latente en la epoca 45 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

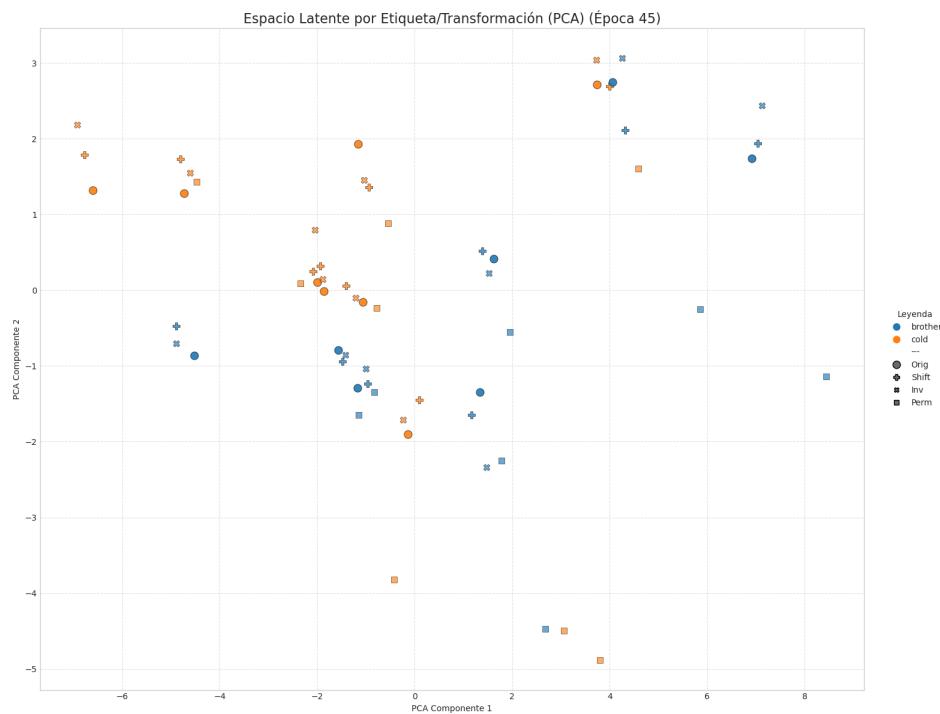
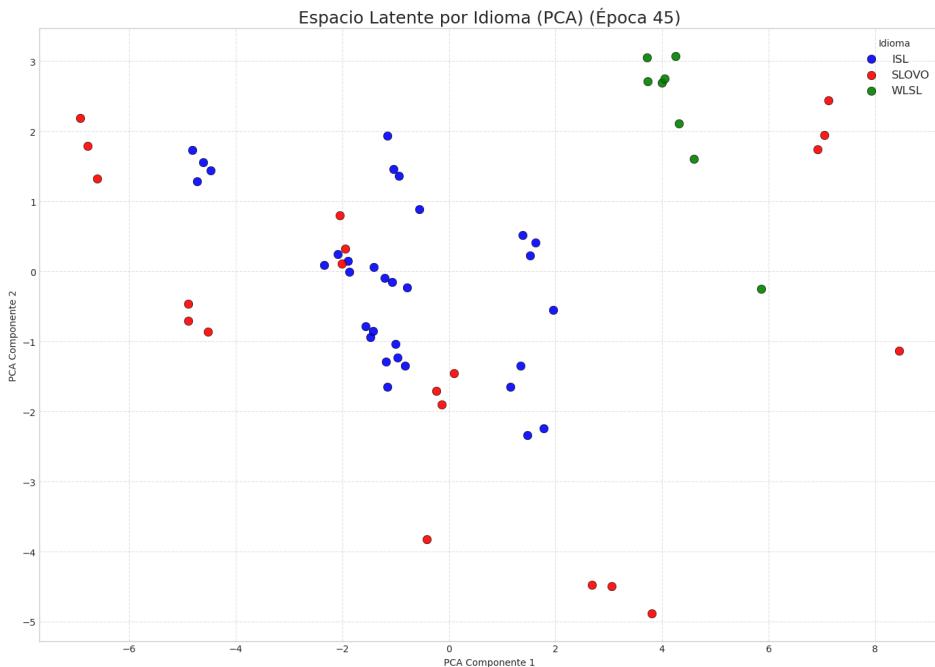


Figura 10.172: Esta grafica muestra el espacio latente en la epoca 45 utilizando pca, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.



1. Gráficas de los Experimentos Realizados

Figura 10.173: Esta grafica muestra el espacio latente en la epoca 45 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

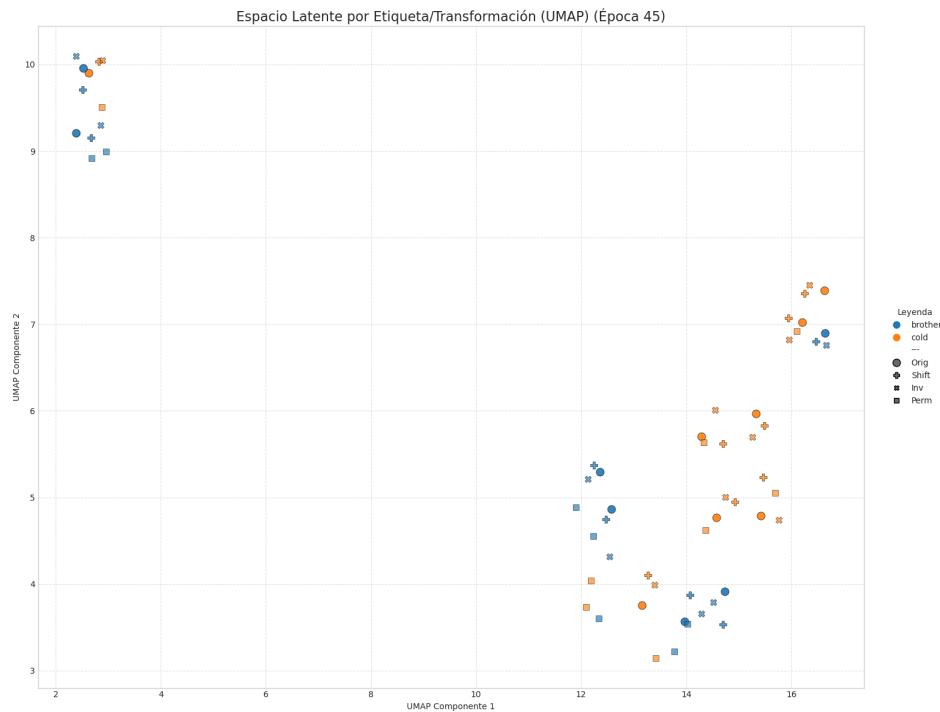
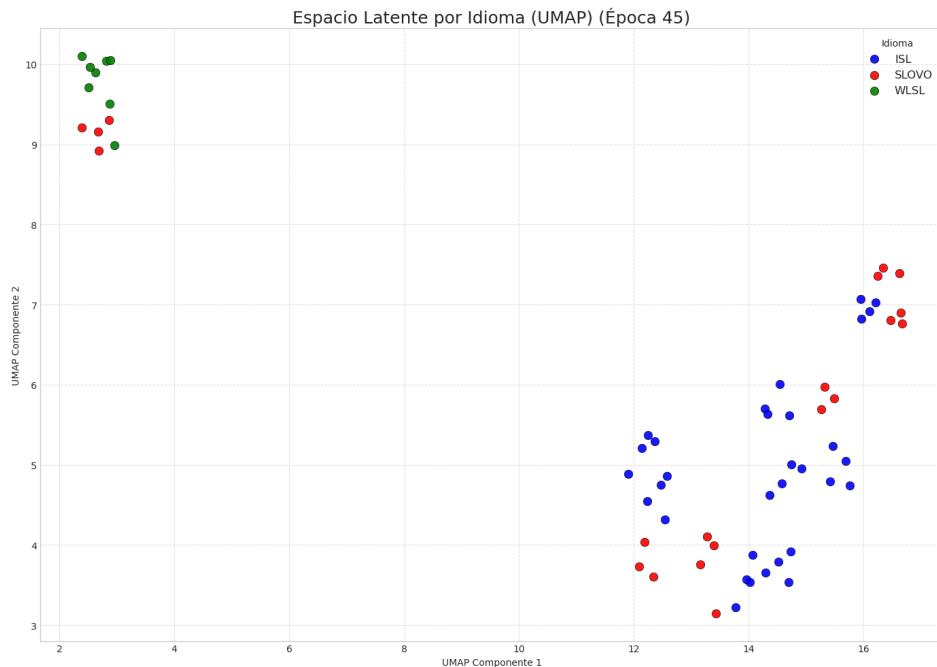


Figura 10.174: Esta grafica muestra el espacio latente en la epoca 45 utilizando umap, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.



1. Gráficas de los Experimentos Realizados

Figura 10.175: Esta grafica muestra el espacio latente en la epoca 65 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

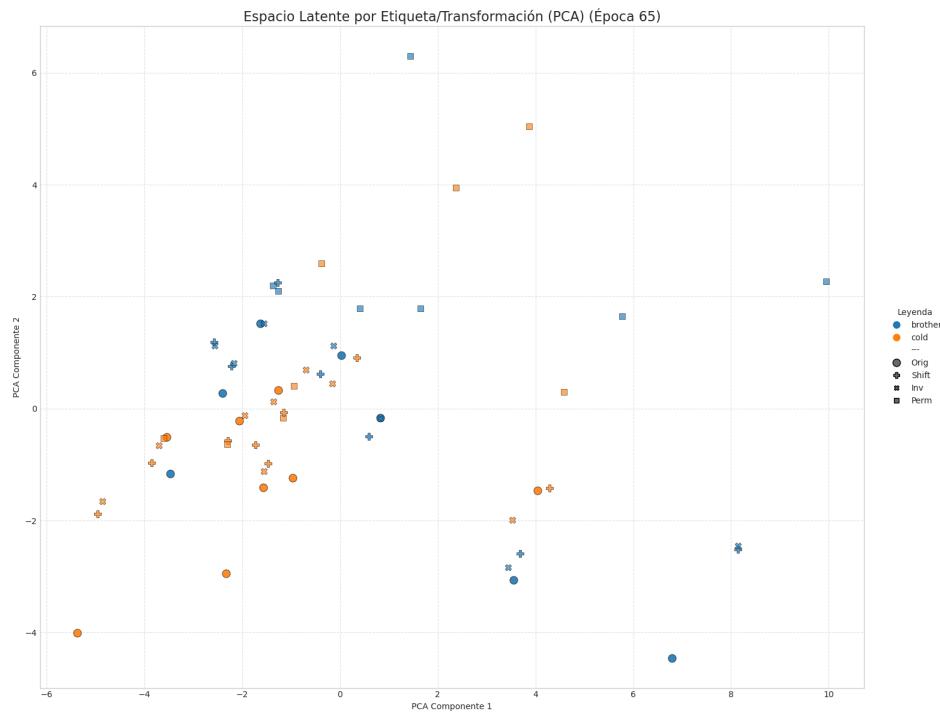
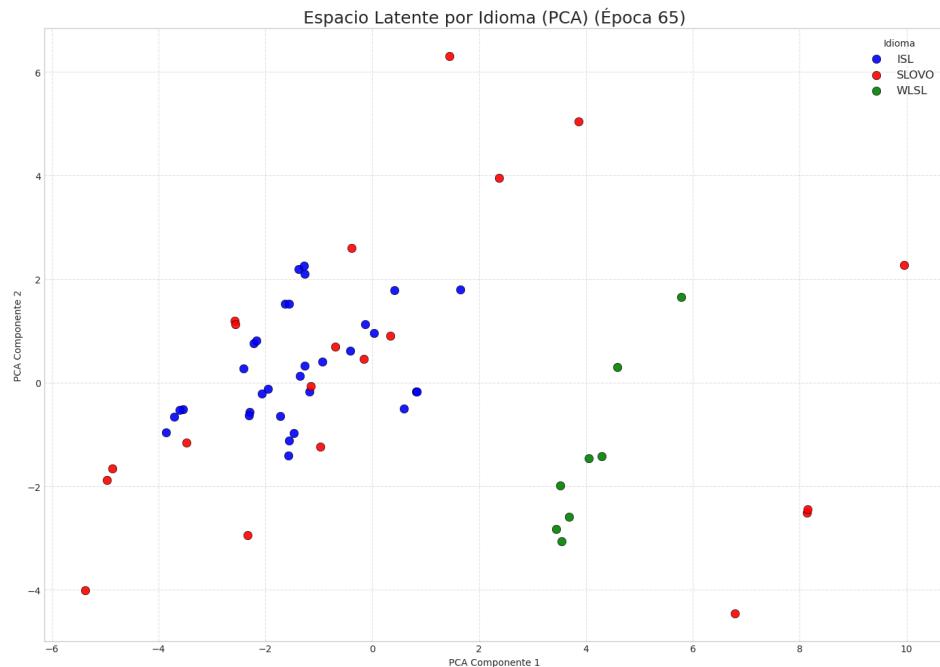


Figura 10.176: Esta grafica muestra el espacio latente en la epoca 65 utilizando pca, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.



1. Gráficas de los Experimentos Realizados

Figura 10.177: Esta grafica muestra el espacio latente en la epoca 65 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

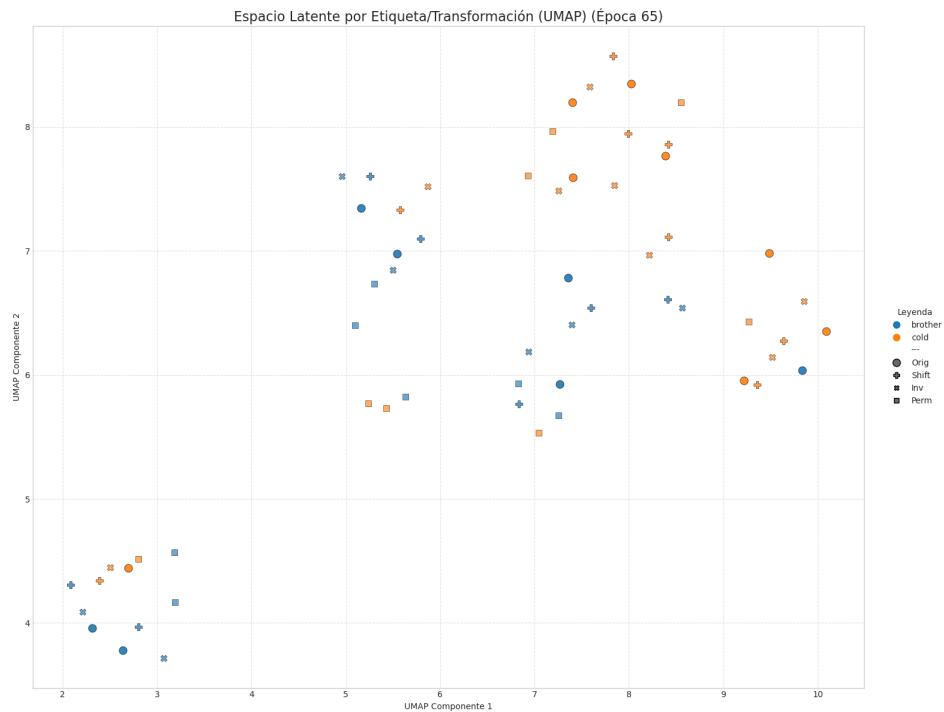
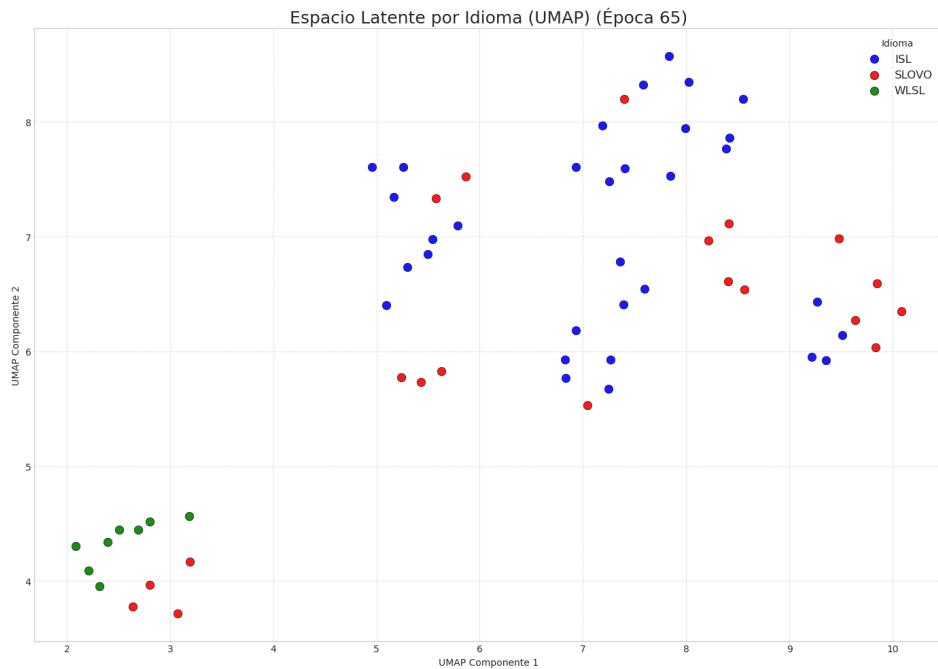


Figura 10.178: Esta grafica muestra el espacio latente en la epoca 65 utilizando umap, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.



1. Gráficas de los Experimentos Realizados

Figura 10.179: Esta grafica muestra el espacio latente en la epoca 85 utilizando pca, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

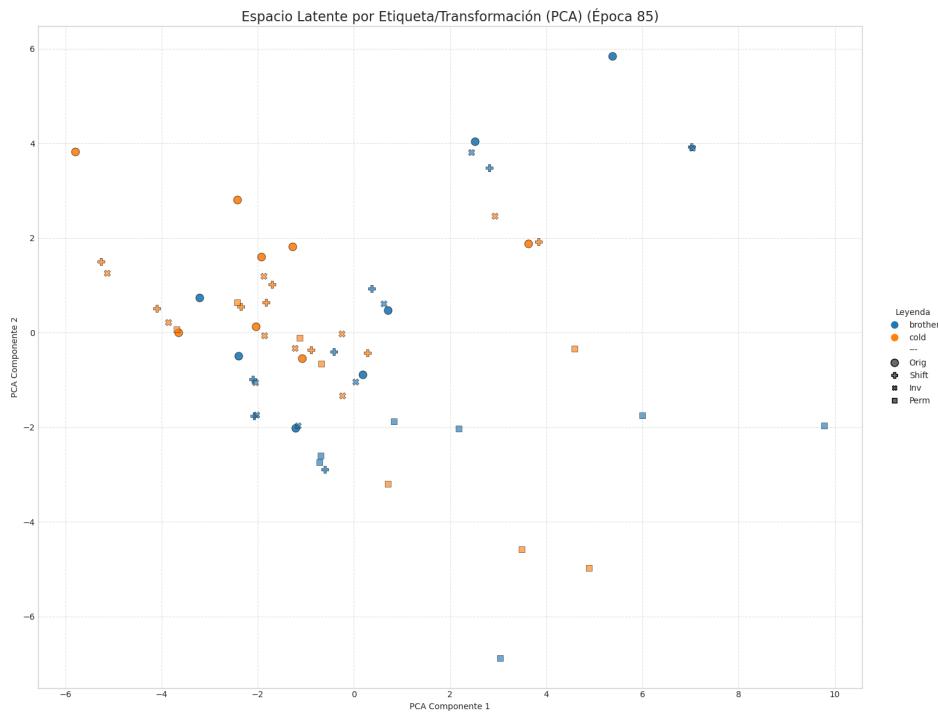
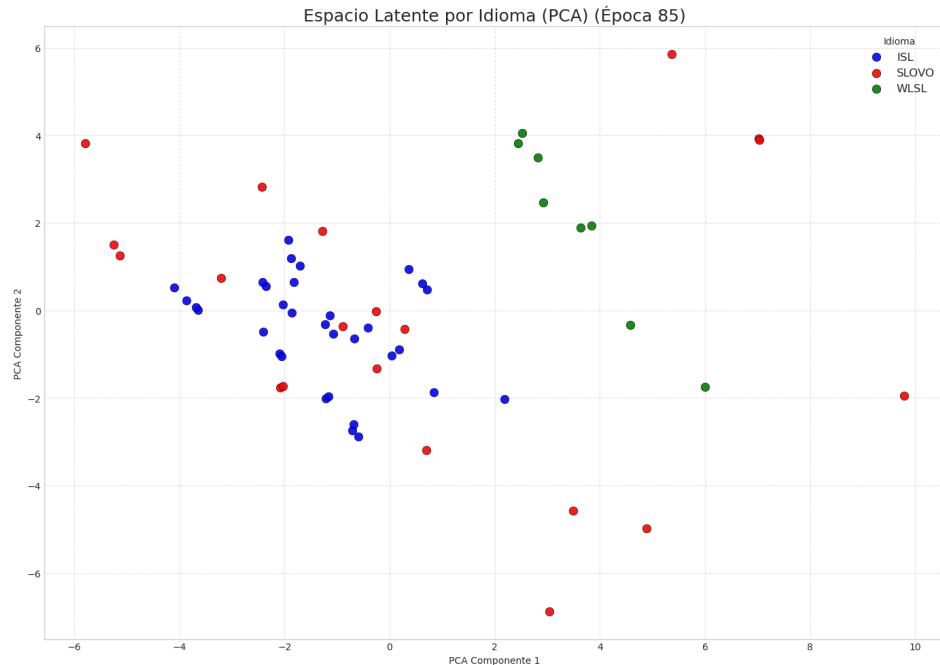


Figura 10.180: Esta grafica muestra el espacio latente en la epoca 85 utilizando pca, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.



1. Gráficas de los Experimentos Realizados

Figura 10.181: Esta grafica muestra el espacio latente en la epoca 85 utilizando umap, donde las señas «brother» estan con color azul y «cold» en naranja. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.

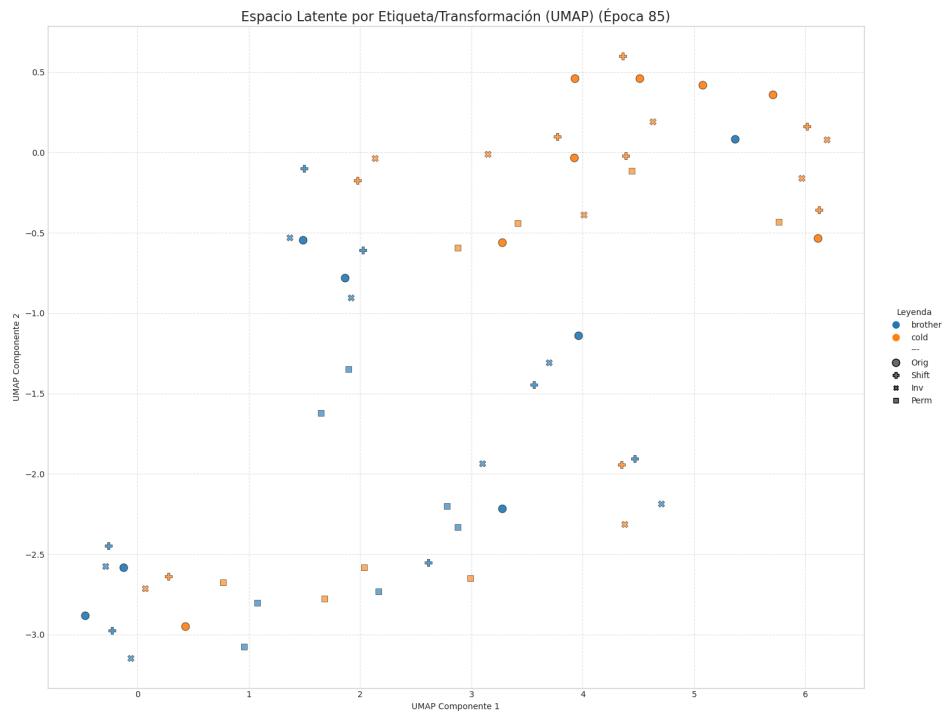
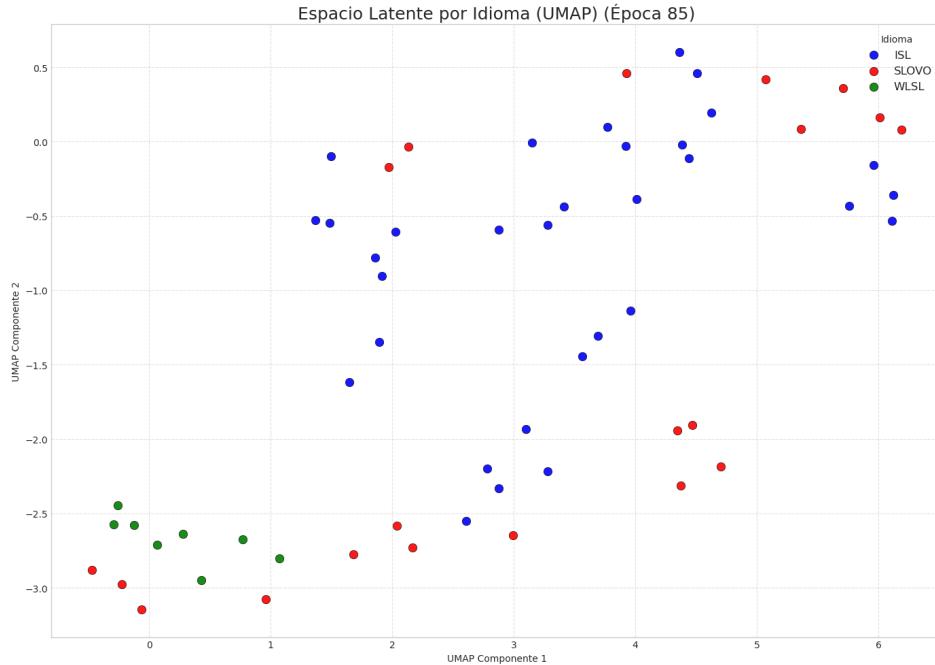
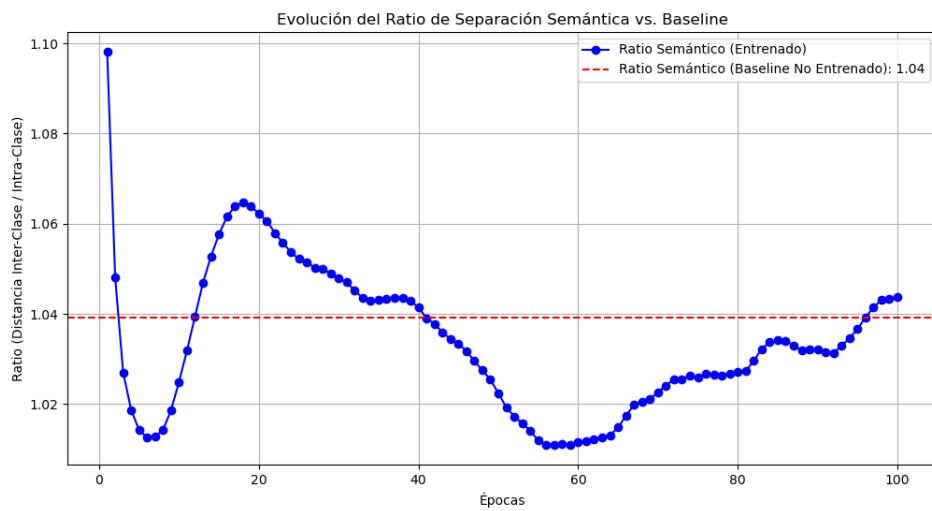


Figura 10.182: Esta grafica muestra el espacio latente en la epoca 85 utilizando umap, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.



Con 3 etiquetas

Figura 10.183: Esta gráfica compara el ratio Semántico del modelo comparado con un «baseline» o punto de referencia. Siendo este el modelo sin entrenar. El eje Y representa el ratio semántico, que mide la capacidad del modelo para diferenciar entre palabras diferentes, mientras que el eje X representa las épocas de entrenamiento.



1. Gráficas de los Experimentos Realizados

Figura 10.184: Esta gráfica evalúa la capacidad del modelo para entender el orden temporal de las secuencias de video comparadas con sus respectivos «baselines» del modelo no entrenado. Muestra la distancia euclídea promedio entre la secuencia original y sus versiones alteradas (shifted, inverted, permuted).

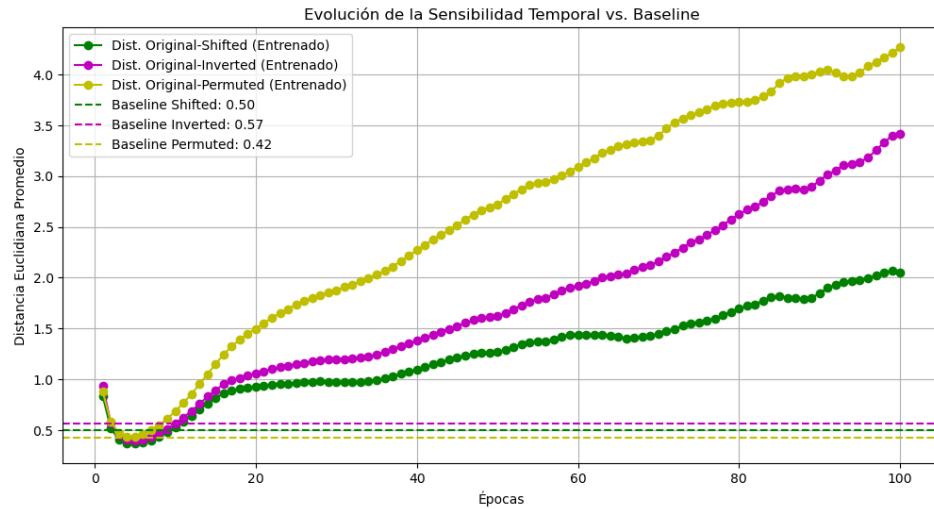


Figura 10.185: Esta gráfica compara el rendimiento del modelo principal con un modelo más simple en términos de la pérdida total de validación. El eje Y representa el valor de la pérdida, una métrica que indica cuán bien el modelo está aprendiendo, donde los valores más bajos son mejores, y el eje X representa las épocas.

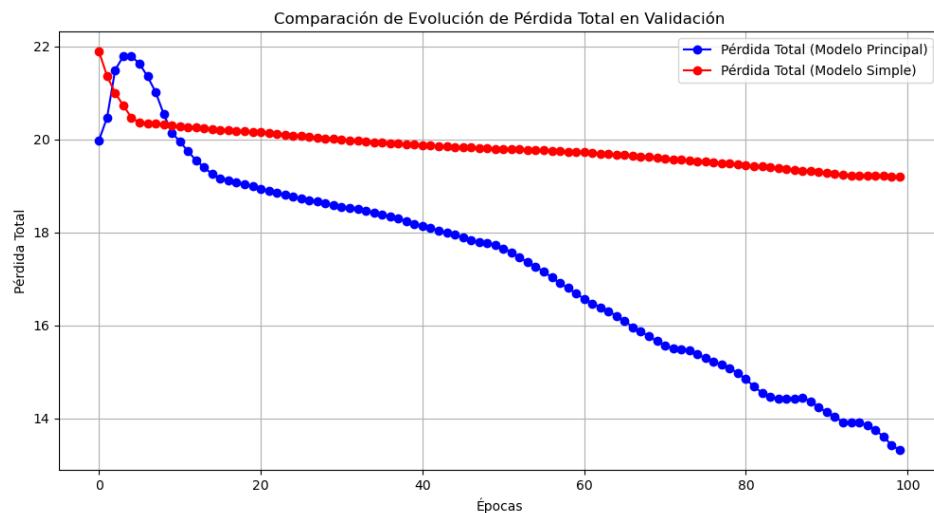


Figura 10.186: Esta gráfica compara el ratio semántico del modelo comparado con dos «baselines» o puntos de referencia. Siendo estos el modelo sin entrenar y la representación de PCA. El eje Y representa el ratio semántico, que mide la capacidad del modelo para diferenciar entre palabras diferentes, mientras que el eje X representa las épocas de entrenamiento.

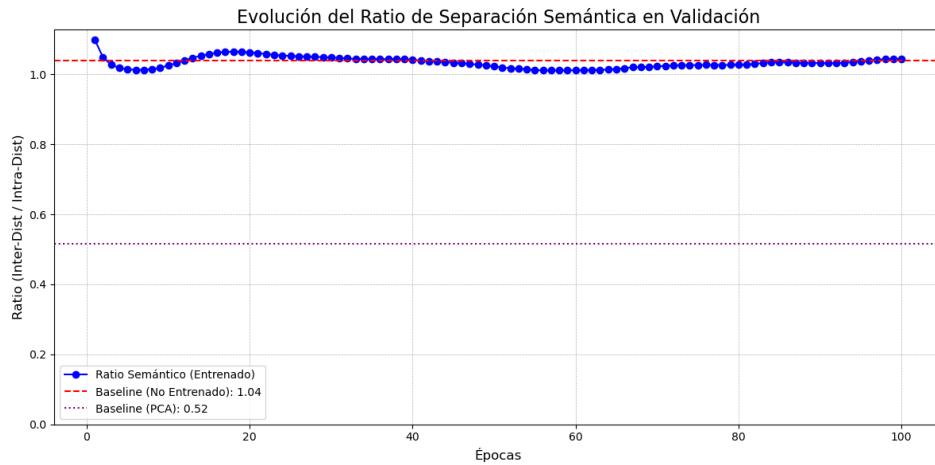
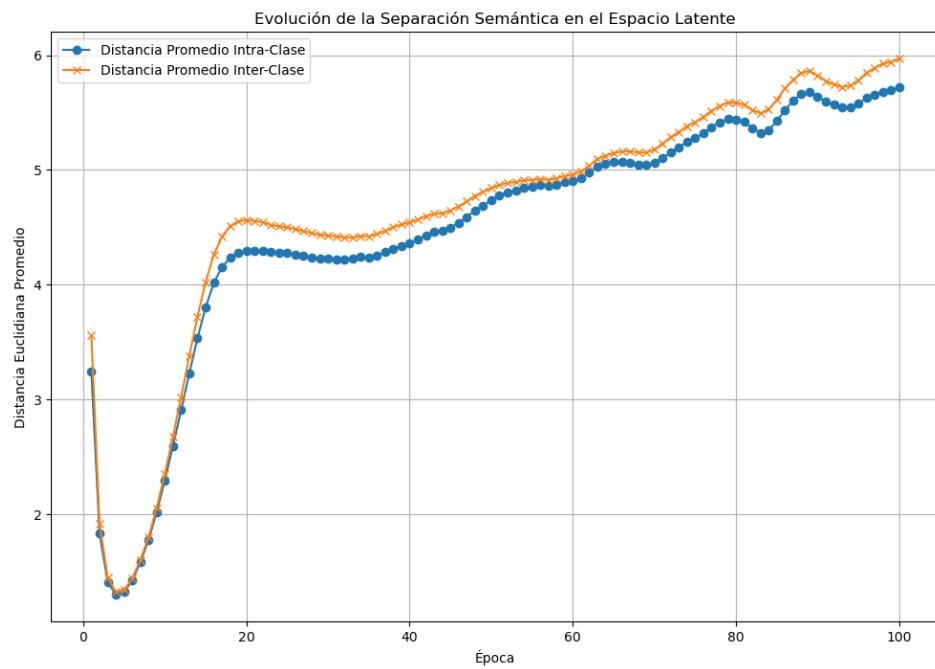


Figura 10.187: Este gráfico mide directamente la calidad de la separación semántica en el espacio latente para palabras con la misma y diferente clase. El eje Y representa la distancia euclídea promedio y el eje X son las épocas.



1. Gráficas de los Experimentos Realizados

Figura 10.188: Este gráfico muestra la evolución de la «Pérdida Total» a lo largo de 100 épocas de entrenamiento. El eje Y representa el valor de la pérdida, una métrica que indica cuán bien el modelo está aprendiendo donde valores más bajos son mejores. El eje X representa las épocas, es decir, cada ciclo completo de entrenamiento sobre el conjunto de datos.

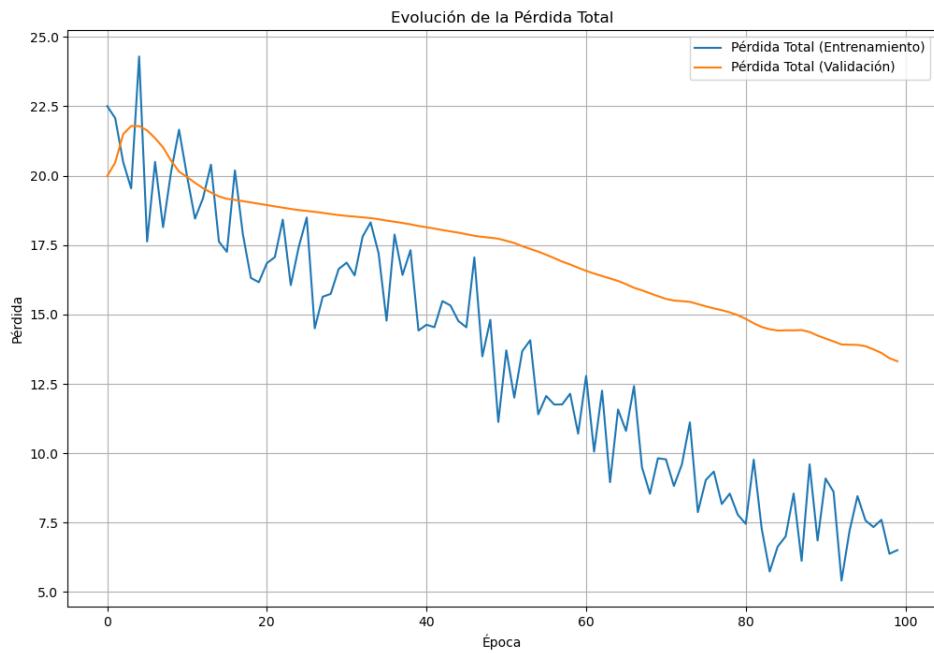
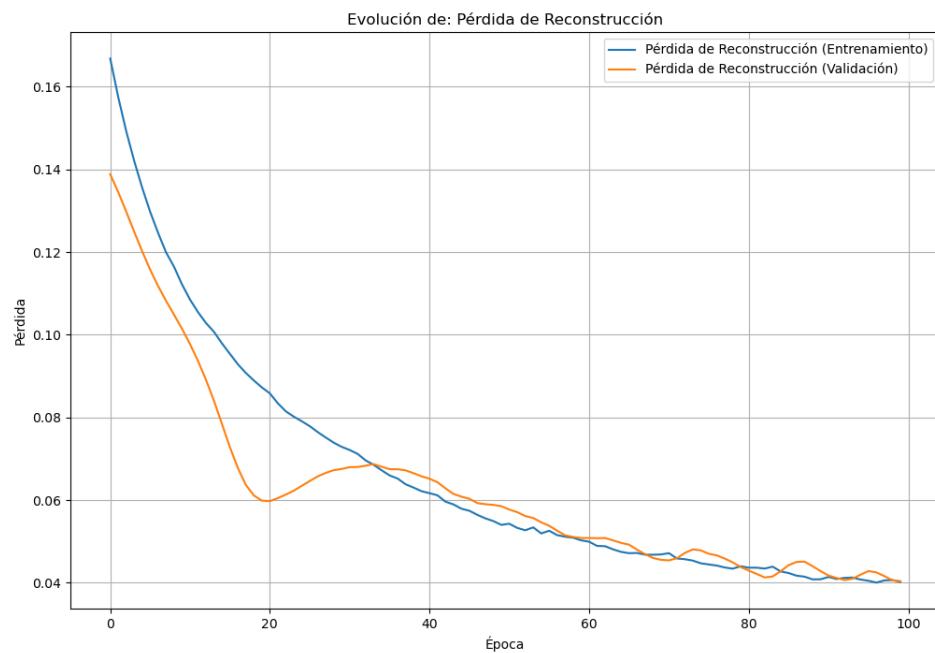


Figura 10.189: Esta gráfica ilustra la «Pérdida de Reconstrucción», que mide qué tan bien el autoencoder del modelo puede reconstruir la entrada original después de haberla comprimido en un espacio latente. Al igual que en la gráfica anterior, el eje Y es el valor de la pérdida y el eje X son las épocas.



1. Gráficas de los Experimentos Realizados

Figura 10.190: Este gráfico muestra la «Pérdida Triplet Semántica», una métrica clave que evalúa si el modelo puede diferenciar entre distintos glosarios, en este caso, las señas «brother», «cold» y «man». El objetivo es que las representaciones de un mismo glosario estén más cerca entre sí que las de glosarios diferentes. Igualmente, el eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.

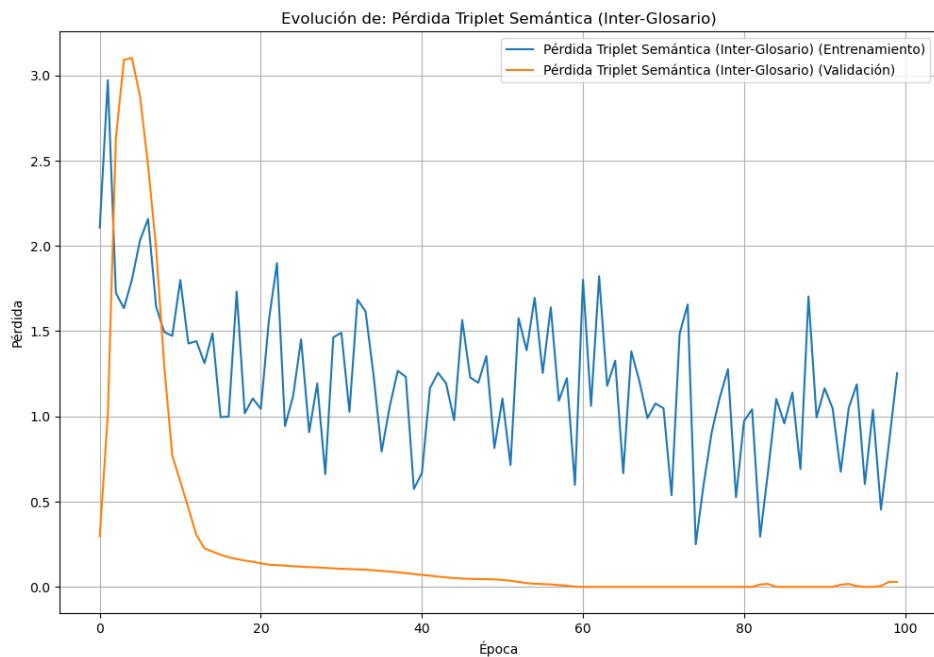
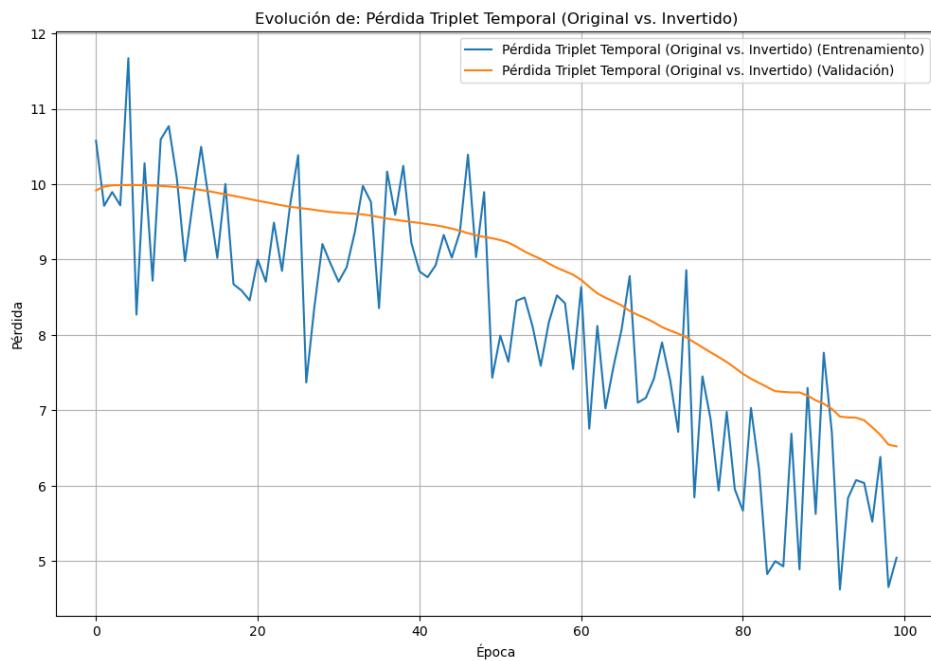


Figura 10.191: Esta visualización se enfoca en la sensibilidad temporal del modelo, puesto que, mide la diferencia entre la secuencia original de un video y su versión invertida. Esto quiere decir que el modelo debe aprender que una secuencia invertida es significativamente diferente de la original. El eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.



1. Gráficas de los Experimentos Realizados

Figura 10.192: Similar a la gráfica anterior, esta también evalúa la sensibilidad temporal, pero en este caso, compara la secuencia original con una versión donde los fotogramas han sido desordenados aleatoriamente. Donde el objetivo es que el modelo reconozca que una secuencia permutada es muy diferente de la original. El eje Y representa el valor de esta pérdida, mientras que el eje X indica las épocas de entrenamiento.

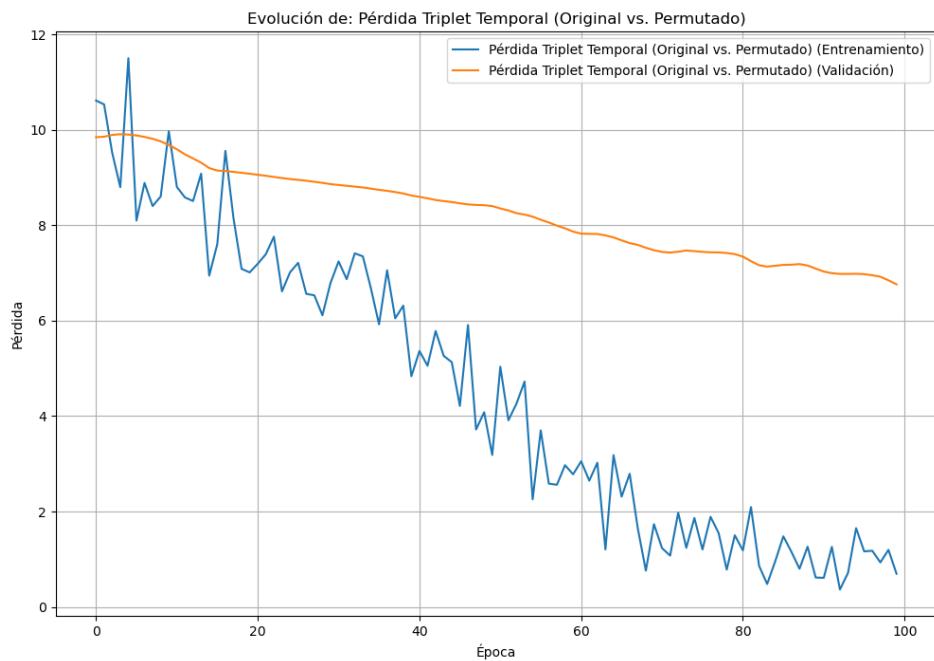
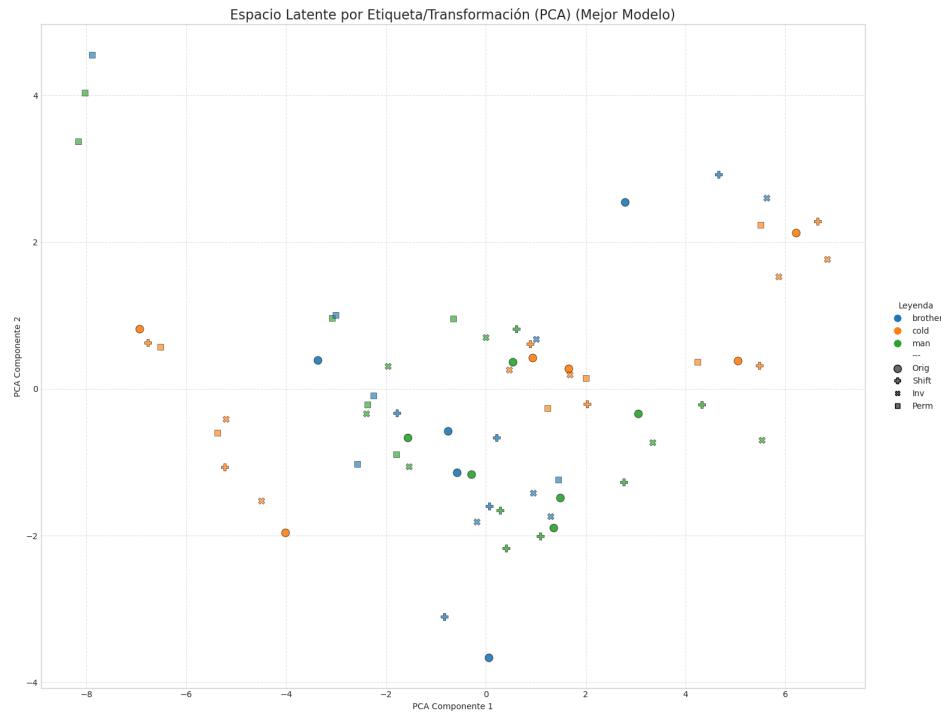


Figura 10.193: Esta grafica muestra el espacio latente en la mejor epoca (100) utilizando pca, donde las señas «brother» estan con color azul, las de «cold» en naranja y las de «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.194: Esta grafica muestra el espacio latente en la mejor epoca (100) utilizando pca, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.

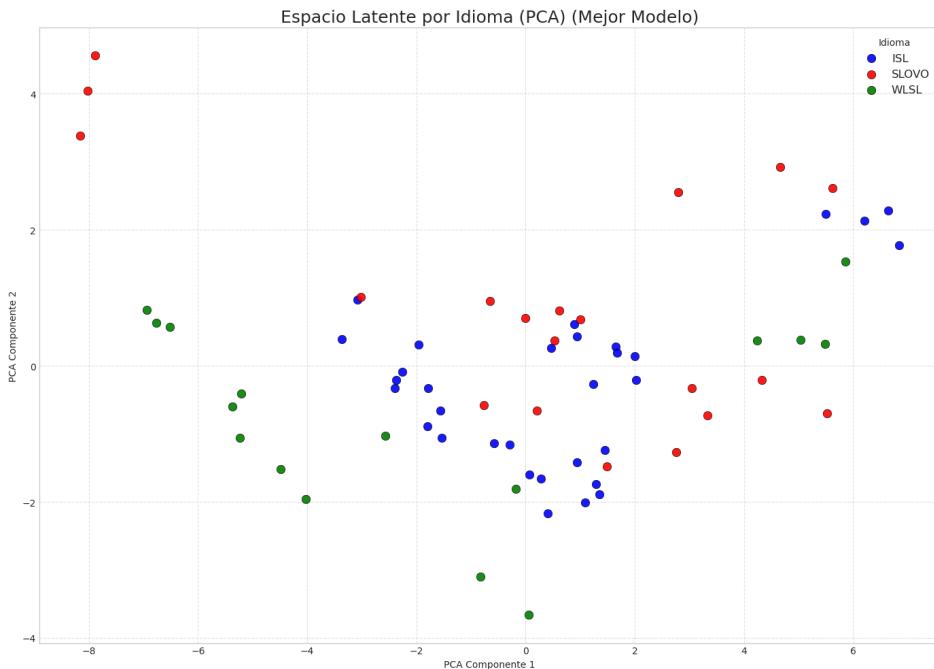
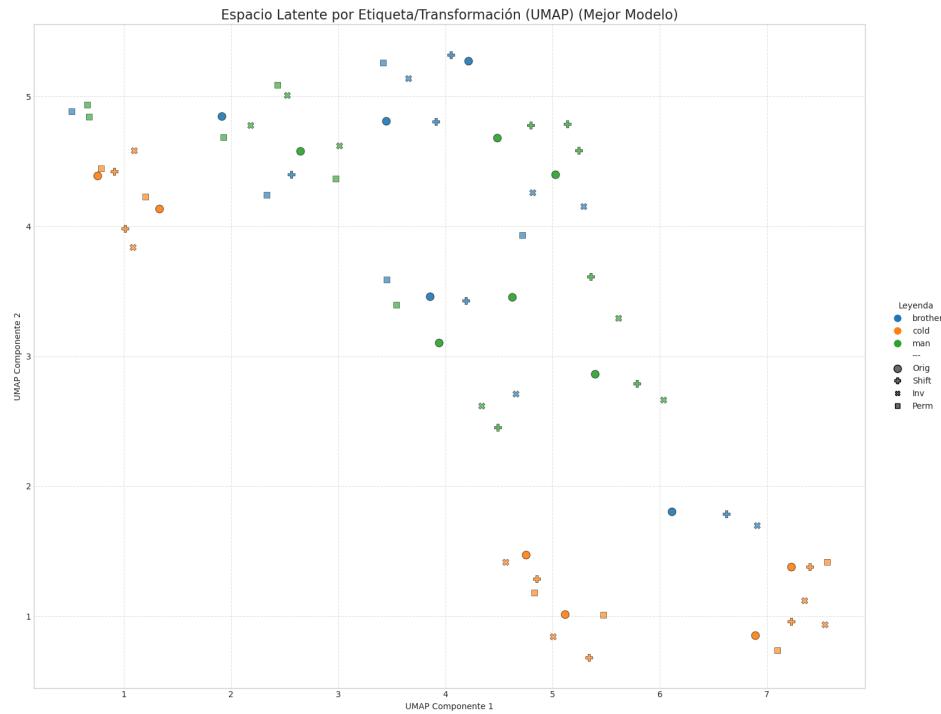


Figura 10.195: Esta grafica muestra el espacio latente en la mejor epoca (100) utilizando umap, donde las señas «brother» estan con color azul, las de «cold» en naranja y las de «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.196: Esta grafica muestra el espacio latente en la mejor epoca (100) utilizando umap, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.

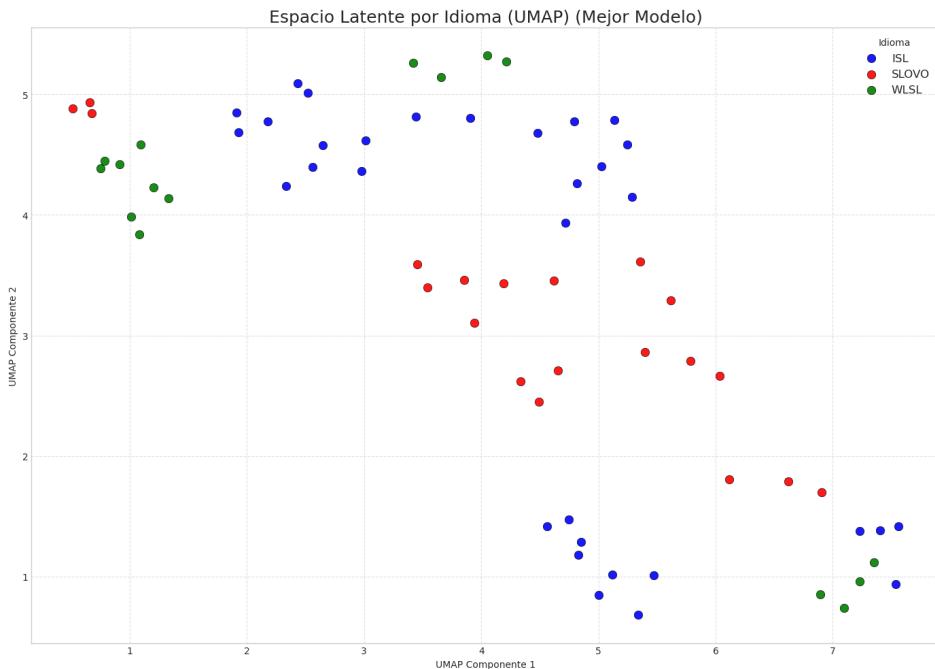
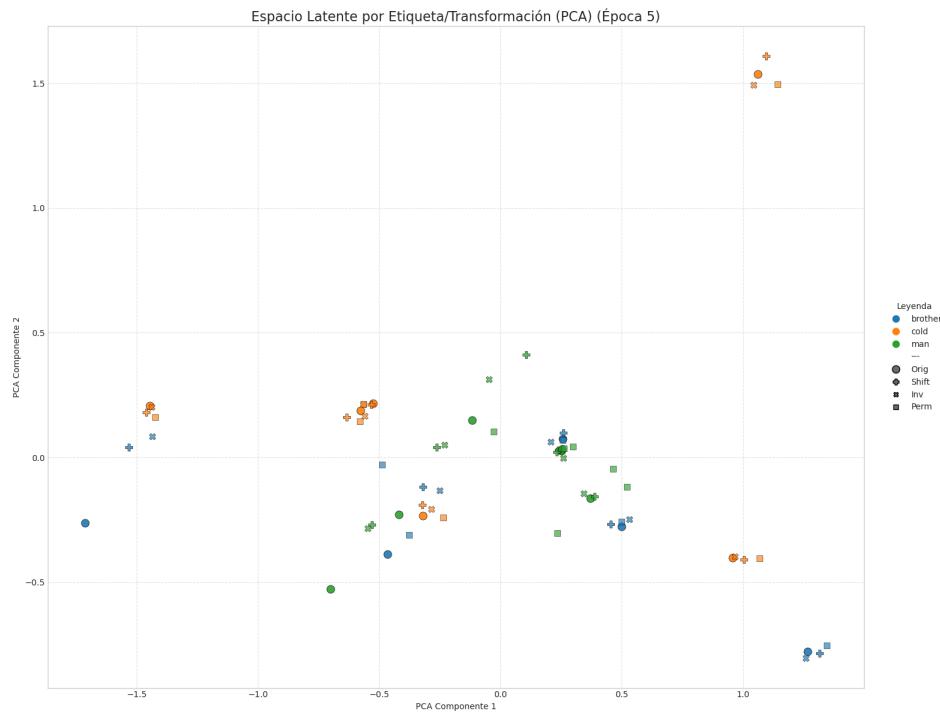


Figura 10.197: Esta grafica muestra el espacio latente en la epoca 5 utilizando pca, donde las señas «brother» estan con color azul, las de «cold» en naranja y las de «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.198: Esta grafica muestra el espacio latente en la epoca 5 utilizando pca, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.

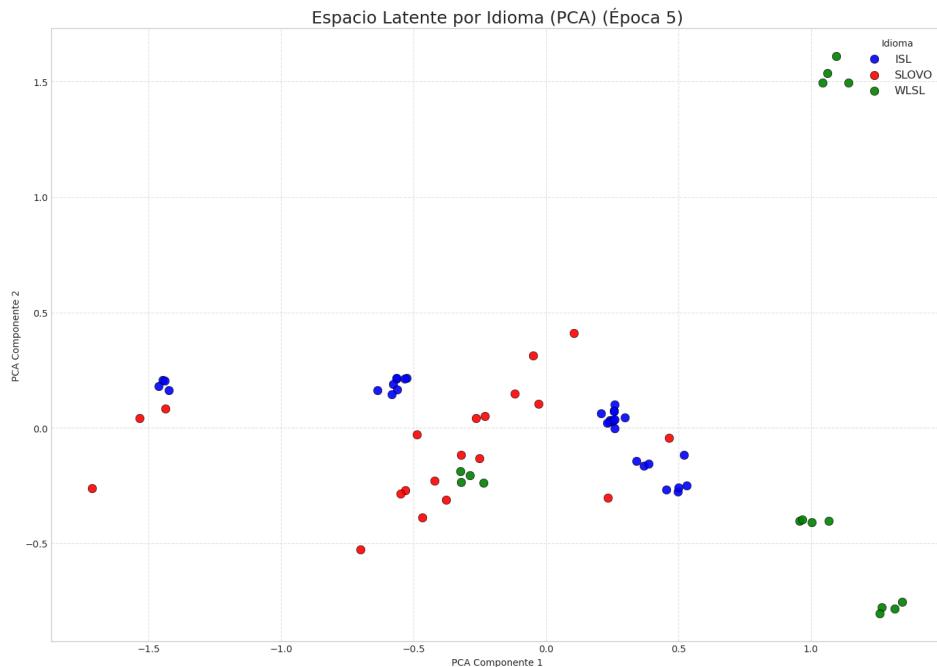
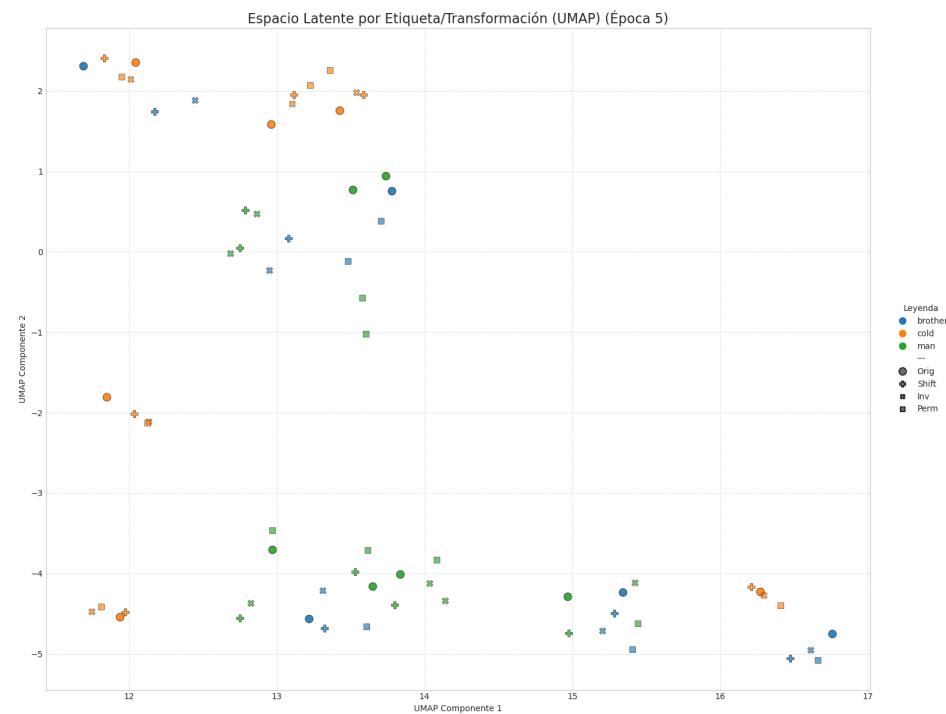


Figura 10.199: Esta grafica muestra el espacio latente en la epoca 5 utilizando umap, donde las señas «brother» estan con color azul, las de «cold» en naranja y las de «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.200: Esta grafica muestra el espacio latente en la epoca 5 utilizando umap, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.

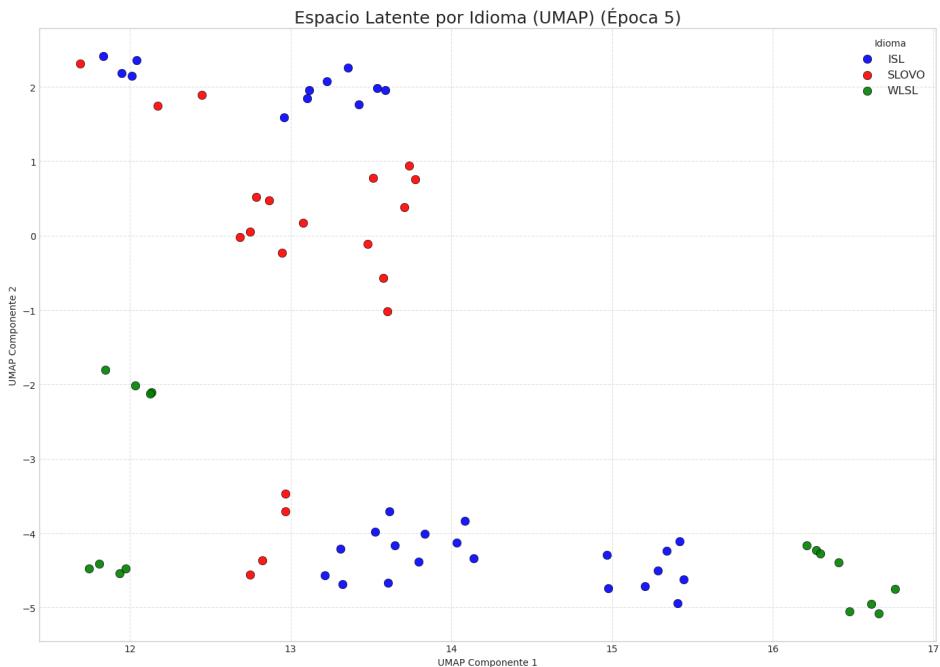
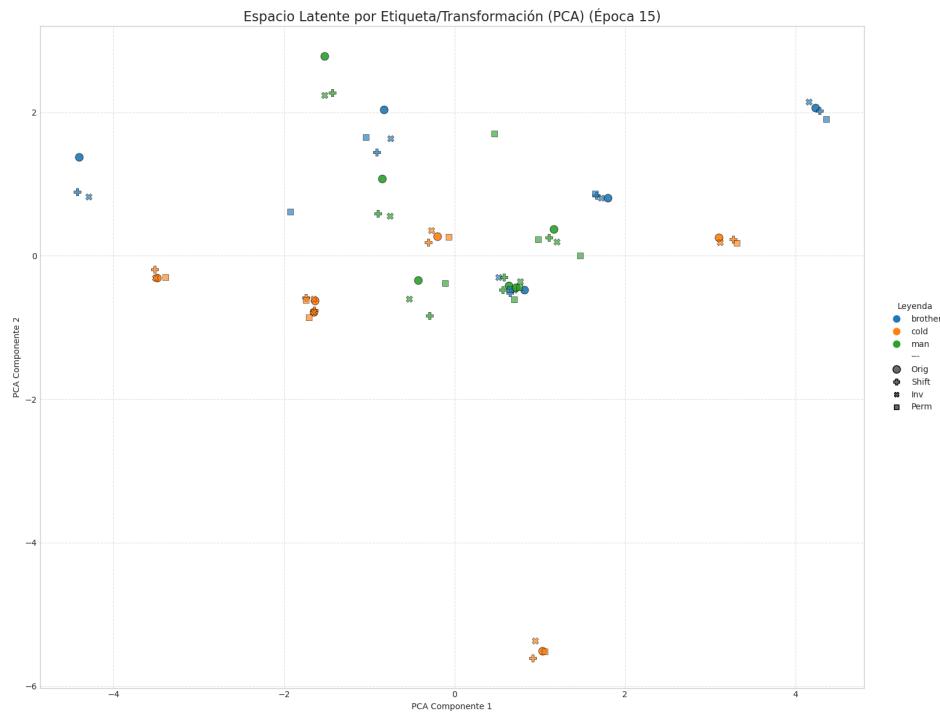


Figura 10.201: Esta grafica muestra el espacio latente en la epoca 15 utilizando pca, donde las señas «brother» estan con color azul, las de «cold» en naranja y las de «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.202: Esta grafica muestra el espacio latente en la epoca 15 utilizando pca, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.

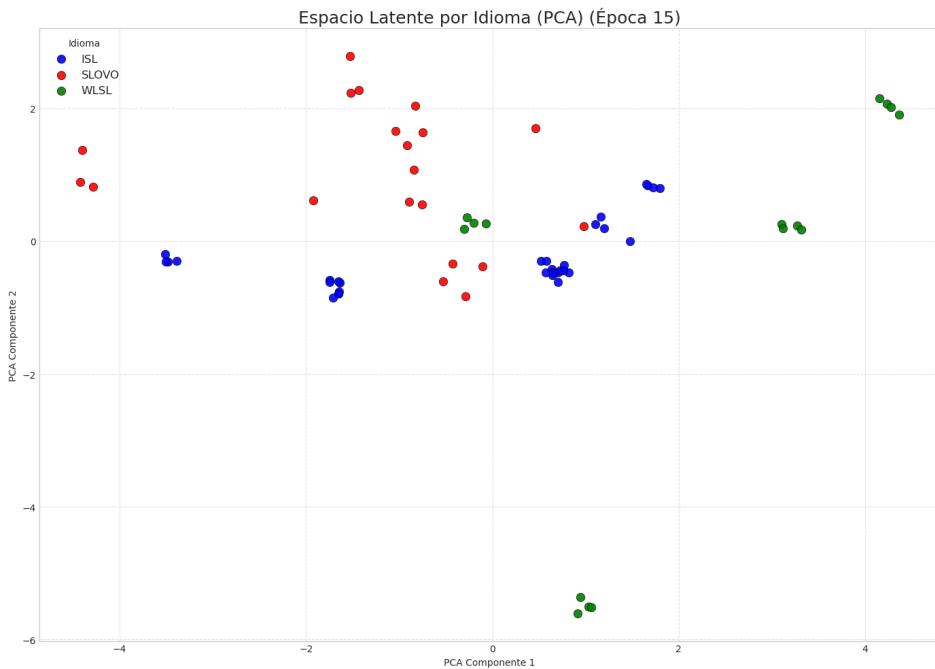
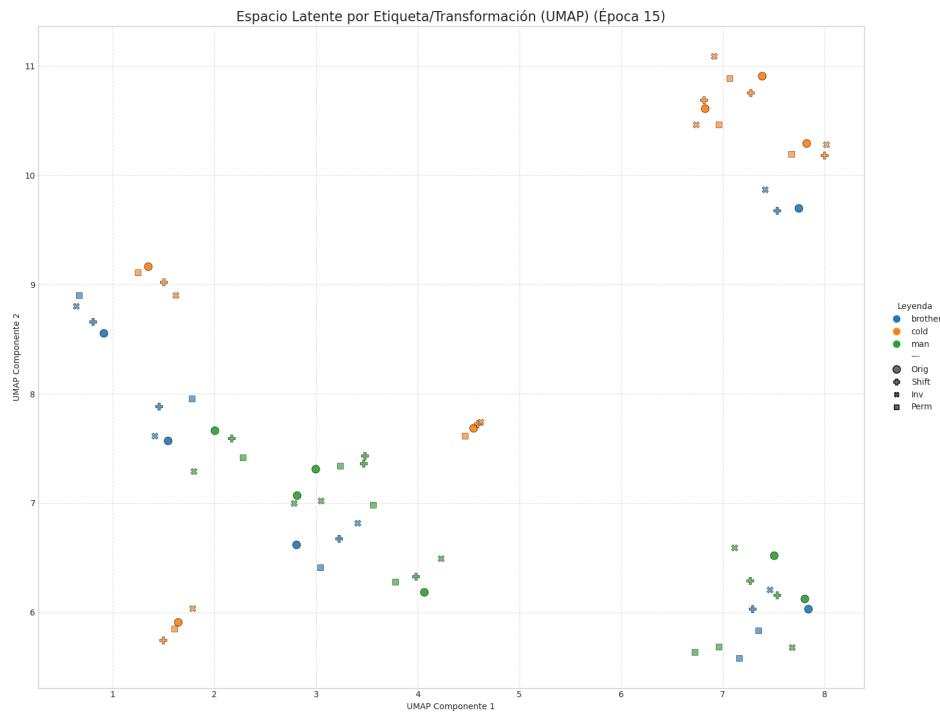


Figura 10.203: Esta grafica muestra el espacio latente en la epoca 15 utilizando umap, donde las señas «brother» estan con color azul, las de «cold» en naranja y las de «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.204: Esta grafica muestra el espacio latente en la epoca 15 utilizando umap, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.

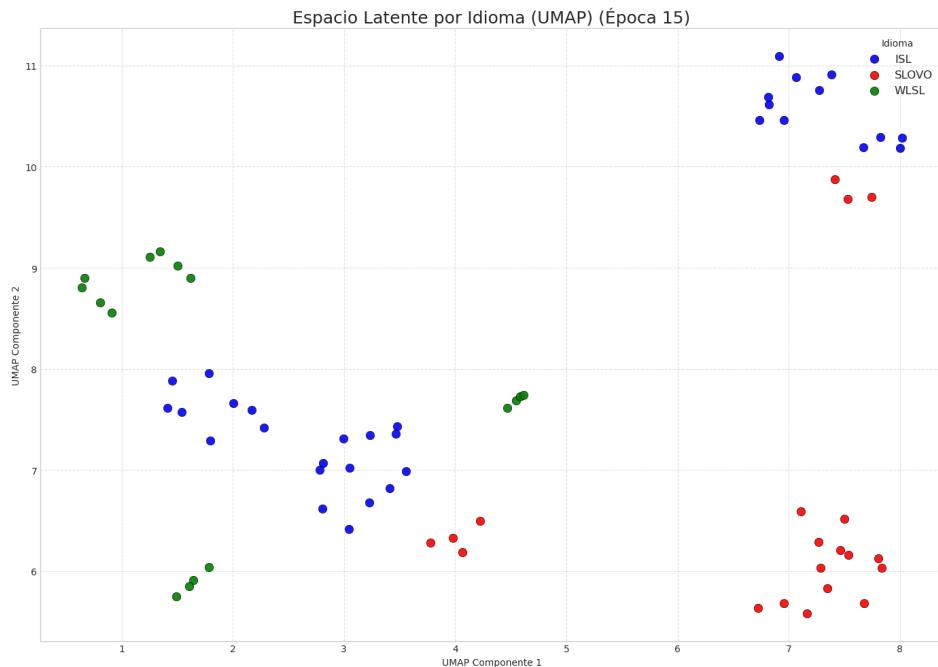
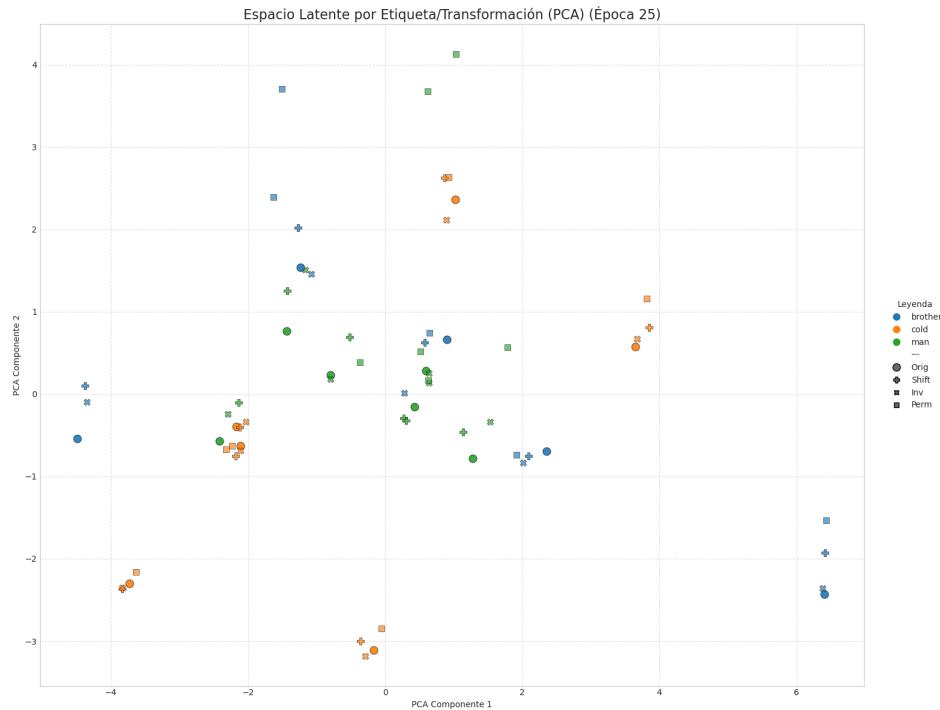


Figura 10.205: Esta grafica muestra el espacio latente en la epoca 25 utilizando pca, donde las señas «brother» estan con color azul, las de «cold» en naranja y las de «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.206: Esta grafica muestra el espacio latente en la epoca 25 utilizando pca, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.

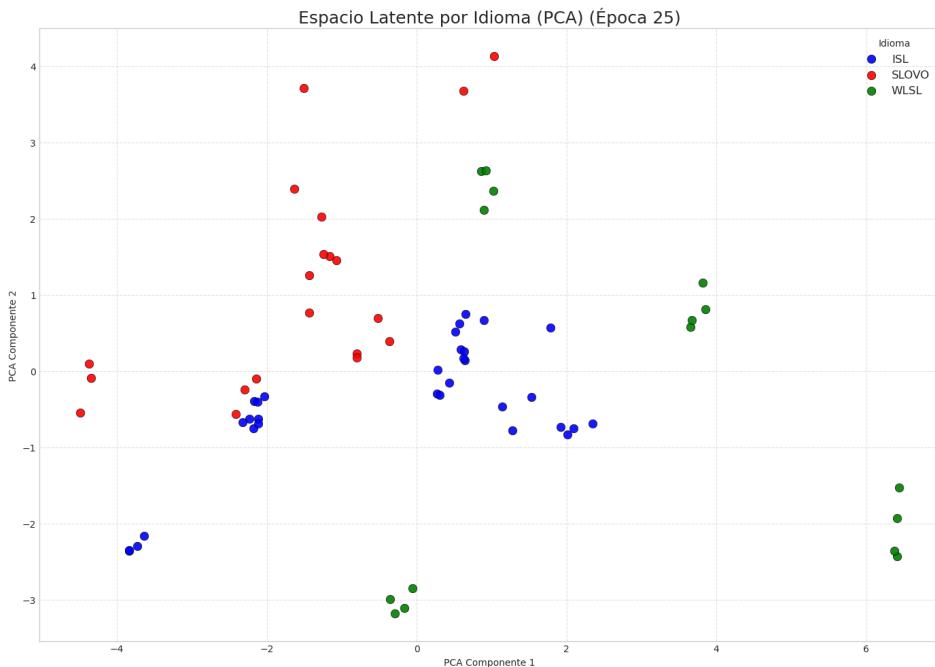
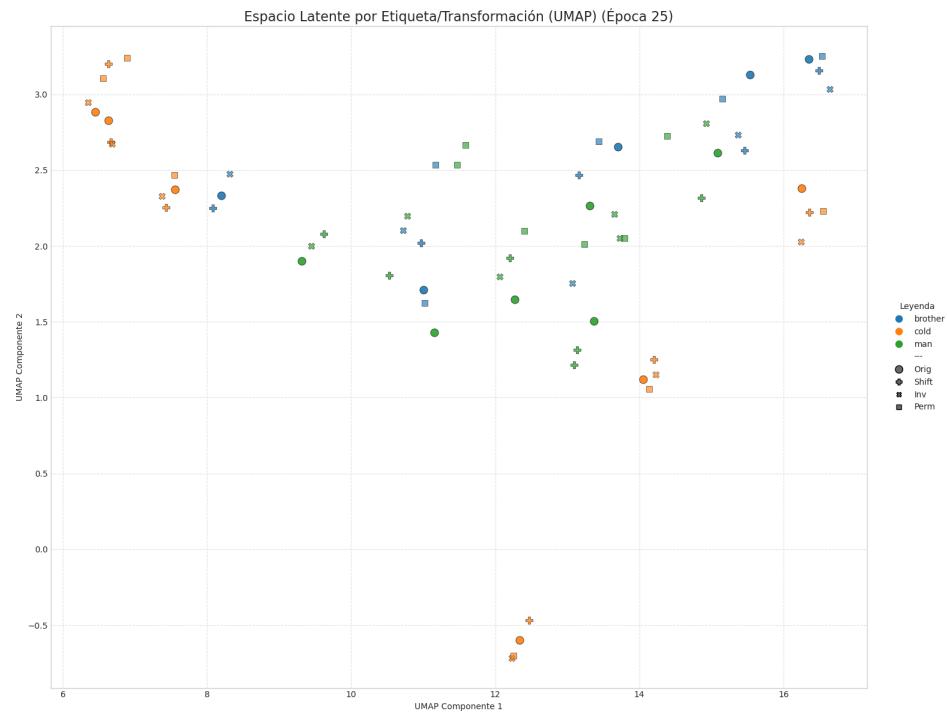


Figura 10.207: Esta grafica muestra el espacio latente en la epoca 25 utilizando umap, donde las señas «brother» estan con color azul, las de «cold» en naranja y las de «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.208: Esta grafica muestra el espacio latente en la epoca 25 utilizando umap, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.

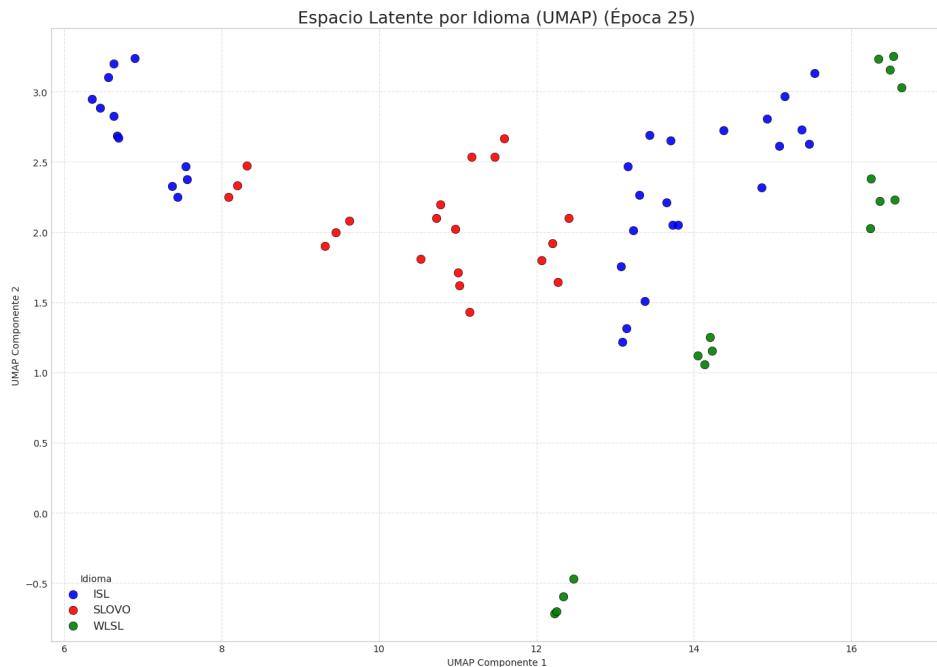
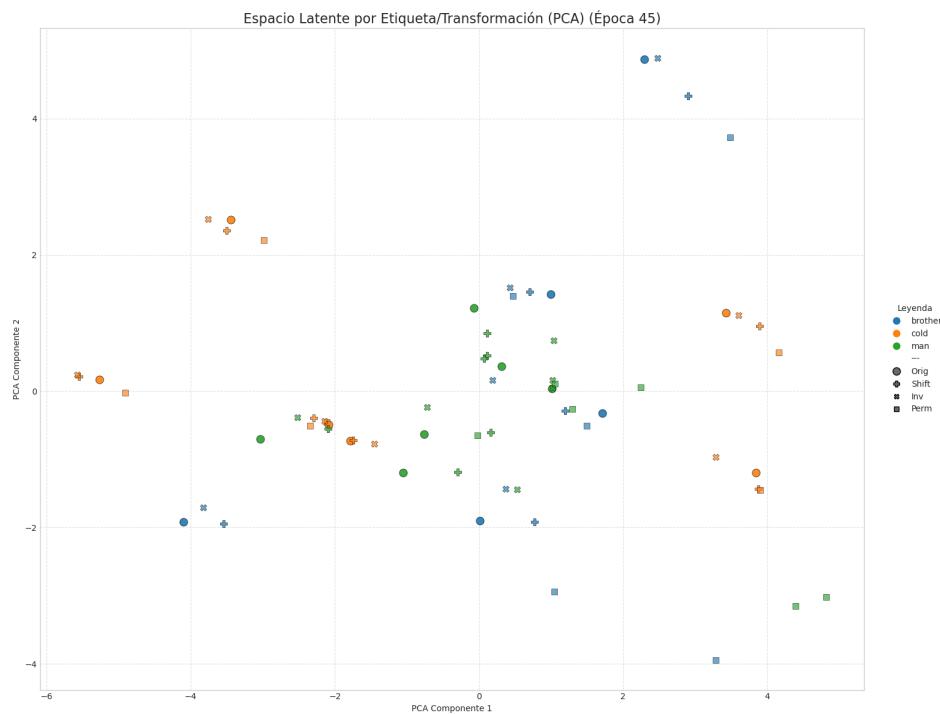


Figura 10.209: Esta grafica muestra el espacio latente en la epoca 45 utilizando pca, donde las señas «brother» estan con color azul, las de «cold» en naranja y las de «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.210: Esta grafica muestra el espacio latente en la epoca 45 utilizando pca, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.

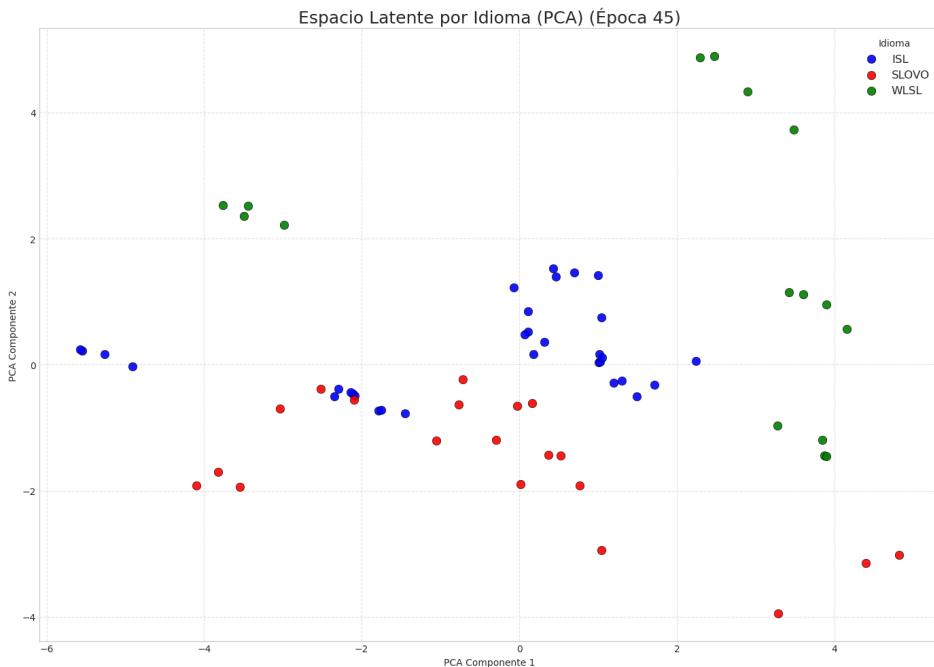
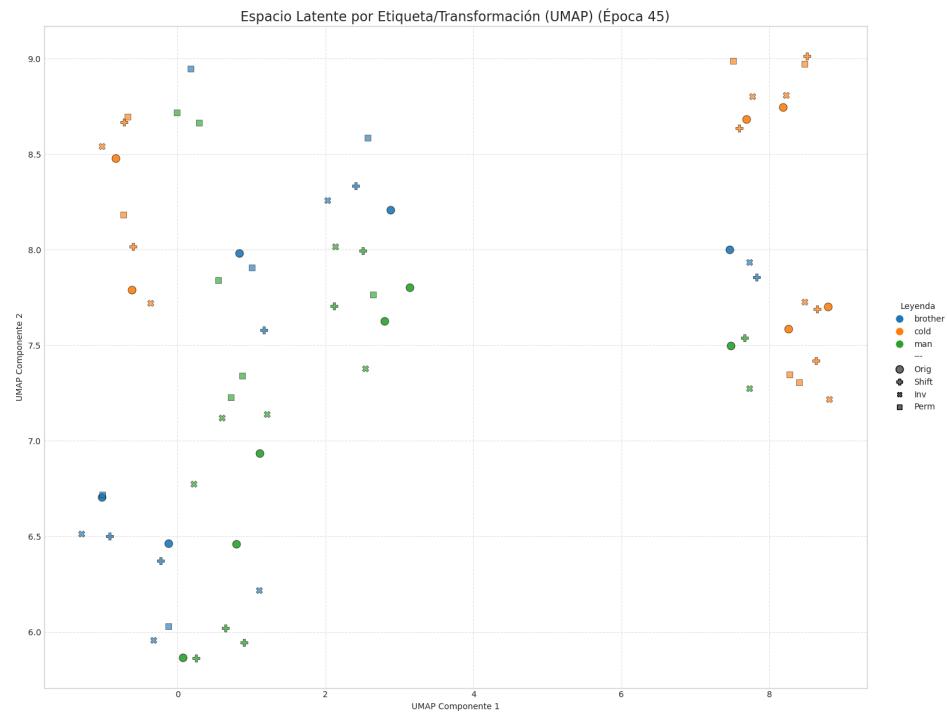


Figura 10.211: Esta grafica muestra el espacio latente en la epoca 45 utilizando umap, donde las señas «brother» estan con color azul, las de «cold» en naranja y las de «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.212: Esta grafica muestra el espacio latente en la epoca 45 utilizando umap, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.

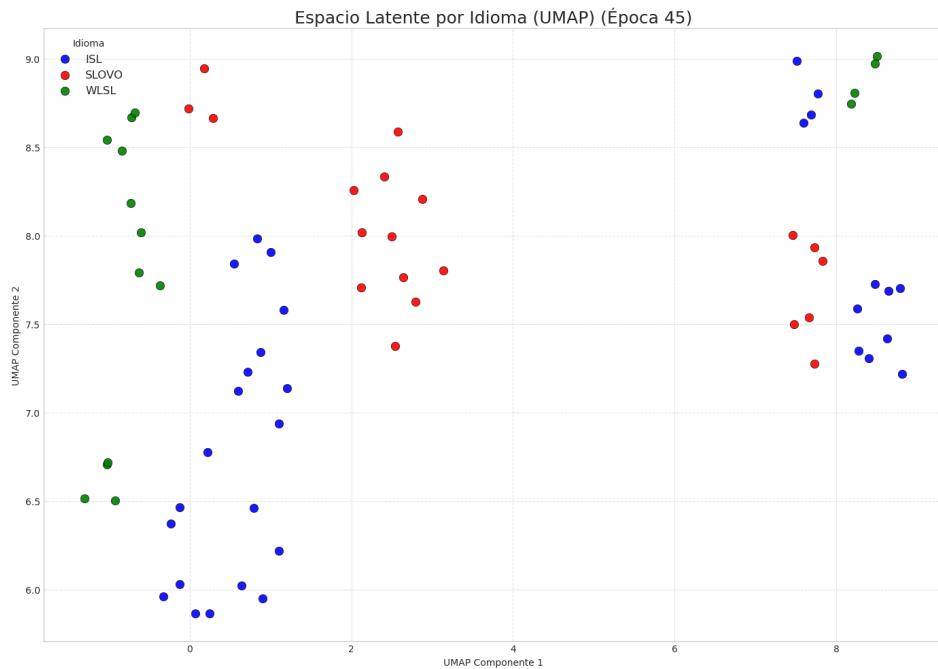
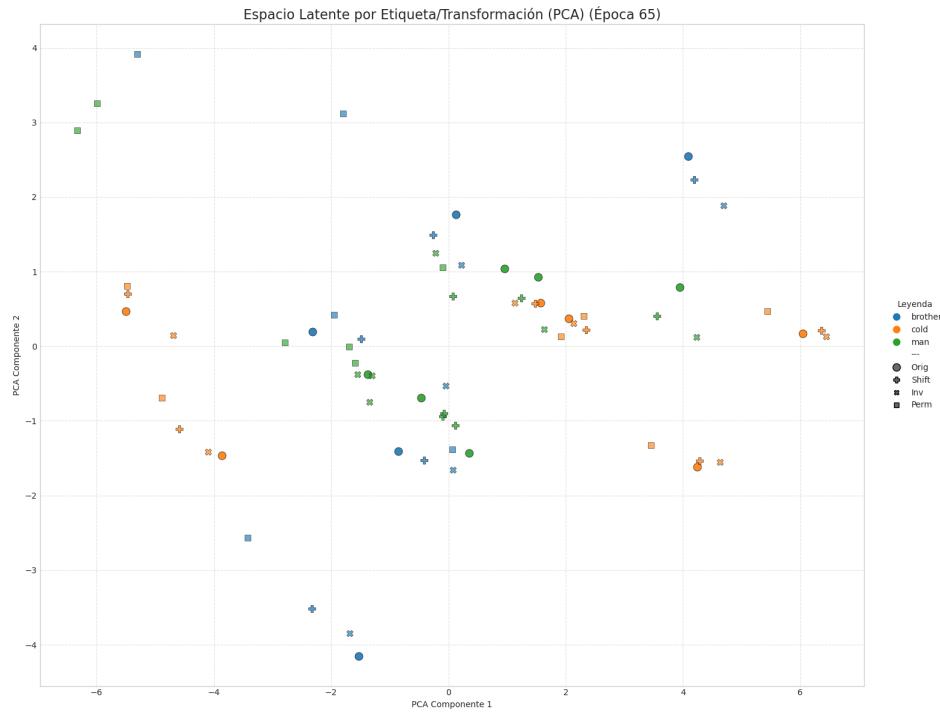


Figura 10.213: Esta grafica muestra el espacio latente en la epoca 65 utilizando pca, donde las señas «brother» estan con color azul, las de «cold» en naranja y las de «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.214: Esta grafica muestra el espacio latente en la epoca 65 utilizando pca, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.

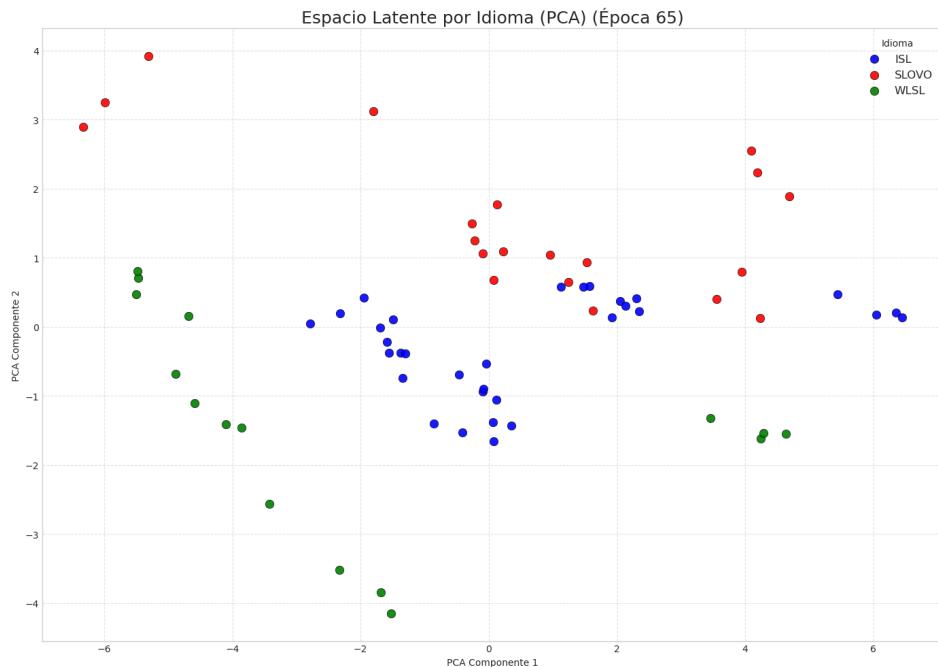
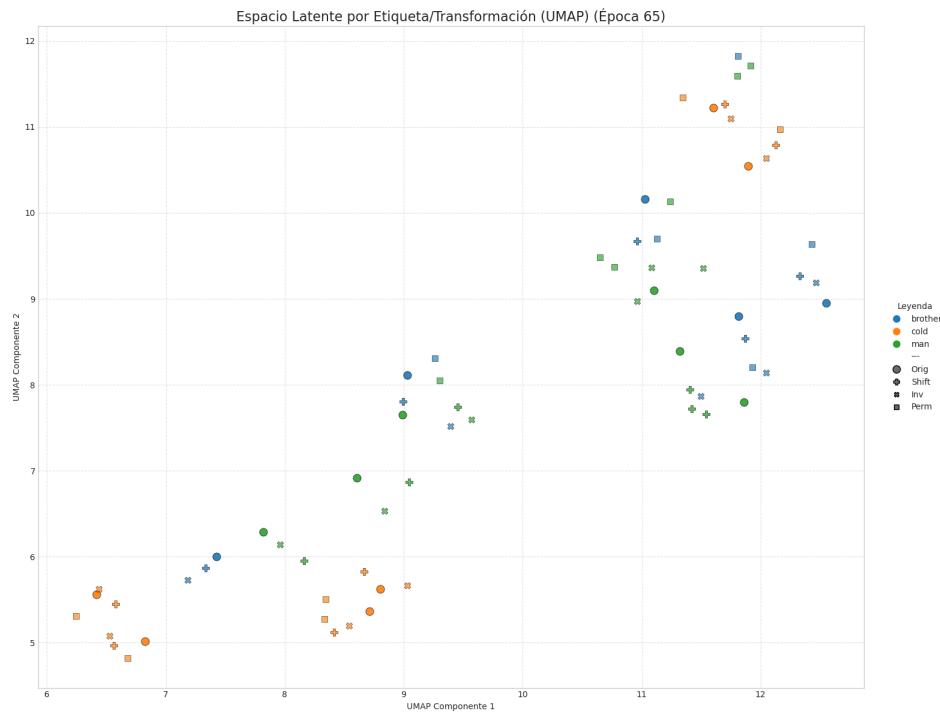


Figura 10.215: Esta grafica muestra el espacio latente en la epoca 65 utilizando umap, donde las señas «brother» estan con color azul, las de «cold» en naranja y las de «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.216: Esta grafica muestra el espacio latente en la epoca 65 utilizando umap, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.

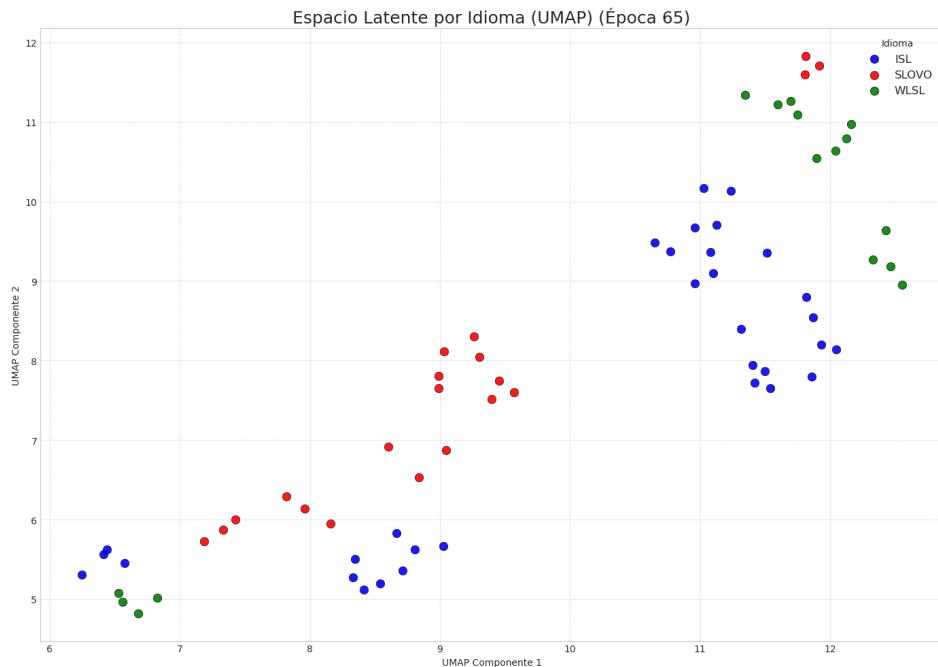
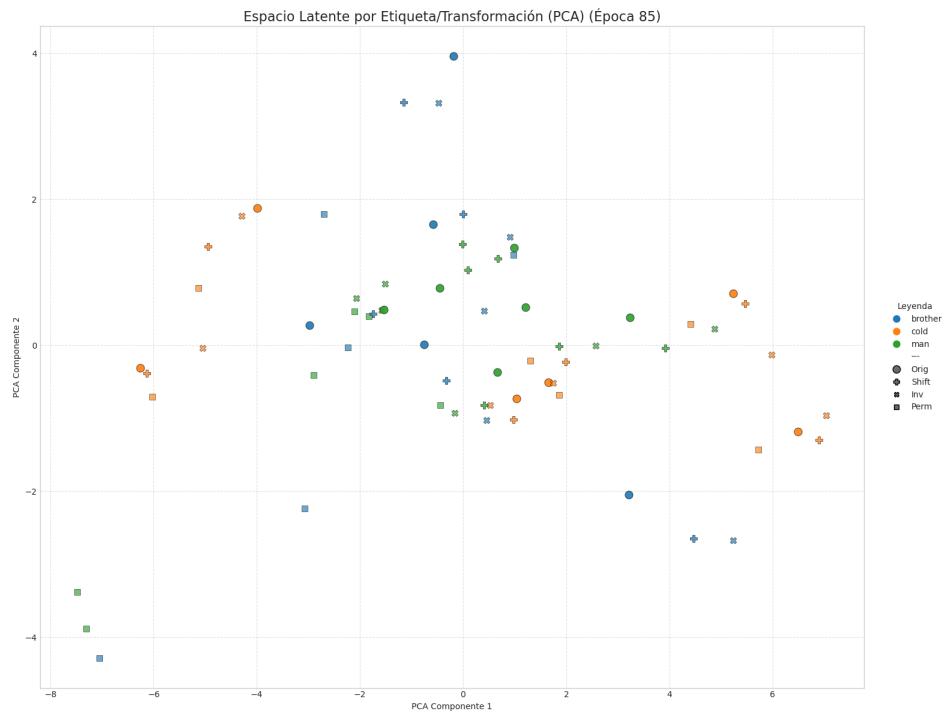


Figura 10.217: Esta grafica muestra el espacio latente en la epoca 85 utilizando pca, donde las señas «brother» estan con color azul, las de «cold» en naranja y las de «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.218: Esta grafica muestra el espacio latente en la epoca 85 utilizando pca, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.

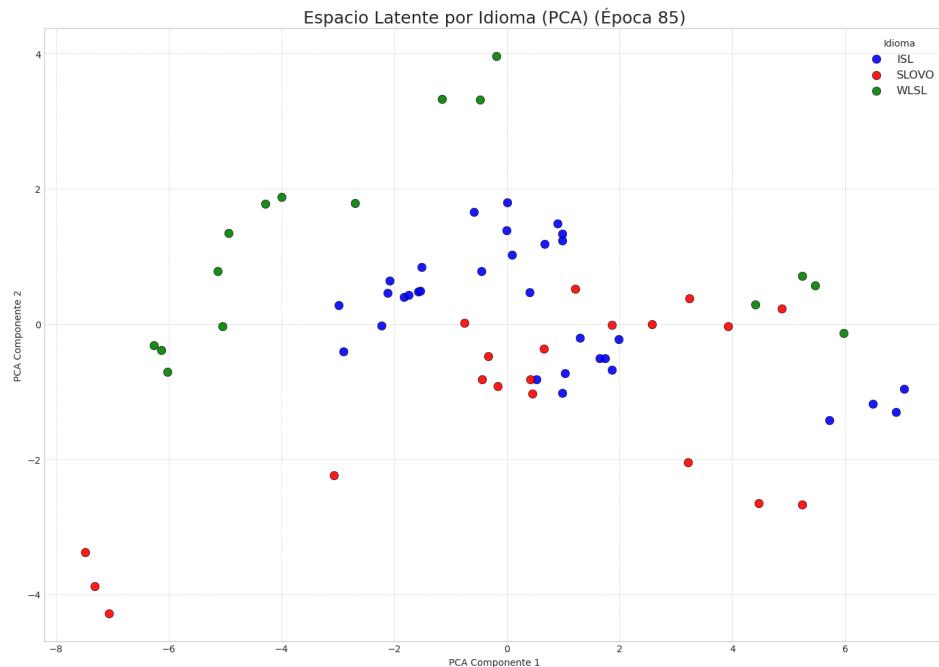
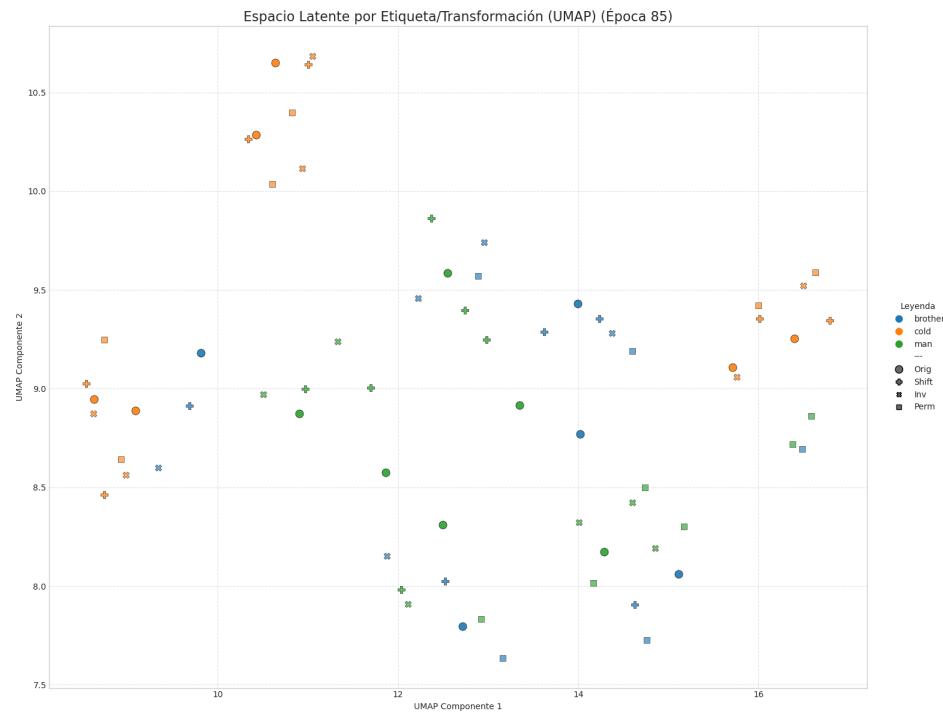
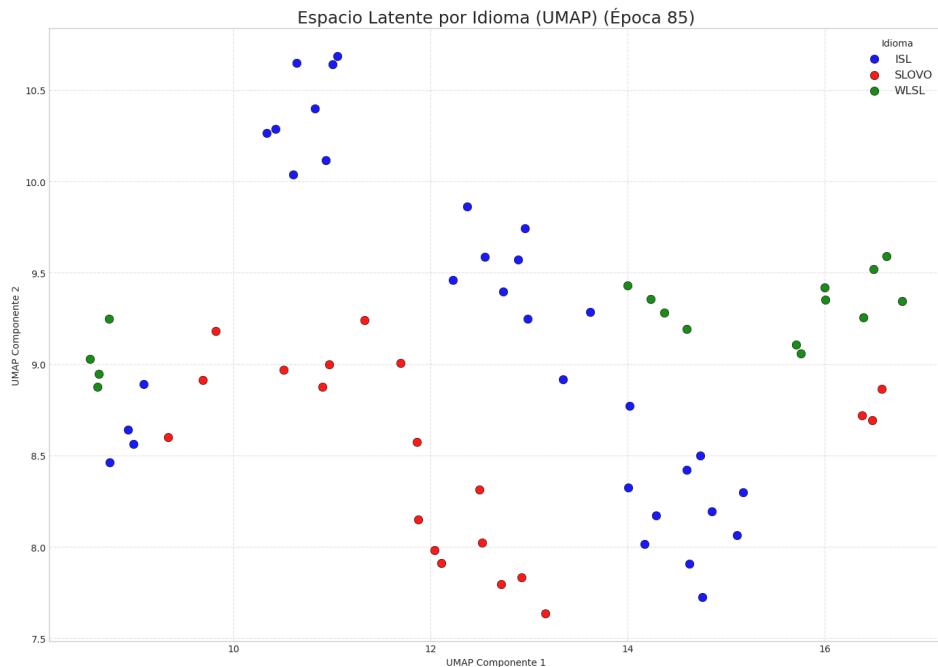


Figura 10.219: Esta grafica muestra el espacio latente en la epoca 85 utilizando umap, donde las señas «brother» estan con color azul, las de «cold» en naranja y las de «man» en verde. Las diferentes formas de los puntos representan una variante diferente del video, siendo el circulo el original, la cruz el desplazado, el cuadrado el desordenado, y la equis la invertida.



1. Gráficas de los Experimentos Realizados

Figura 10.220: Esta grafica muestra el espacio latente en la epoca 85 utilizando umap, donde las señas del lenguaje «WLSL» estan color verde, las del «SLOVO» en rojo y las del «ISL» en azul.



Bibliografía

- Adler, H. J. «Language complexities for deaf and hard of hearing individuals in their pursuit of a career in science, technology, engineering, mathematics, and medicine: Perspectives from an LSL/ASL user». En: *Ear And Hearing* (2025), págs. 851-855. DOI: [10.1097/aud.0000000000001637](https://doi.org/10.1097/aud.0000000000001637).
- Akarsh, A. et al. «Sign Language Recognition: Advancing Human-Computer Interaction through Machine Learning». En: *IEEE Xplore* 2 (2024), págs. 1-4. DOI: [10.1109/icetcs61022.2024.10544114](https://doi.org/10.1109/icetcs61022.2024.10544114).
- Amazon Web Services. *Tipos de instancias disponibles para su uso con Studio Classic*. Publicado el 24 de octubre de 2022. Amazon SageMaker. 2022. URL: https://docs.aws.amazon.com/es_es/sagemaker/latest/dg/notebooks-available-instance-types.html (visitado 05-07-2025).
- Bedoin, D. «Exploring identity building, language transmission and educational strategies for immigrant d/Deaf multilingual learners». En: *Journal of Multilingual and Multicultural Development* (2024), págs. 162-175. DOI: [10.1080/01434632.2024.2390570](https://doi.org/10.1080/01434632.2024.2390570).
- Bolton, T. et al. «A novel triplet loss architecture with visual explanation for detecting the unwanted rotation of bolts in safety-critical environments». En: *Engineering Applications of Artificial Intelligence* 156 (2025), págs. 1-16. DOI: [10.1016/j.engappai.2025.111097](https://doi.org/10.1016/j.engappai.2025.111097).
- Dave, I. et al. «TCLR: Temporal contrastive learning for video representation». En: *Computer Vision and Image Understanding* 219 (2022), págs. 1-9. DOI: [10.1016/j.cviu.2022.103406](https://doi.org/10.1016/j.cviu.2022.103406).
- Deng, Z. et al. «VTAN: A novel Video Transformer Attention-Based Network for Dynamic Sign Language Recognition». En: *Computers, Materials & Continua* 82.2 (2024), págs. 2793-2812. DOI: [10.32604/cmc.2024.057456](https://doi.org/10.32604/cmc.2024.057456).
- Dey, A., S. Biswas y D. Le. «Recognition of Wh-Question Sign Gestures in Video Streams using an Attention Driven C3D-BiLSTM Network». En: *Procedia Computer Science* 235 (2024), págs. 2920-2931. DOI: [10.1016/j.procs.2024.04.276](https://doi.org/10.1016/j.procs.2024.04.276).
- EBSCO Information Services, Inc. *Garbage in, garbage out (GIGO)*. s.f. EBSCO. URL: <https://www.ebsco.com/research-starters/computer-science/garbage-garbage-out-gigo> (visitado 04-07-2025).
- Fischer, R. y H. L. Lane. *Looking back: a reader on the history of deaf communities and their sign languages*. 1993. URL: <http://ci.nii.ac.jp/ncid/BA29032250>.
- Geng, X. et al. «No blind alignment but generation: A different view of continuous sign language recognition based on diffusion». En: *Pattern Recognition* (2025), págs. 1-11. DOI: [10.1016/j.patcog.2025.111960](https://doi.org/10.1016/j.patcog.2025.111960).

Bibliografía

- Gu, Y., H. Oku y M. Todoh. «American Sign Language Recognition and translation using Perception Neuron Wearable Inertial Motion Capture System». En: *Sensors* 24.2 (2024), págs. 1-15. DOI: [10.3390/s24020453](https://doi.org/10.3390/s24020453).
- Hashi, A. O., S. Z. M. Hashim y A. B. Asamah. «A Systematic Review of Hand Gesture Recognition: An update from 2018 to 2024». En: *IEEE Access* 12 (2024), págs. 1-35. DOI: [10.1109/access.2024.3421992](https://doi.org/10.1109/access.2024.3421992).
- Inamdar, R. et al. «Lips Reading Using 3D Convolution and LSTM». En: *Vellore Institute of Technology* (2023), págs. 1-6. DOI: [10.20944/preprints202312.0928.v1](https://doi.org/10.20944/preprints202312.0928.v1).
- Innocente, C. et al. «Deep Learning-Based Lip-Reading for Vocal Impaired Patient Rehabilitation». En: *Computer Modeling in Engineering & Sciences* 143.2 (2025), págs. 1355-1379. DOI: [10.32604/cmes.2025.063186](https://doi.org/10.32604/cmes.2025.063186).
- Jiang, X. et al. «A survey on Chinese sign language recognition: From traditional methods to Artificial Intelligence». En: *Computer Modeling in Engineering & Sciences* 140.1 (2024), págs. 1-40. DOI: [10.32604/cmes.2024.047649](https://doi.org/10.32604/cmes.2024.047649).
- Kapitanov, Alexander et al. *Slovo: Russian Sign Language Dataset*. arXiv:2305.14527. 2023. URL: <https://arxiv.org/abs/2305.14527> (visitado 05-07-2025).
- Khan, A. et al. «Deep Learning Approaches for Continuous Sign Language Recognition: A Comprehensive review». En: *IEEE Access* 13 (2025), págs. 1-21. DOI: [10.1109/access.2025.3554046](https://doi.org/10.1109/access.2025.3554046).
- Li, Dongxu et al. *WLASL: Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset*. s.f. WLASL. URL: <https://dxli94.github.io/WLASL/> (visitado 05-07-2025).
- Liu, J. et al. «TB-Net: Intra- and inter-video correlation learning for continuous sign language recognition». En: *Information Fusion* 109 (2024), págs. 1-10. DOI: [10.1016/j.inffus.2024.102438](https://doi.org/10.1016/j.inffus.2024.102438).
- Obi, Y. et al. «Sign language recognition system for communicating to people with disabilities». En: *Procedia Computer Science* 216 (2023), págs. 13-20. DOI: [10.1016/j.procs.2022.12.106](https://doi.org/10.1016/j.procs.2022.12.106).
- Parmar, B. et al. «I always feel like I'm the first deaf person they have ever met: "Deaf Awareness, Accessibility and Communication in the United Kingdom's National Health Service (NHS): How can we do better?"» En: *PLoS ONE* 20.5 (2025), págs. 1-18. DOI: [10.1371/journal.pone.0322850](https://doi.org/10.1371/journal.pone.0322850).
- Pathan, R. K. et al. «Sign language recognition using the fusion of image and hand landmarks through multi-headed convolutional neural network». En: *Scientific Reports* 13.1 (2023), págs. 1-11. DOI: [10.1038/s41598-023-43852-x](https://doi.org/10.1038/s41598-023-43852-x).
- Pham, D. N. y T. Rahne. «Entwicklung und Evaluation eines Deep-Learning-Algorithmus für die Worterkennung aus Lippenbewegungen für die deutsche Sprache». En: *HNO* 70.6 (2022), págs. 456-465. DOI: [10.1007/s00106-021-01143-9](https://doi.org/10.1007/s00106-021-01143-9).
- Probierz, B. et al. «Sign language interpreting - relationships between research in different areas - overview». En: *Annals of Computer Science and Information Systems* 35 (2023), págs. 213-223. DOI: [10.15439/2023f2503](https://doi.org/10.15439/2023f2503).
- Risang Baskoro. *WLASL Processed Dataset*. s.f. WLASL. URL: <https://www.kaggle.com/datasets/risangbaskoro/wlasl-processed> (visitado 05-05-2025).

- Shaw, E. e Y. Delaporte. *A Historical and Etymological Dictionary of American Sign Language: the origin and evolution of more than 500 signs*. 2015. URL: <http://ci.nii.ac.jp/ncid/BB1918069X>.
- Sridhar, A. et al. *INCLUDE: A Large Scale Dataset for Indian Sign Language Recognition*. Data set, ACM Multimedia 2020 (ACMMM2020). Zenodo. 2020. DOI: [10.1145/3394171.3413528](https://doi.org/10.1145/3394171.3413528). URL: <https://zenodo.org/records/4010759> (visitado 05-07-2025).
- Supalla, T. y P. Clark. *Sign Language Archaeology*. 2015. DOI: [10.2307/j.ctv2rcng45](https://doi.org/10.2307/j.ctv2rcng45).
- Villares, M., C. M. Saunders y N. Fey. «Comparison of dimensionality reduction techniques for the visualisation of chemical space in organometallic catalysis». En: *Artificial Intelligence Chemistry* 2.1 (2024), págs. 1-10. DOI: [10.1016/j.aichem.2024.100055](https://doi.org/10.1016/j.aichem.2024.100055).
- Wang, Z. et al. «PAC-Bayes information bottleneck». En: *Proceedings of the International Conference on Learning Representations*. 2022. URL: <https://openreview.net/forum?id=iLHOIDsPv1P>.
- Wu, J., Z. Huang y C. Liu. «Advancing video self-supervised learning via image foundation models». En: *Pattern Recognition Letters* 192 (2025), págs. 22-28. DOI: [10.1016/j.patrec.2025.03.015](https://doi.org/10.1016/j.patrec.2025.03.015).
- Xie, P. et al. «G2P-DDM: Generating Sign Pose Sequence from Gloss Sequence with Discrete Diffusion Model». En: *Proceedings of the AAAI Conference on Artificial Intelligence* 38.6 (2024), págs. 6234-6242. DOI: [10.1609/aaai.v38i6.28441](https://doi.org/10.1609/aaai.v38i6.28441).
- Xu, M. et al. «ATCM-Net: A deep learning method for phase unwrapping based on perception optimization and learning enhancement». En: *Optics & Laser Technology* 190 (2025), págs. 1-20. DOI: [10.1016/j.optlastec.2025.113185](https://doi.org/10.1016/j.optlastec.2025.113185).
- Yi, S., Z. Fan y D. Wu. «Batch feature standardization network with triplet loss for weakly-supervised video anomaly detection». En: *Image and Vision Computing* 120 (2022), págs. 1-9. DOI: [10.1016/j.imavis.2022.104397](https://doi.org/10.1016/j.imavis.2022.104397).
- Zeng, W. et al. «Clustering-Guided Pairwise Metric Triplet Loss for Person Reidentification». En: *IEEE Internet of Things Journal* 9.16 (2022), págs. 1-11. DOI: [10.1109/jiot.2022.3147950](https://doi.org/10.1109/jiot.2022.3147950).