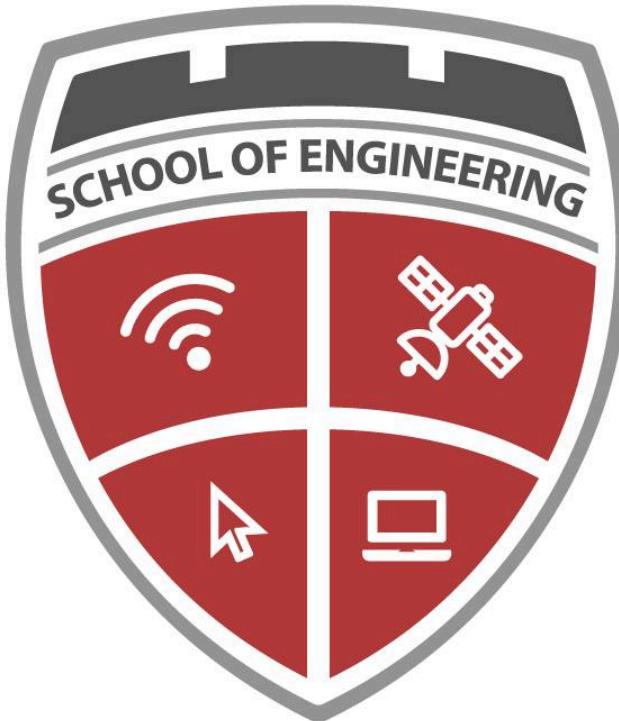


UNIVERSIDAD AUTÓNOMA GABRIEL RENÉ MORENO
SCHOOL OF ENGINEERING SOE - UNIDAD DE POSGRADO
MAESTRÍA EN CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL V1E2



**CLASIFICACIÓN ZERO-SHOT CON CLIP PARA INVENTARIO Y
ORGANIZACIÓN DOMÉSTICA**

**CASO: Objetos de Living Room (Dataset Personalizado)
LABORATORIO 3 - GRUPO 3**

Participantes: ♦ Yesika Luna
♦ Elvis Miranda
♦ Ivan Mamani Condori

Docente: Msc. Danny Luis Huanca Sevilla

Módulo: GENERATIVE MODELS AND COMPUTER VISION

Febrero, 2026

ÍNDICE

INTRODUCCIÓN.....	3
1. ANTECEDENTES.....	4
2. OBJETIVOS.....	4
3. DATASET.....	4
4. IMPLEMENTACIÓN Y RESULTADOS.....	6
4.1. Construcción de un dataset de imágenes domésticas organizado por ambiente.....	6
4.2. Definición de un catálogo de objetos y un conjunto de prompts textuales para CLIP.....	7
4.3. Implementación de inferencia zero-shot con CLIP, calculando similitudes imagen-texto.	9
4.4. Generación de gráficas de probabilidades que permitan interpretar los resultados del modelo.....	9
4.5. Creación de un reporte de inventario doméstico, indicando qué objetos se detectan con mayor confianza.....	10
5. DISCUSIÓN.....	10
7. CONCLUSIONES.....	11
8. ANEXOS.....	12

INTRODUCCIÓN

En los últimos años, las soluciones de hogar inteligente han evolucionado hacia sistemas capaces de comprender el entorno doméstico de forma automática, permitiendo mejorar la organización, seguridad y eficiencia de los espacios habitacionales. En este contexto, HomeVision surge como una startup enfocada en el desarrollo de tecnologías basadas en visión por computadora que faciliten la gestión inteligente del hogar. Sin embargo, muchas de las soluciones actuales requieren sensores específicos, configuraciones manuales o modelos supervisados que demandan grandes volúmenes de datos etiquetados, lo que incrementa los costos y tiempos de implementación.

Como alternativa a estos enfoques tradicionales, HomeVision propone el uso de modelos multimodales fundacionales capaces de interpretar imágenes y lenguaje natural de manera conjunta. En particular, el modelo CLIP permite realizar clasificación zero-shot, es decir, identificar objetos en imágenes sin haber sido entrenado explícitamente para dichas clases. Esta capacidad abre la posibilidad de construir sistemas flexibles que puedan adaptarse a distintos ambientes del hogar sin necesidad de procesos extensivos de entrenamiento o etiquetado manual.

En este laboratorio se desarrolla una prueba de concepto orientada a la construcción de un sistema de inventario visual doméstico para ambientes como la sala de estar. A partir de un dataset propio generado con imágenes reales capturadas desde un dispositivo móvil (iPhone X) y gestionadas mediante Roboflow, se implementa un pipeline de clasificación zero-shot utilizando CLIP. El objetivo es evaluar la viabilidad de este enfoque para identificar objetos presentes en el entorno doméstico mediante descripciones textuales, facilitando así la organización automática del hogar sin recurrir a modelos supervisados tradicionales.

1. ANTECEDENTES

El avance reciente en modelos de aprendizaje profundo ha permitido el desarrollo de enfoques capaces de realizar tareas de clasificación sin requerir entrenamiento específico sobre cada conjunto de datos. En este contexto, el paradigma zero-shot learning ha emergido como una alternativa eficiente frente a los métodos tradicionales supervisados, permitiendo que los modelos generalicen hacia clases no vistas previamente mediante el uso de representaciones semánticas. Esta capacidad resulta especialmente relevante en escenarios donde la recolección y anotación de grandes volúmenes de datos etiquetados es costosa o limitada.

Paralelamente, el crecimiento en la disponibilidad de dispositivos móviles con cámaras de alta calidad ha facilitado la generación de datasets personalizados en contextos reales. En lugar de depender exclusivamente de conjuntos de datos estandarizados, es posible construir colecciones de imágenes adaptadas a dominios específicos, como ambientes domésticos. Este enfoque permite evaluar el desempeño de modelos modernos bajo condiciones más cercanas al uso práctico, integrando datos capturados con dispositivos cotidianos, como smartphones, y plataformas de gestión de datasets, contribuyendo así a estudios aplicados en visión computacional y clasificación semántica.

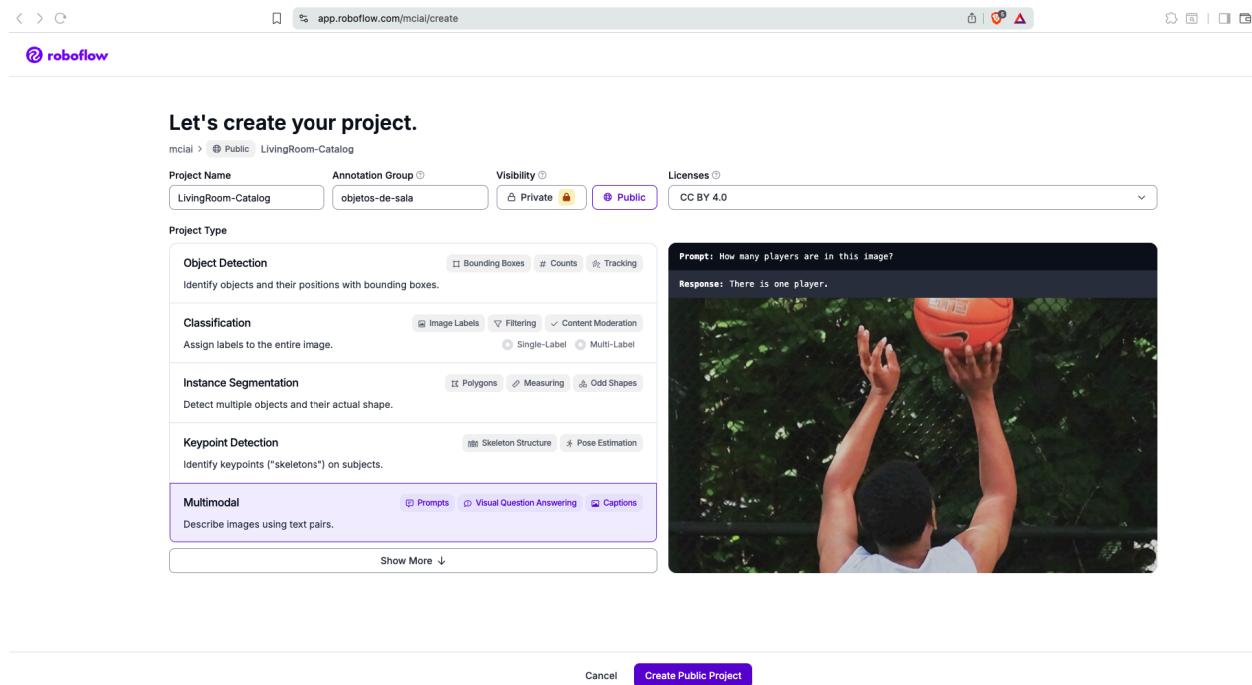
2. OBJETIVOS

Evaluar la capacidad de un modelo de clasificación basado en zero-shot learning para identificar objetos presentes en imágenes de un entorno doméstico, utilizando un dataset propio generado a partir de fotografías capturadas con un iPhone X, con el fin de analizar su desempeño en un escenario real sin entrenamiento específico sobre las clases evaluadas.

3. DATASET

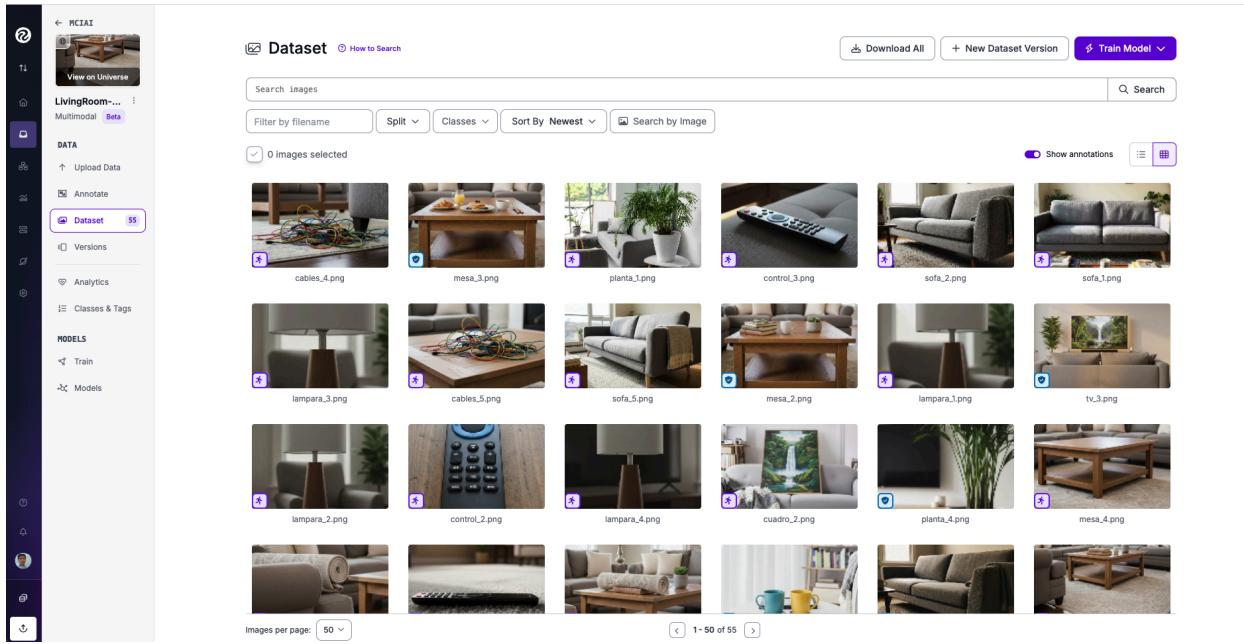
Para el desarrollo de este laboratorio se construyó un dataset propio orientado a representar un entorno doméstico real. Las imágenes fueron capturadas utilizando un iPhone X, lo que permitió obtener datos visuales en condiciones naturales de

iluminación, perspectiva y disposición de objetos. Este enfoque buscó alejarse de datasets sintéticos o altamente curados, con el objetivo de evaluar el comportamiento del modelo en escenarios más cercanos a su aplicación práctica.



The screenshot shows the Roboflow web interface for creating a new project. At the top, it says "Let's create your project." and displays the URL "app.roboflow.com/mcial/create". Below this, there are fields for "Project Name" (set to "LivingRoom-Catalog"), "Annotation Group" (set to "objetos-de-sala"), "Visibility" (set to "Public"), and "Licenses" (set to "CC BY 4.0"). On the right, there's a preview window showing a basketball player from behind, reaching up to shoot a basketball. A prompt at the top of the preview window asks "Prompt: How many players are in this image?", and the response "Response: There is one player." is shown below it. The interface also includes sections for "Object Detection", "Classification", "Instance Segmentation", "Keypoint Detection", and "Multimodal". The "Multimodal" section is currently selected and shows options for "Prompts", "Visual Question Answering", and "Captions". At the bottom right of the interface, there is a "Create Public Project" button.

El conjunto de datos fue estructurado y organizado utilizando la plataforma Roboflow, lo que permitió gestionar las imágenes, sus etiquetas y su posterior uso dentro del flujo experimental. A través de esta herramienta se consolidó el dataset denominado ***LivingRoom Catalog***, que contiene diferentes objetos presentes en una sala de estar, tales como muebles, dispositivos electrónicos y elementos decorativos.



El dataset generado se encuentra disponible públicamente en el siguiente enlace, lo que garantiza la reproducibilidad del experimento y la transparencia del proceso:

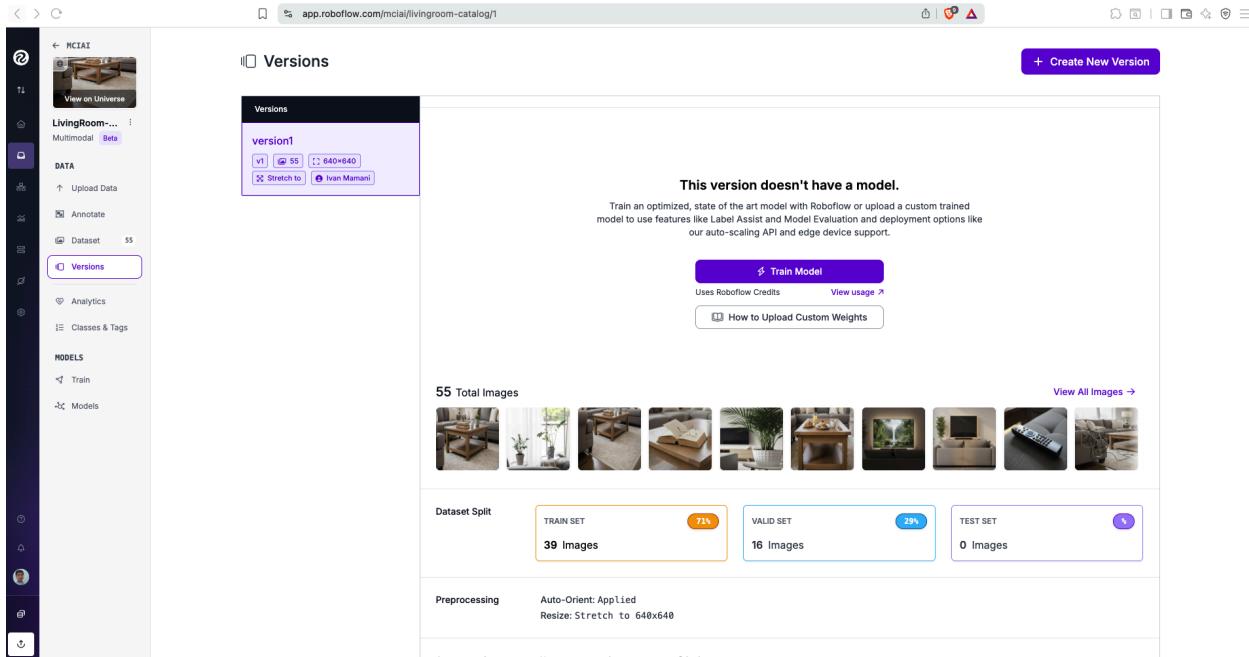
<https://app.roboflow.com/mciai/livingroom-catalog/1>

4. IMPLEMENTACIÓN Y RESULTADOS

En esta sección se describe el proceso técnico seguido para la aplicación del enfoque zero-shot learning mediante el modelo CLIP sobre el dataset construido. Se detalla la preparación de los datos, la definición del espacio semántico textual, la ejecución del proceso de inferencia y la interpretación de los resultados obtenidos a partir de las similitudes imagen-texto. Finalmente, se presenta la forma en que estos resultados permiten generar un inventario doméstico basado en reconocimiento visual sin necesidad de entrenamiento supervisado adicional.

4.1. Construcción de un dataset de imágenes domésticas organizado por ambiente

Se construyó un dataset compuesto por imágenes reales capturadas en entornos domésticos, específicamente un ambiente de sala o living room. Esta organización permitió estructurar la información visual de manera contextual, facilitando la posterior interpretación semántica por parte del modelo.

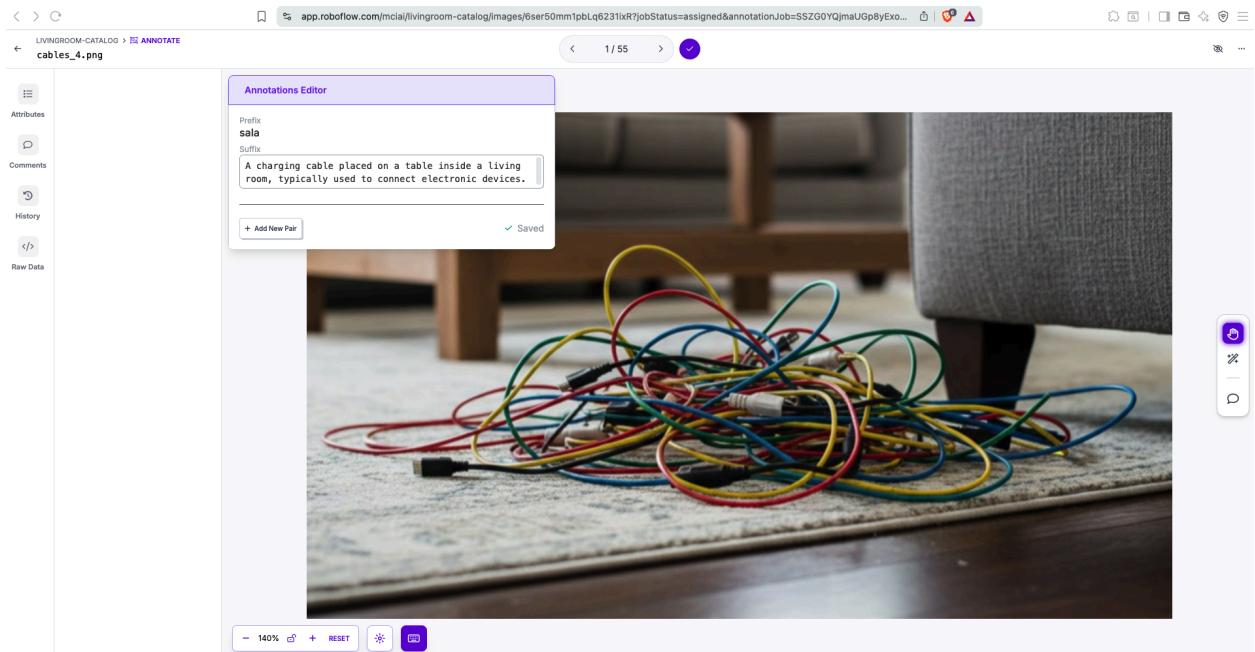


Las imágenes fueron previamente organizadas dentro del flujo de trabajo del dataset generado en Roboflow, asegurando consistencia en nombres, formato y accesibilidad para su posterior procesamiento.

4.2. Definición de un catálogo de objetos y un conjunto de prompts textuales para CLIP

A partir de la inspección del dataset, se definió un catálogo de objetos representativos del entorno doméstico, tales como muebles, elementos decorativos y objetos funcionales presentes en las escenas como ser: Televisor, control remoto, cuadros, plantas, mesas, libros, tazas, lámparas, alfombras, sofá, cables con ***un total de 55 imágenes, 5 fotos por objeto.***

Posteriormente, se generó un conjunto de descripciones textuales (prompts) que representan posibles clases semánticas. Estas descripciones fueron utilizadas como candidatas de clasificación en el modelo CLIP, permitiendo establecer correspondencias entre el contenido visual y su representación lingüística sin necesidad de entrenamiento específico.



Versions

version1

Generated on Feb 15, 2026 by Ivan Mamani

Download

Jupyter Terminal Raw URL

Paste this snippet into a [notebook from our model library](#) to download and unzip your dataset:

```
!pip install roboflow
from roboflow import Roboflow
rf = Roboflow(api_key="REDACTED")
project = rf.workspace("mciai").project("livingroom-catalog")
version = project.version(1)
dataset = version.download("openai")
```

Warning: Do not share this snippet beyond your team, it contains a private key that is tied to your Roboflow account.
Acceptable use policy applies.

Done

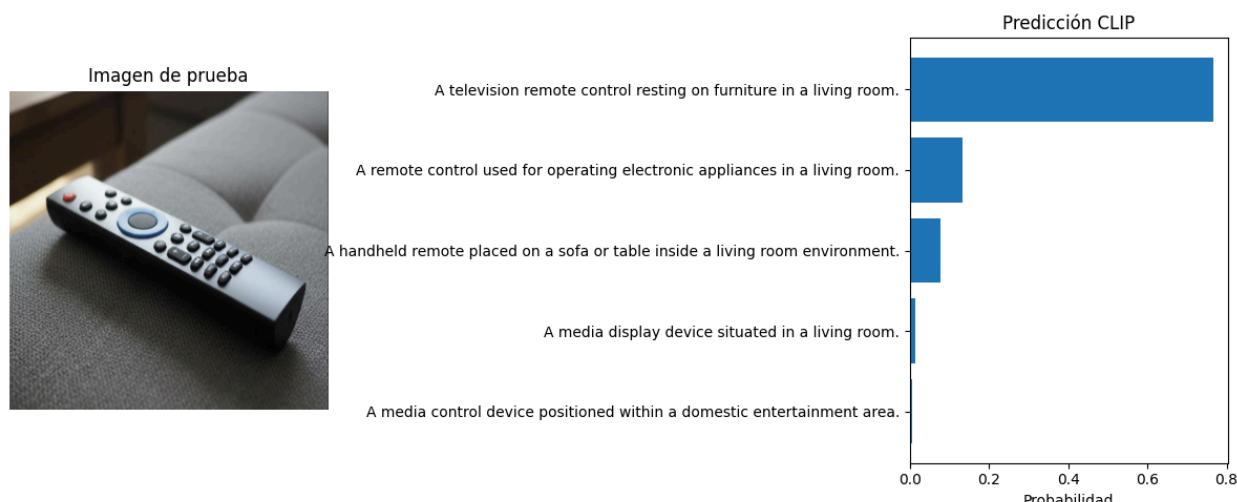
4.3. Implementación de inferencia zero-shot con CLIP, calculando similitudes imagen-texto

Se implementó un proceso de inferencia zero-shot utilizando el modelo CLIP, en donde cada imagen fue comparada contra el conjunto de prompts textuales definidos.

El modelo codificó tanto la imagen como los textos en un espacio latente compartido, permitiendo calcular medidas de similitud entre ambos. A partir de estas similitudes, se obtuvieron probabilidades que reflejan qué tan alineada está cada imagen con las distintas descripciones textuales propuestas.

4.4. Generación de gráficas de probabilidades que permitan interpretar los resultados del modelo

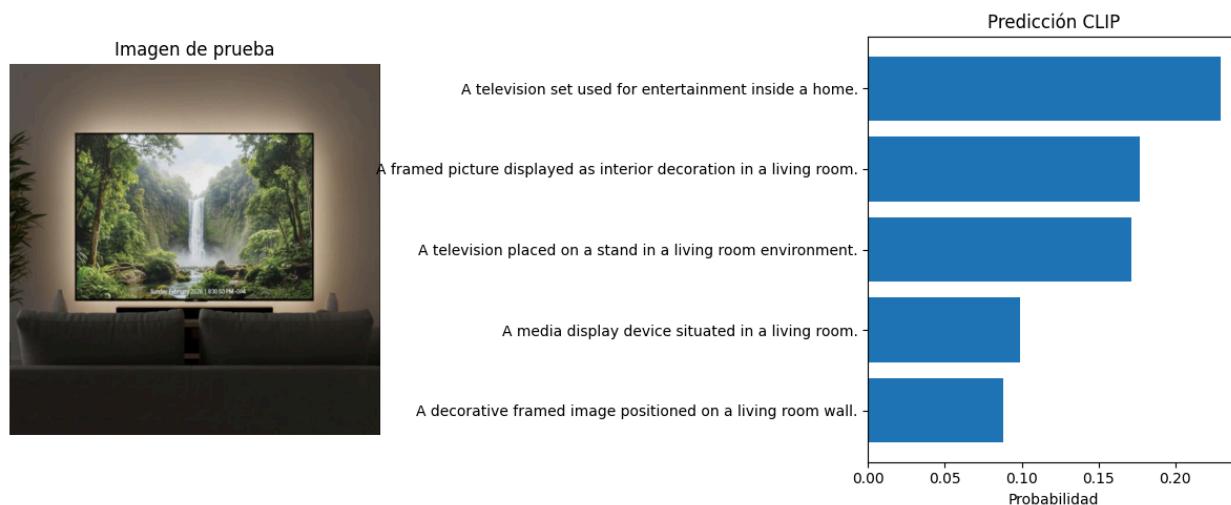
Para facilitar la interpretación de los resultados, se generaron visualizaciones que muestran las probabilidades asociadas a las predicciones más relevantes del modelo.



Estas gráficas permiten observar de forma clara cuáles son las descripciones textuales más cercanas semánticamente a cada imagen, proporcionando una interpretación visual del proceso de clasificación zero-shot.

4.5. Creación de un reporte de inventario doméstico, indicando qué objetos se detectan con mayor confianza

A partir de las predicciones generadas, se construyó un reporte de inventario doméstico que identifica los objetos detectados con mayor nivel de confianza.



Este reporte demuestra el potencial del enfoque zero-shot para tareas de catalogación automática en entornos reales, permitiendo identificar elementos presentes en una escena sin requerir entrenamiento supervisado previo sobre el dataset específico.

5. DISCUSIÓN

Los resultados obtenidos evidencian que el modelo CLIP posee una capacidad significativa para identificar objetos domésticos dentro de distintos ambientes sin necesidad de entrenamiento específico sobre el dataset utilizado. A partir de la inferencia zero-shot, fue posible establecer correspondencias semánticas coherentes entre las imágenes capturadas y los prompts definidos, lo cual valida el uso de modelos multimodales como una alternativa viable frente a enfoques tradicionales que requieren grandes volúmenes de datos etiquetados.

Asimismo, se observó que la calidad de las predicciones depende en gran medida de la formulación de los prompts textuales. Descripciones más claras y contextualmente alineadas con el contenido visual tienden a generar mejores niveles de similitud, lo que

sugiere que el diseño del lenguaje utilizado juega un rol crítico en el desempeño del sistema. Esto resalta la importancia de una adecuada definición del catálogo de objetos y de la representación textual cuando se trabaja con modelos de comprensión multimodal.

El enfoque implementado demuestra su utilidad como prueba de concepto para sistemas de inventario visual doméstico, permitiendo identificar objetos relevantes en ambientes como patio y sala con un nivel de confianza interpretable. Si bien el modelo no reemplaza completamente a soluciones supervisadas en escenarios de alta precisión, sí representa una herramienta eficiente para validaciones tempranas, reducción de costos de etiquetado y desarrollo rápido de soluciones basadas en visión por computadora.

7. CONCLUSIONES

El presente laboratorio permitió validar la viabilidad del uso de modelos multimodales para la clasificación de objetos domésticos sin necesidad de entrenamiento supervisado. A través de la implementación de un pipeline basado en CLIP y el uso de un dataset propio construido en Roboflow a partir de imágenes reales capturadas con un dispositivo móvil, se demostró que es posible identificar objetos presentes en ambientes como sala y patio mediante inferencia zero-shot. Este enfoque simplifica significativamente el proceso de desarrollo al eliminar la dependencia de grandes volúmenes de datos etiquetados.

Asimismo, los resultados obtenidos confirman que la combinación de representaciones visuales y textuales en un espacio latente compartido permite establecer relaciones semánticas útiles para tareas de clasificación. La generación de probabilidades y su visualización facilitaron la interpretación del comportamiento del modelo, permitiendo construir un inventario doméstico preliminar basado en niveles de confianza. Esto demuestra que el uso de prompts bien definidos es un factor clave para mejorar la calidad de las predicciones.

En conclusión, esta prueba de concepto respalda el potencial del uso de modelos fundamentales como CLIP en soluciones de domótica e inventario visual inteligente, alineándose con los objetivos planteados por HomeVision. Si bien aún existen oportunidades de mejora, como la expansión del catálogo de objetos o la incorporación de modelos especializados en etapas futuras, el enfoque desarrollado representa un punto de partida sólido para la construcción de sistemas automatizados de análisis visual en entornos residenciales.

8. ANEXOS

Se adjunta Notebook en Python con el script aplicado al laboratorio en el siguiente repositorio:

<https://github.com/mc-ivan/zero-shot/tree/main/lab3>