UNIVERSIDAD DE CHILE

Inteligencia Artificial Generativa

Let's talk about hype stuff

Valentin Barriere // Clemente Henriquez

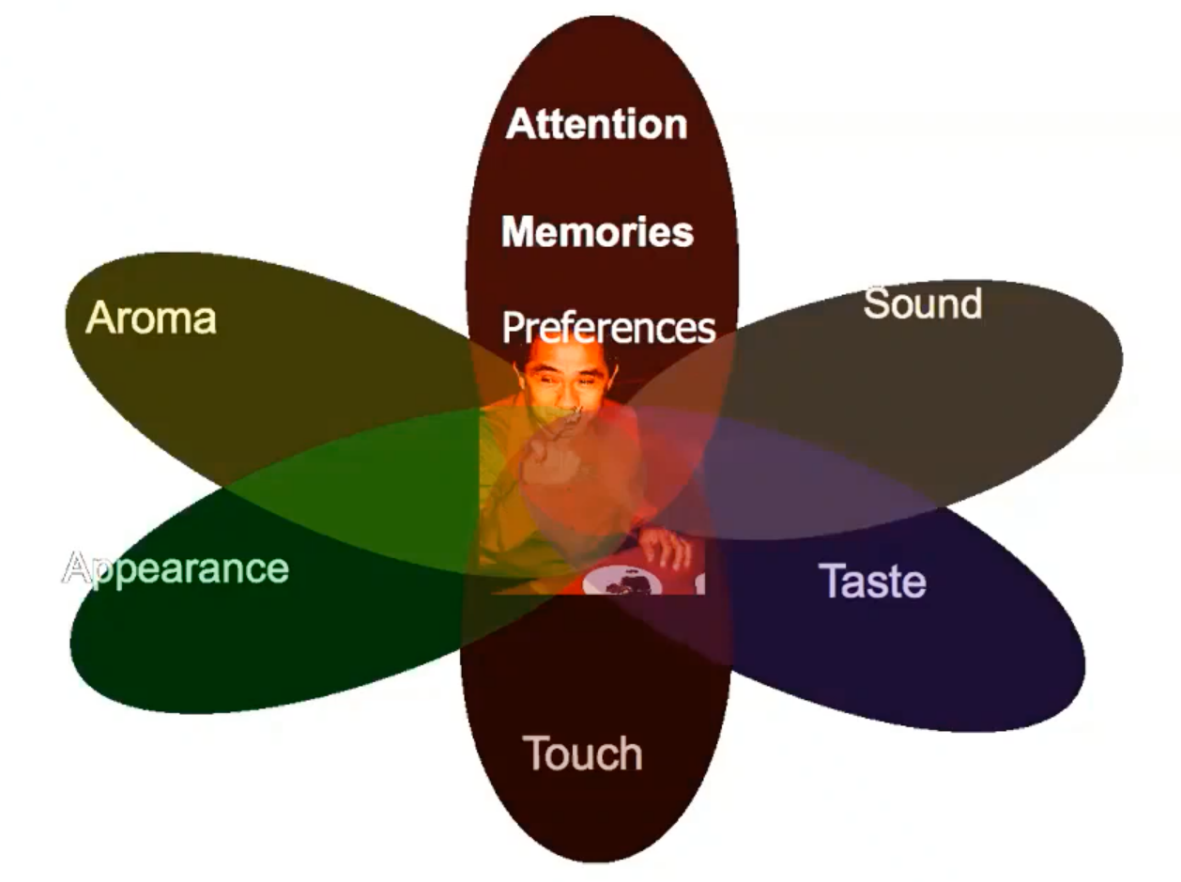
Universidad de Chile – DCC

Diplomado de Postítulo en Inteligencia Artificial, Primavera 2025

Modelos Multimodales

¿Qué es multimodal?

Las modalidades pueden referirse a los sentidos humanos



Antecedentes: Multimodalidad

Inteligencia Humana e Inteligencia Artificial, recogen y procesan la información antes de tomar una decisión

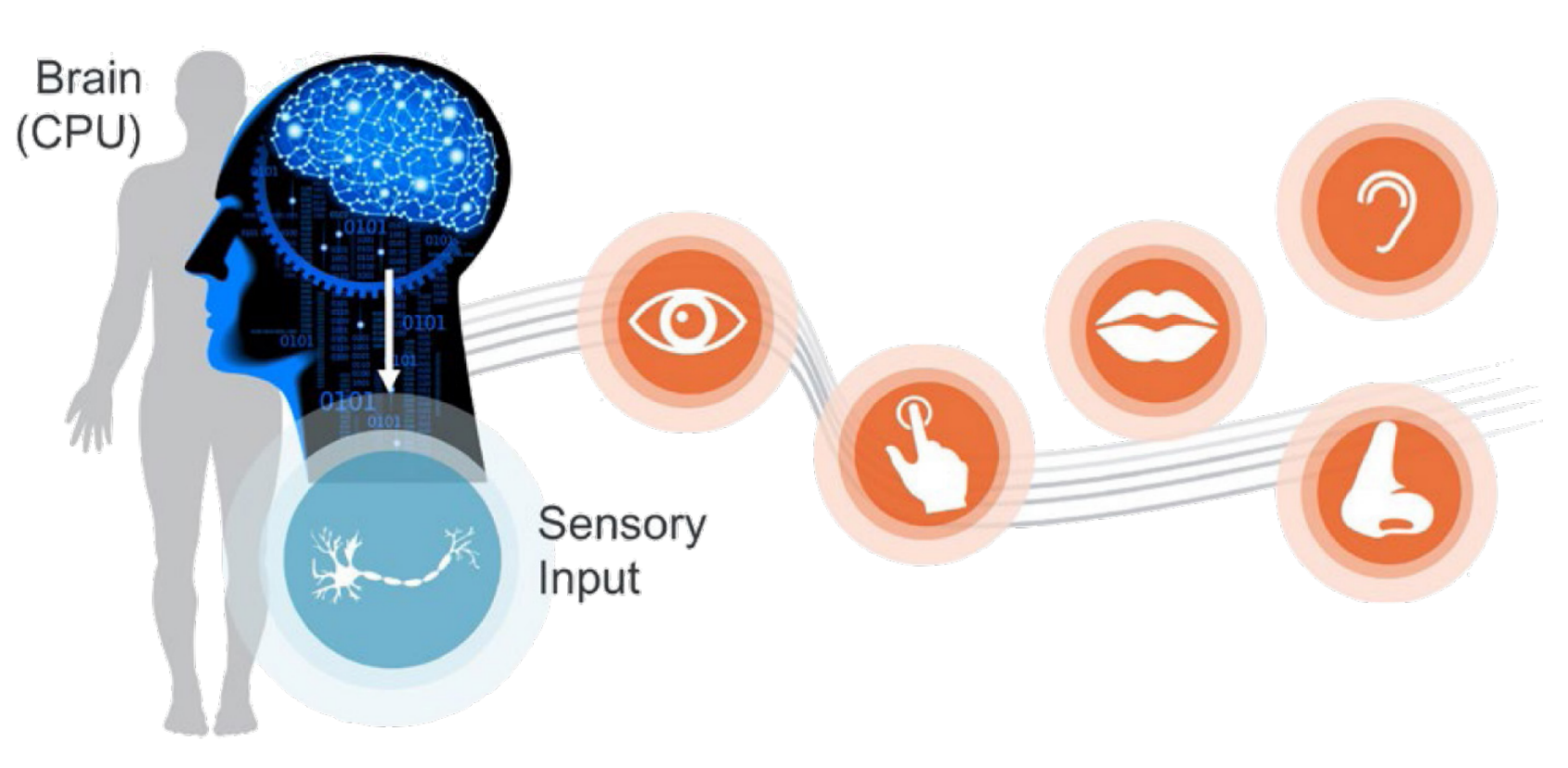


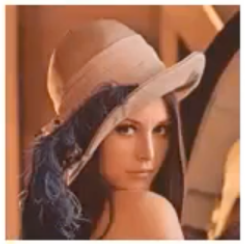
Figure 1: Usamos la multimodalidad en la vida cotidiana, así es como los humanos se comunican

Antecedentes: Multimodalidad

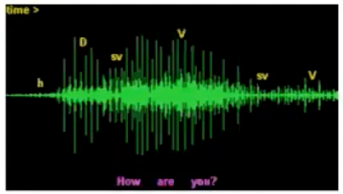
- **Lenguaje**: el contenido verbal contiene palabras, pero también pragmática y sintaxis

Language

- **Visión**: gestos, expresiones faciales, mirada, lenguaje corporal, ...


Vision

- **Habla**: prosodia, expresiones vocales, ...



Speech

Pero también: tacto, fisiológico (ECG), móvil (ubicación), social (amigos en una red)...

Antecedentes: Multimodalidad

Definiciones básicas

Modalidad se refiere a un cierto tipo de información y su formato de representación en el que se almacena. Se transmite a través de un medio, que es un sistema de comunicación/transmisión.

Comunicación Humana

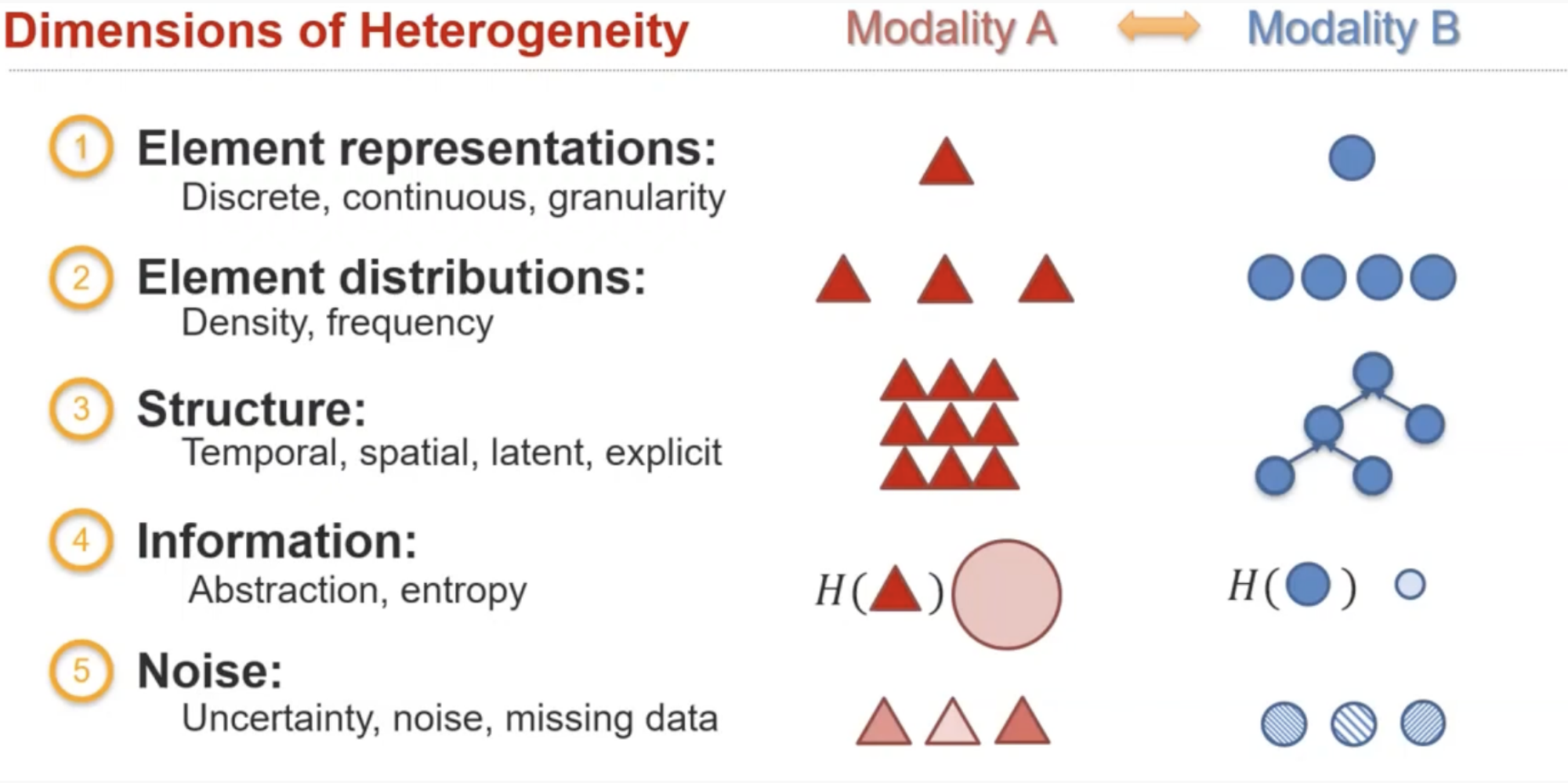
Transmisión de una señal multimodal sujeta a la interpretación de los individuos.

Dependiendo de las diferentes modalidades que transportan la información, hablamos de señales vocales, verbales, faciales, gestuales, ...

Uso de estas señales para detectar diferentes aspectos del hablante: Emoción [22], Rasgos [12], o Empleabilidad [4].

¡Son heterogéneas e interconectadas!

Modalidades

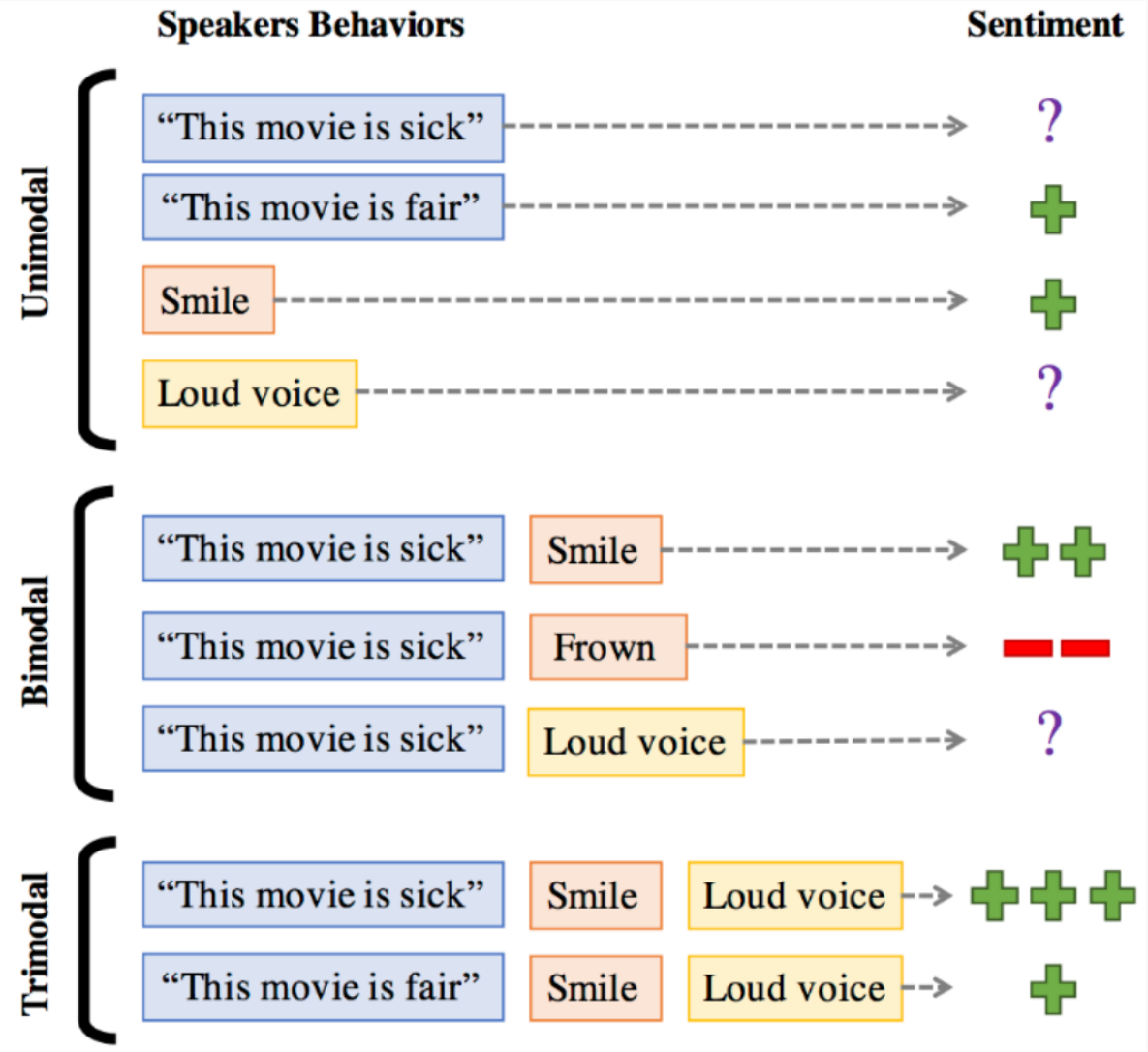


Pero también están interconectadas (Efecto McGurk):

`https://www.youtube.com/watch?v=2k8fHR9jKVM`

Complementariedad

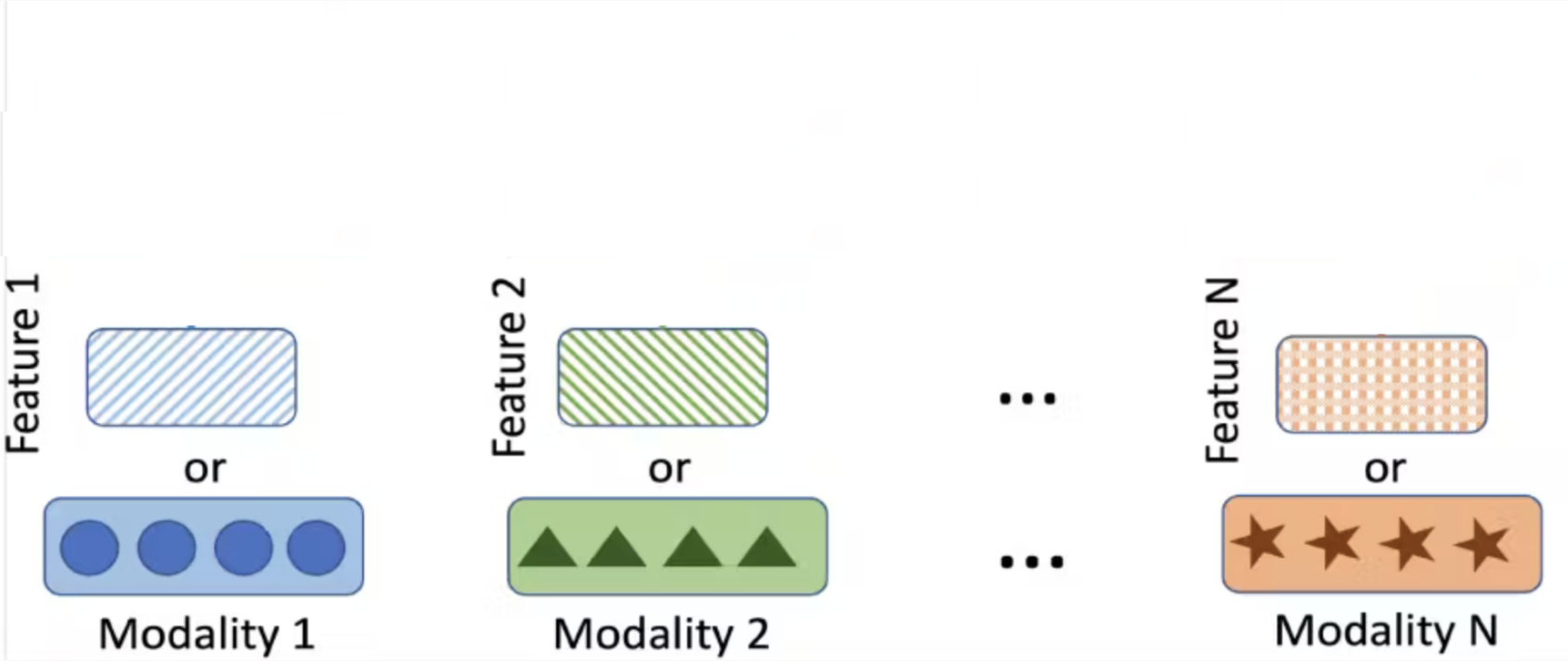
Diferentes modalidades pueden ser consensuales, complementarias, una puede ayudar a reforzar a la otra, etc...



Representación y Codificación

Codificación

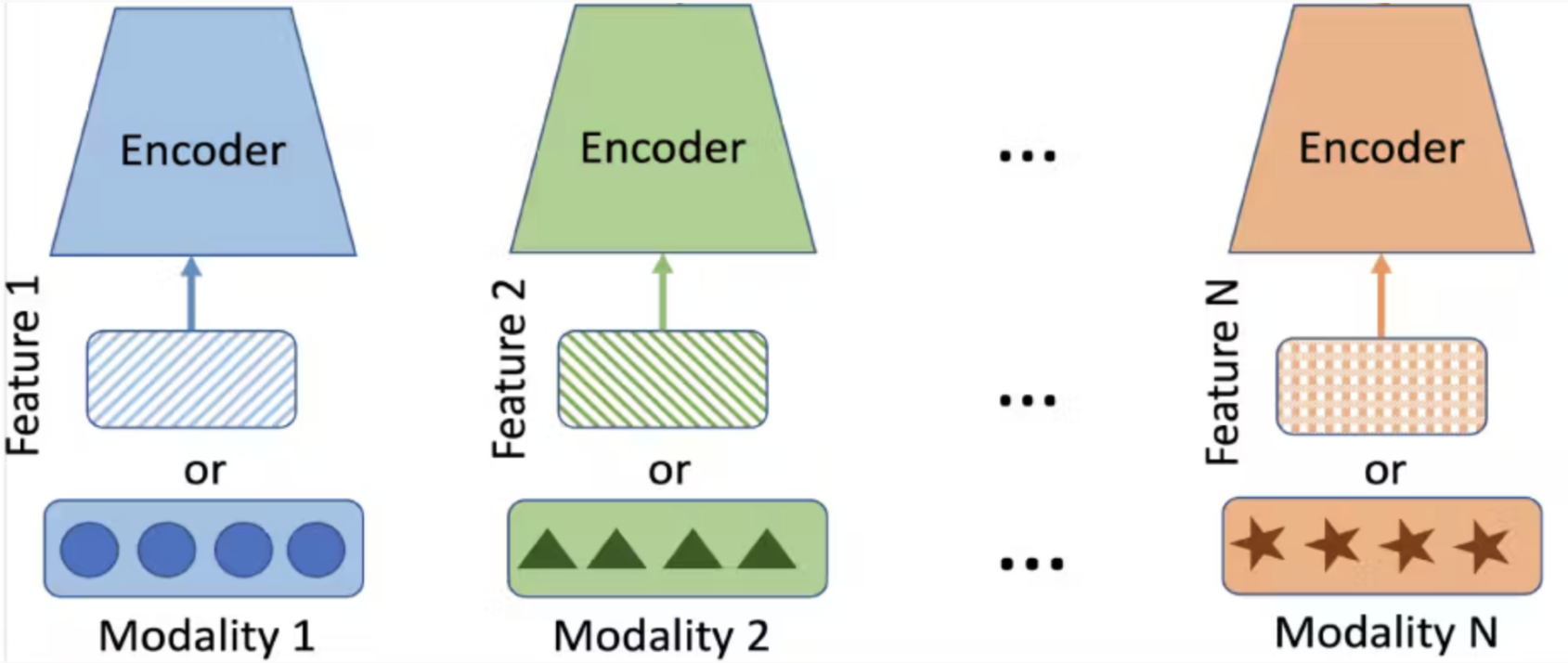
Puedes codificar datos multimodales usando diferentes métodos: extraer descriptores de audio, descriptores de imagen, descriptores de texto, etc...



Representación y Codificación

Codificación

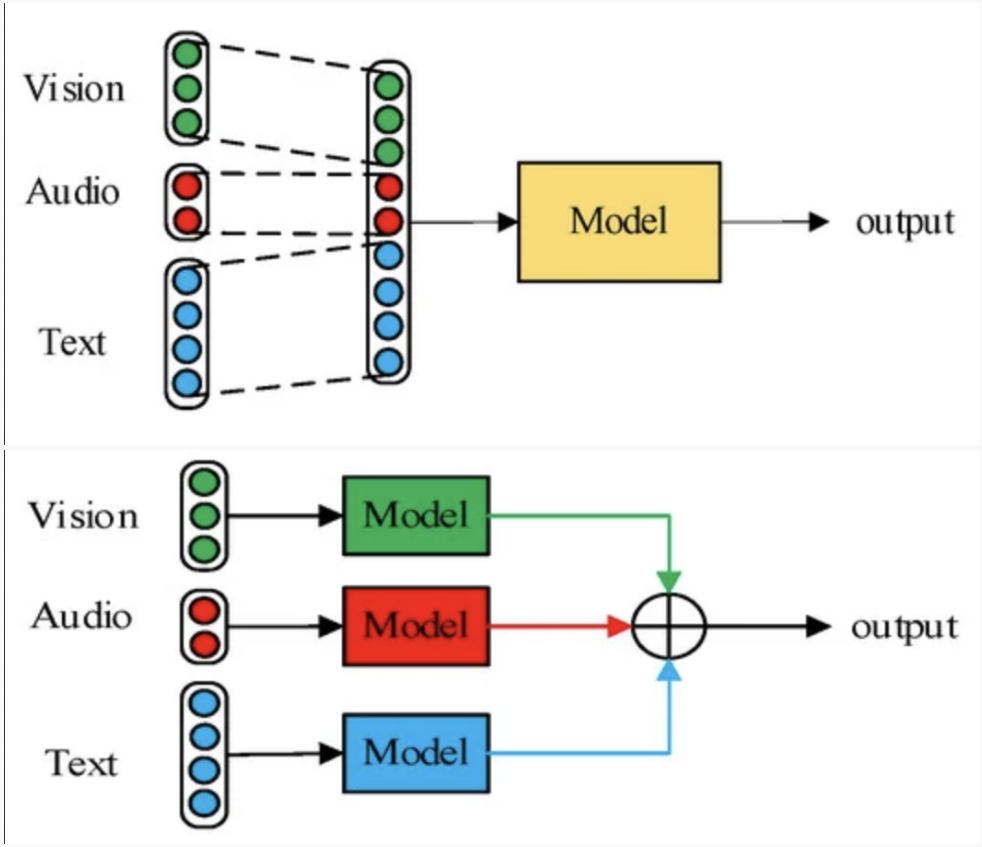
Puedes codificar datos multimodales usando diferentes métodos: extraer descriptores de audio, descriptores de imagen, descriptores de texto, etc...



También es posible usando redes: una CNN preentrenada para codificar la imagen, un transformador o embeddings de palabras preentrenados

Fusiones Básicas

- Fusión temprana: antes del procesamiento
- Fusión tardía: después de la predicción

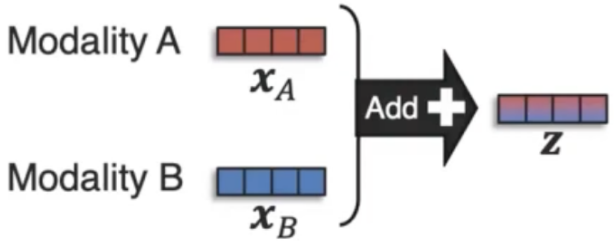


Históricamente

Cuando los descriptores se extraían manualmente, la fusión de la representación unimodal ocurría antes o después del procesamiento por un modelo. Hoy en día, son grandes redes las que extraen las características y las fusionan internamente.

Fusiones basadas en Modelos I: Tipos de fusión

- Fusión aditiva:

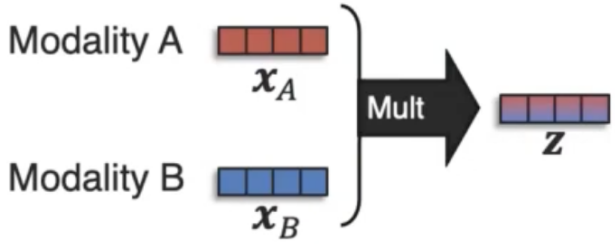


Additive fusion:

$$z = w_1 x_A + w_2 x_B$$

⟹ 1-layer neural network can be seen as additive

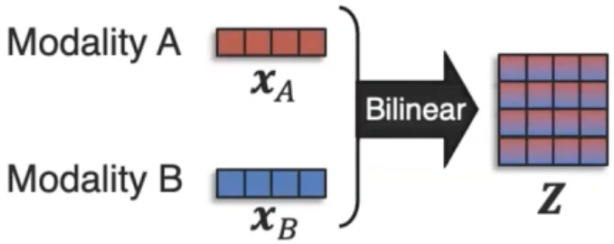
- Fusión multiplicativa:



Simple multiplicative fusion:

$$z = w(x_A \times x_B)$$

- Fusión bilineal:



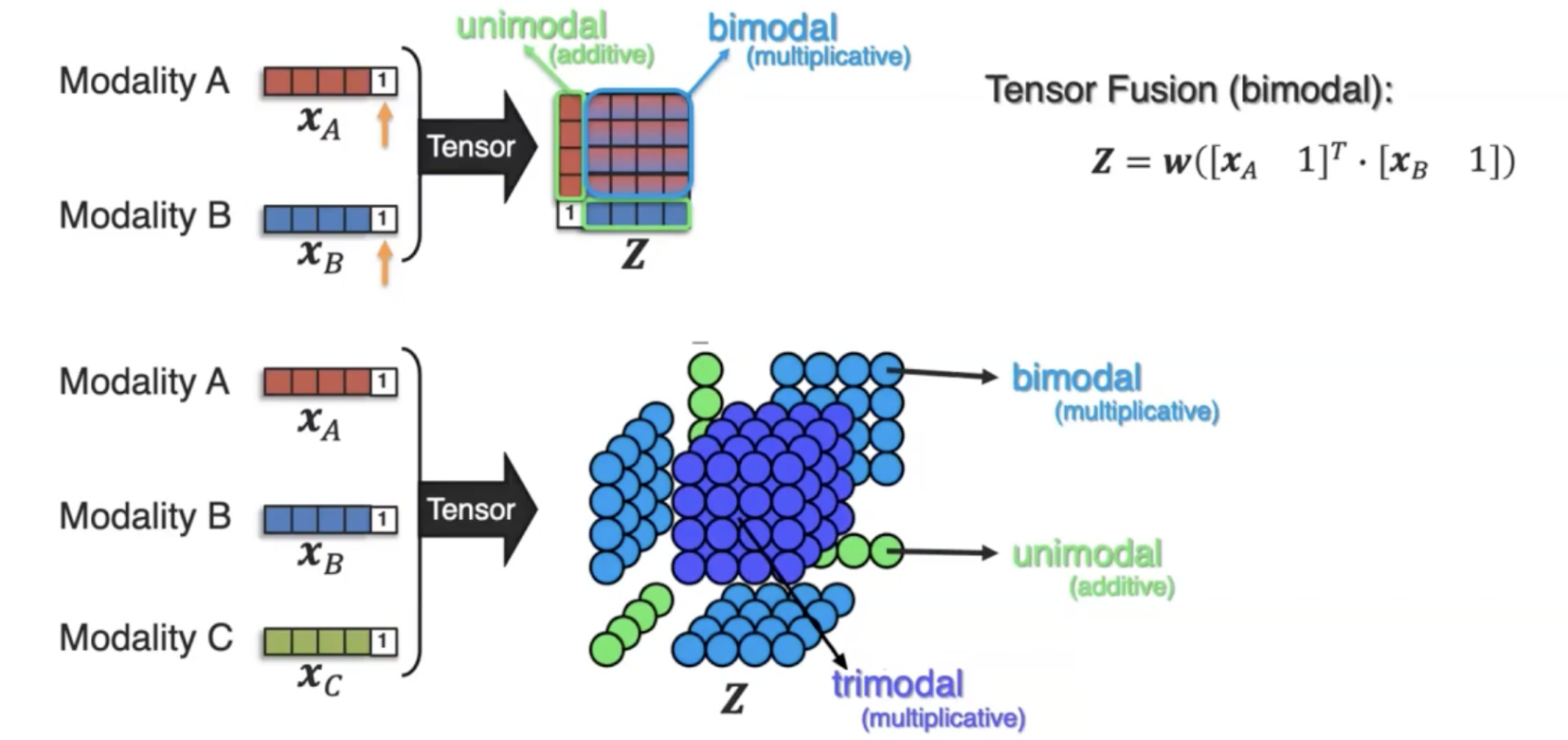
Bilinear Fusion:

$$Z = W(x_A^T \cdot x_B)$$

Recuerda: ¡Los datos ya tienen las interacciones en sí, el modelo solo intenta aprenderlas!

Fusiones basadas en Modelos II

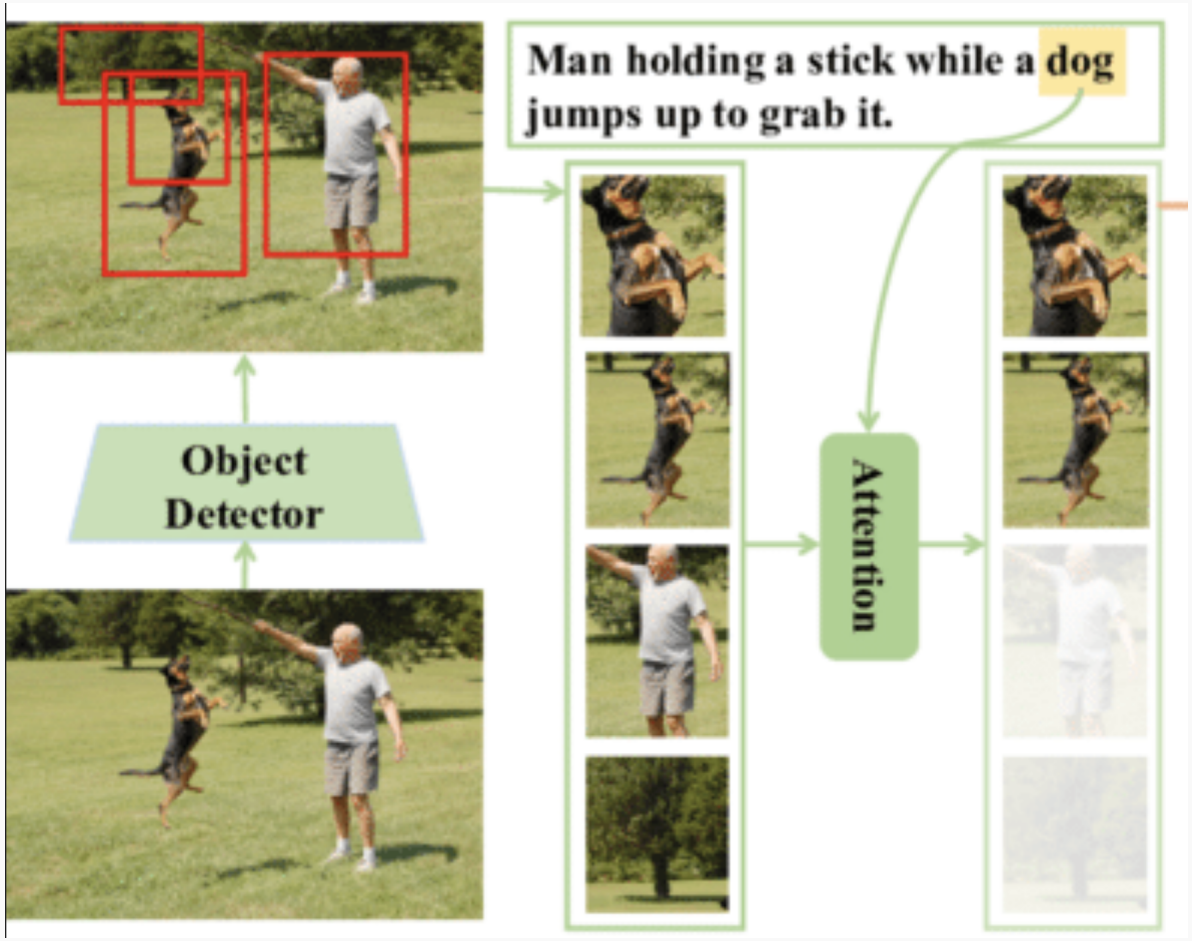
También puedes crear módulos de fusión más complejos, ¡pero cuidado con la complejidad y el número de parámetros!



Más en los artículos [14, 28, 6, 13, 16]

Fusiones basadas en Modelos III

Es posible usar mecanismos de compuerta similares a atención, y en particular atención cruzada multimodal. Esto ayuda a enfocarse en la parte correcta de las modalidades cuando se busca algo específico.



Fusiones basadas en Modelos IV

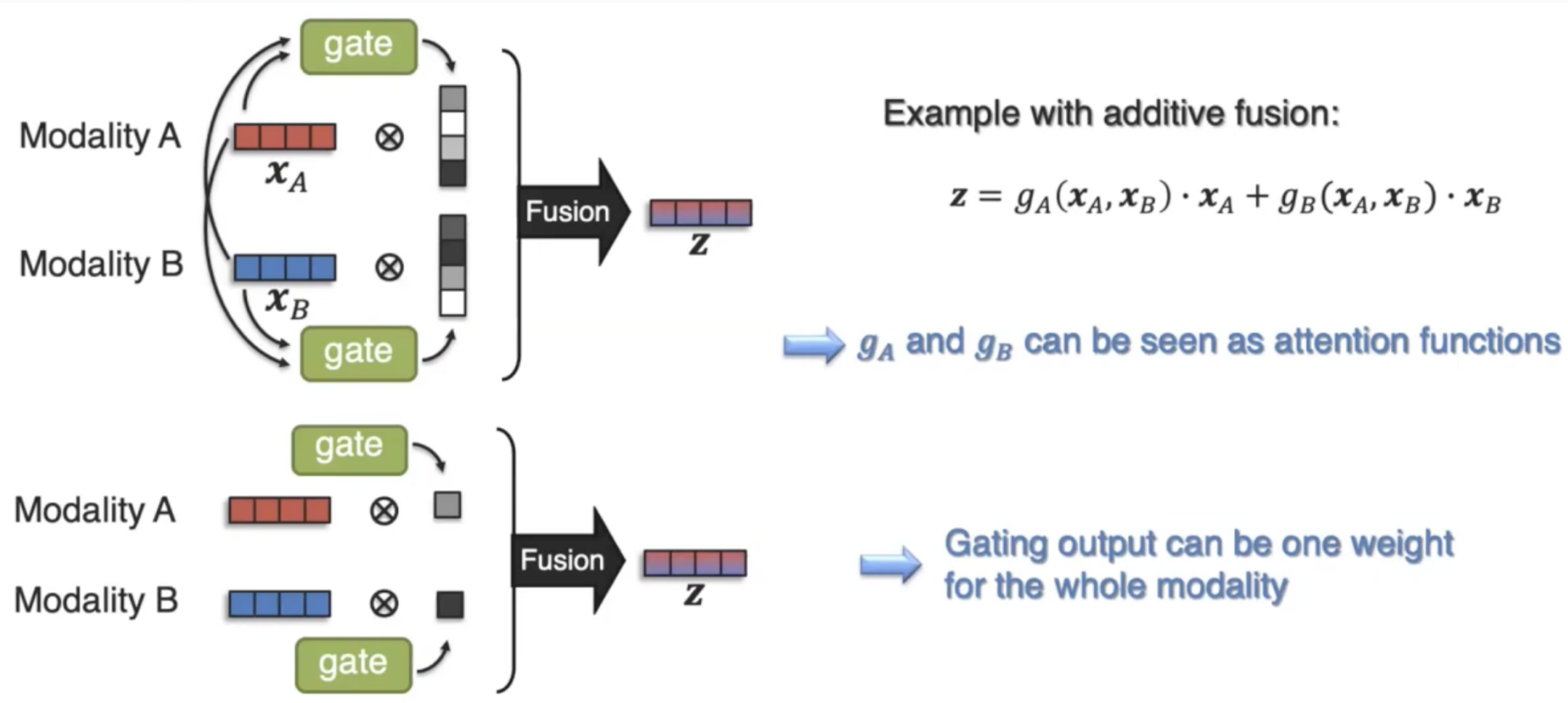


Figure 2: Atención de compuerta cruzada [2]

Se puede ver como una forma de:

- prevenir que señales no deseadas se propaguen hacia adelante (negativo; *compuerta*)
- seleccionar la señal preferida para avanzar (positivo; *atención*)

Tareas

Output \ Input	Image	Text	Image & Text
Image	Vision tasks 	Text-based image generation, image retrieval Baby pandas walking on the grass 	Text-guided image editing, referring segmentation Panda on the far left 
Text	Image captioning  A group of baby pandas on the grass	NLP tasks ... world's rarest mammals, the giant [MASK] . Only about 1,500 of these black-and-white bears...	Visual question answering How many pandas are there?  There are 8 pandas
Image & Text	 How many pandas are there? There are 8 pandas Show me the image of a movie with a panda 	Visual Dialogue	

Cortesía de Javiera Castillo-Navaro

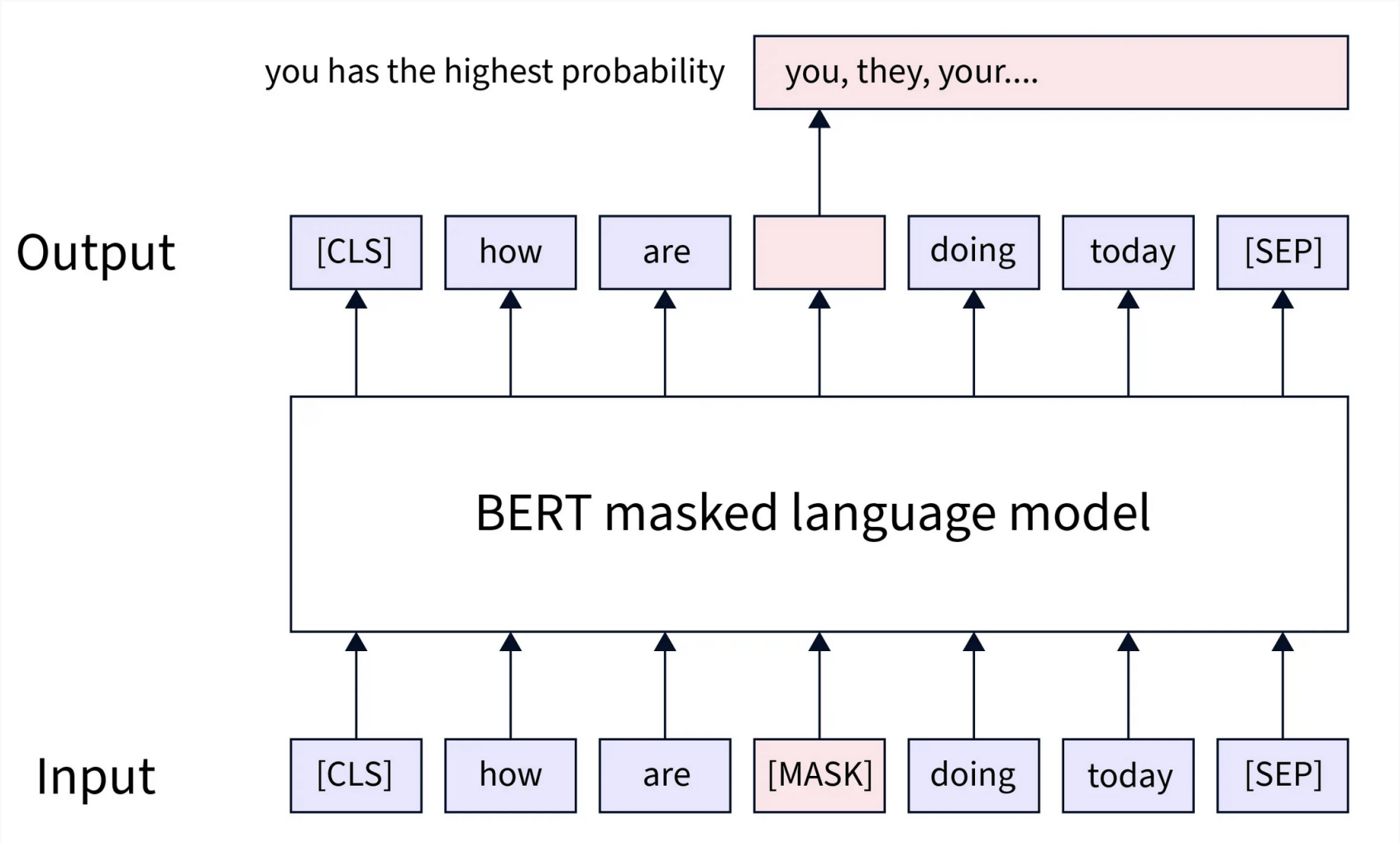
"Old School" Datasets

Figure 3: Estos datasets se usaban para entrenar modelos multimodales en el pasado

Para modelos generativos, veremos que hoy en día hay nuevos datos disponibles.

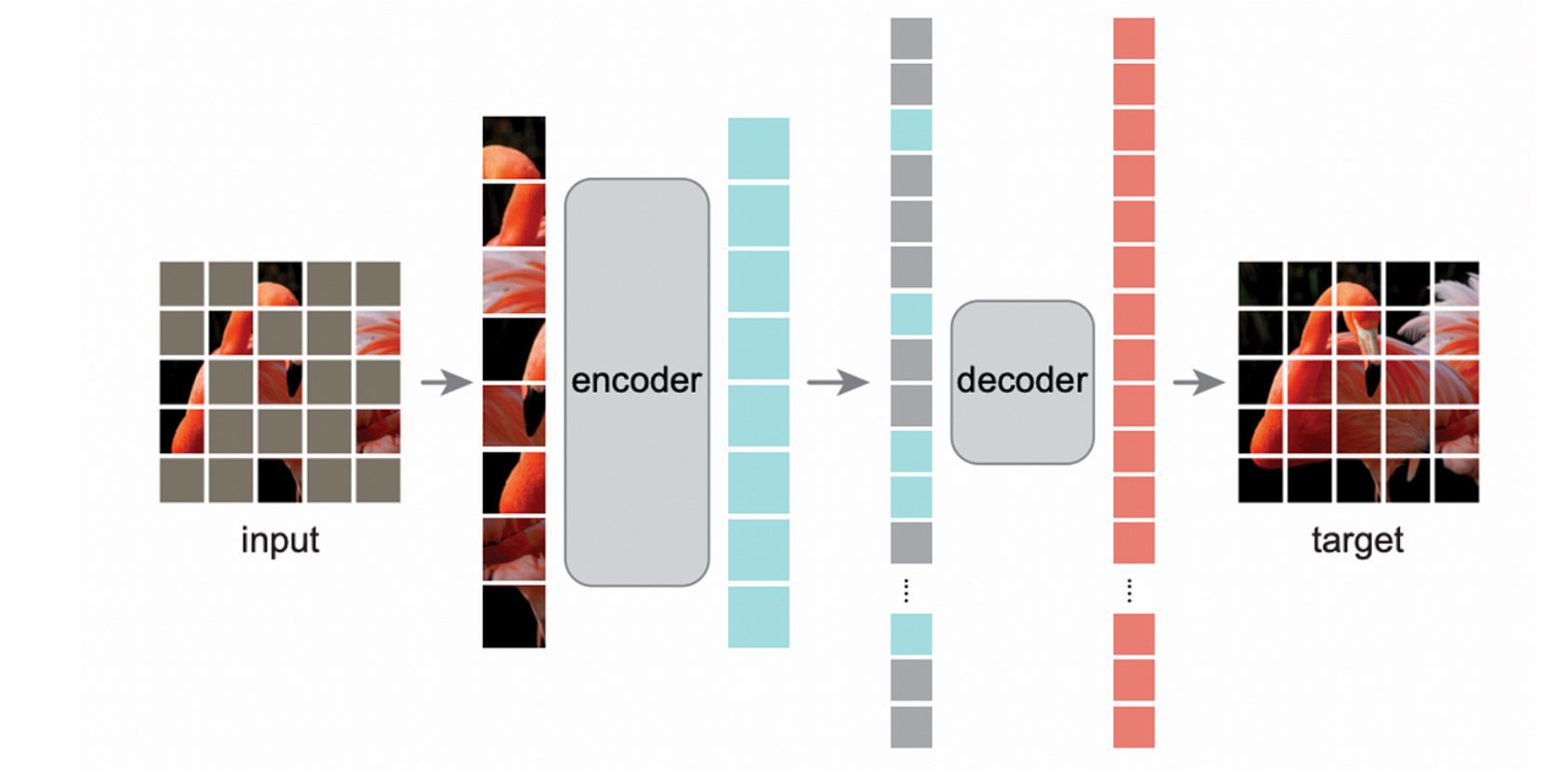
Fusión usando Transformers

Al final, la arquitectura de transformers funciona bien para fusionar las diferentes modalidades, incluso si sus datos en bruto son muy diferentes.



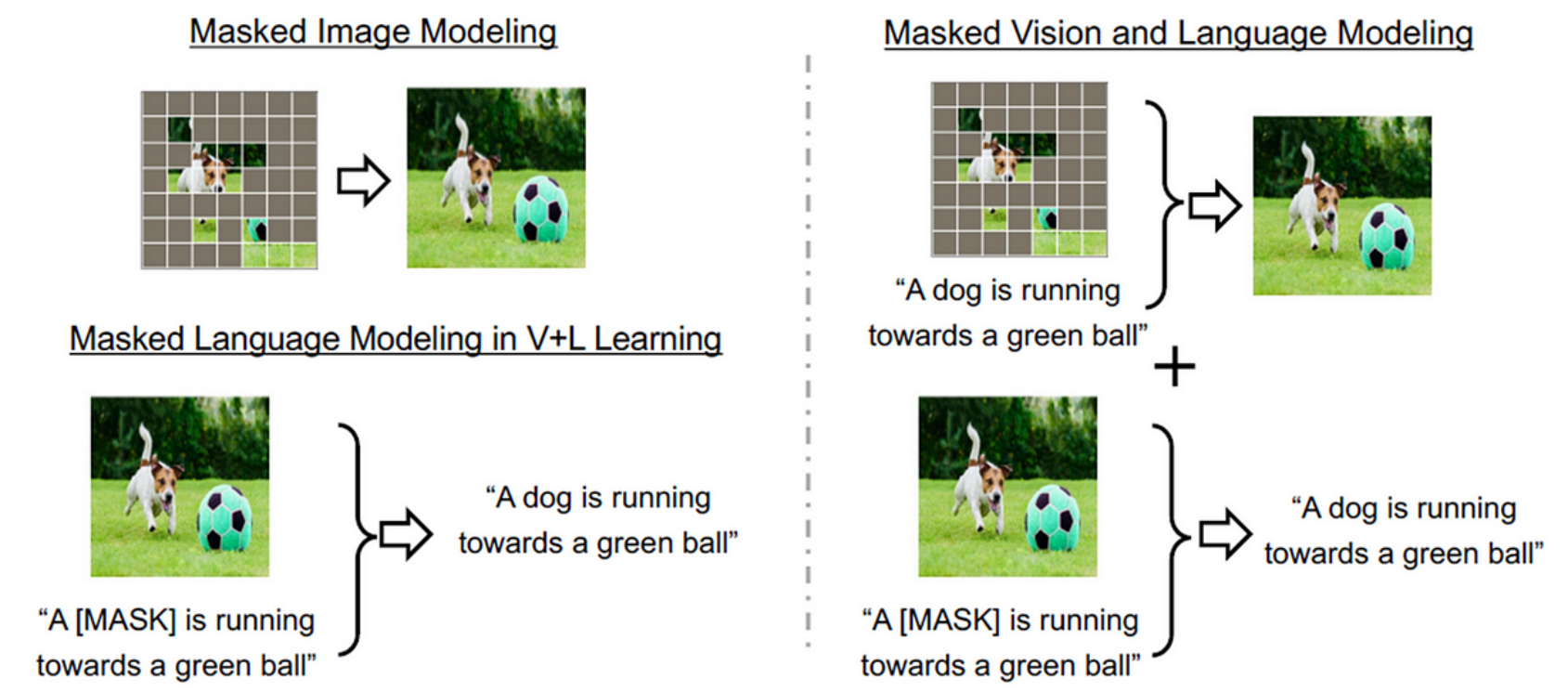
Fusión usando Transformers

Al final, la arquitectura de transformers funciona bien para fusionar las diferentes modalidades, incluso si sus datos en bruto son muy diferentes.



Fusión usando Transformers

Al final, la arquitectura de transformers funciona bien para fusionar las diferentes modalidades, incluso si sus datos en bruto son muy diferentes.



First LMMs: LXMERT

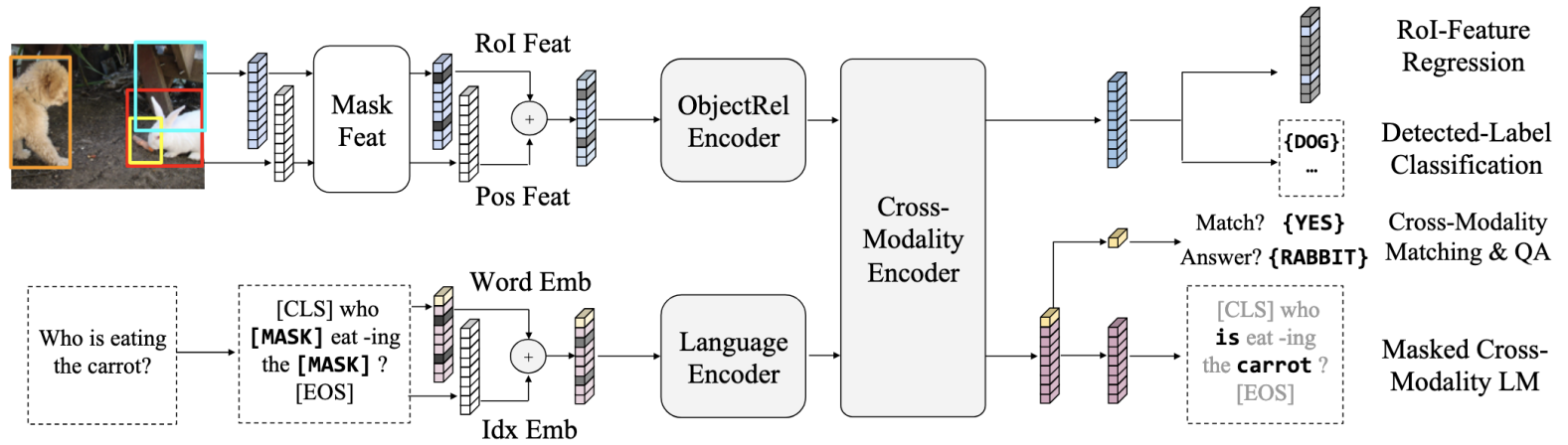
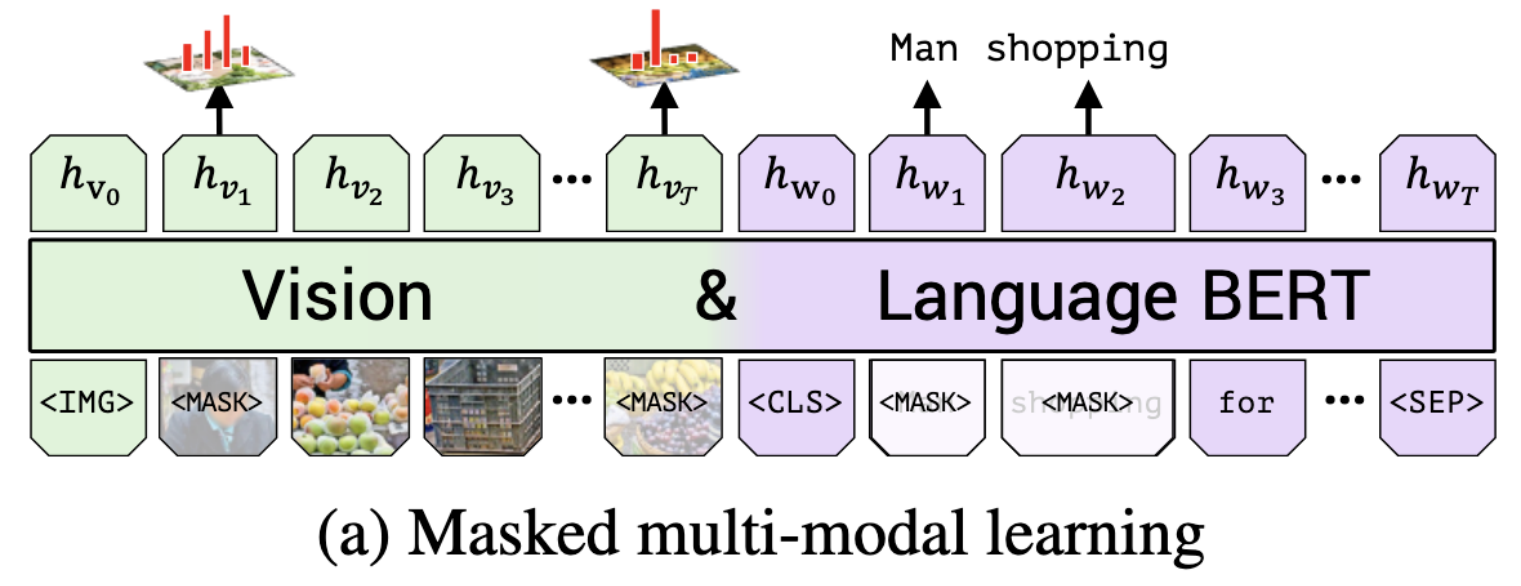


Figure 2: Pre-training in LXMERT. The object RoI features and word tokens are masked. Our five pre-training tasks learn the feature representations based on these masked inputs. Special tokens are in brackets and classification labels are in braces.

Los primeros Transformers multimodales [25]

- Usan un faster-RCNN entrenado para extraer embeddings de Regiones de Interés (RoIs) y aprender a clasificarlas
- Enmascaran palabras para el lenguaje, o dimensiones para la visión
- Tareas de emparejamiento cruzado multimodal y VQA

First LMMs: ViLBERT



First LMMs: ViLBERT

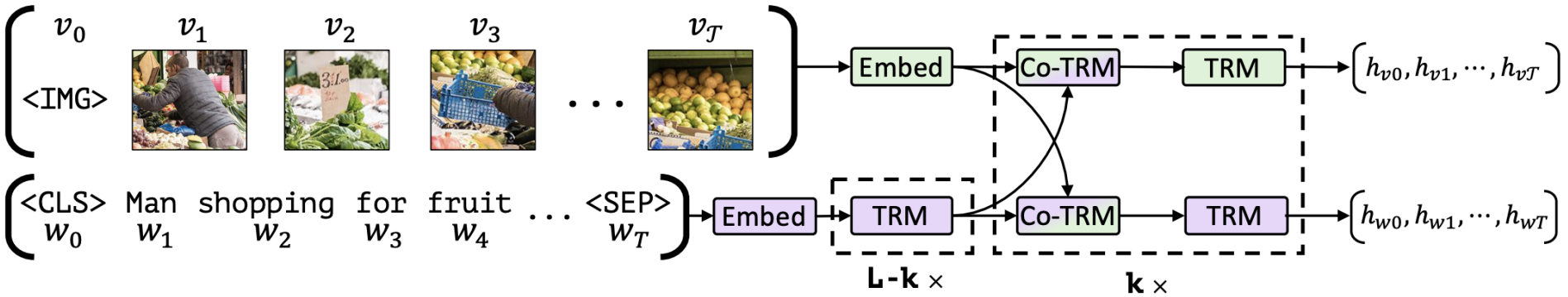
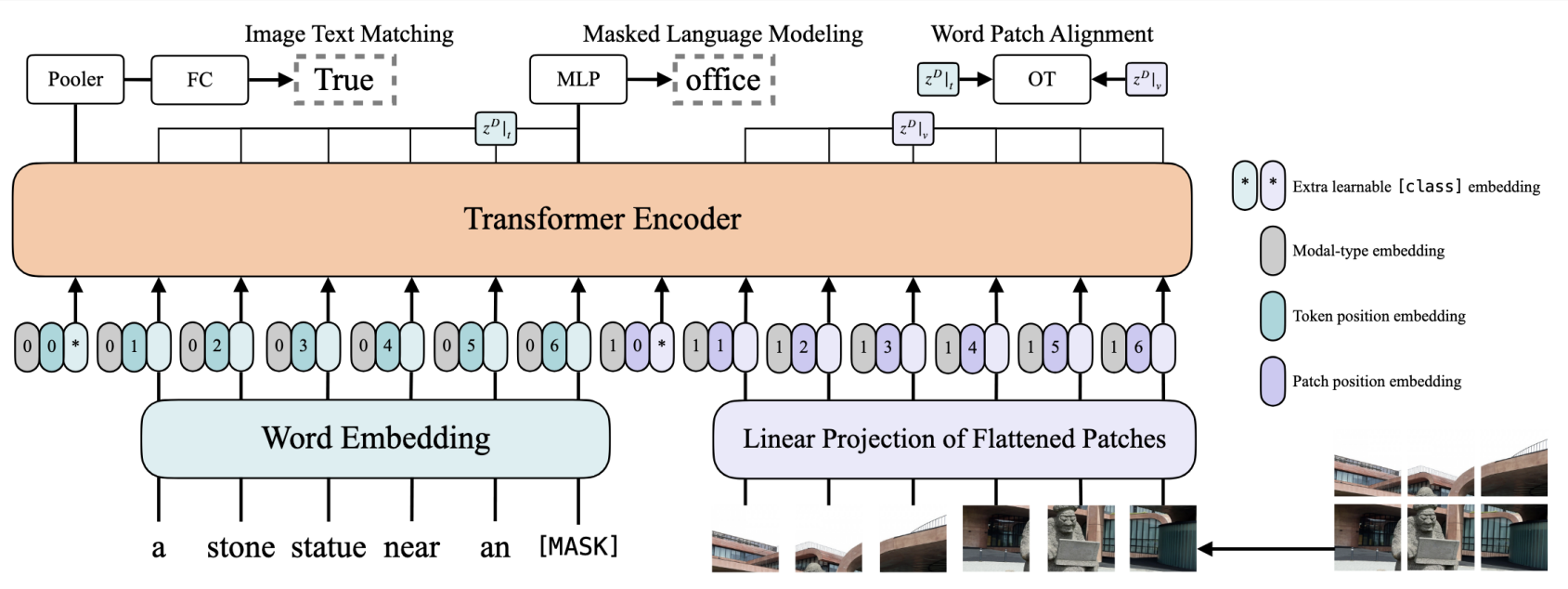


Figure 1: Our ViLBERT model consists of two parallel streams for visual (green) and linguistic (purple) processing that interact through novel co-attentional transformer layers. This structure allows for variable depths for each modality and enables sparse interaction through co-attention. Dashed boxes with multiplier subscripts denote repeated blocks of layers.

El modelo ViLBERT tiene capas de transformador con atención cruzada, usando queries y keys/values de diferentes modalidades, permitiendo interacción entre las modalidades dentro de las capas. [17]

First LMMs: ViLT

Vision Language Transformer no usa ninguna CNN para codificar la imagen, mostrando que ¡aplanar la imagen antes de las capas del transformador es suficientemente bueno! [7]



Usa tres pérdidas, similar a lo que vimos antes.

Coordinación

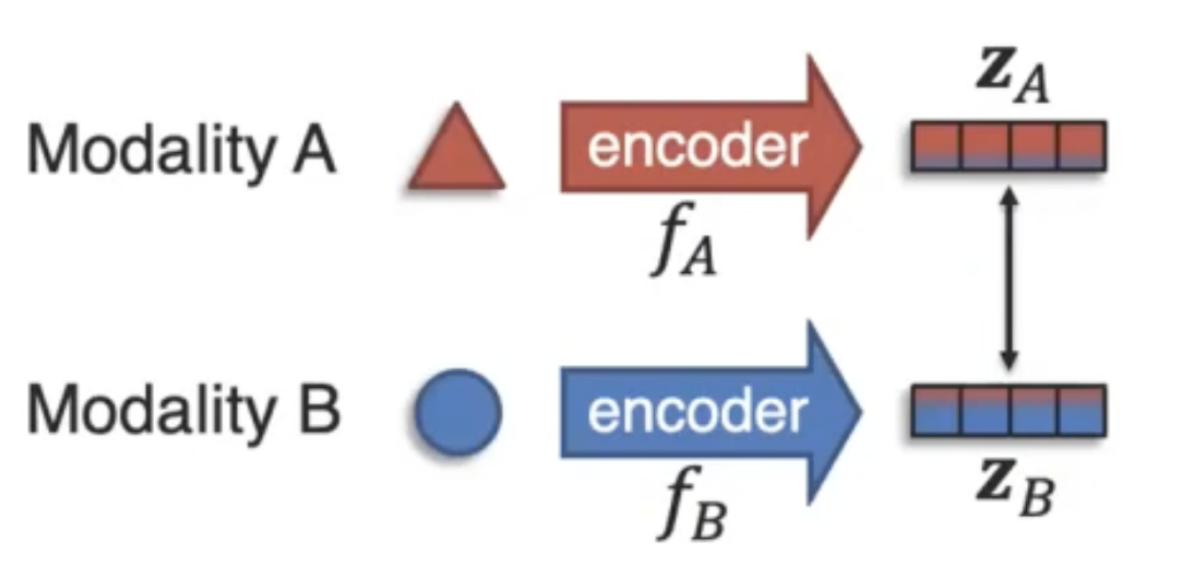


Figure 4: Coordinación: alinear las modalidades en un espacio latente común es permite entrenar un modelo multimodal de manera auto-supervisada.

Coordinación

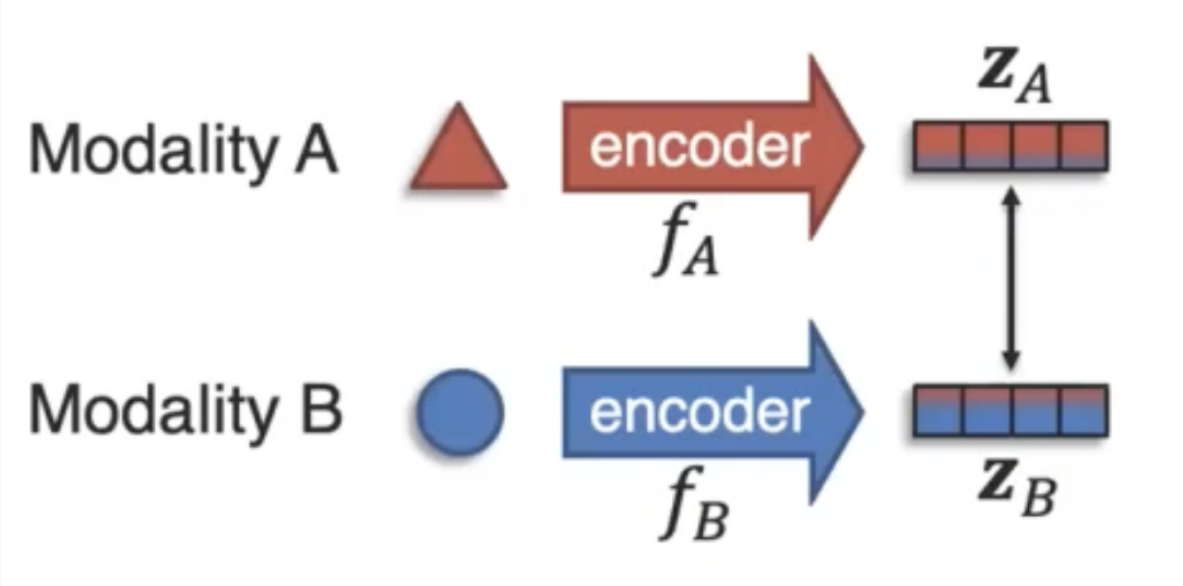
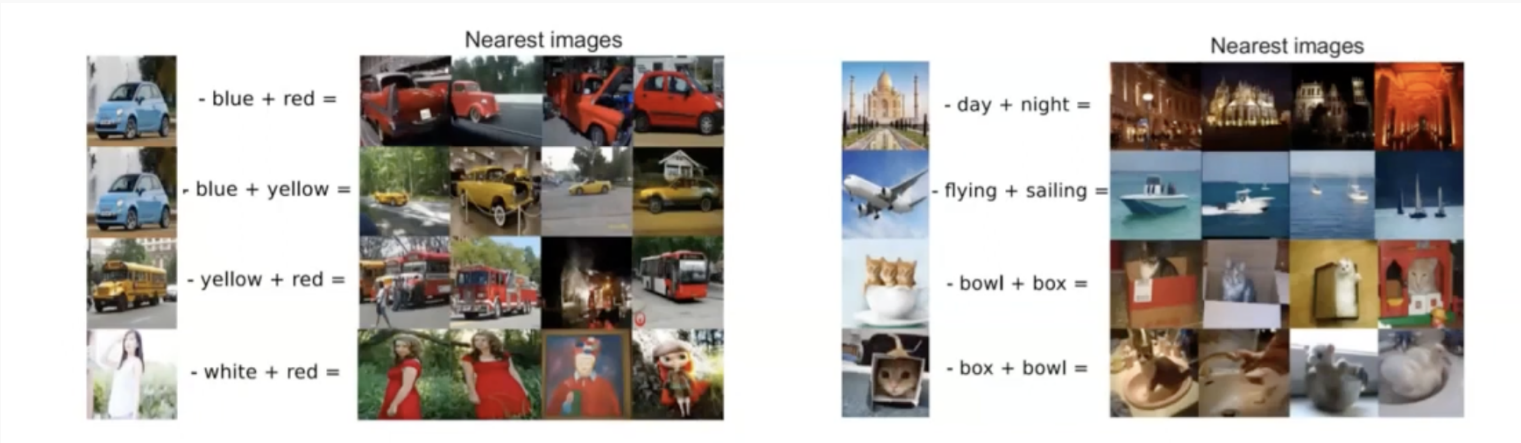


Figure 4: Coordinación: alinear las modalidades en un espacio latente común es permite entrenar un modelo multimodal de manera auto-supervisada.



Aprendizaje contrastivo: Ejemplo en imágenes

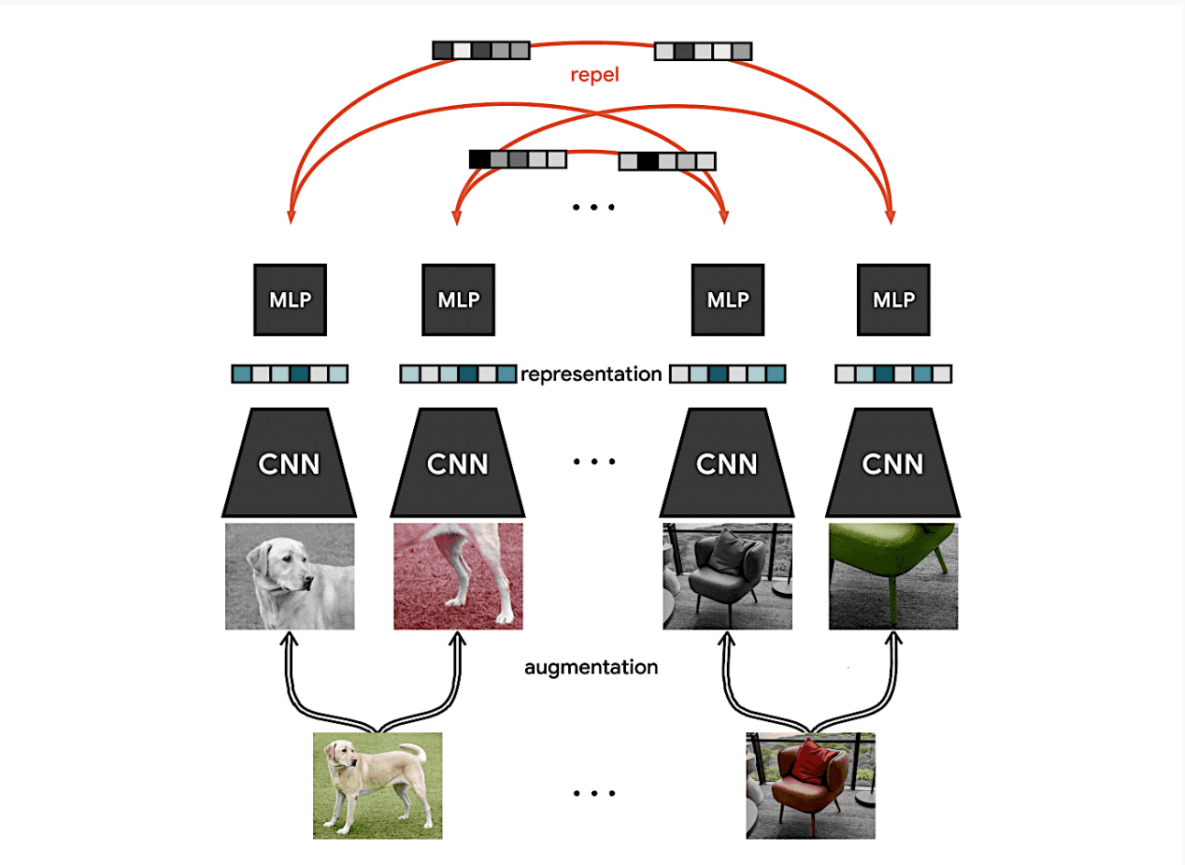


Figure 5: SimCLR es un ejemplo de pre-entrenamiento en ImageNet

Principio

Las representaciones de partes de la misma imagen se acercan, mientras que las representaciones de parches de imágenes diferentes se alejan. ¡También puede funcionar con clases!

CLIP: Contrastive Language-Image Pre-training

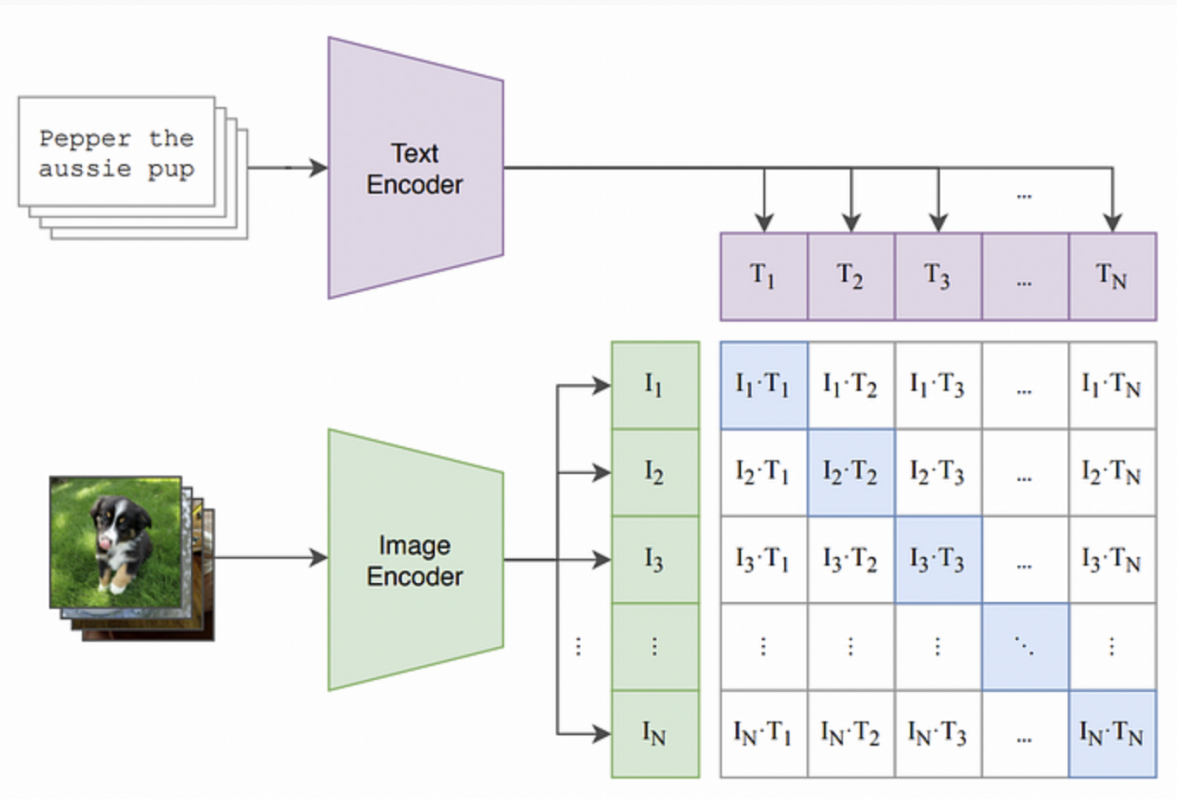


Figure 6: CLIP ha sido entrenado sobre 400M pares imagen-texto

Principio

Aprender un espacio de embeddings multimodal compartido, y maximizar/minimizar la distancia entre los embeddings que pertenecen juntos. Pérdida equivalente a aprender una función que maximiza la información mutua entre las modalidades [20].

CLIP: zero-shot

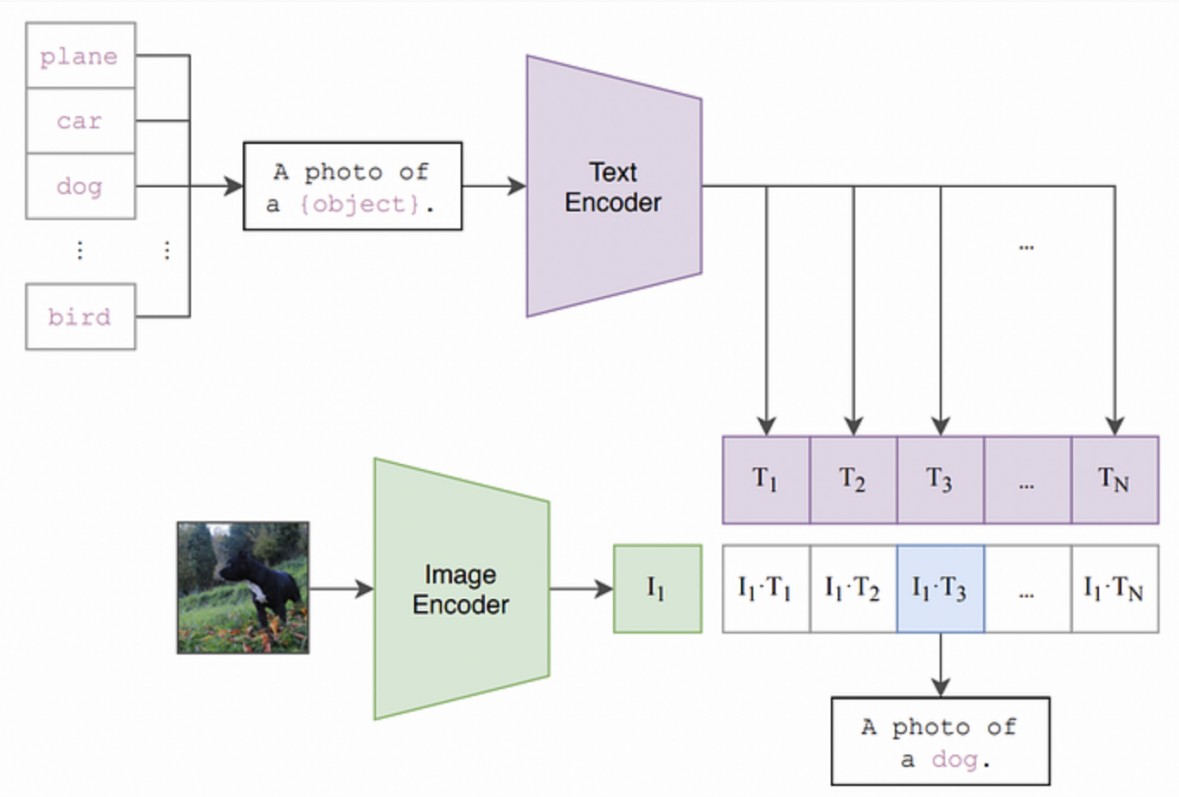


Figure 7: CLIP es muy bueno en zero-shot usando plantillas de texto

Sin embargo, esto depende mucho de los datos y las clases (es malo para OOD como datos satelitales).

Más sobre CLIP aquí

BLIP: Bootstrapping Language-Image Pre-training

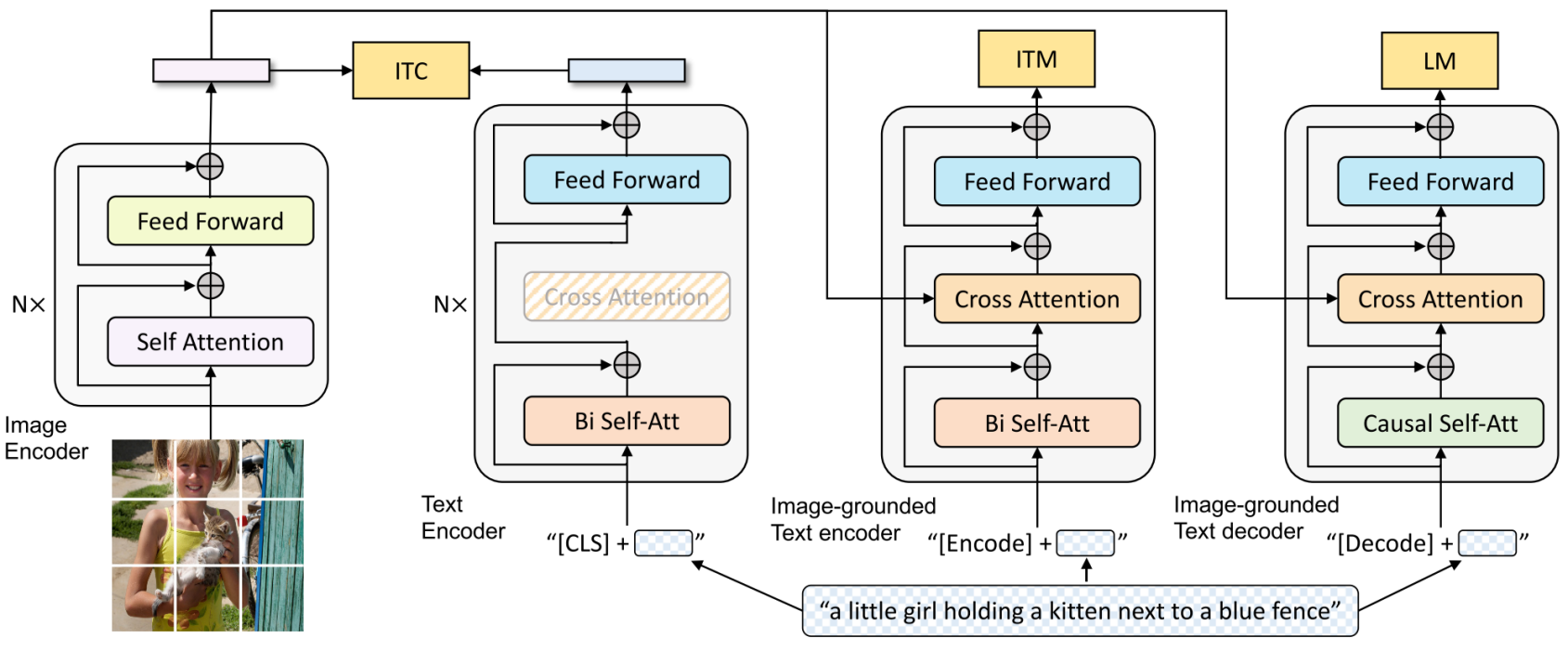


Figure 8: Unimodal encoders + multimodal image-grounded text-encoders [10]

Tres pérdidas

Pérdida contrastiva imagen-texto (ITC) para los encoders unimodales. Para los encoders multimodales, un emparejamiento imagen-texto (ITM) para distinguir pares positivos y negativos, y una pérdida de Modelo de Lenguaje (LM) para generar texto. **Además limpia el dataset para filtrar ejemplos ruidosos.**

Generador de texto a imagen: CLIP + VQ-GAN

Figure 9: Generación de imágenes artificiales usando CLIP + VQ-GAN

Generador de texto a imagen: CLIP + VQ-GAN

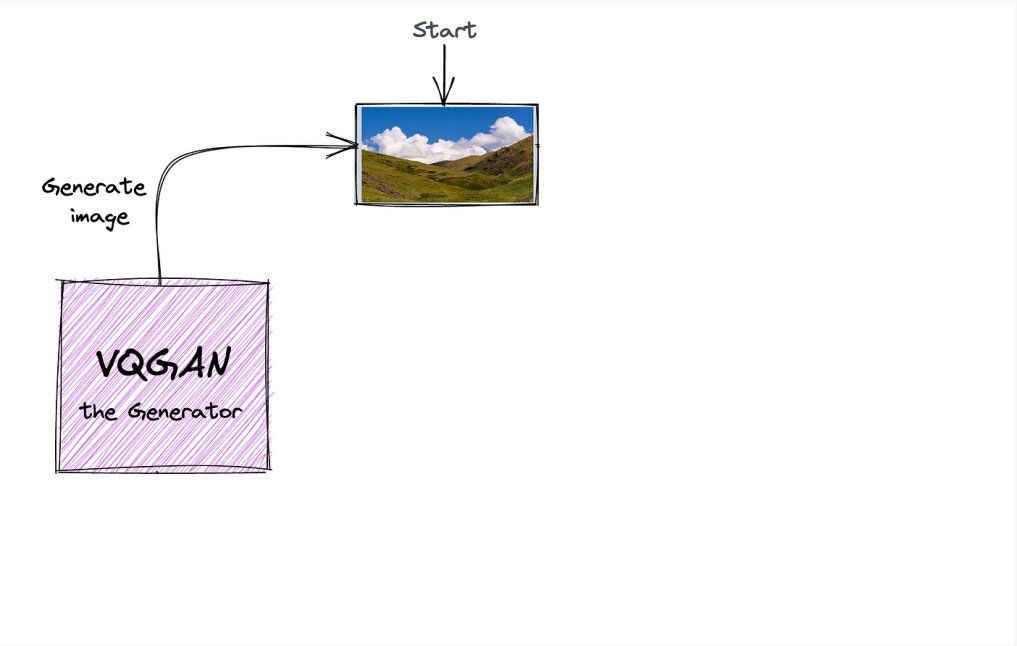


Figure 10: Generación de imágenes artificiales condicionadas por texto

Generador de texto a imagen: CLIP + VQ-GAN

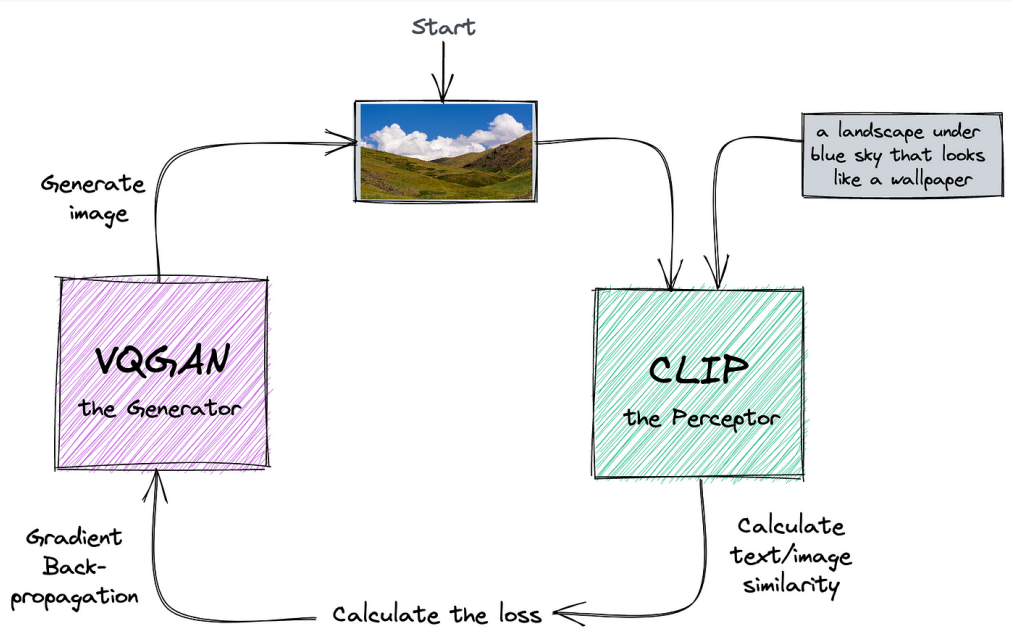


Figure 10: Generación de imágenes artificiales condicionadas por texto

Generador de texto a imagen: CLIP + VQ-GAN

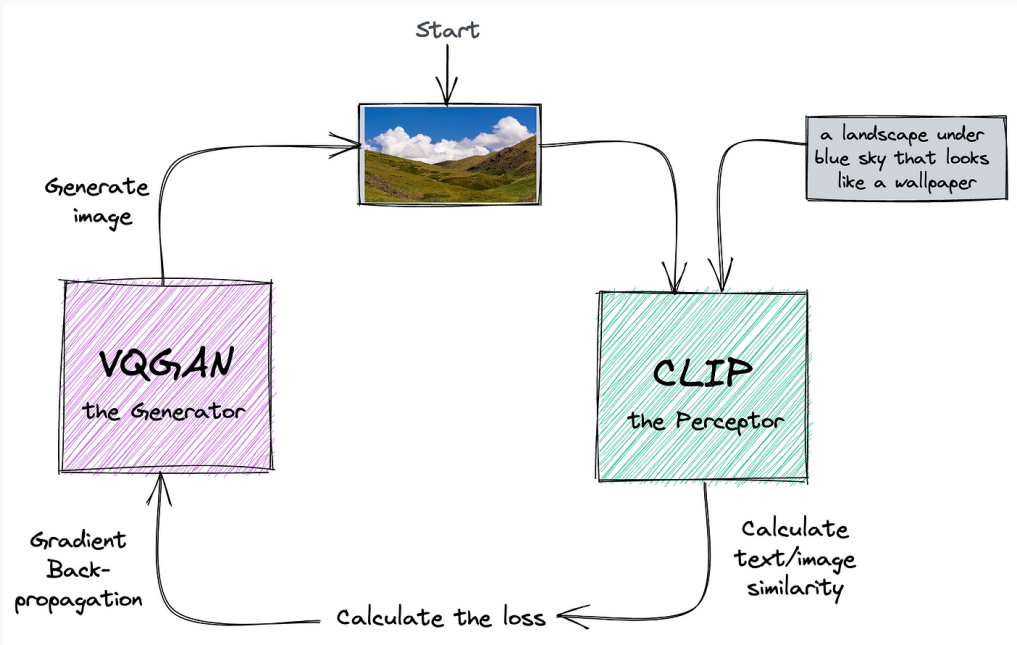
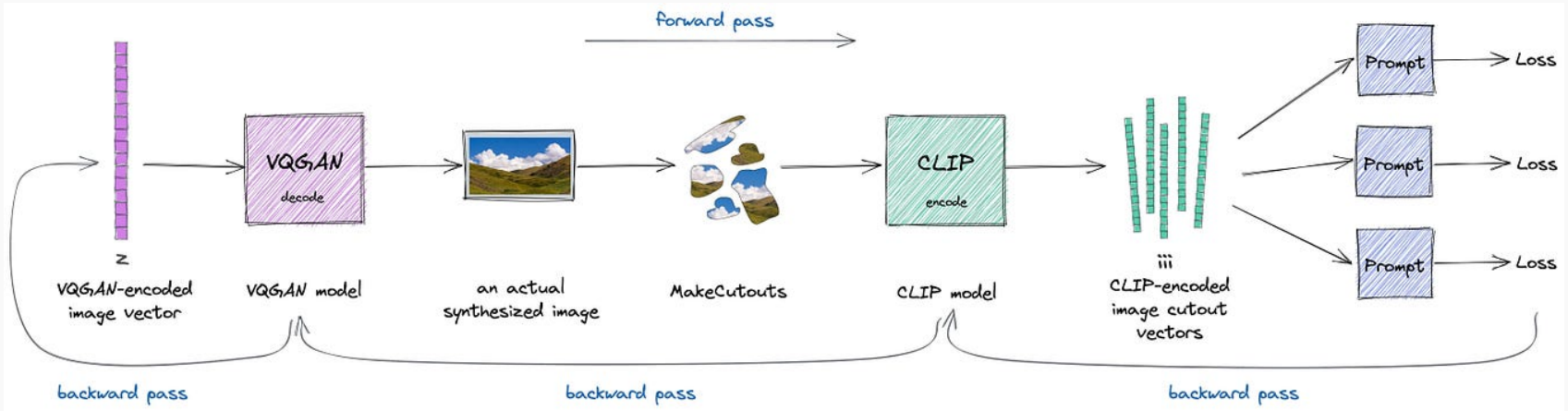


Figure 10: Generación de imágenes artificiales condicionadas por texto



Explicaciones interesantes y un notebook con explicaciones de código

Diffusion: Nuevos avances en texto-a-imagen I

Modelos de difusión permiten generar imágenes condicionadas por texto

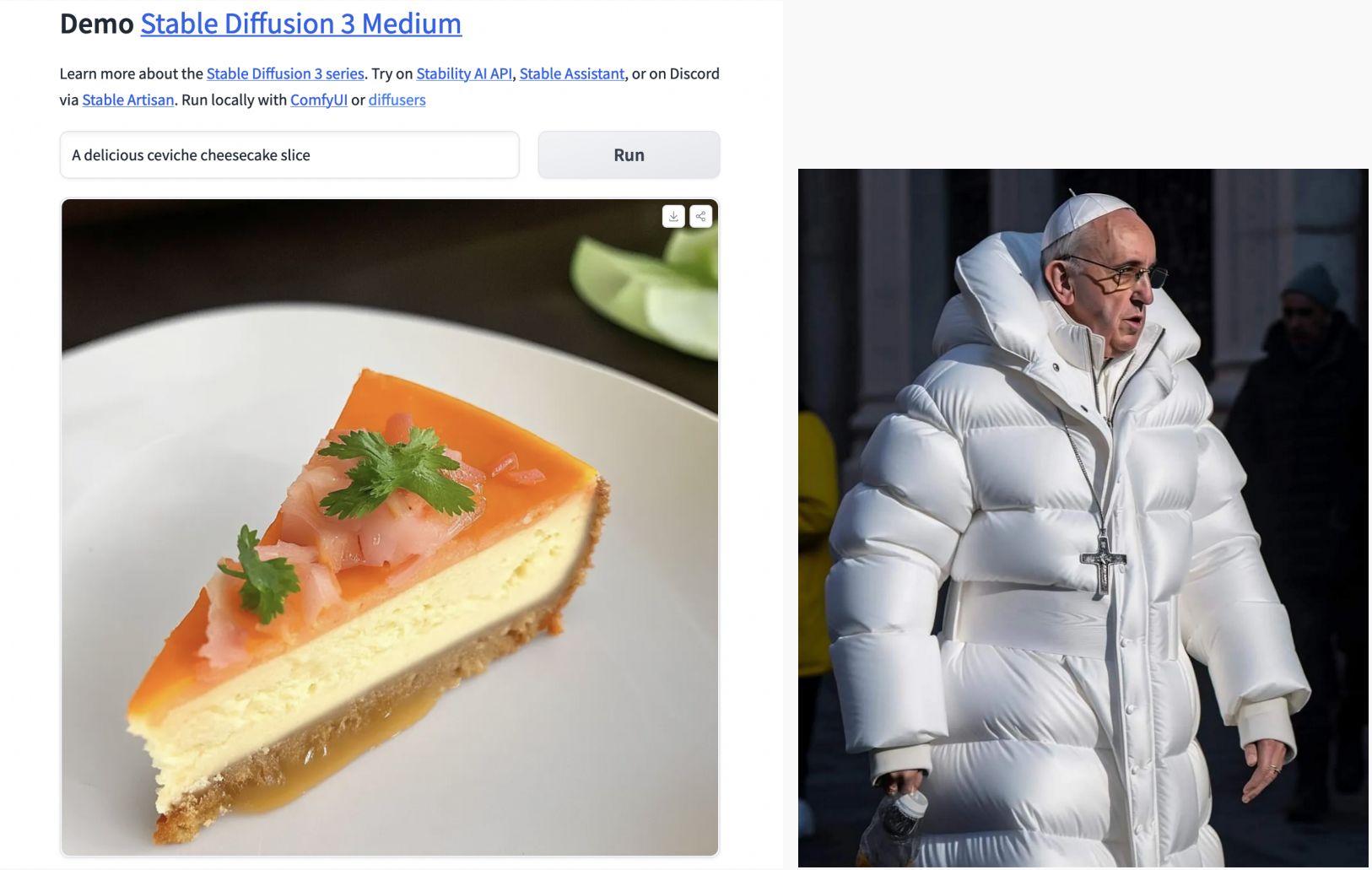


Figure 11: Ejemplos de imágenes divertidas obtenidas con modelos de difusión

Diffusion: Nuevos avances en texto-a-imagen II

Definición

Modelos probabilísticos que aprenden la distribución de los datos modelando la reversa de un proceso de difusión (añadir ruido paso a paso) para generar nuevos puntos de datos.

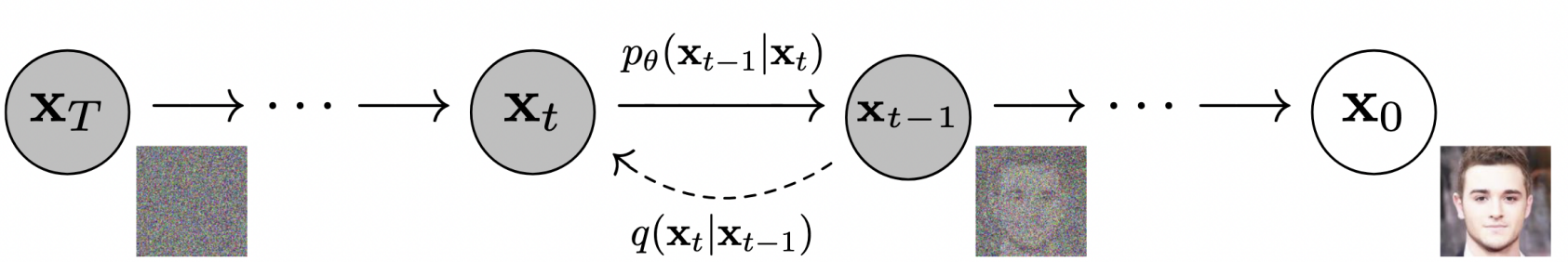


Figure 12: Desruído de una imagen [5]

Proceso

- **Forward**: Añadir pequeñas cantidades de ruido a los datos, que gradualmente se vuelve más ruidoso con los pasos, hasta asemejar ruido aleatorio.
- **Reverse**: Aprender a quitar gradualmente el ruido para recuperar los datos originales, empezando desde ruido puro y generando datos realistas.

Diffusion: Stable Diffusion

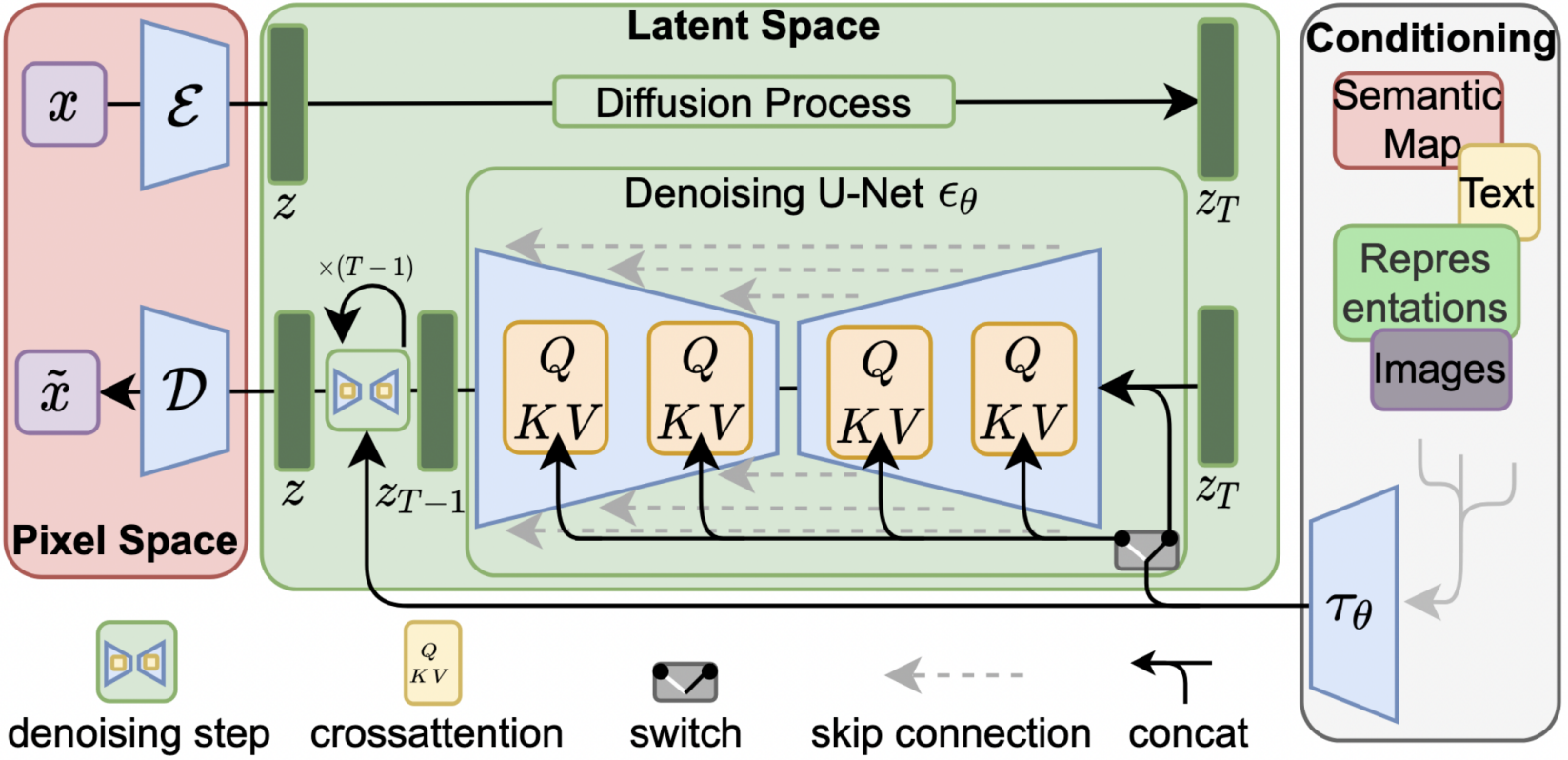


Figure 13: Arquitectura de Stable Diffusion [21]

- Usar una Red de Desruído (por ejemplo, U-Net) para reducir el ruido en cada paso.
- Condicionamiento (por ejemplo, embedding de texto) ayuda a guiar al modelo para generar una salida específica usando un mecanismo de x-attention.

Basados en LLM: Encoders congelados

- Aprender embeddings de imagen alineados con un **modelo de lenguaje congelado**
- Objetivo: codificar imágenes **en el espacio de embedding de palabras de un LLM**
- El LLM debe generar captions para esas imágenes
- Muy bueno en few-shot

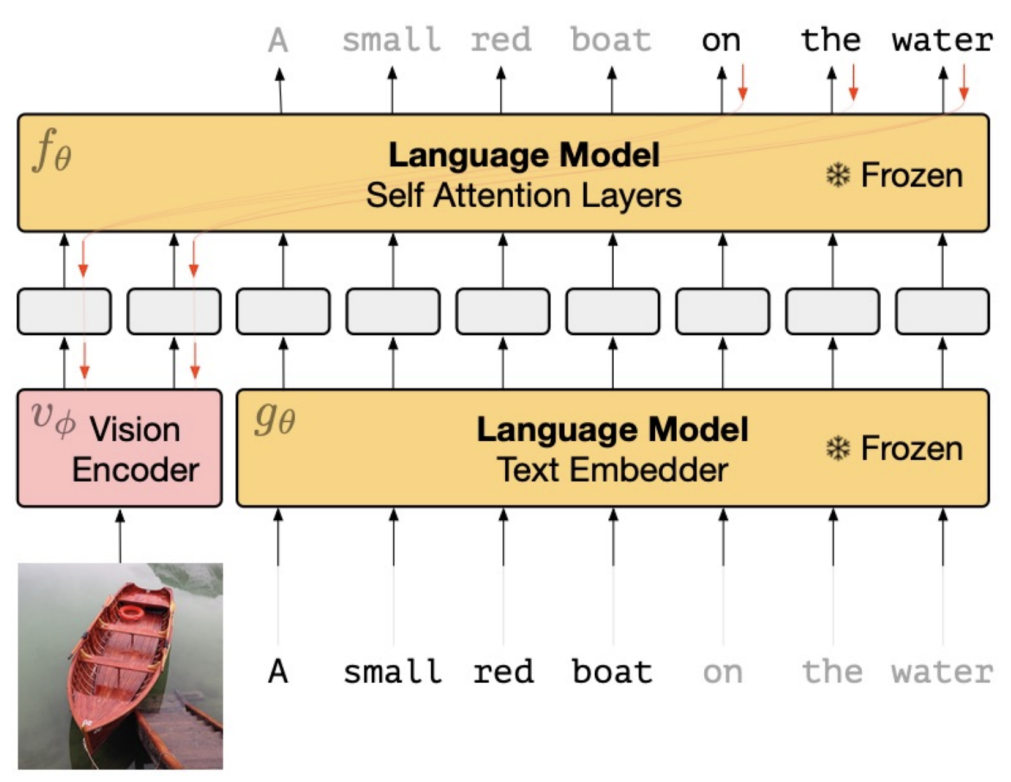


Figure 14: Ejemplos few-shot del modelo llamado Frozen [26]

Basados en LLM: BLIP2 y BLIP3

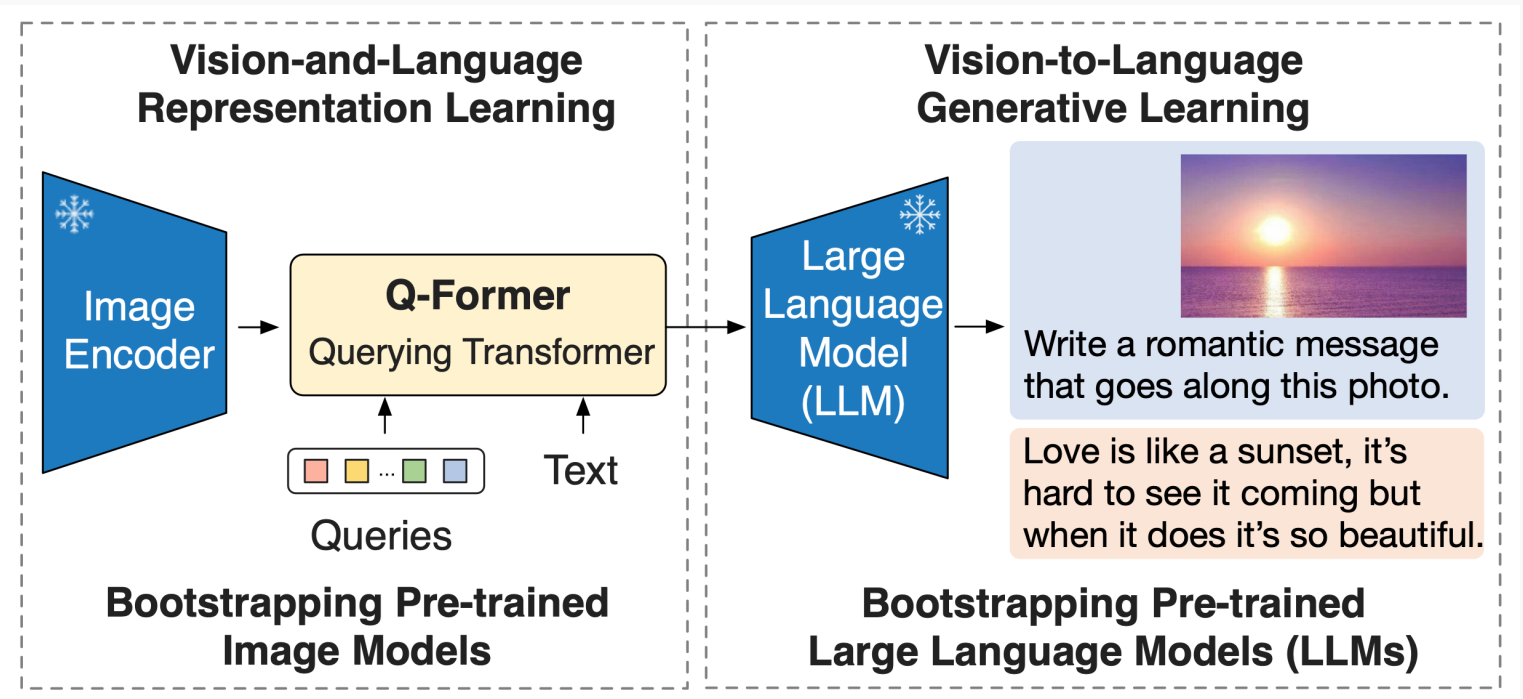


Figure 15: BLIP 2/3 [9, 27] usan encoders congelados, aprendiendo solo el Q-Former

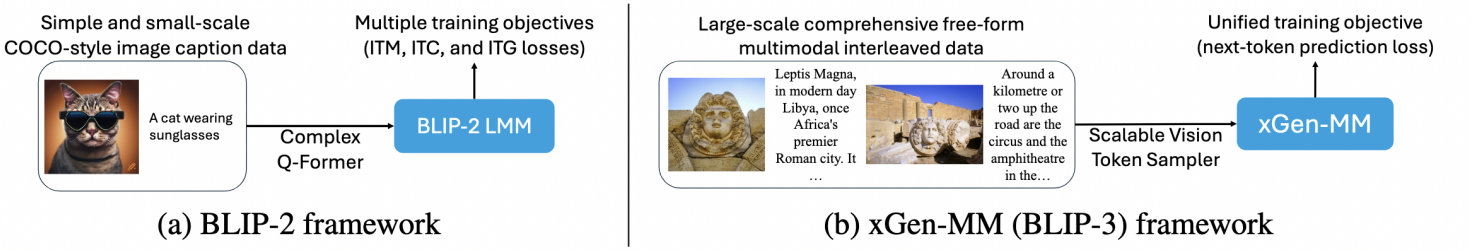


Figure 1: **We introduce xGen-MM (BLIP-3)**, a framework (b) for developing Large Multimodal Models (LMMs). Our framework improves upon BLIP-2 (a) [1] by (1) increasing the richness, scale, and diversity of training data, (2) replacing the Q-Former layers with a more scalable vision token sampler, and (3) simplifying the training process via the unification of the training objectives to a single loss at every training stage. The resulting suite of LMMs can perform various visual language tasks and achieve competitive performance across benchmarks.

Basados en LLM: xGen-MM (BLIP3) I

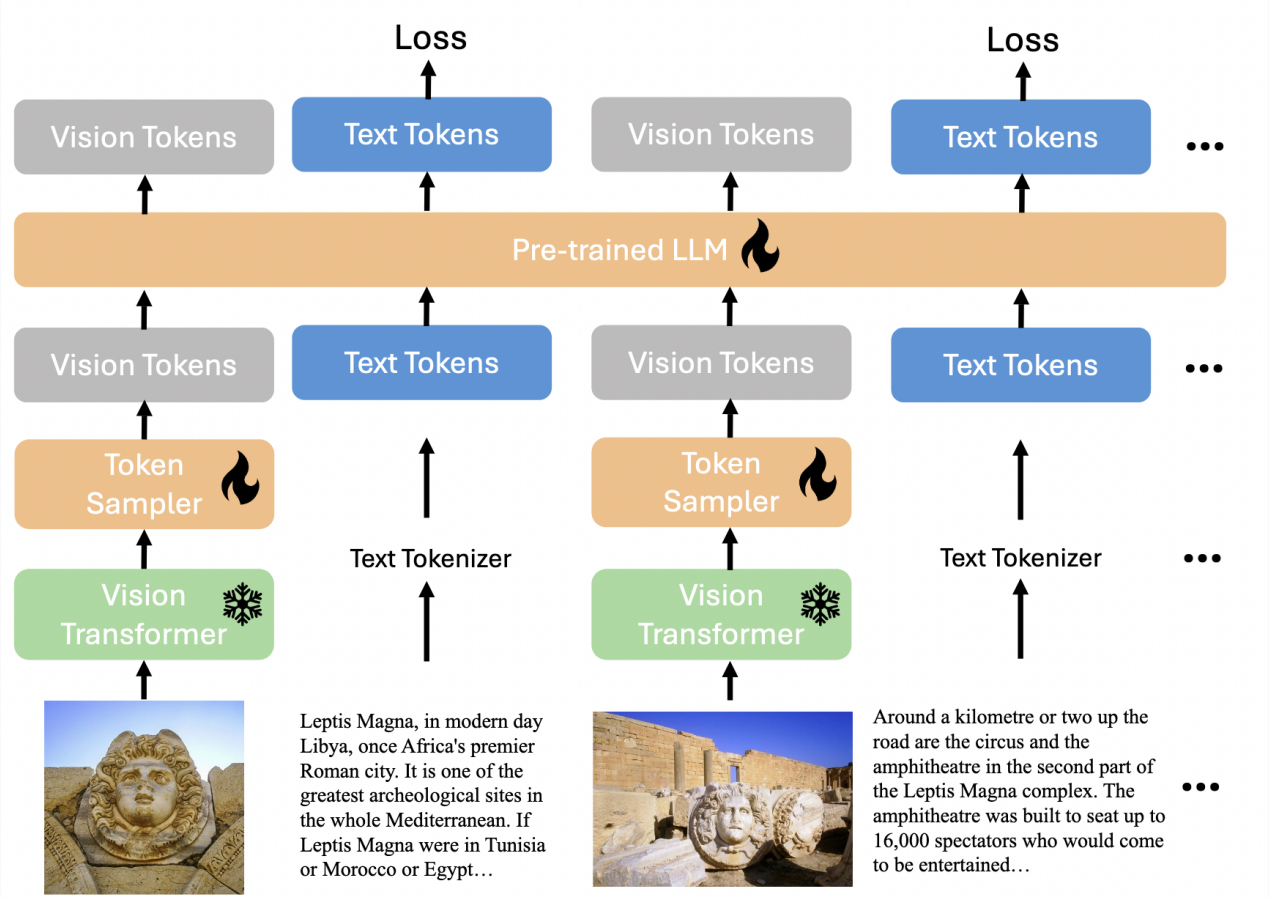


Figure 2: **Overview of the xGen-MM (BLIP-3) framework.** Free-form interleaved images and texts from the ensembled interleaved and caption datasets are input into the framework, with each modality undergoing a separate tokenization process to be fed into the pre-trained LLM in natural order. A standard auto-regressive loss is then applied to the text tokens. The Vision Transformer is kept frozen during training, while all other parameters, including the token sampler and the pre-trained LLM, are trained.

Figure 16: BLIP 3 usa encoders congelados, pero ahora se apoya únicamente en la tarea de Modelado de Lenguaje. Modelos disponibles aquí

Basados en LLM: xGen-MM (BLIP3) II

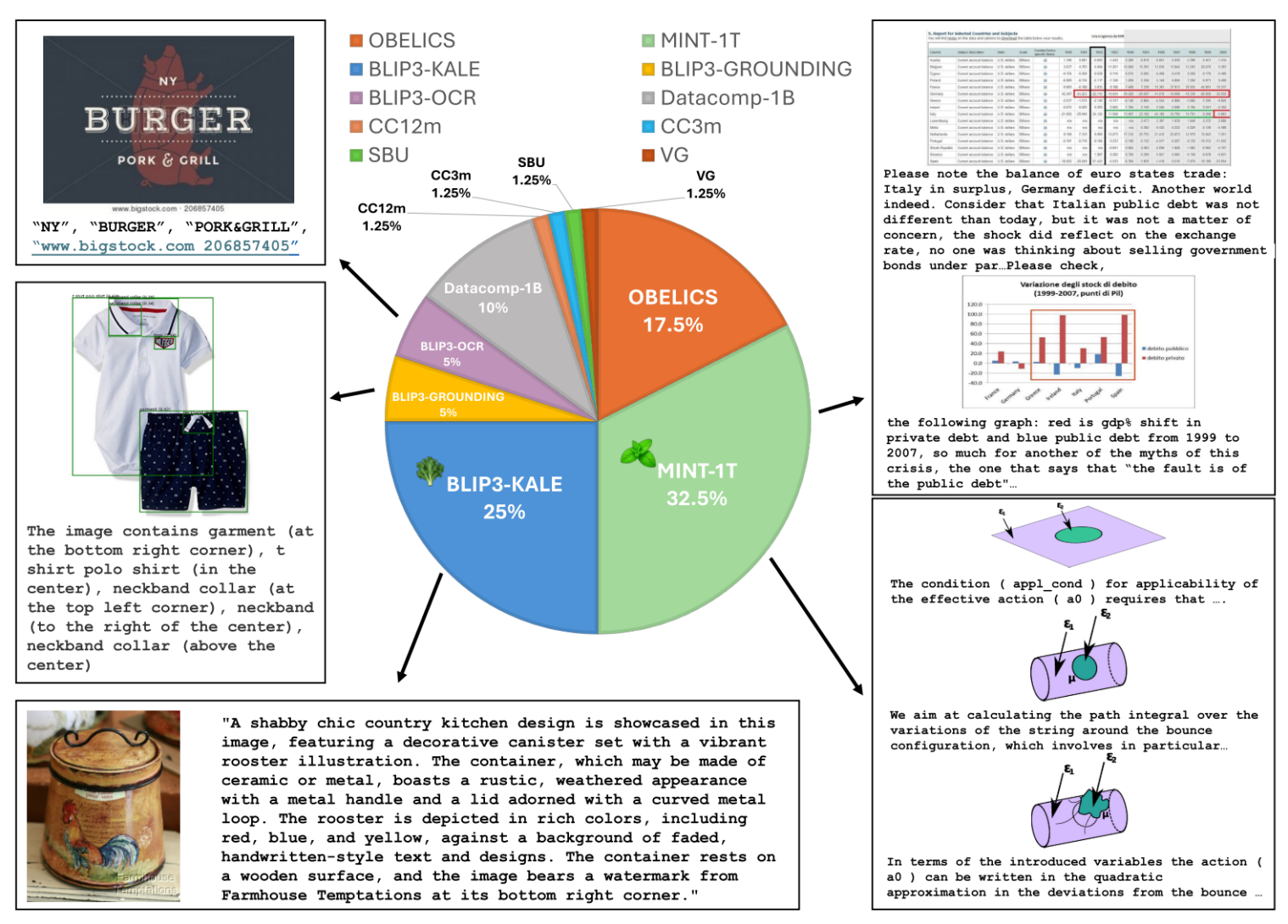


Figure 17: El conjunto de entrenamiento de BLIP3 contiene datos multimodales entrelazados (mezcla de secuencias de imágenes y texto) de contenido muy diverso

Basados en LLM: Flamingo

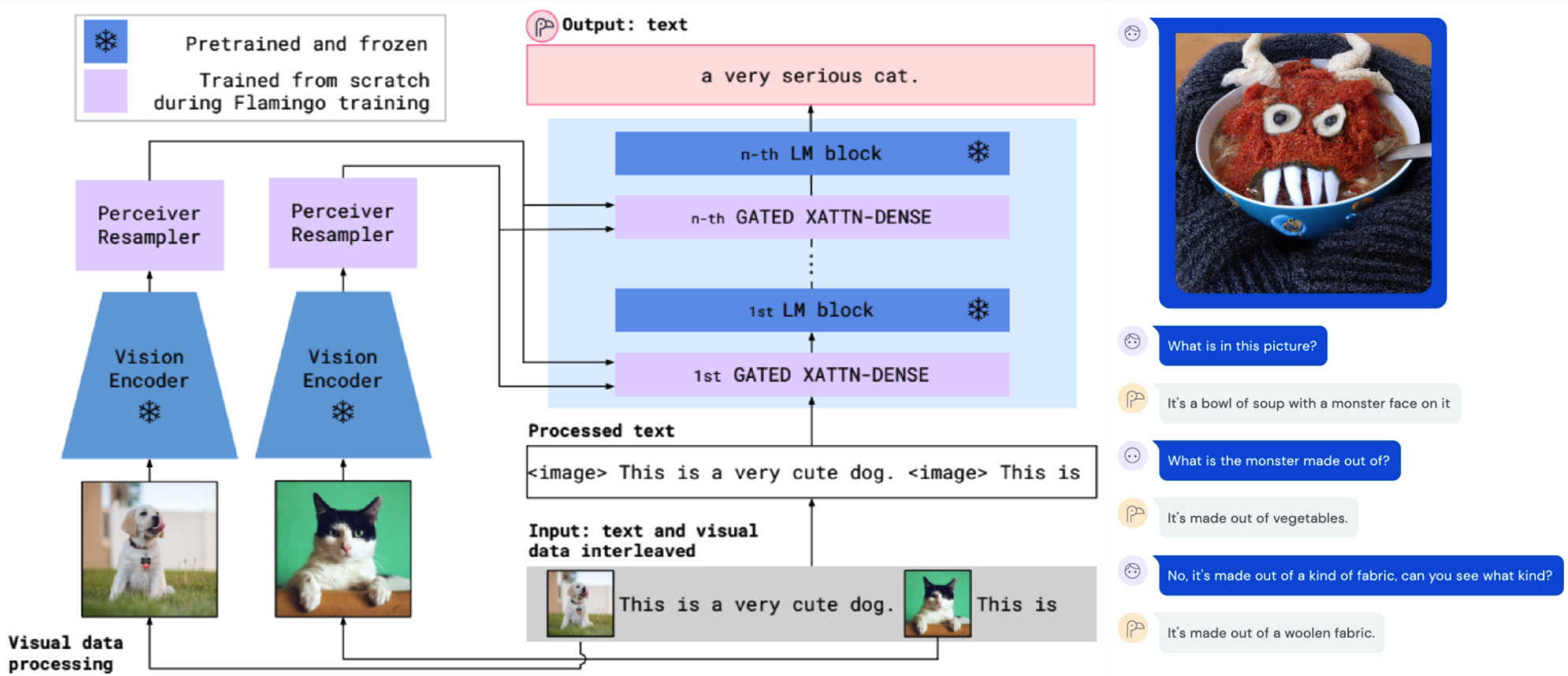
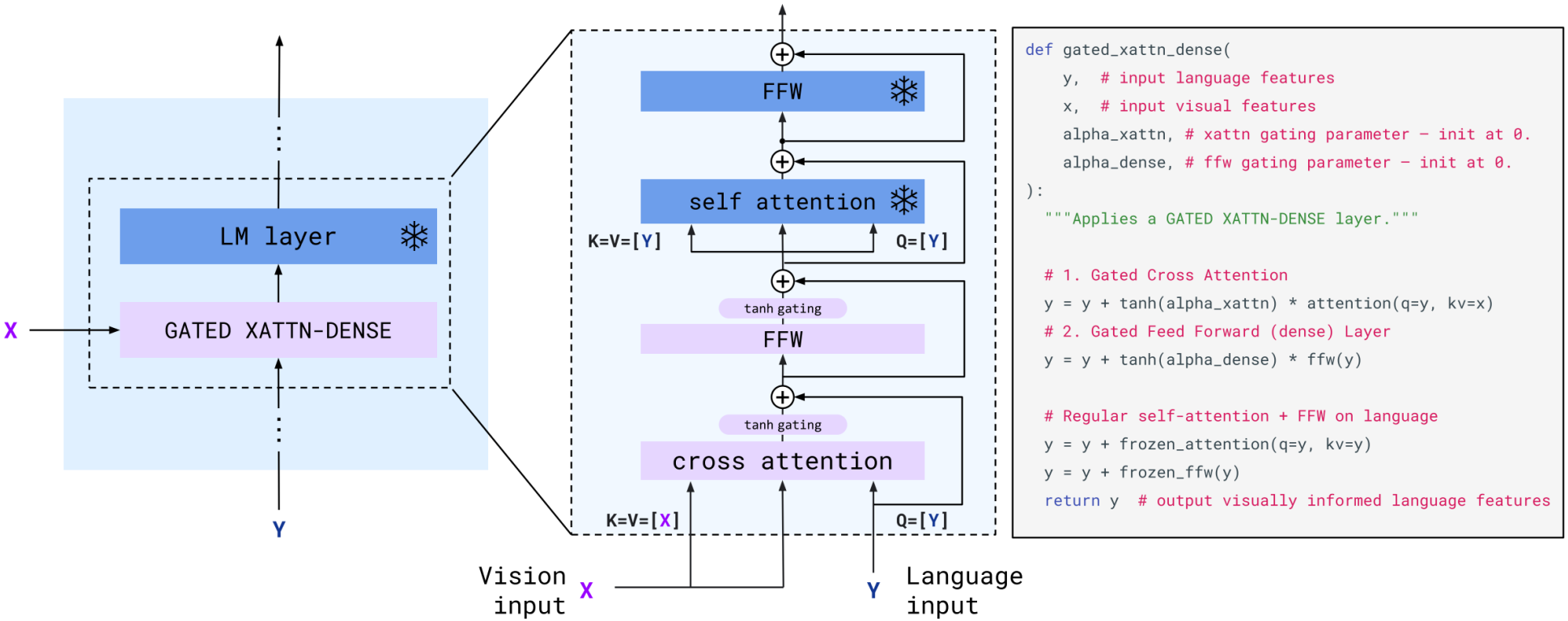


Figure 18: Flamingo fue el primer LMM que puede conversar con un humano [1]

- Partes entrenadas desde cero, otras preentrenadas y congeladas
- Mecanismo de compuerta cruzada
- Texto intercalado con imágenes

Basados en LLM: Flamingo II



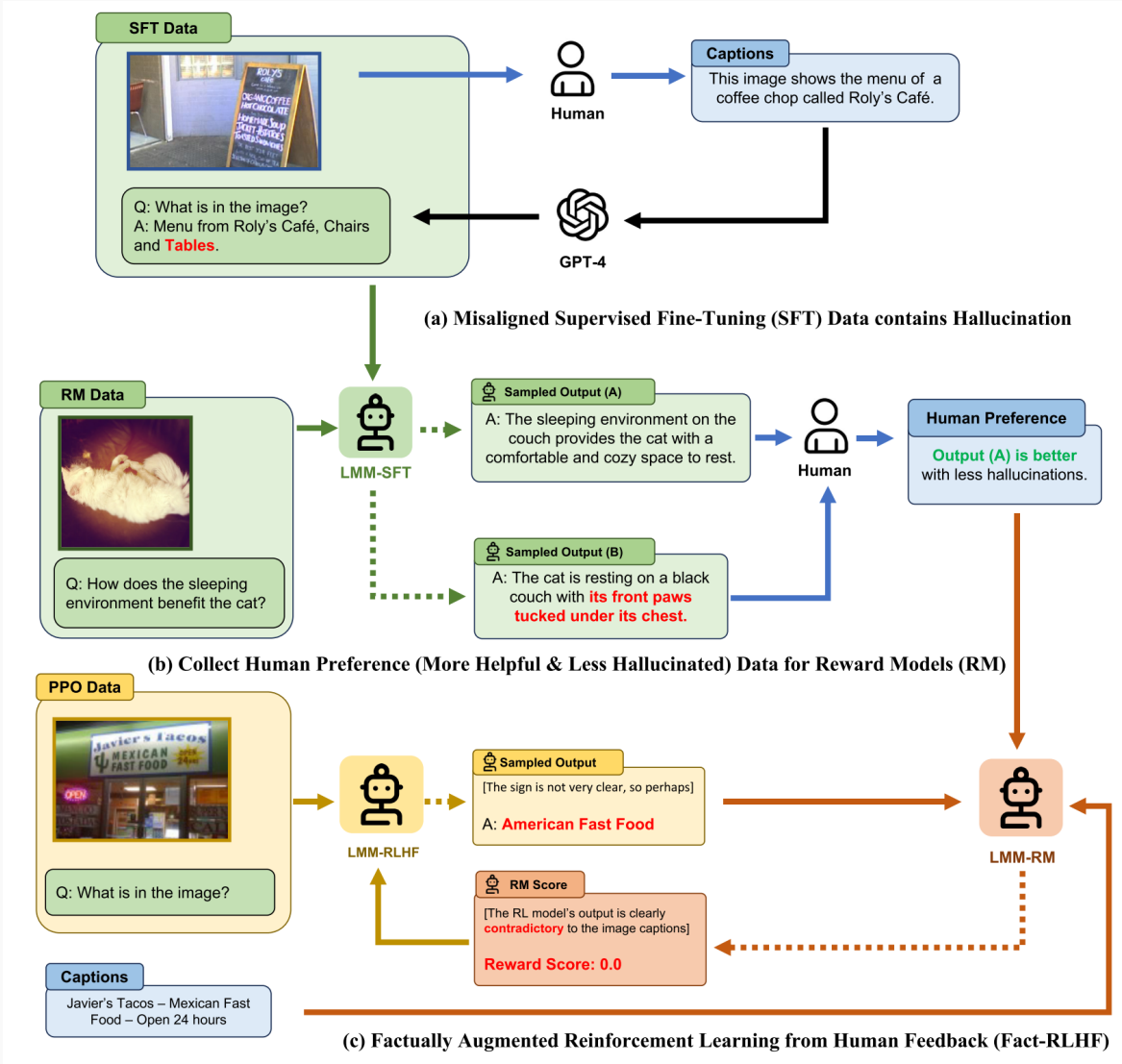

```
def gated_xattn_dense(
    y,  # input language features
    x,  # input visual features
    alpha_xattn, # xattn gating parameter – init at 0.
    alpha_dense, # ffw gating parameter – init at 0.
):
  """Applies a GATED XATTN-DENSE layer."""

  # 1. Gated Cross Attention
  y = y + tanh(alpha_xattn) * attention(q=y, kv=x)
  # 2. Gated Feed Forward (dense) Layer
  y = y + tanh(alpha_dense) * ffw(y)

  # Regular self-attention + FFW on language
  y = y + frozen_attention(q=y, kv=y)
  y = y + frozen_ffw(y)
  return y  # output visually informed language features
```

Figure 19: Mecanismo de atención con compuertas de Flamingo

Las capas del LM están congeladas para preservar la calidad obtenida de un pre-entrenamiento muy grande.

Basados en LLM: Large Language and Vision Assistant (LLaVA)

- Modelo open-source basado en LLaMA, adaptándolo a datos multimodales [15]
- Matriz de proyección desde la imagen suministrada usando CLIP y alimenta a LLaMA
- Pre-entrenamiento de la matriz de proyección, luego fine-tuning añadiendo los pesos del LLM
- Posible usar RLHF [24]



Datasets open-source: OBELICS

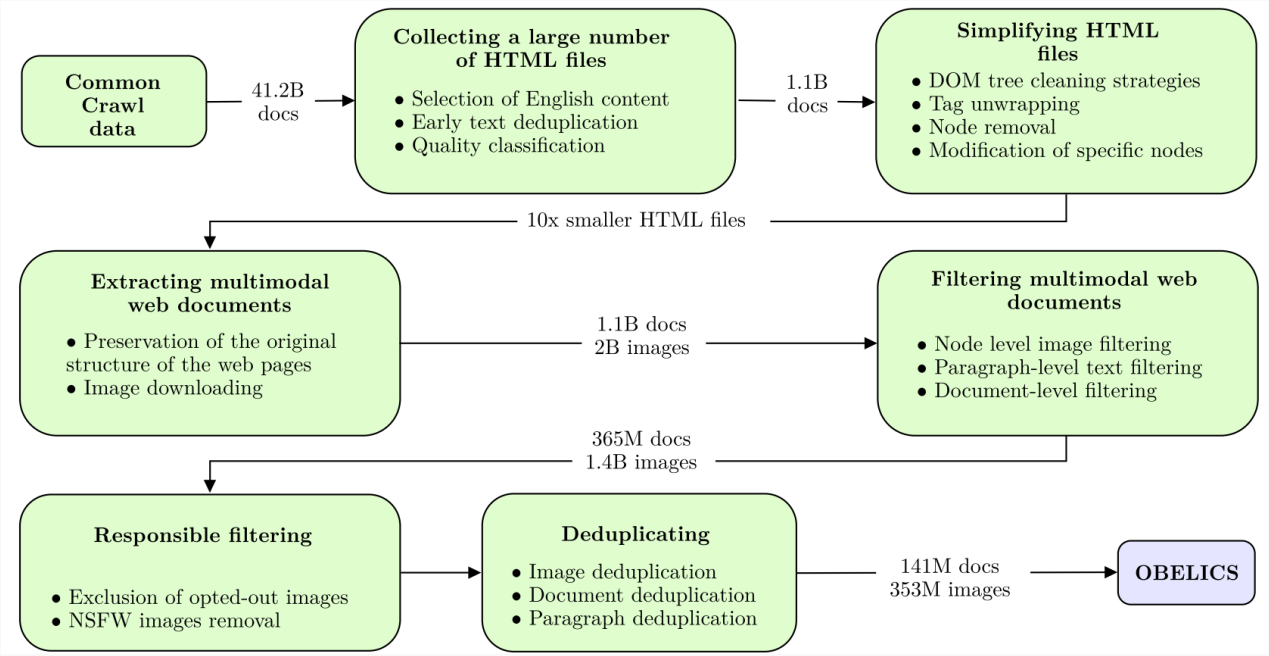


Figure 20: OBELICS [8] es una colección abierta, masiva y curada de documentos web imagen-texto entrelazados, con 141M documentos, 115B tokens de texto y 353M imágenes

Una serie de modelos llamados IDEFICS entrenados en este dataset están disponibles open-source aquí.

Hacia video

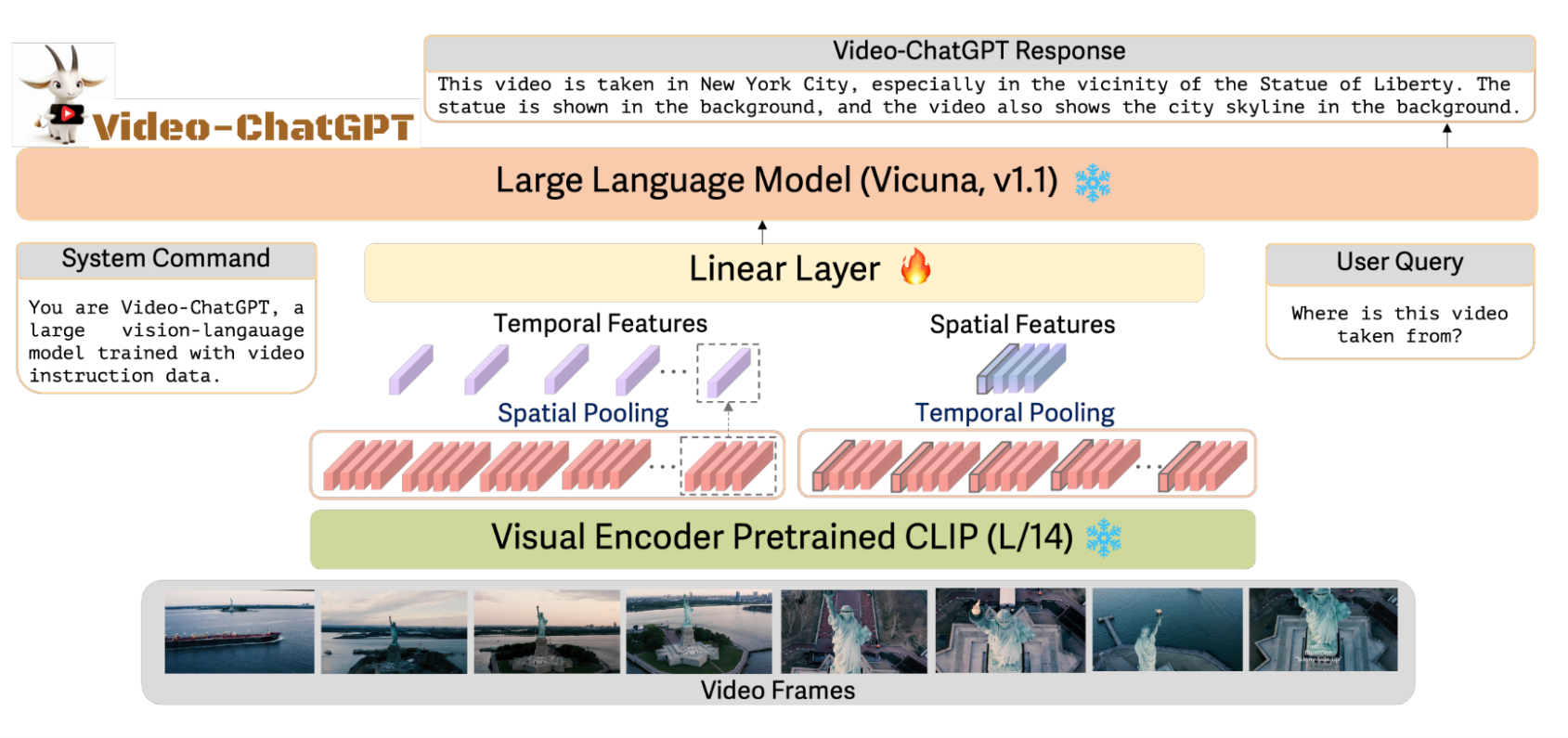
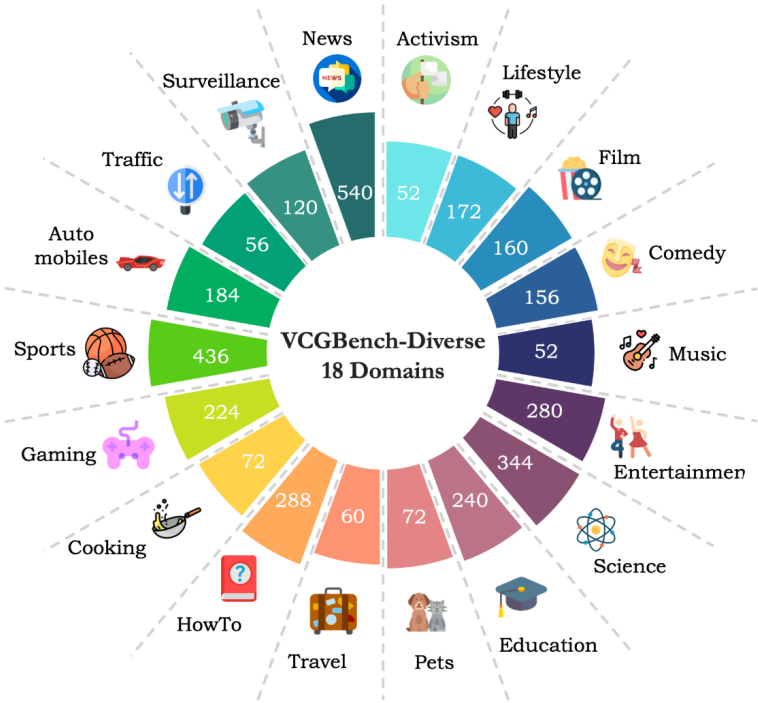


Figure 21: Arquitectura de Video-ChatGPT: el encoder visual extrae características que se agrupan espacial y temporalmente [19]

¿Cómo evaluar el LLM generativo de video?



Category	Description	#	Domains
	5 Video Capturing Methods		
Stable Settings	Videos shot in stable, predictable environments with minimal camera movement.	1200	Cooking, How-to, Education
Dynamic Settings	Videos with significant camera movement requiring adaptation to rapid context shifts.	448	Sports, Traffic, Travel
Fixed Cameras	Videos recorded from stationary cameras, providing consistent viewpoints for monitoring purposes.	124	Surveillance, Automobile
Professional Quality	Professionally produced videos with high visual and audio quality, and controlled lighting.	1608	News, Film
Variable Quality	Informal videos with varying quality, often using handheld devices, captured in spontaneous settings.	124	Lifestyle, Pets
	6 Reasoning Complexities		
Sequential Understanding	Requires comprehension and following of a series of steps or actions in order.	828	Cooking, How-to, Education
Predictive Reasoning	Involves understanding and predicting outcomes of dynamic, intricate action sequences.	180	Sports, Gaming
World Knowledge	Demands integration of broader contextual information and world knowledge to interpret video content.	848	Science, News
Causal Reasoning	Focuses on understanding cause-and-effect relationships within the video.	340	Surveillance, Activism
Emotional Reasoning	Involves interpreting stories, character motivations, and emotional subtexts.	1080	Entertainment, Film, Comedy
Analytical Reasoning	Requires critical analysis and interpretation of complex information or situations.	228	Traffic, Automobile

Figure 22: VCGBench-Diverse benchmark conversacional de video [18]

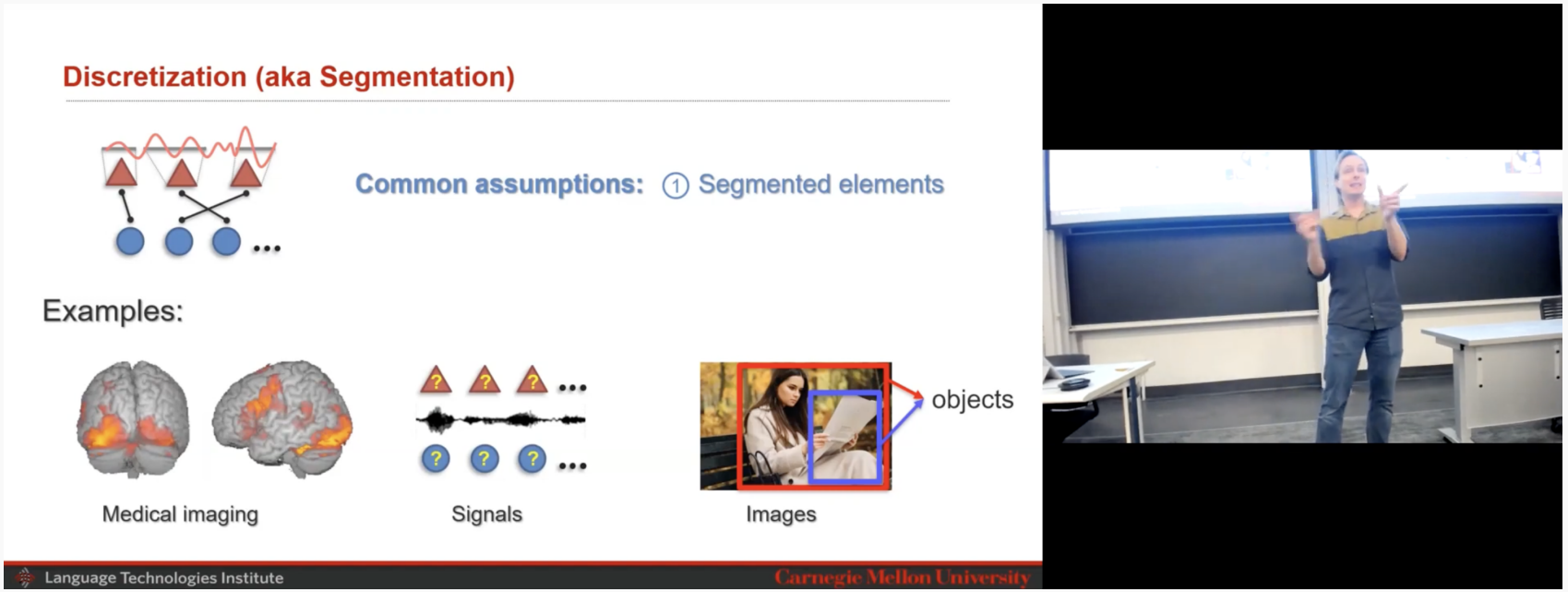
- 18 categorías amplias de video, con 4,354 pares pregunta-respuesta
- tareas: captioning denso de video, comprensión espacial y temporal, y razonamiento complejo
- cinco métodos de captura de video, asegurando diversidad y generalización robusta y seis niveles de complejidad de razonamiento

Hoy, ¡todo es un token!

- Tokenización de imagen con SEED: [3]
- SpeechTokenizer de [30]
- AnyGPT: [29]
- OmniVec: [23]

¿Quieres aprender más sobre ML Multimodal?

El excelente curso Multimodal Machine Learning de CMU hecho por LP Morency y Paul Piu Liang te llevará lo más lejos posible!



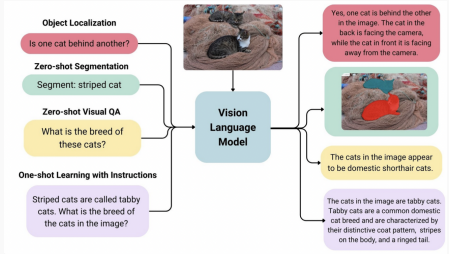
`https://www.youtube.com/@LPMorency/videos`

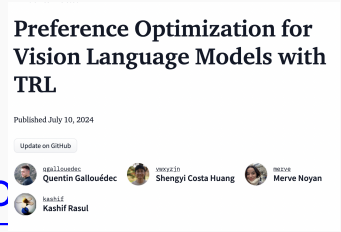
También puedes revisar estos artículos increíbles [14, 11].

Y estas charlas sobre Fundamentos de la representación multimodal y Fundamentos de AI multisensorial de alta modalidad

¿Quieres aprender más sobre ML Multimodal?

- Puedes revisar este [tutorial de LMM](#) de `huggingface`
- Este otro sobre [Optimización de preferencias para LMMs usando](#)
- Otro tutorial sobre cómo [usar PaliGemma](#), con [una demo](#)





Questions?

References i

J. B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. C. T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan.
Flamingo: a Visual Language Model for Few-Shot Learning.
In Advances in Neural Information Processing Systems, volume 35, 2022.

J. Arevalo, T. Solorio, M. Montes-Y-Gómez, and F. A. González.
Gated multimodal units for information fusion.
In 5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings, 2017.

References ii

Y. Ge, Y. Ge, Z. Zeng, X. Wang, and Y. Shan.
MAKING LLAMA SEE AND DRAW WITH SEED TOKENIZER.
In ICLR, pages 1–26, 2024.

L. Hemamou, G. Felhi, V. Vandenbussche, J.-c. Martin, and C. Clavel.
HireNet : a Hierarchical Attention Model for the Automatic Analysis of Asynchronous Video Job Interviews.
In AAAI, 2019.

J. Ho, A. Jain, and P. Abbeel.
Denoising diffusion probabilistic models.
In Advances in Neural Information Processing Systems, volume 2020-Decem, pages 1–25, 2020.

References iii

Y.-H. Hubert Tsai, P. Pu Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov.
Learning Factorized Multimodal Representations.
In ICLR, number Pas publie, 2019.

W. Kim, B. Son, and I. Kim.
ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision.
In Proceedings of Machine Learning Research, volume 139, pages 5583–5594, 2021.

References iv

H. Laurençon, L. Saulnier, L. Tronchon, S. Bekman, A. Singh, A. Lozhkov, T. Wang, S. Karamcheti, A. M. Rush, D. Kiela, M. Cord, and V. Sanh.
OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents.
In Advances in Neural Information Processing Systems, volume 36, pages 1–20, 2023.

J. Li, D. Li, S. Savarese, and S. Hoi.
BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models.
2023.

J. Li, D. Li, C. Xiong, and S. Hoi.
BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation.
(2), 2022.

References

P. P. Liang, Y. Cheng, X. Fan, C. K. Ling, S. Nie, R. J. Chen, Z. Deng, N. Allen, R. Auerbach, F. Mahmood, R. Salakhutdinov, and L. P. Morency.
Quantifying & Modeling Multimodal Interactions: An Information Decomposition Framework.
Advances in Neural Information Processing Systems, 36(NeurIPS):1–43, 2023.

P. P. Liang, Z. Liu, A. Zadeh, and L.-P. Morency.
Multimodal Language Analysis with Recurrent Multistage Fusion.
In EMNLP, 2018.

References vi


P. P. Liang, A. Zadeh, and L.-P. Morency.
Multimodal Local-Global Ranking Fusion for Emotion Recognition.
2018.


P. P. Liang, A. Zadeh, and L.-P. Morency.
Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions.
(1):1–65, 2022.


H. Liu, C. Li, Y. Li, and Y. Jae.
Improved Baselines with Visual Instruction Tuning.
In Neurips, 2023.

Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency.
Efficient Low-rank Multimodal Fusion with Modality-Specific Factors.
In ACL, 2018.

J. Lu, D. Batra, D. Parikh, and S. Lee.
ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks.
In NIPS'19, number NeurIPS, pages 1–11, 2019.

M. Maaz, H. Rasheed, S. Khan, and F. Khan.
VideoGPT+: Integrating Image and Video Encoders for Enhanced Video Understanding.
pages 1–18, 2024.

References viii

M. Maaz, H. Rasheed, S. Khan, and F. S. Khan.
Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models.
In ACL, volume 1, pages 12585–12602, 2024.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever.
Learning Transferable Visual Models From Natural Language Supervision.
2021.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer.
High-Resolution Image Synthesis with Latent Diffusion Models.
CVPR, 2022.

S. Sahay, S. H. Kumar, R. Xia, J. Huang, and L. Nachman.
Multimodal Relational Tensor Network for Sentiment and Emotion Classification.
2018.

S. Srivastava and G. Sharma.
OmniVec: Learning robust representations with cross modal sharing.
In Proceedings - 2024 IEEE Winter Conference on Applications of Computer Vision, WACV 2024, pages 1225–1237, 2024.

Z. Sun, S. Shen, S. Cao, H. Liu, C. Li, Y. Shen, C. Gan, L.-Y. Gui, Y.-X. Wang, Y. Yang, K. Keutzer, and T. Darrell.
Aligning Large Multimodal Models with Factually Augmented RLHF.
pages 1–26, 2023.

References x


H. Tan and M. Bansal.
LXMERT: Learning Cross-Modality Encoder Representations from Transformers.
In EMNLP, pages 5099–5110, 2019.


M. Tsimpoukelli, J. Menick, S. Cabi, S. M. A. Eslami, O. Vinyals, and F. Hill.
Multimodal Few-Shot Learning with Frozen Language Models.
2021.


L. Xue, M. Shu, A. Awadalla, J. Wang, and A. Yan.
xGen-MM (BLIP-3): A Family of Open Large Multimodal Models.
pages 7–11, 2024.


A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency.
Tensor Fusion Network for Multimodal Sentiment Analysis.
In EMNLP, 2017.


J. Zhan, J. Dai, J. Ye, Y. Zhou, D. Zhang, Z. Liu, X. Zhang, R. Yuan, G. Zhang, L. Li, H. Yan, J. Fu, T. Gui, T. Sun, Y.-g. Jiang, and X. Qiu.
AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling.
In ACL, volume 1, pages 9637–9662, 2024.


X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu.
SPEECHTOKENIZER: UNIFIED SPEECH TOKENIZER FOR SPEECH LANGUAGE MODELS.
In 12th International Conference on Learning Representations, ICLR 2024, pages 1–21, 2024.