De la simple fusión a las nuevas fronteras: una introducción a la IA multimodal



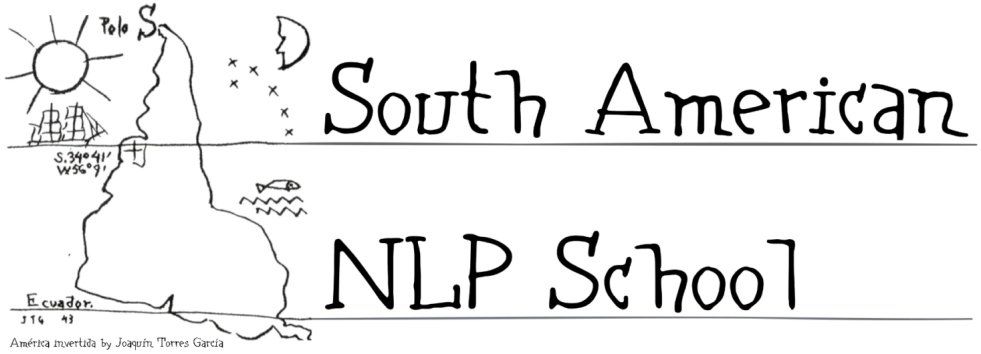
UNIVERSIDAD DE CHILE

2nd South American NLP School, Buenos Aires

Valentin Barriere

August 3rd 2026

Universidad de Chile – DCC



Índice

¿Por qué Multimodal?

Fusión: de Temprana/Tardía a Compuertas Aprendidas

Fusion basada en transformers

Cuantificando Interacciones: Descomposición de Información Parcial

Generalistas Multimodales Centrados en LLM

Tokenizacion

Ampliando la Frontera de las Modalidades

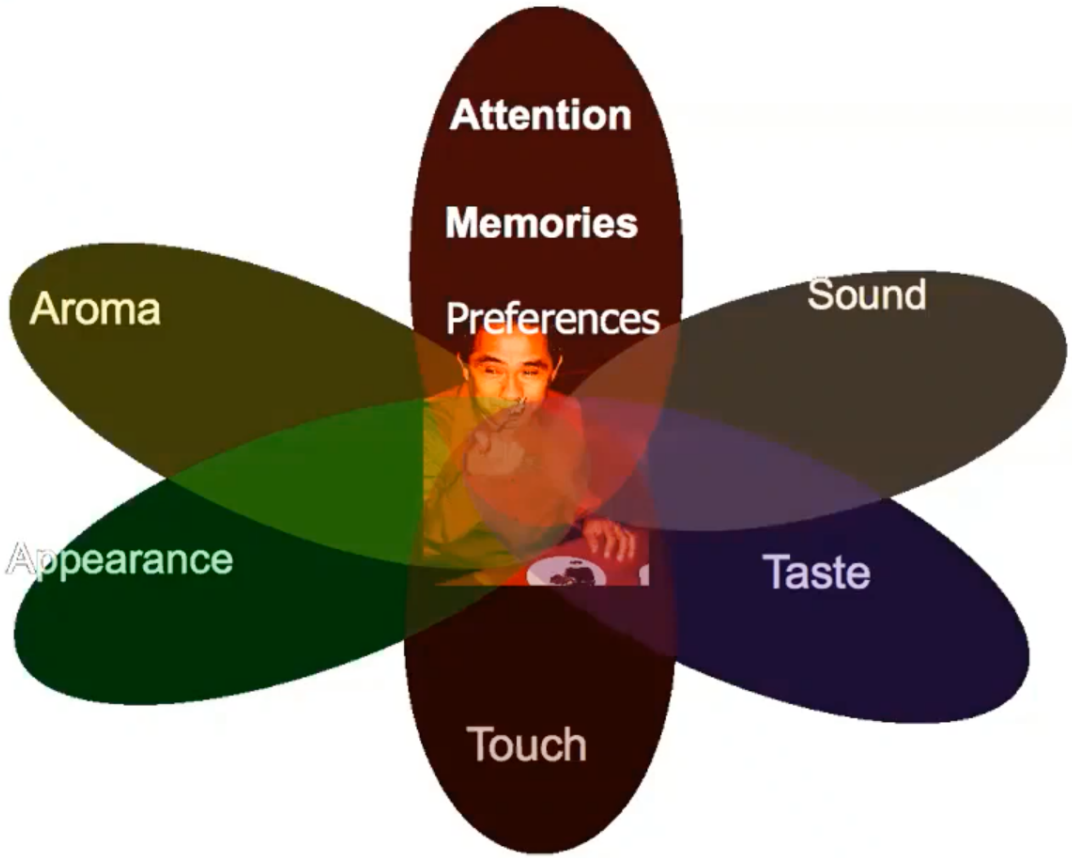
Nuestros trabajos en curso

Cierre

¿Por qué Multimodal?

¿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

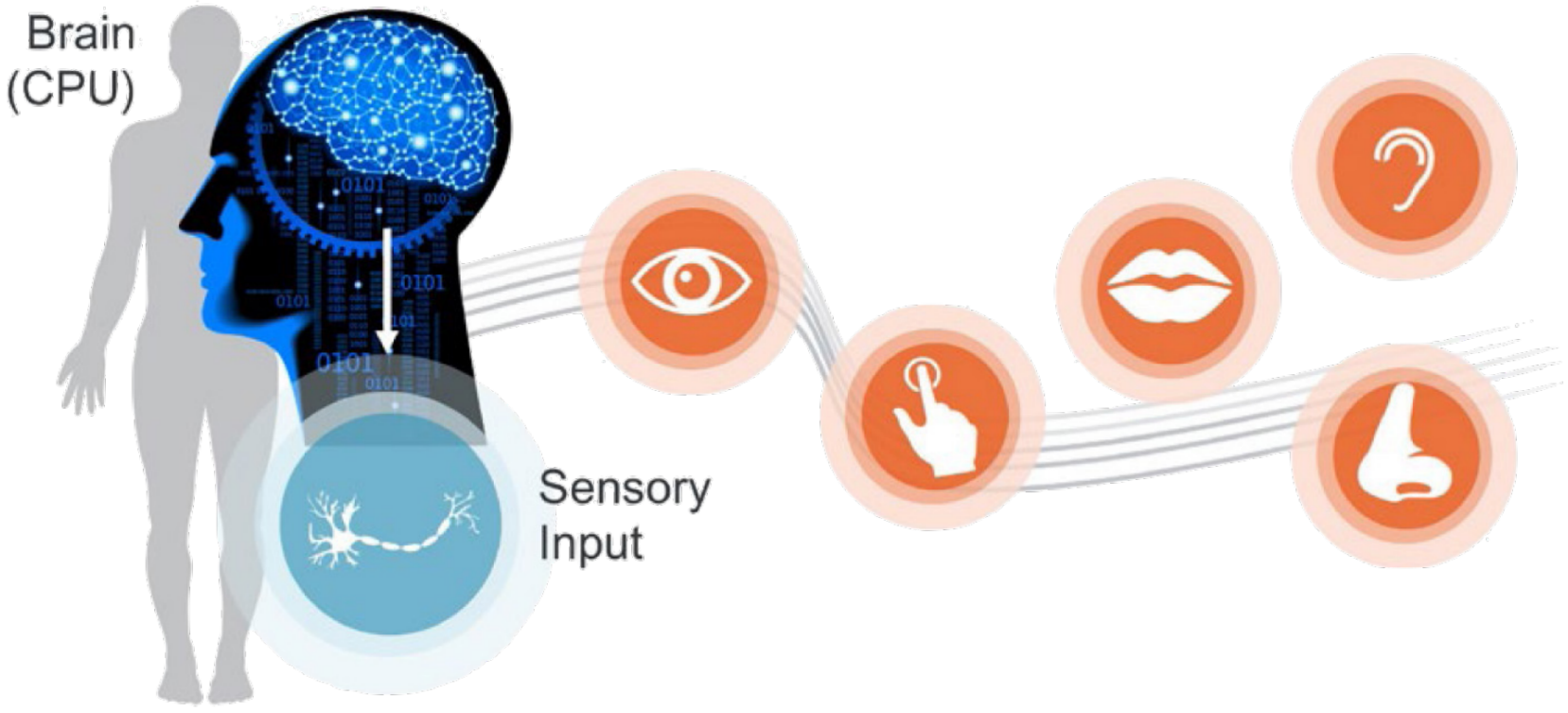


Figure 1: Usamos la multimodalidad en la vida cotidiana, así es como los humanos se comunican

¿Por qué debería importarle a una audiencia de NLP?

La versión corta

Todo LLM reciente considerado un "avance" se está convirtiendo silenciosamente en un modelo multimodal: GPT-4o, Gemini, Claude, Llama 4, Qwen-VL/Audio, ...

- El texto ya no es la única entrada/salida de un modelo de lenguaje
- Los trucos que nos trajeron hasta aquí (tokenización, alineación, instruction-tuning) son los mismos trucos de NLP, solo aplicados a otras señales
- Los agentes multimodales, los sistemas de RAG y de razonamiento necesitan cada vez más anclar el lenguaje en visión, audio, sensores, ...

Hoy: cómo se fusionan las modalidades, cómo medimos lo que realmente comparten, dos maneras de conectar una modalidad a un LLM, y hasta dónde se puede estirar la definición de 'modalidad'

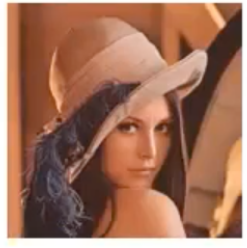
Antecedentes: Multimodalidad

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



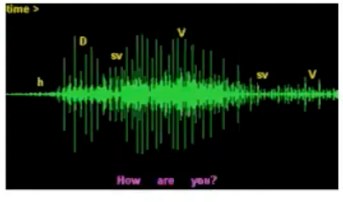
Language

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



Vision

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

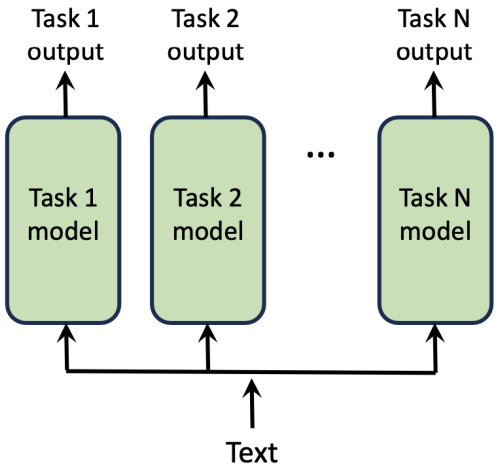


Speech

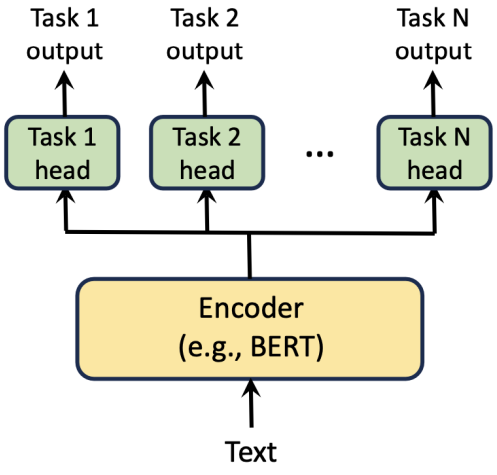
Pero también: tacto, fisiológico (ECG), móvil (ubicación), social (amigos en una red)...

Evolución de los modelos fundacionales de texto

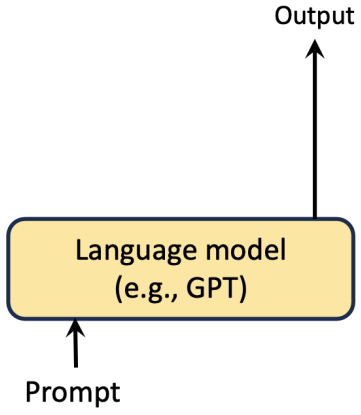
The task-specific model era (- 2018)



The encoder era (2018 - 2022)

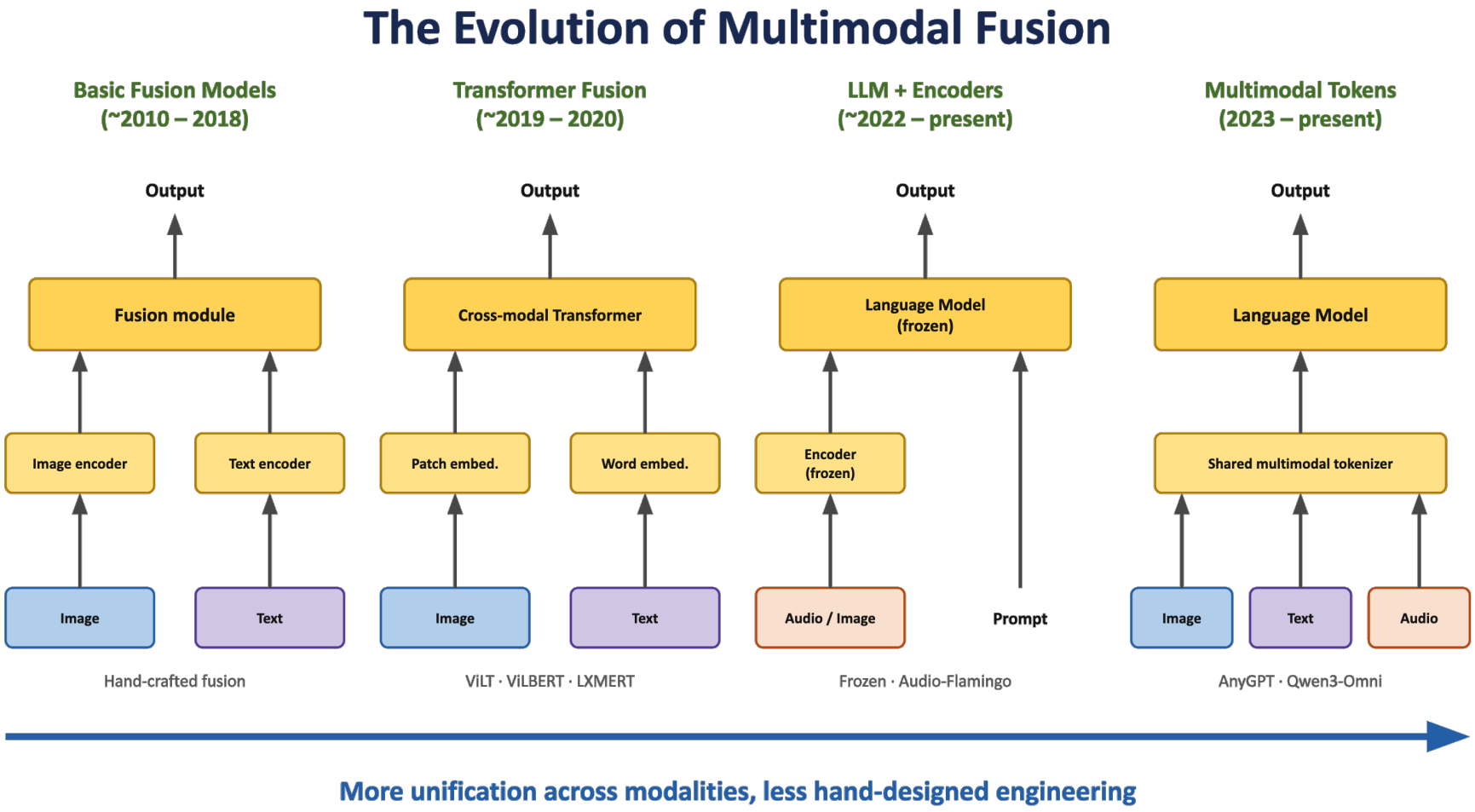


The large language model era (2022 -)

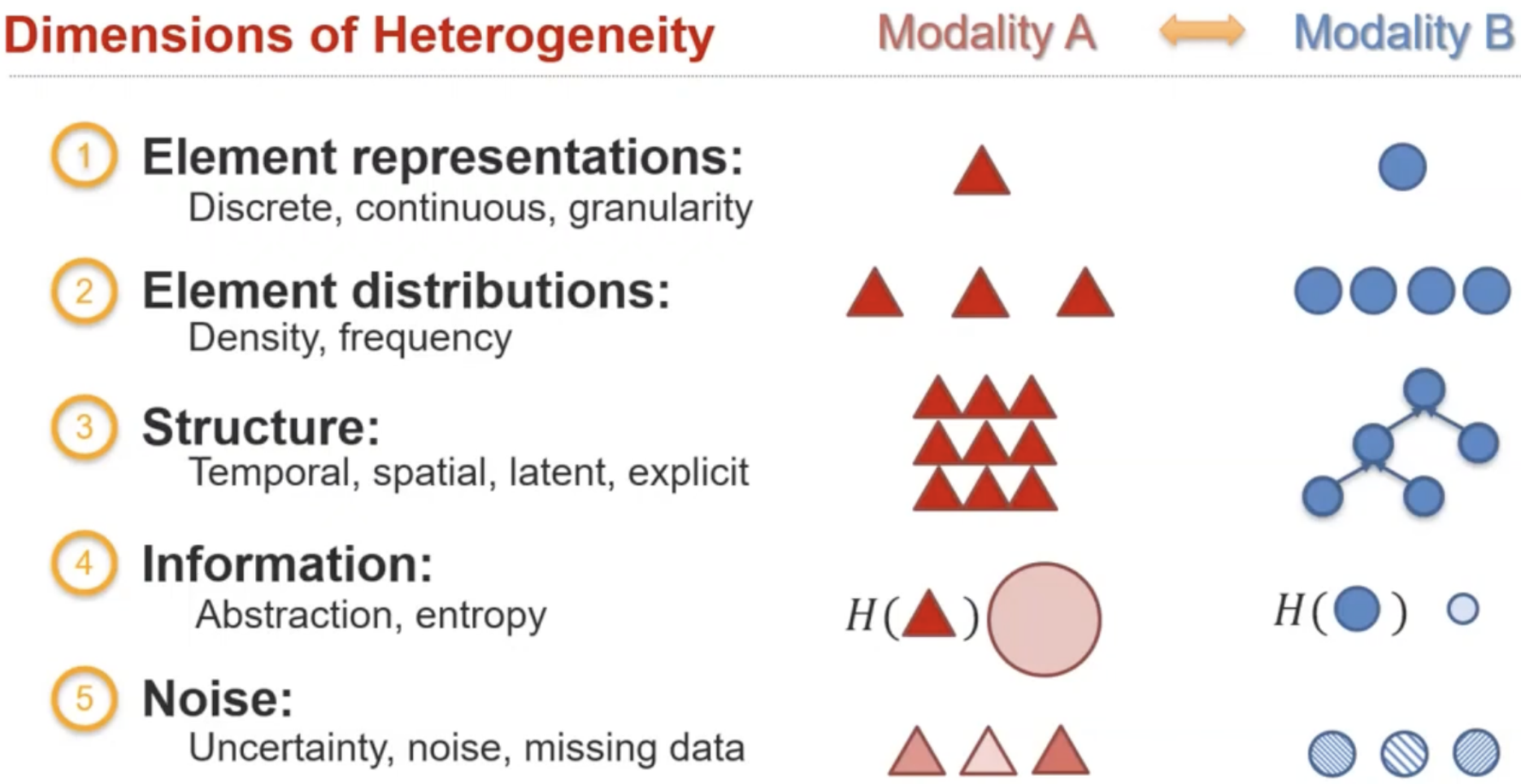


More task-universality, less human effort

Evolución de los modelos fundacionales de texto y mas



Las modalidades son heterogéneas . . . pero interconectadas

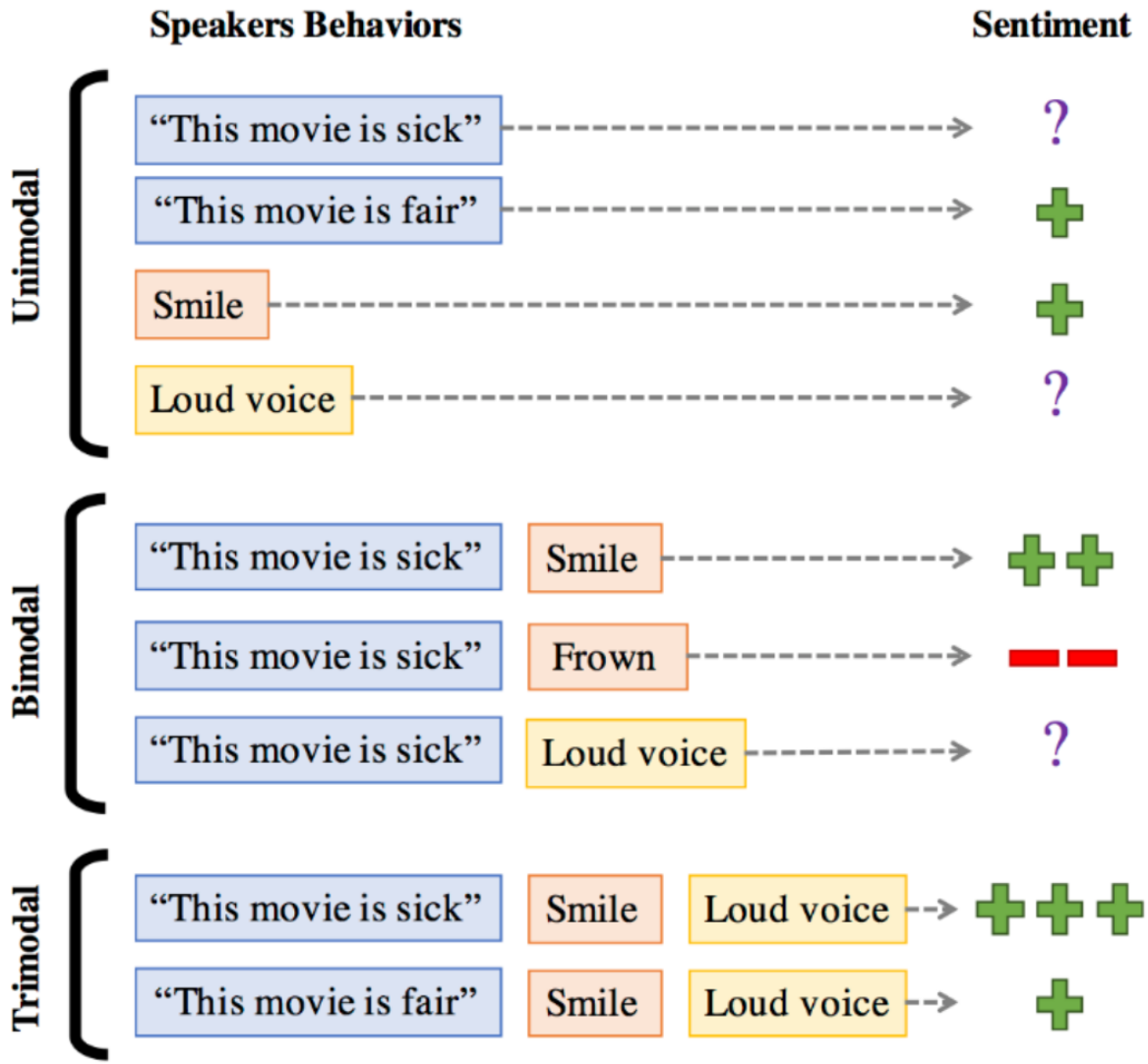


Pero también están interconectadas (Efecto McGurk):

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

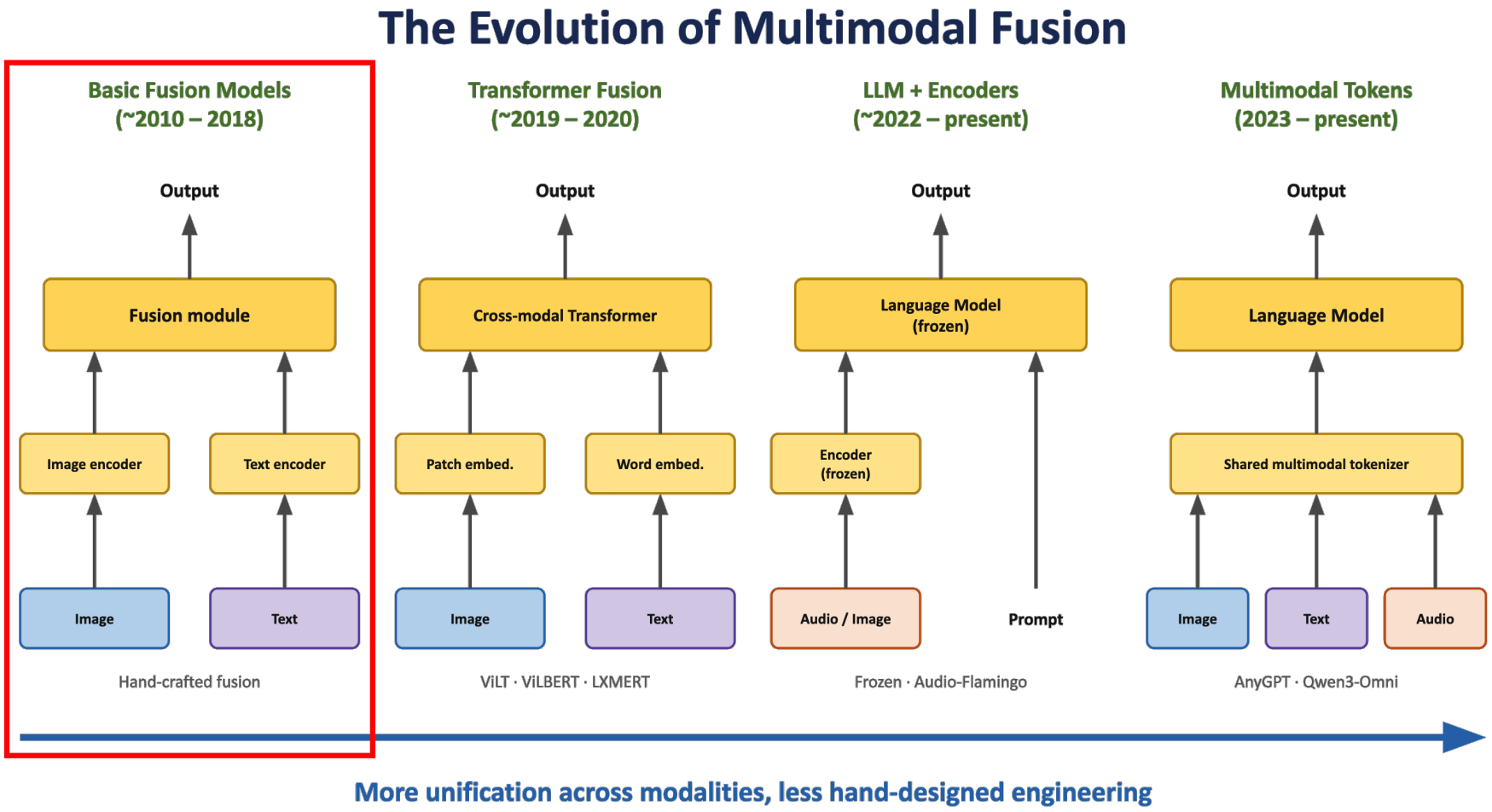
Complementariedad

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



Fusión: de Temprana/Tardía a Compuertas Aprendidas

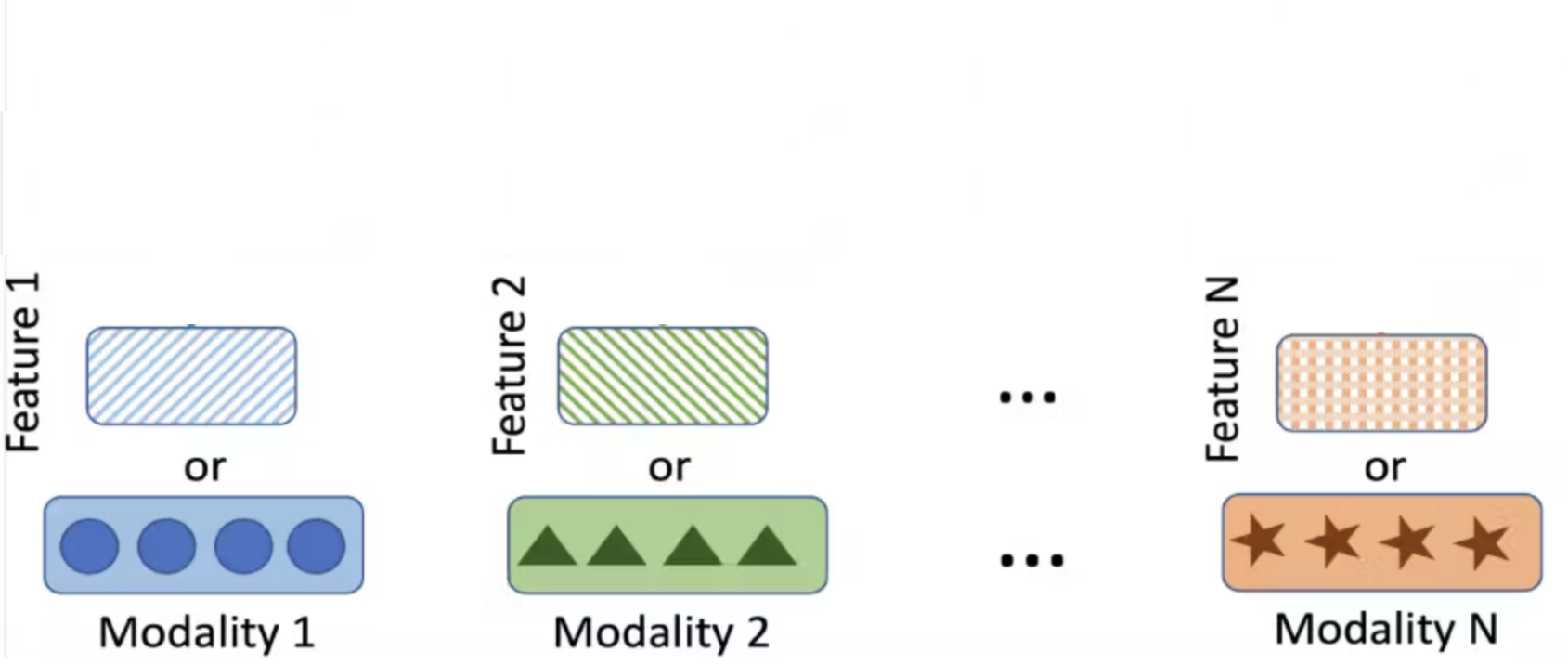
Evolución de los modelos fundacionales de fusion



Representación y Codificación

Encoding

Historicamente: extraer descriptores expertos de audio/imagen/texto.

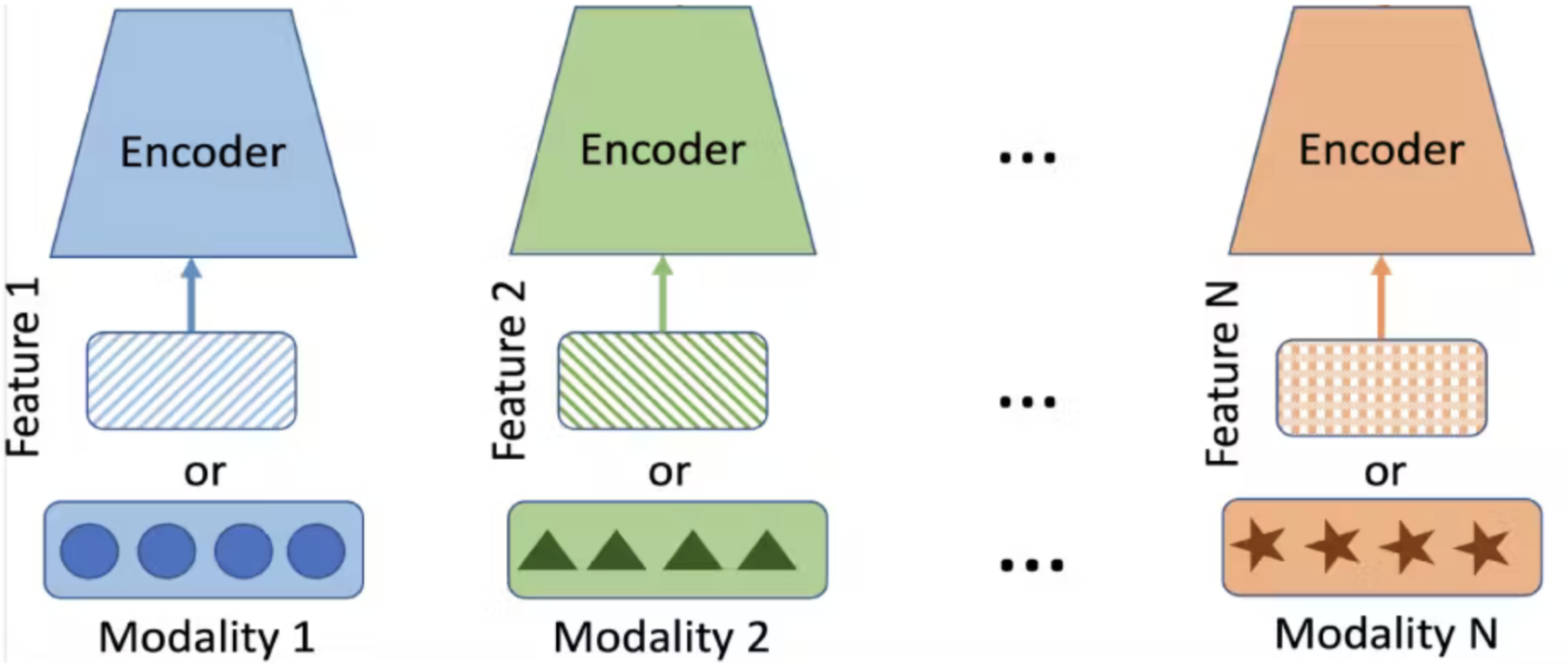


Representación y Codificación

Encoding

Historicamente: extraer descriptores expertos de audio/imagen/texto.

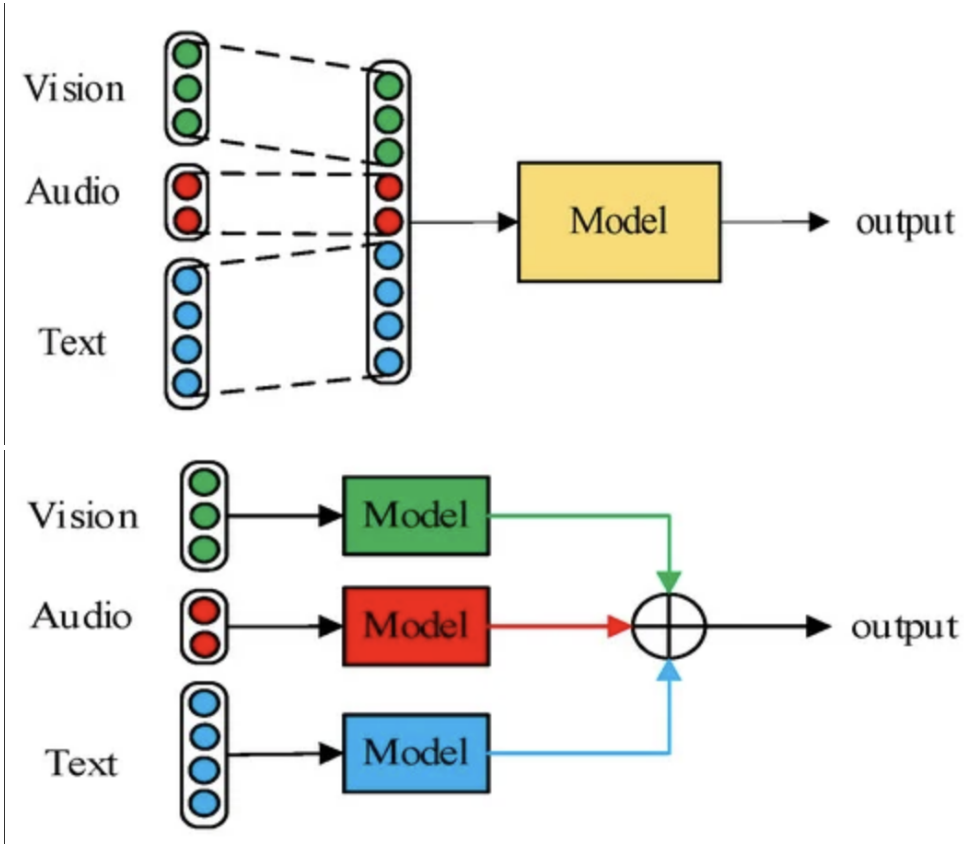
Hoy dia: a red pre-entrenado codifica cada modalidad (CNN por imagenes, word embedding por el texto, ...)



Fusions Basicas: Early vs Late

- Fusión temprana: antes del procesamiento

- Fusión tardía: después de la predicción

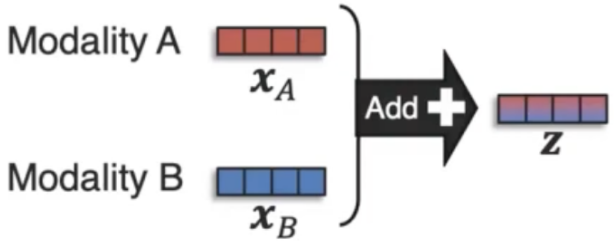


Históricamente

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: Fusión Basica

- Fusión aditiva:

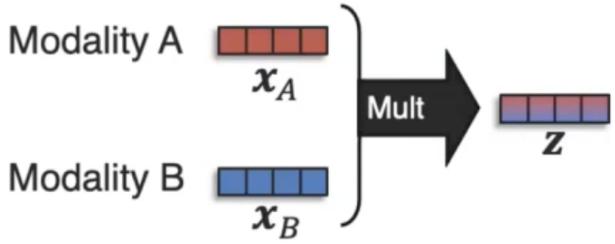


Additive fusion:

$$z = w_1 x_A + w_2 x_B$$

⇒ 1-layer neural network can be seen as additive

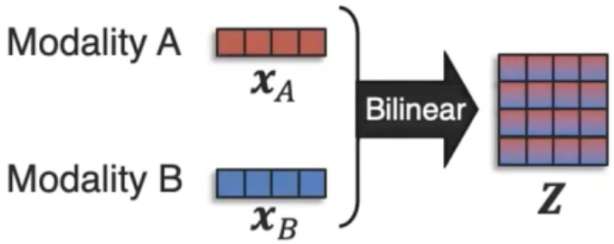
- Fusión multiplicativa:



Simple multiplicative fusion:

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

- Fusión bilineal:



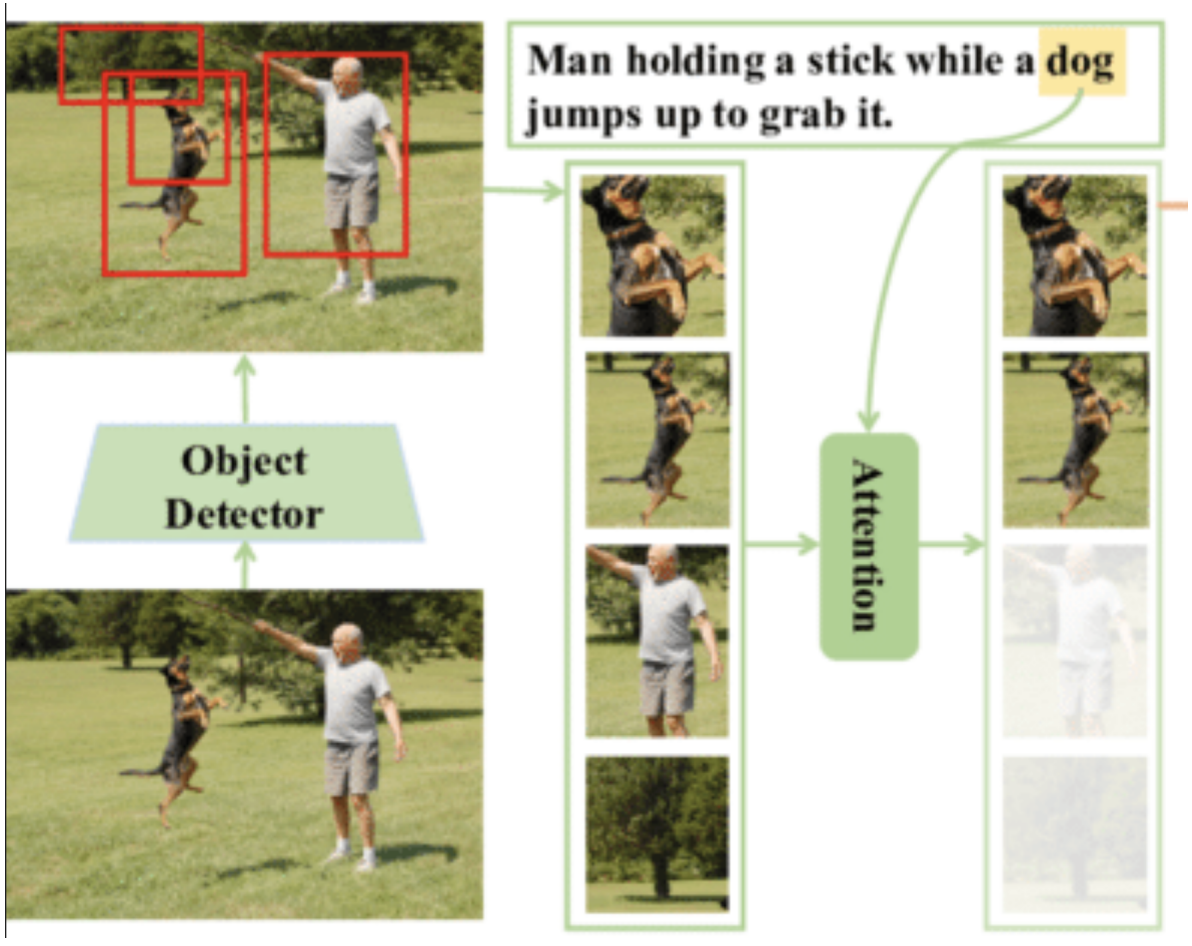
Bilinear Fusion:

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

Recuerda: ¡Los datos ya tienen las interacciones en sí, el modelo solo intenta aprenderlas! More complex fusion modules exist (tensor fusion),

Cross-modal Attention & Gating

Attention-like gating mechanisms, and in particular cross-modal attention, let a model focus on the right part of each modality when looking for something specific.



Cross-modal Attention & Gating

Attention-like gating mechanisms, and in particular cross-modal attention, let a model focus on the right part of each modality when looking for something specific.

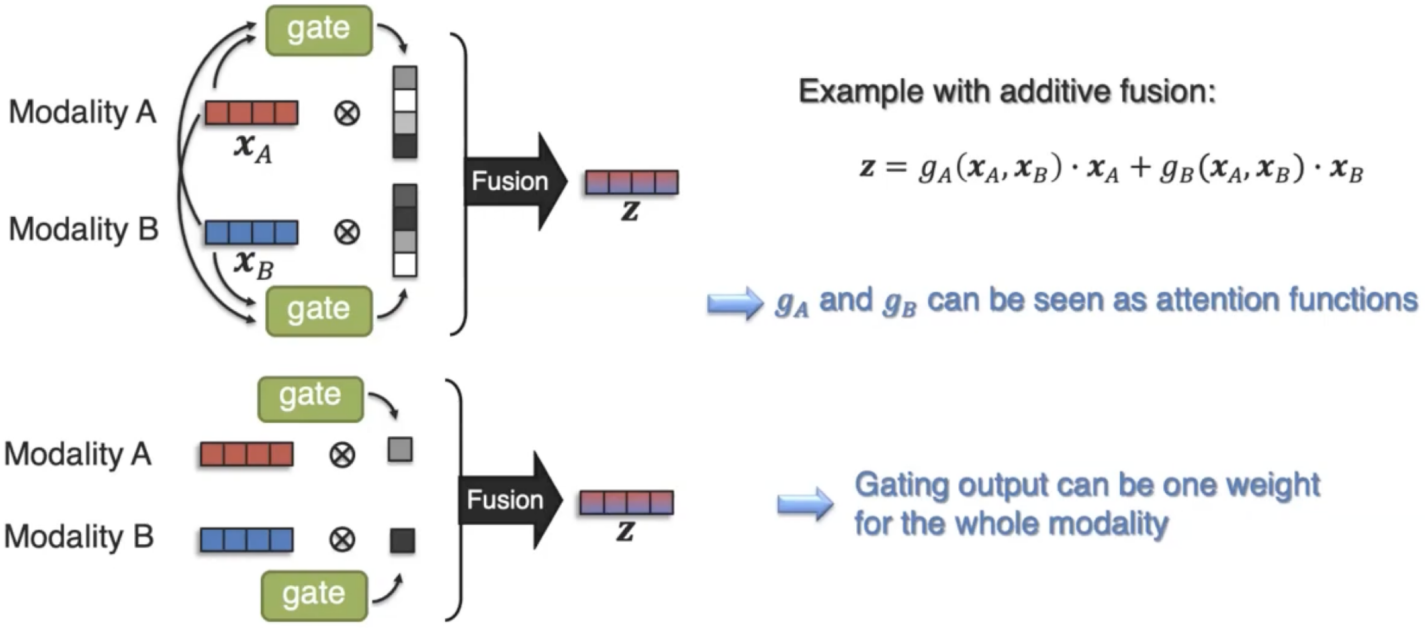


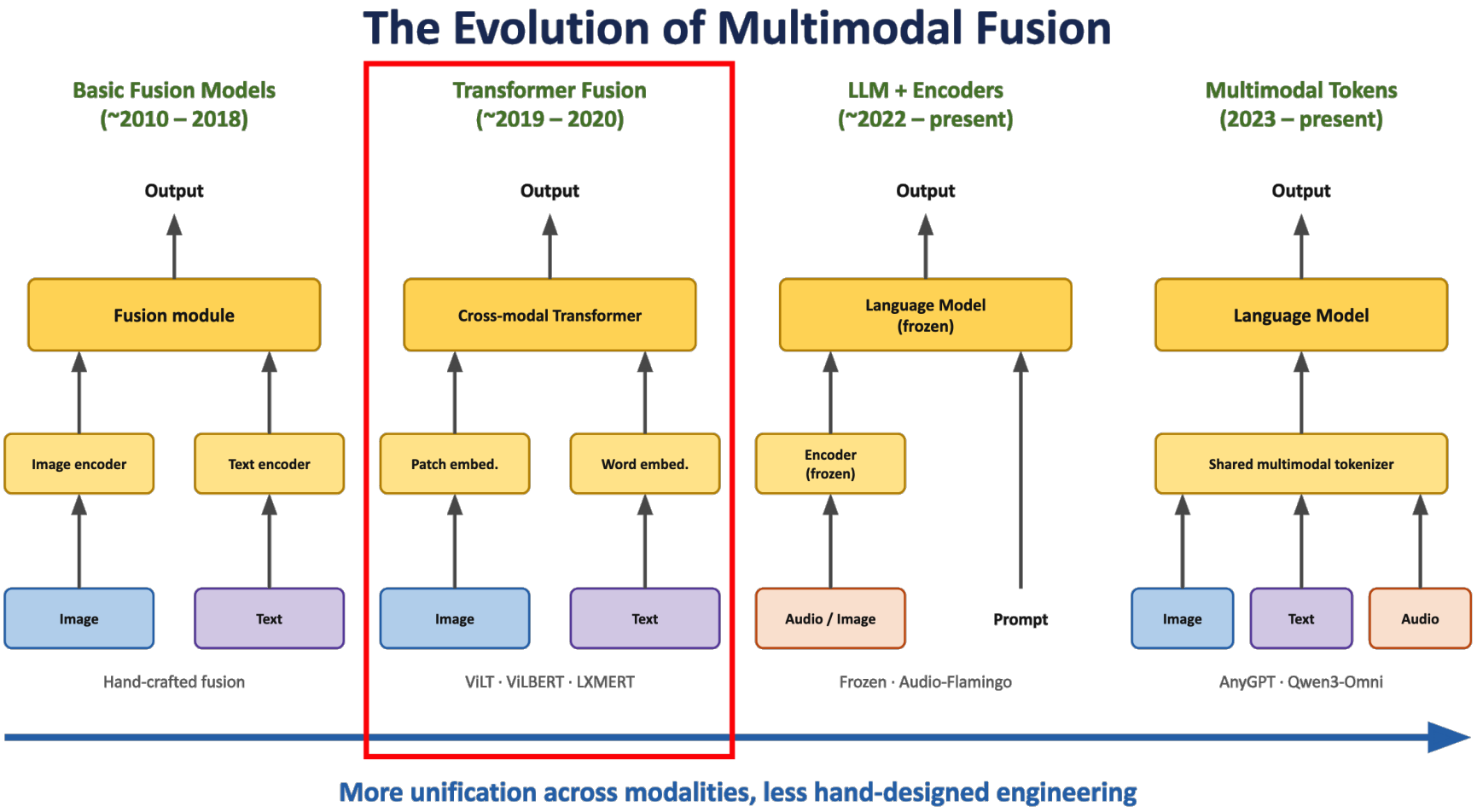
Figure 2: Cross-modal gating attention [2]

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*)

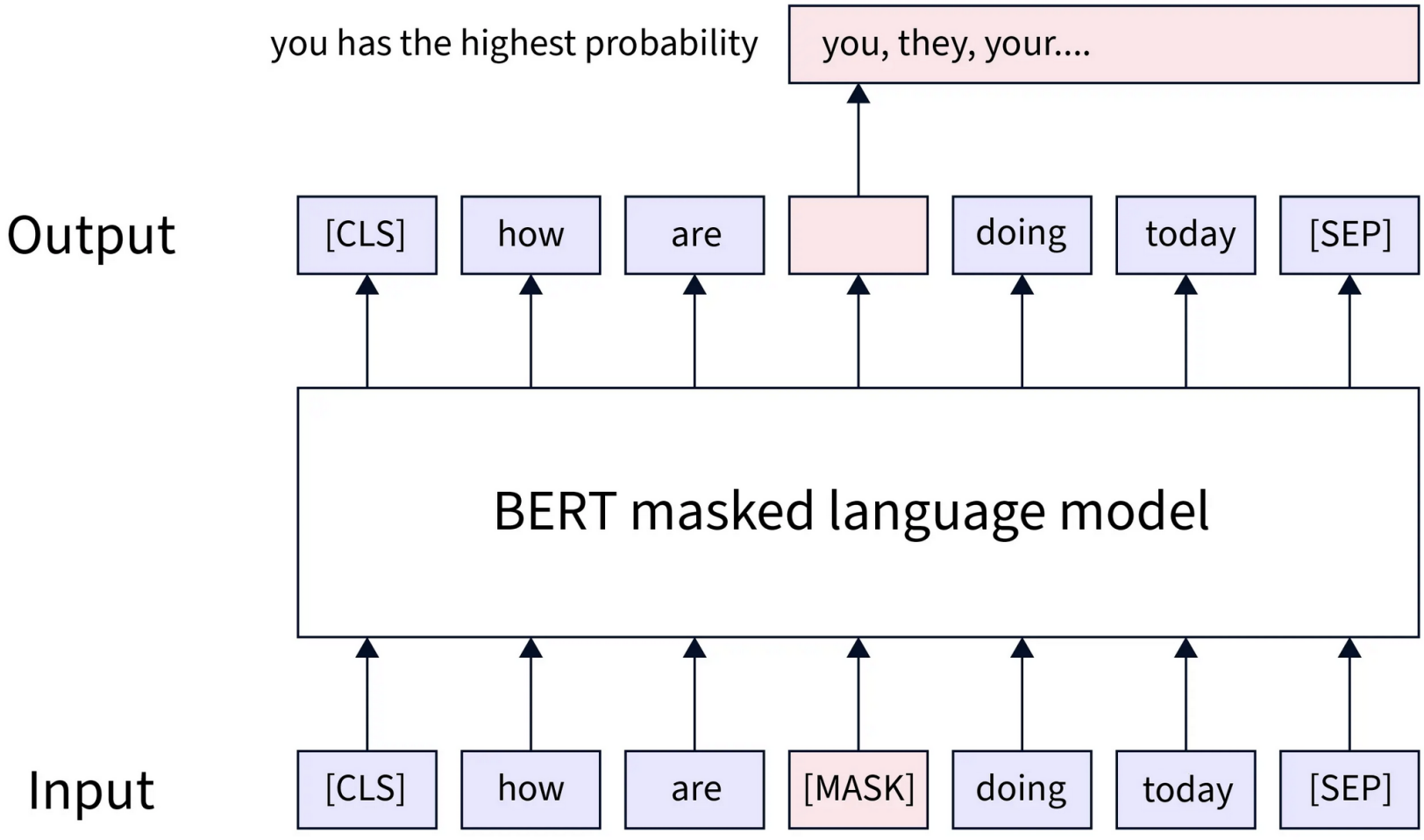
Fusion basada en transformers

Evolución de los modelos fundacionales de fusion



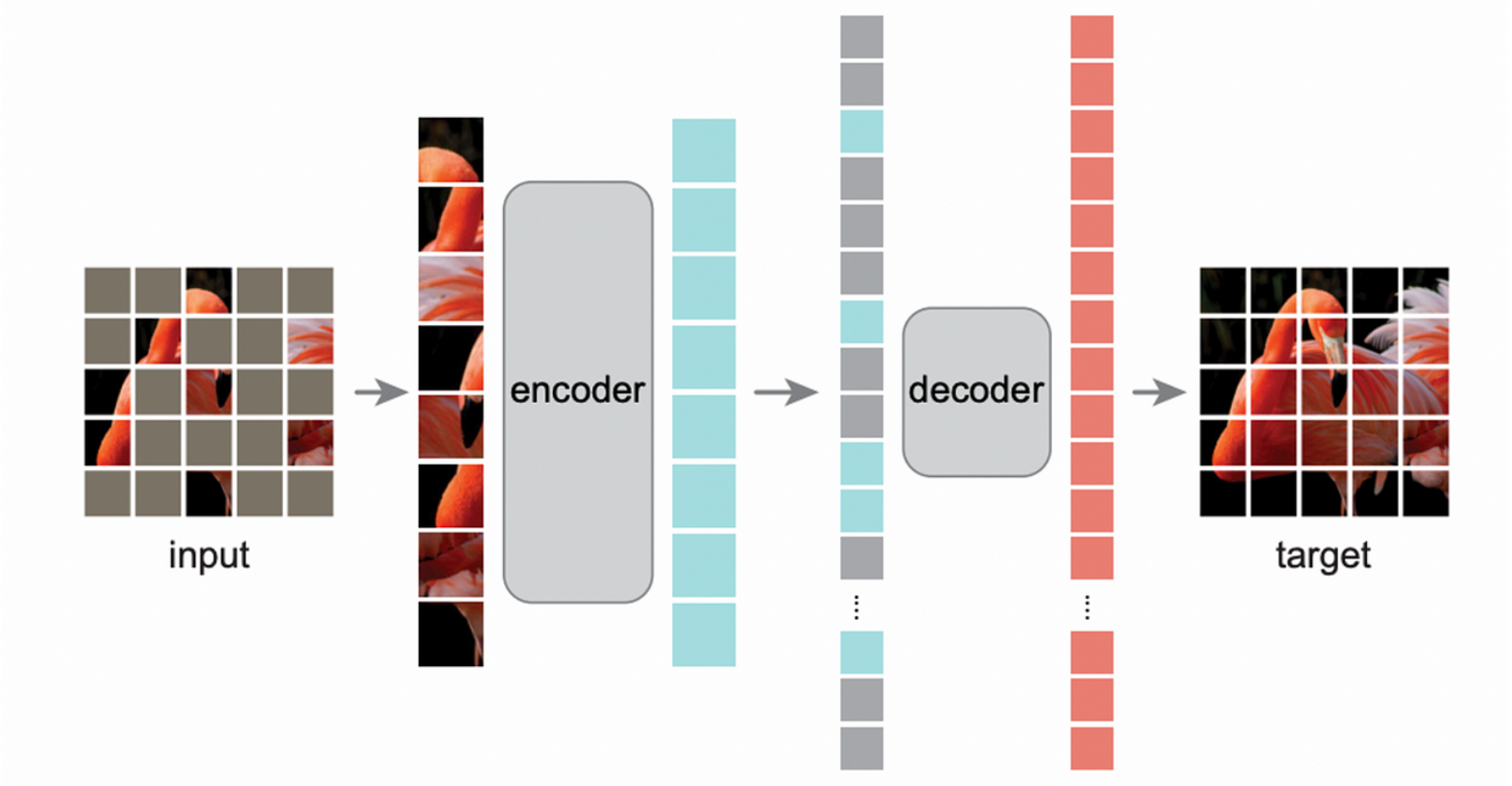
Fusión usando Transformers

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



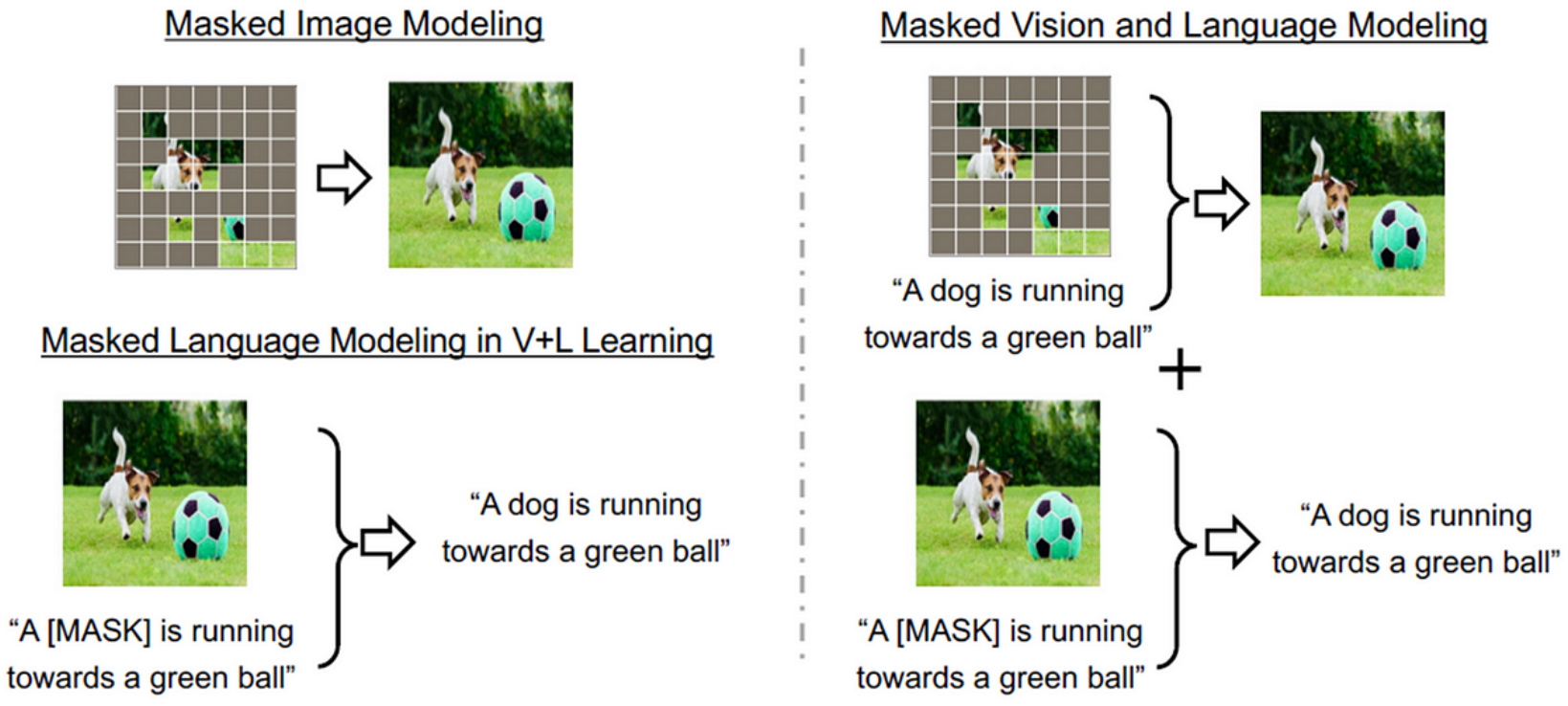
Fusión usando Transformers

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



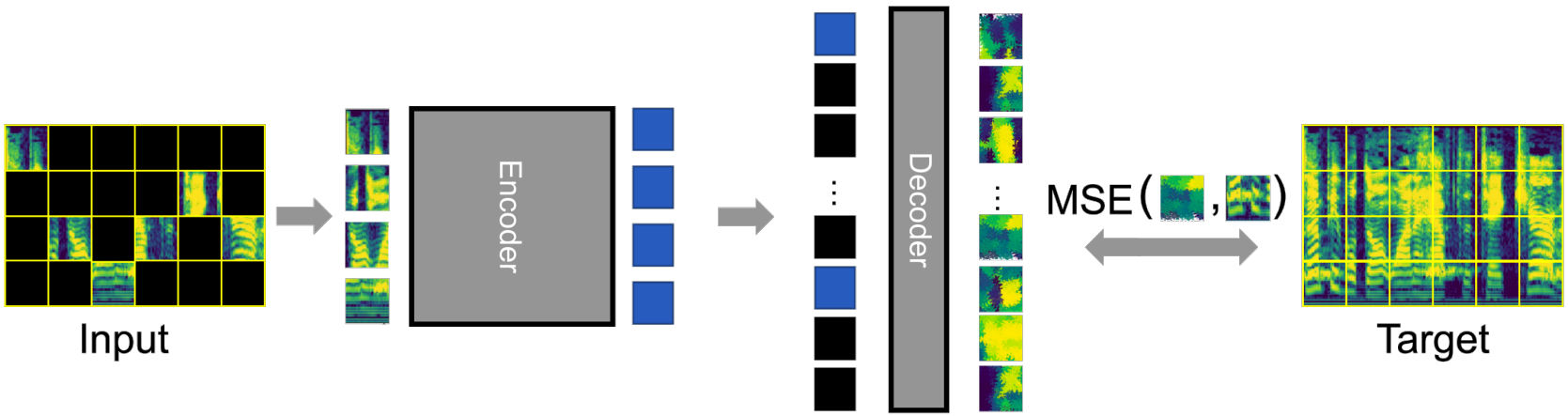
Fusión usando Transformers

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

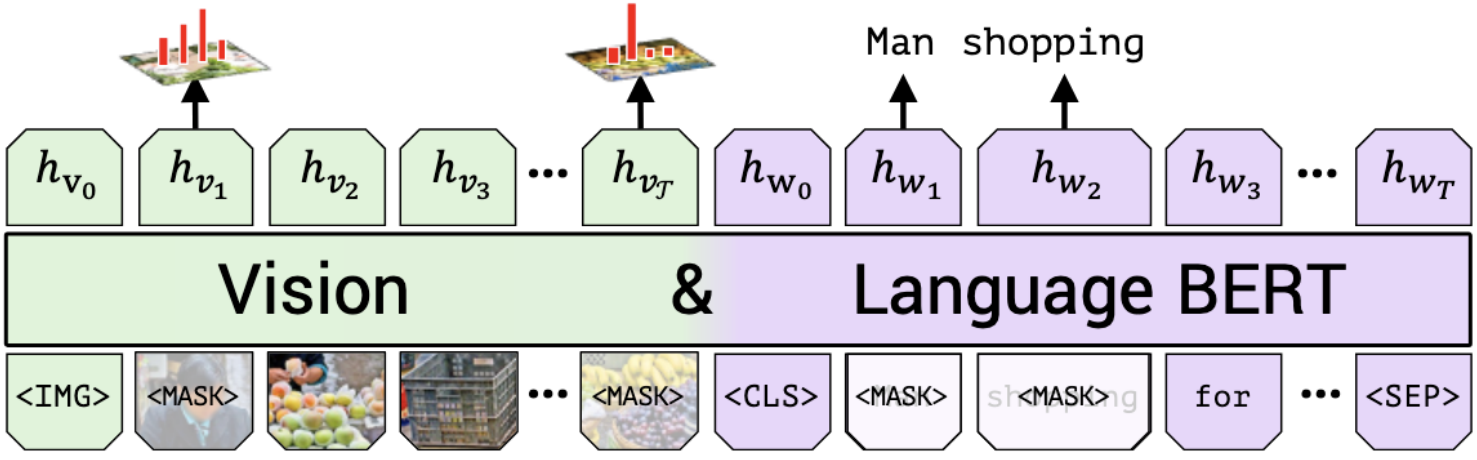


Fusión usando Transformers

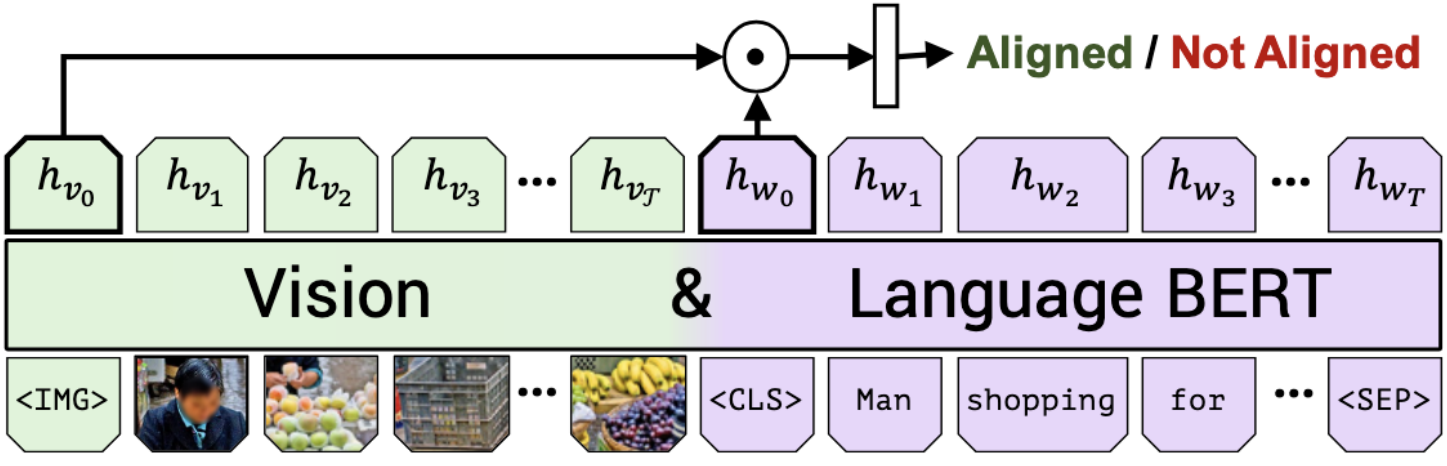
Al final, la arquitectura de transformers funciona bien para fusionar las diferentes modalidades, incluso si sus datos en bruto son muy diferentes.



First LMMs: ViLBERT



(a) Masked multi-modal learning



(b) Multi-modal alignment prediction

First LMMs: ViLBERT

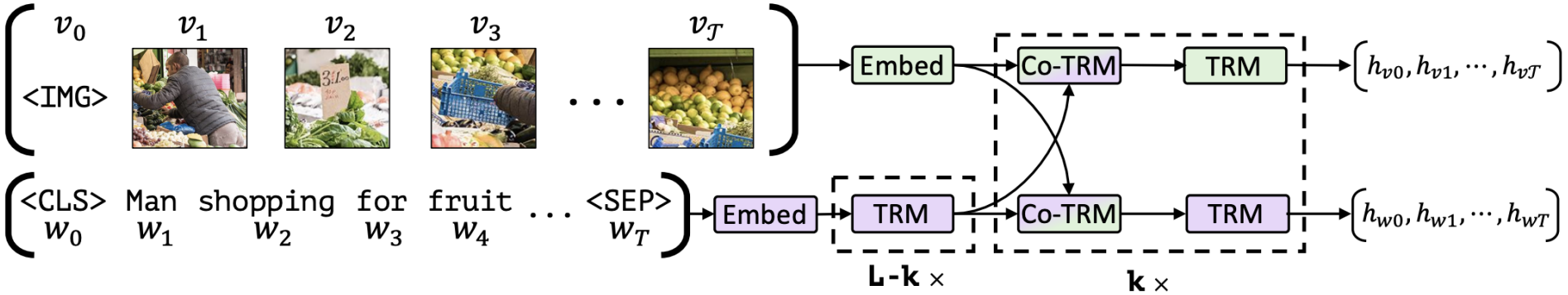
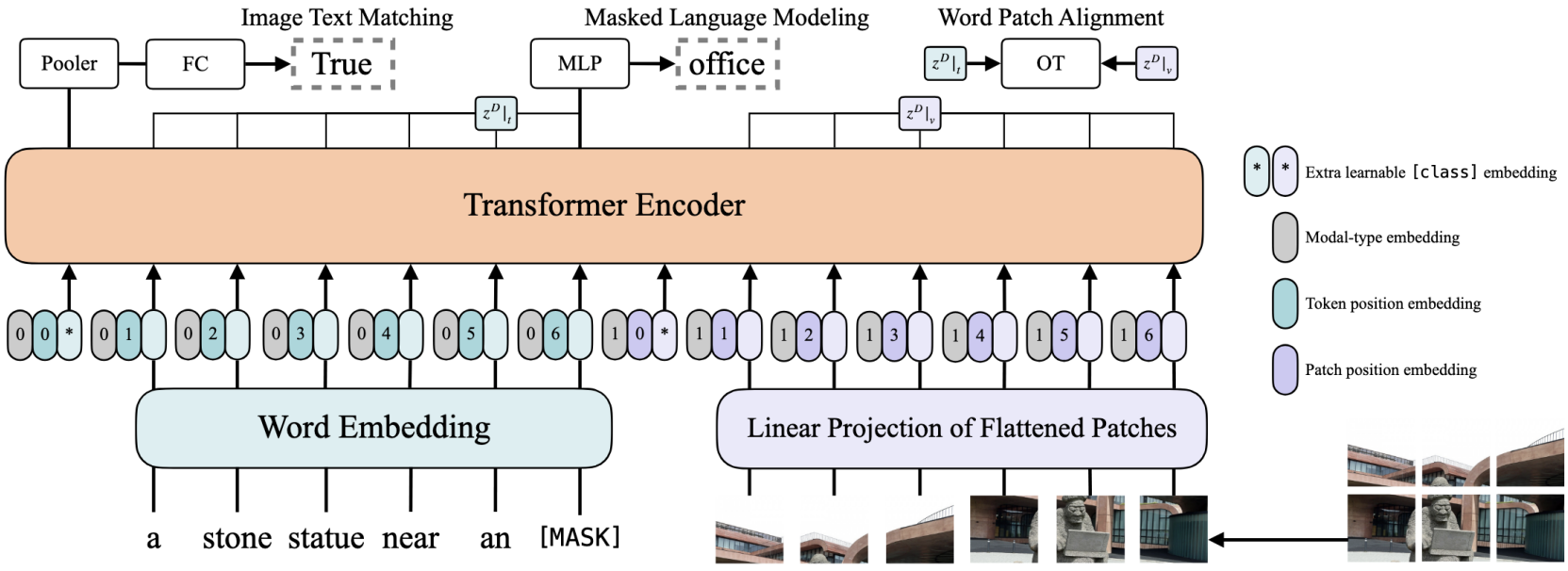


Figure 1: Our ViLBERT model consists of two parallel streams for visual (green) and linguistic (purple) processing that interact through novel co-attentional transformer layers. This structure allows for variable depths for each modality and enables sparse interaction through co-attention. Dashed boxes 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. [18]

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! [10]



Usa tres pérdidas, similar a lo que vimos antes.

Cuantificando Interacciones: Descomposición de Información Parcial

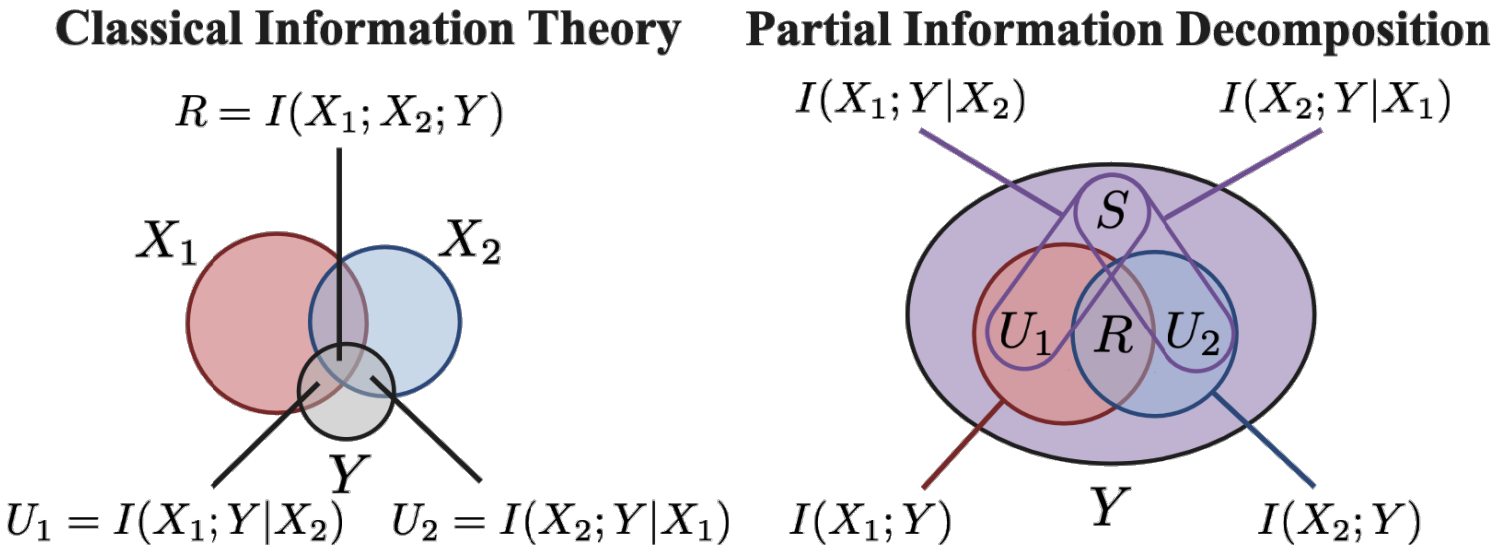
¿Cuánto comparten realmente las modalidades?

Acabamos de ver cómo fusionar modalidades – pero ¿cómo sabemos cuánta interacción hay para explotar, en primer lugar?

Descomposición de Información Parcial (PID)

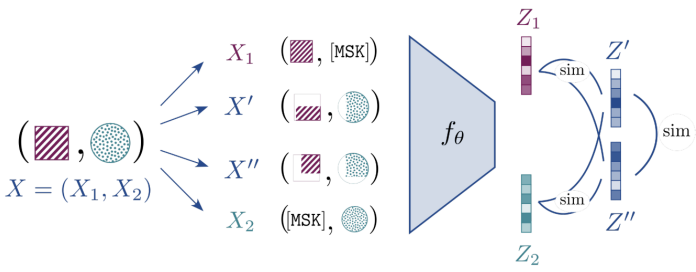
Descompone la información que dos modalidades X_1, X_2 dan sobre una tarea Y en cuatro componentes no negativos [12]: $I = R + S + U_1 + U_2$

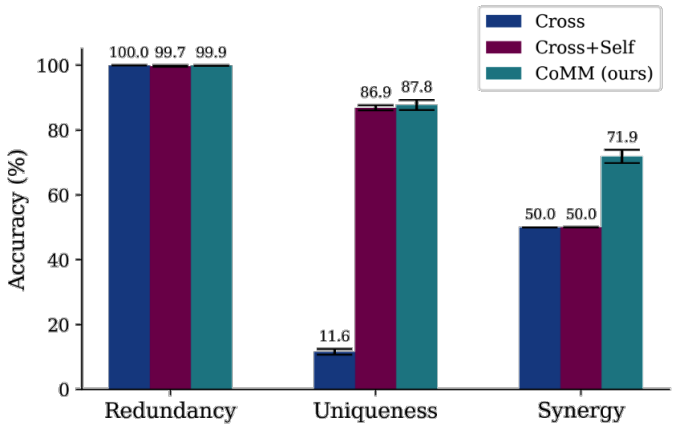
- **Redundancia** R: información disponible desde cualquiera modalidad
- **Unicidad** U_1, U_2: información disponible desde una sola modalidad
- **Sinergia** S: información emergente si se leen de forma conjunta



Lo que PID nos dice en la práctica

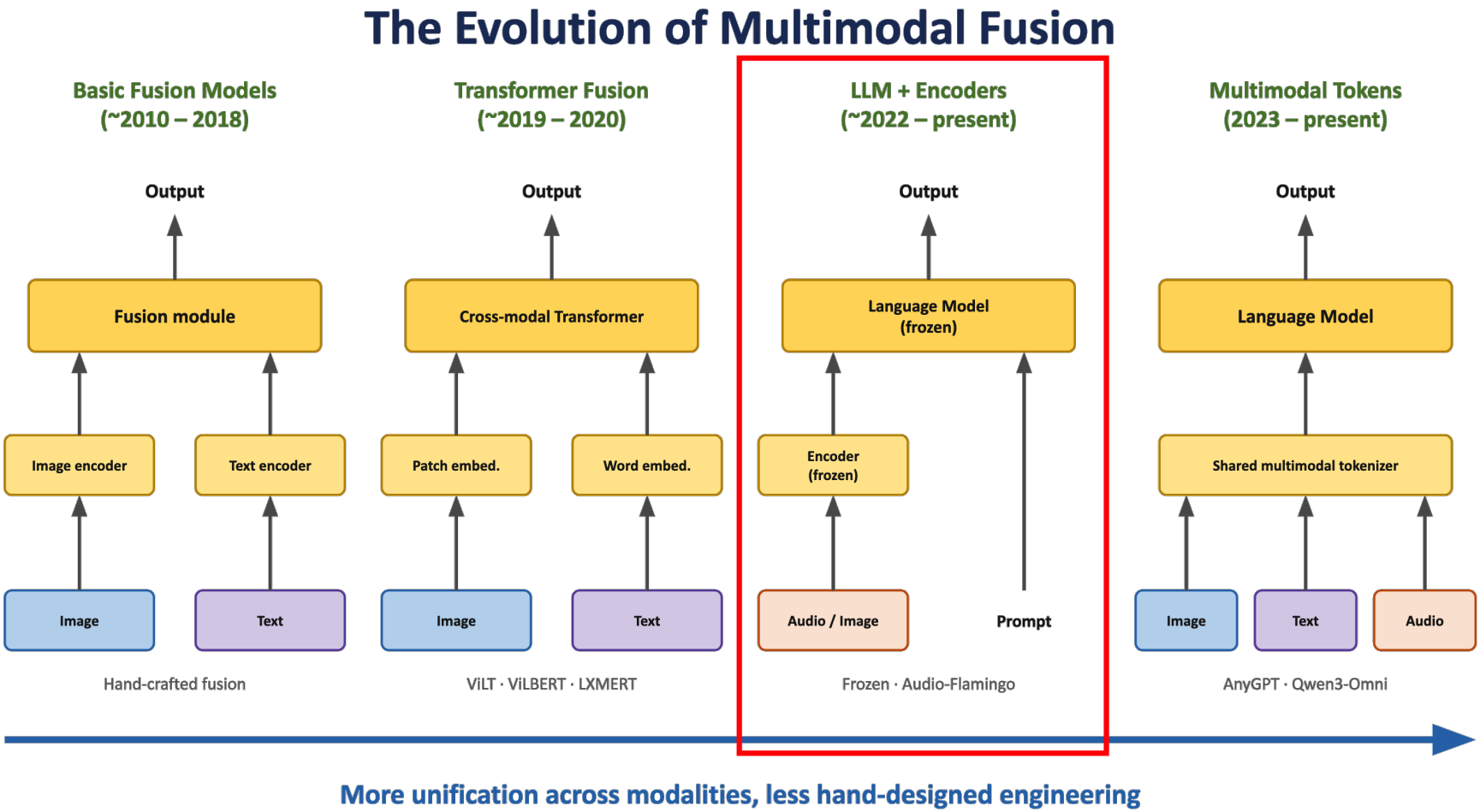
- Ahora existen estimadores escalables para representaciones continuas de alta dimensión [12]
- Empíricamente: la mayoría de los métodos de fusión actuales son mucho mejores explotando la redundancia que la sinergia: CLIP optimiza la Información Mutua [21]
- CoMM [**?**] mejora CLIP, permitiendo modelar mejor la sinergia





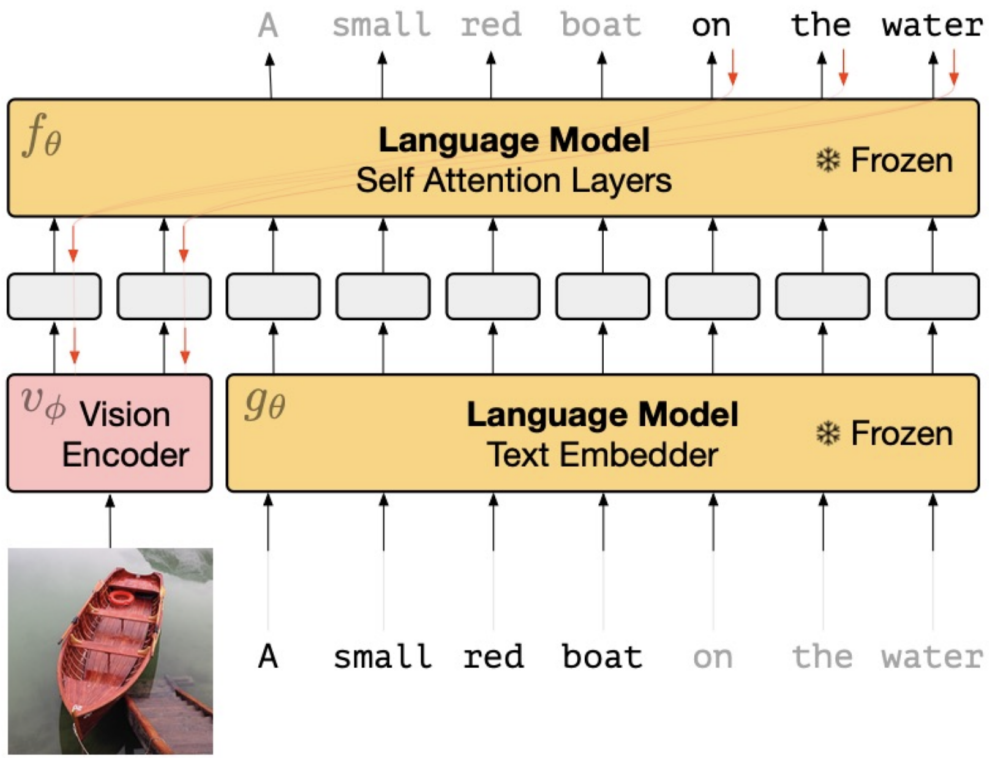
Generalistas Multimodales Centrados en LLM

Evolución de los modelos fundacionales de fusion



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



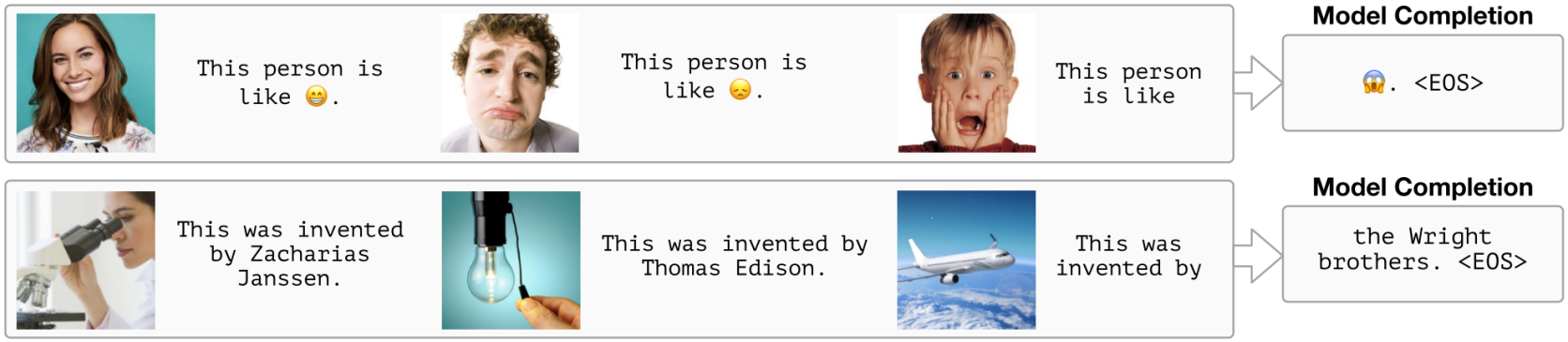


Figure 3: Ejemplos few-shot del modelo llamado Frozen [26]

Basados en LLM: Flamingo

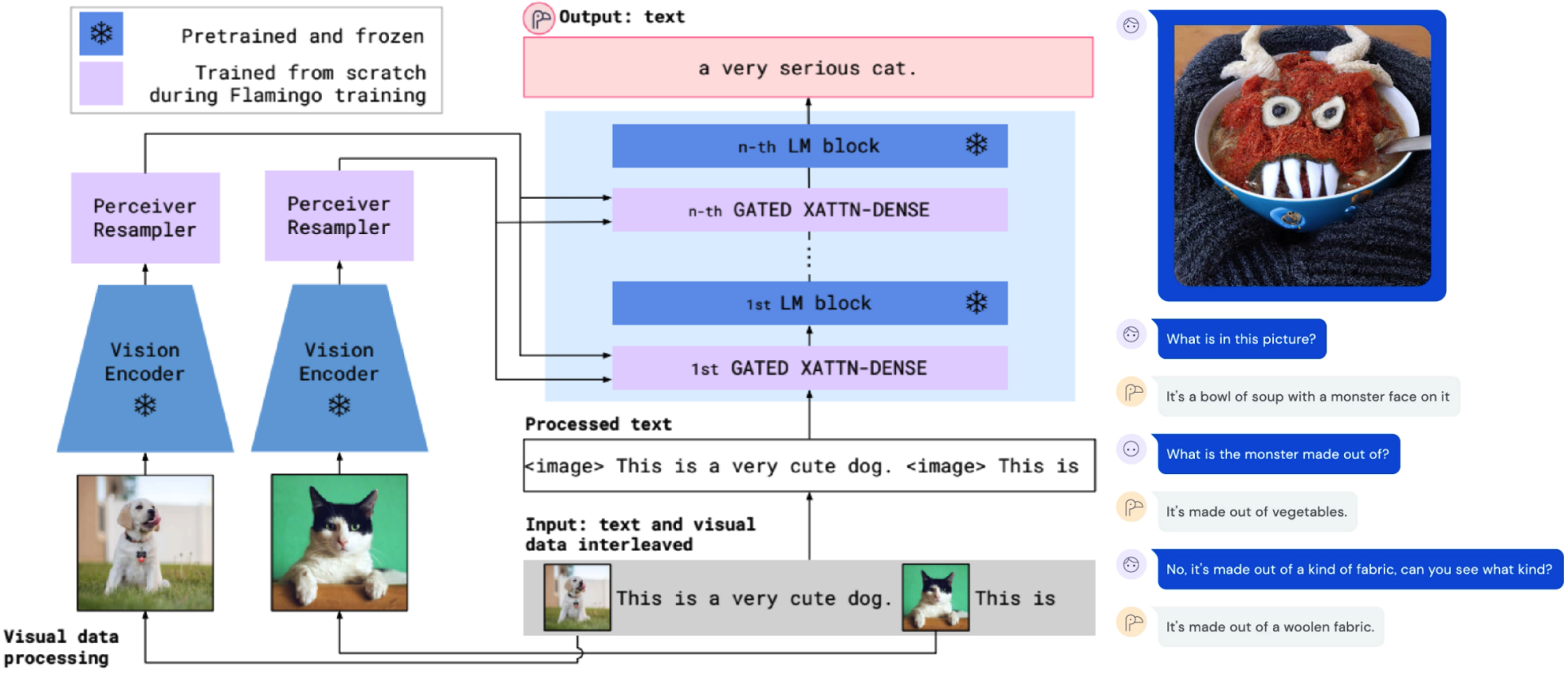
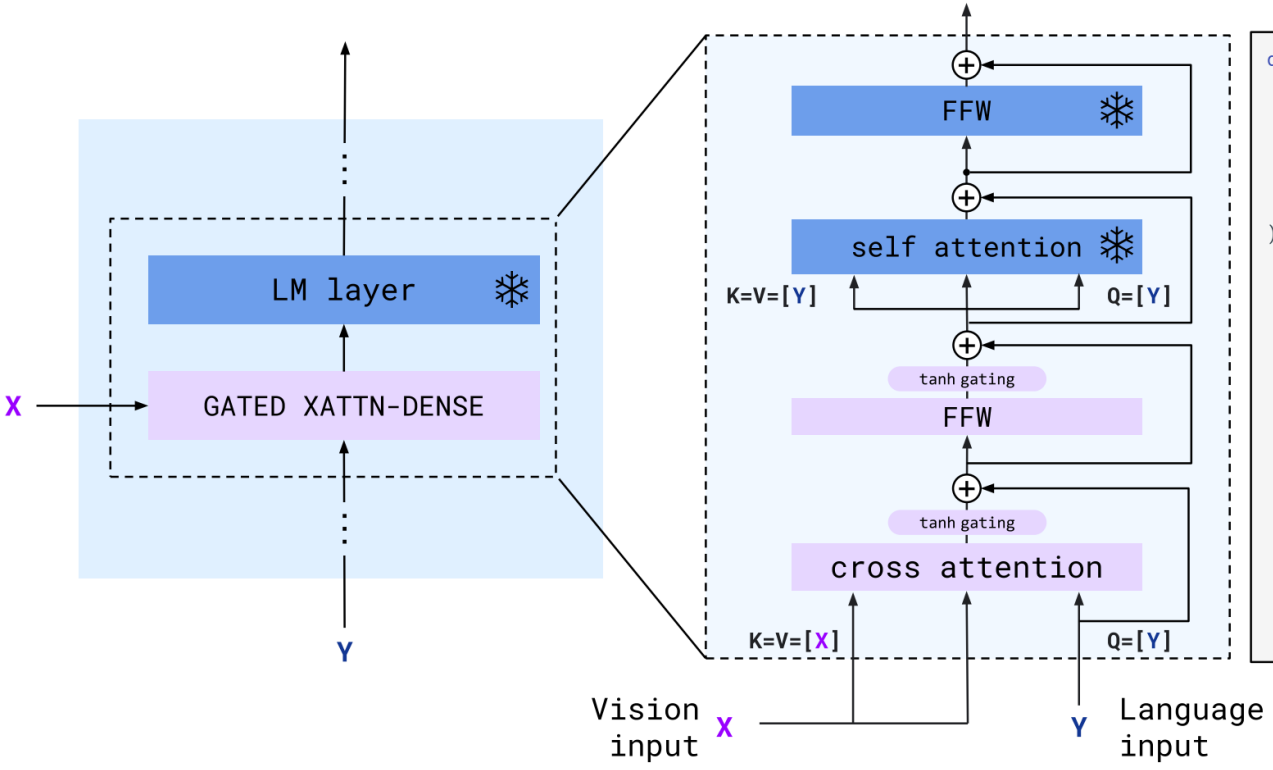


Figure 4: Flamingo fue el primer LMM que puede conversar con un humano [1]

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

Basados en LLM: Flamingo II




```python
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 5: 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: xGen-MM (BLIP3) I

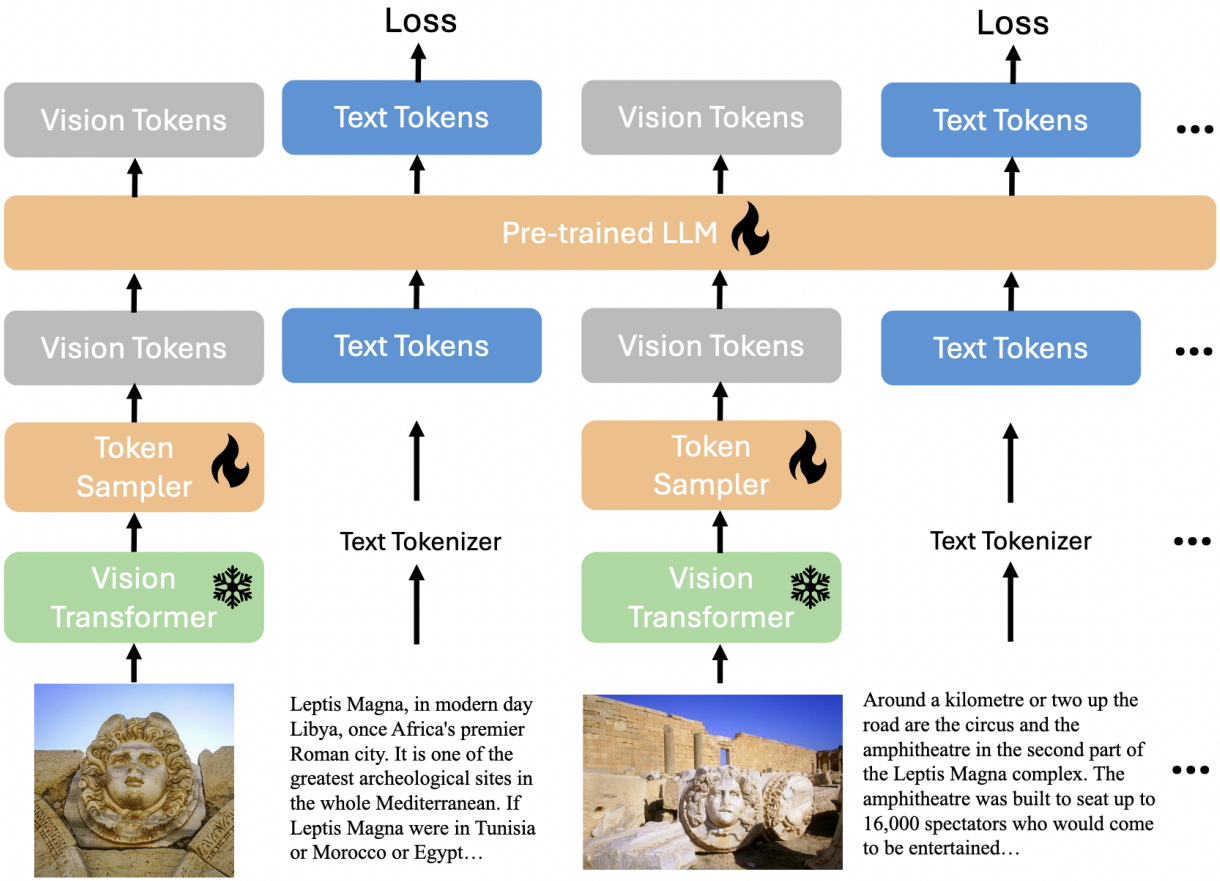


Figure 2: **Overview of the xGen-MM (BLIP-3) framework.** Free-form interleaved images and texts from the ensembled interleaved and caption datasets are input into the framework, with each modality undergoing a separate tokenization process to be fed into the pre-trained LLM in natural order. A standard 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 6: 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

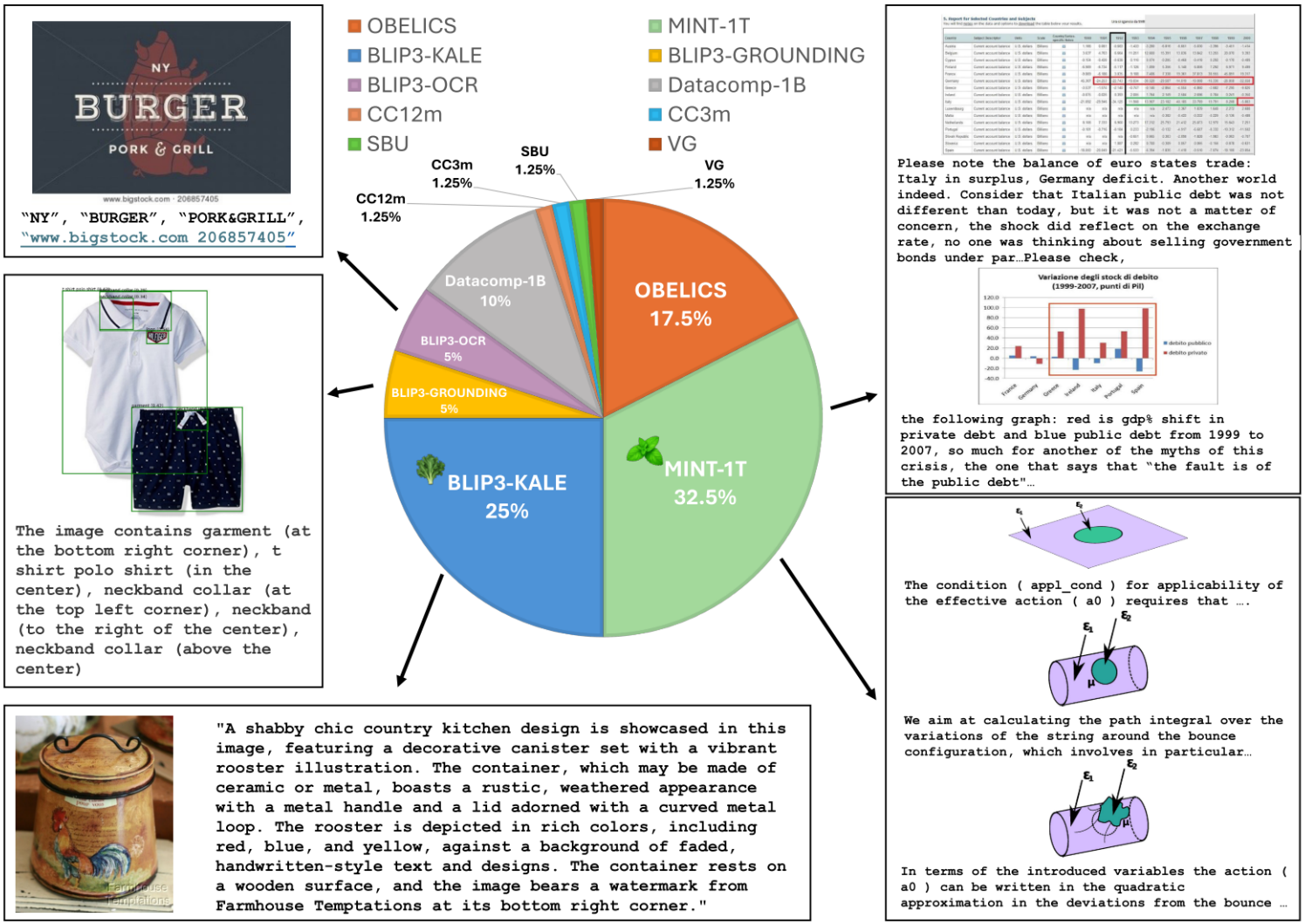
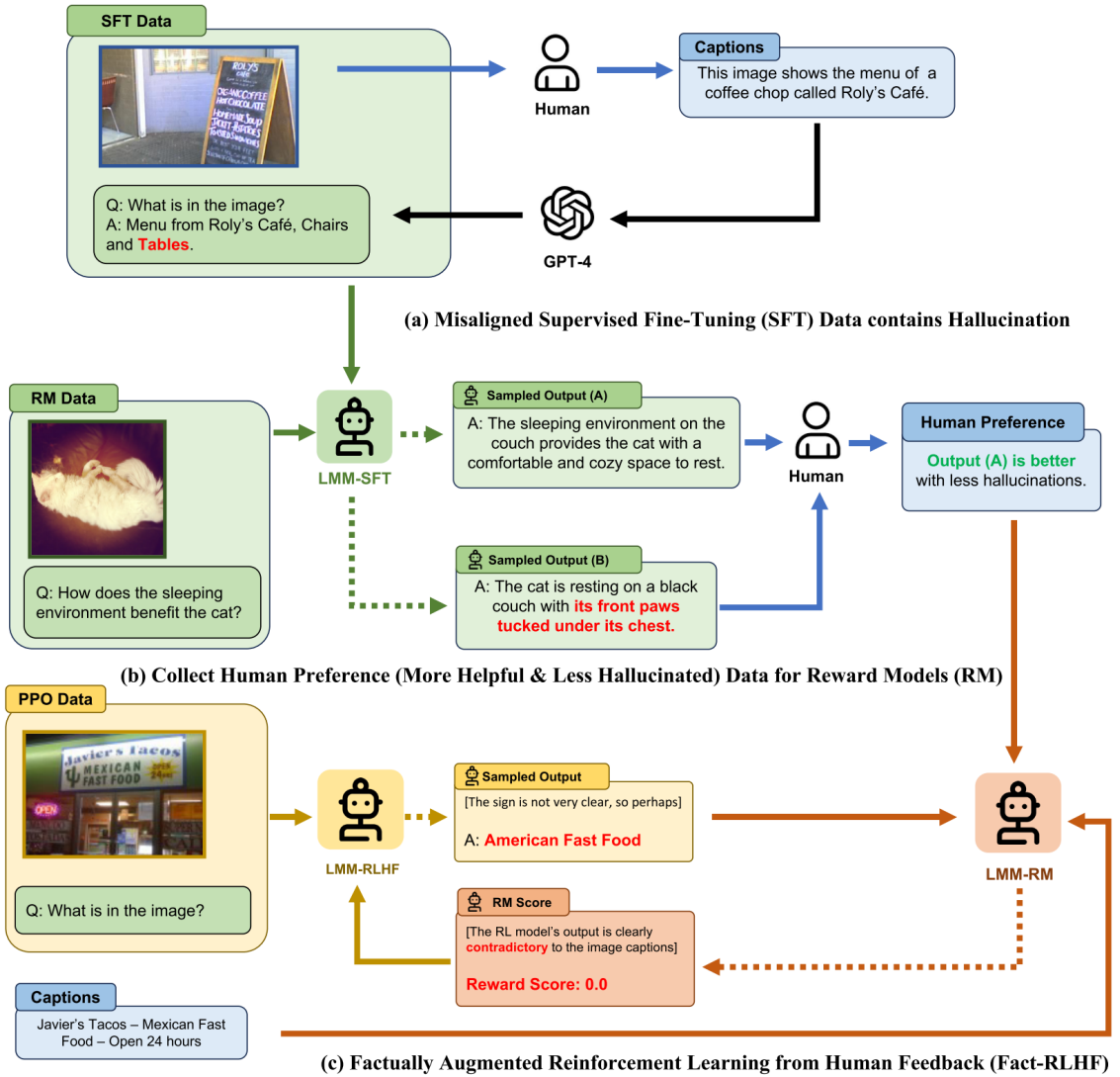


Figure 7: El conjunto de entrenamiento de BLIP3 contiene datos multimodales entrelazados (mezcla de secuencias de imágenes y texto) de contenido muy diverso

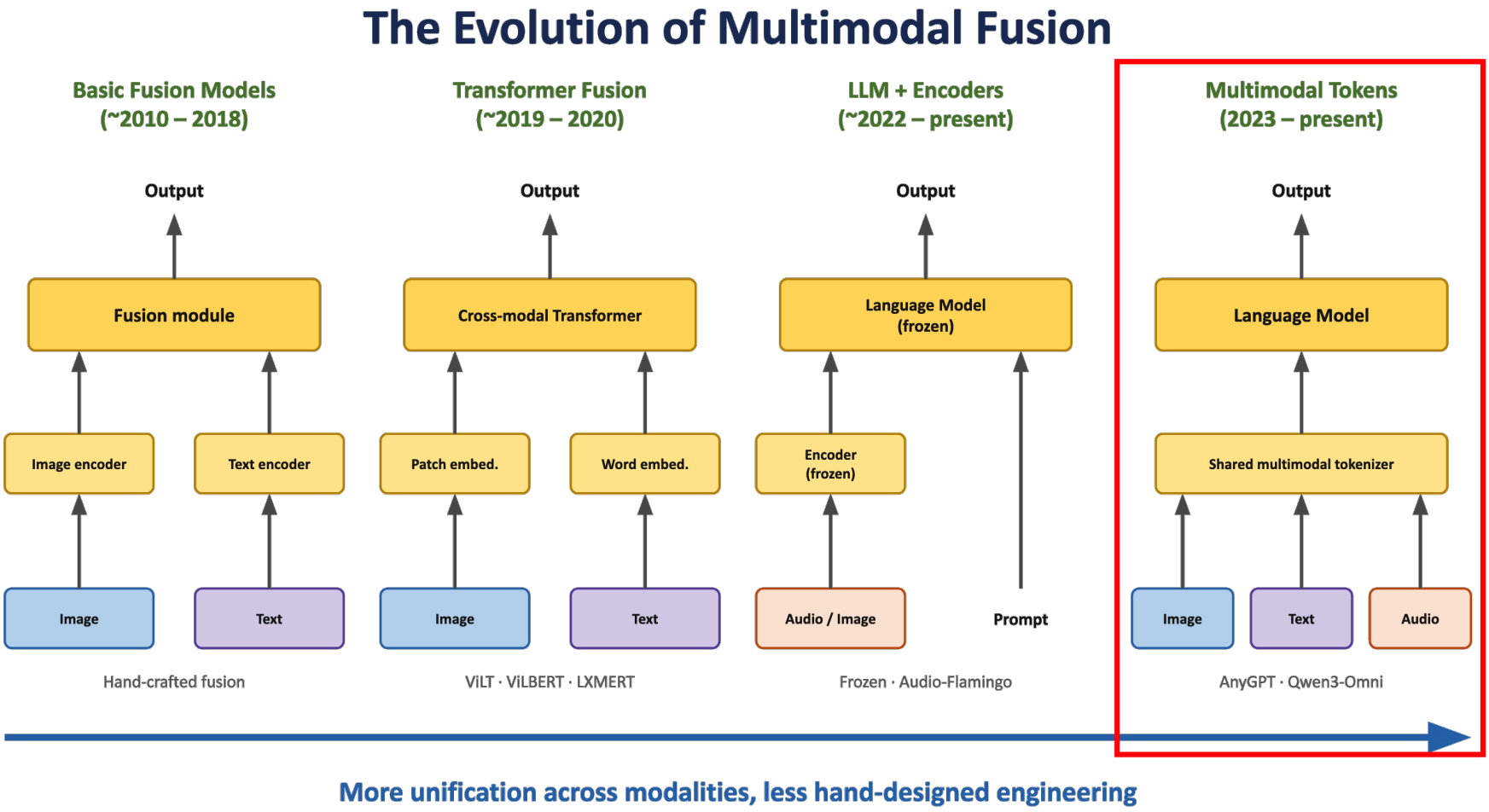
Basados en LLM: Large Language and Vision Assistant (LLaVA)

- Modelo open-source basado en LLaMA, adaptándolo a datos multimodales [16]
- 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]



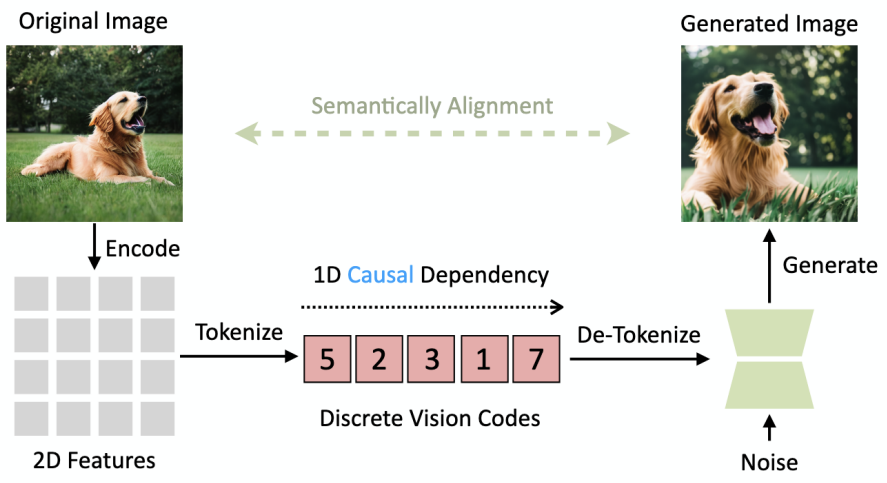
Tokenizacion

Evolución de los modelos fundacionales de fusion

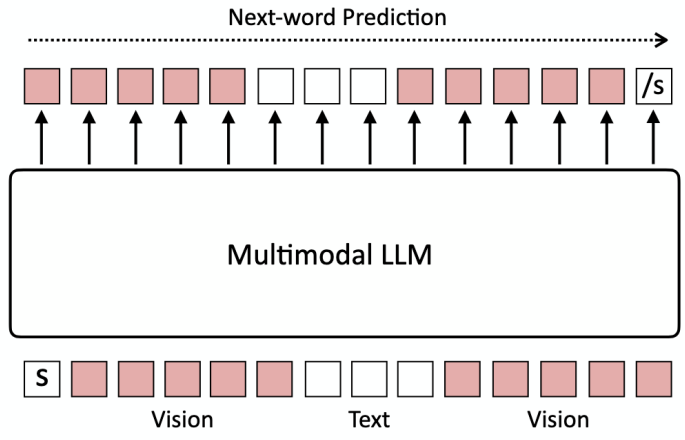


Hoy, todo es un token discreto!

Tokenes de Vision [5]:

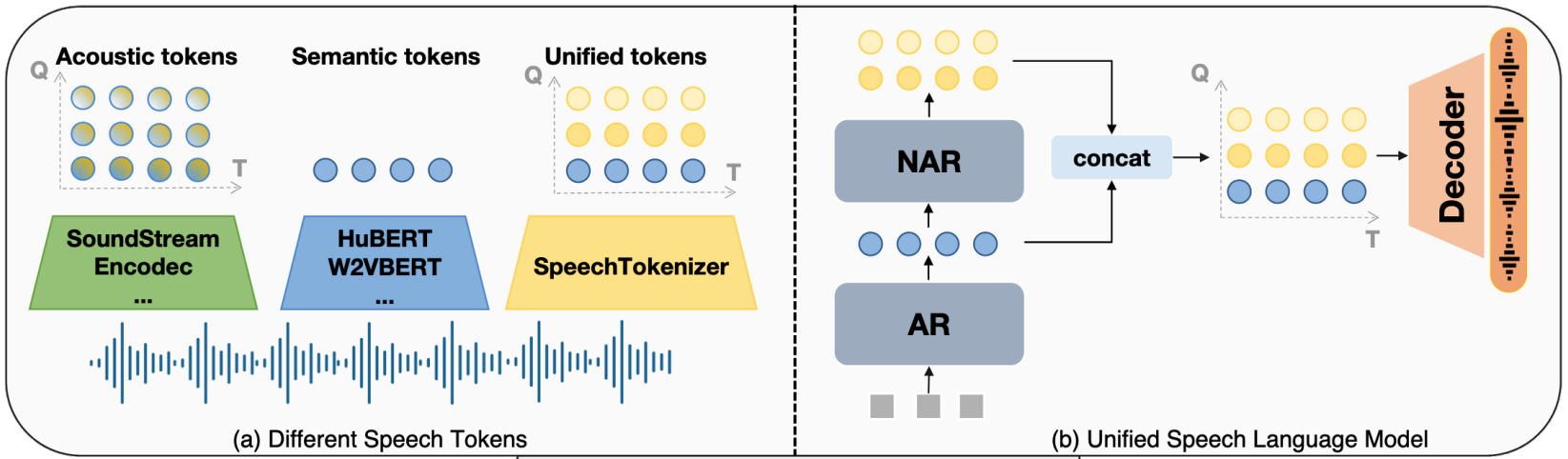


(a) SEED Visual Tokenizer



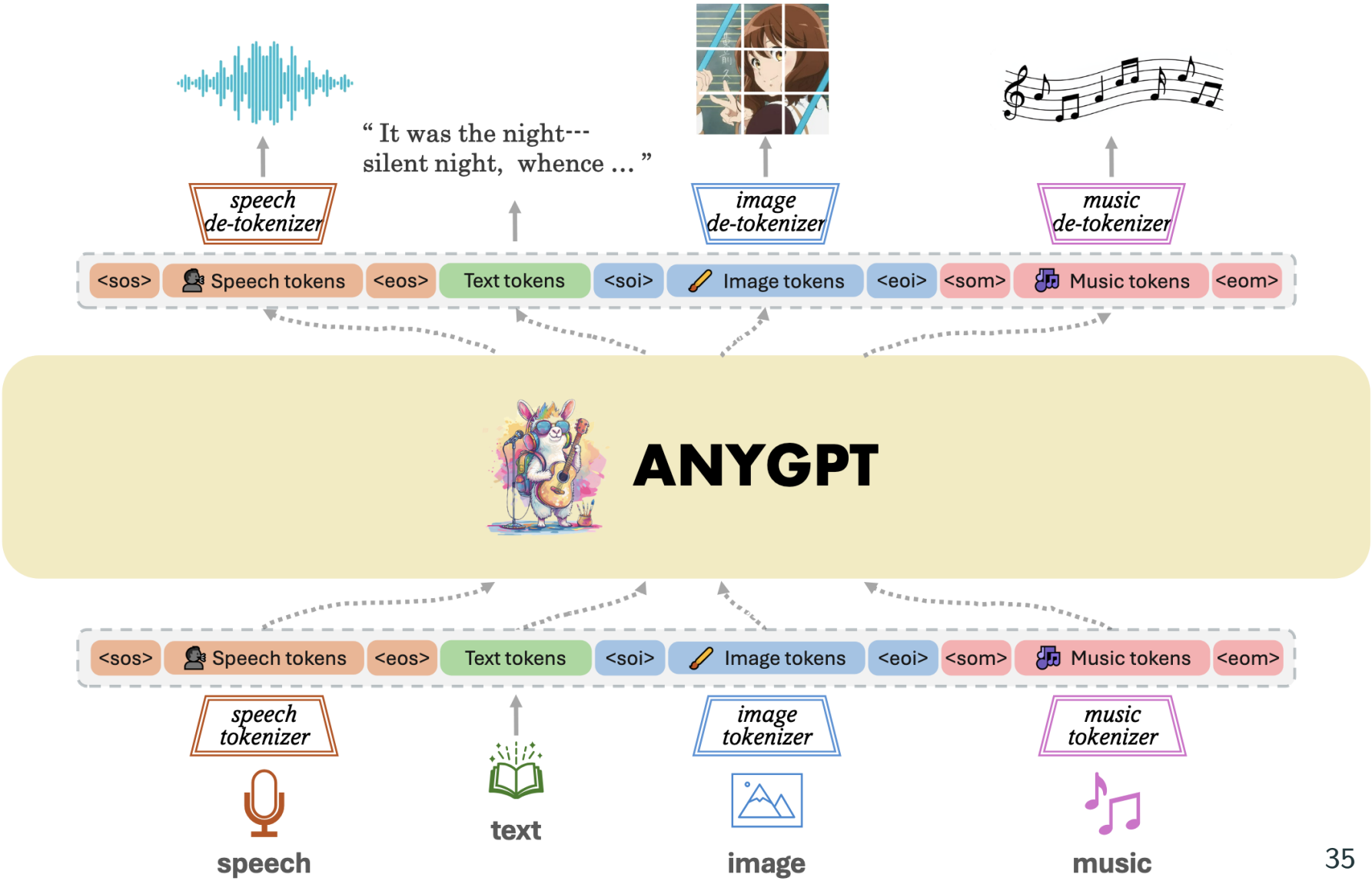
(b) Multimodal Autoregression with SEED visual tokens

Tokenes de Habla [30]:



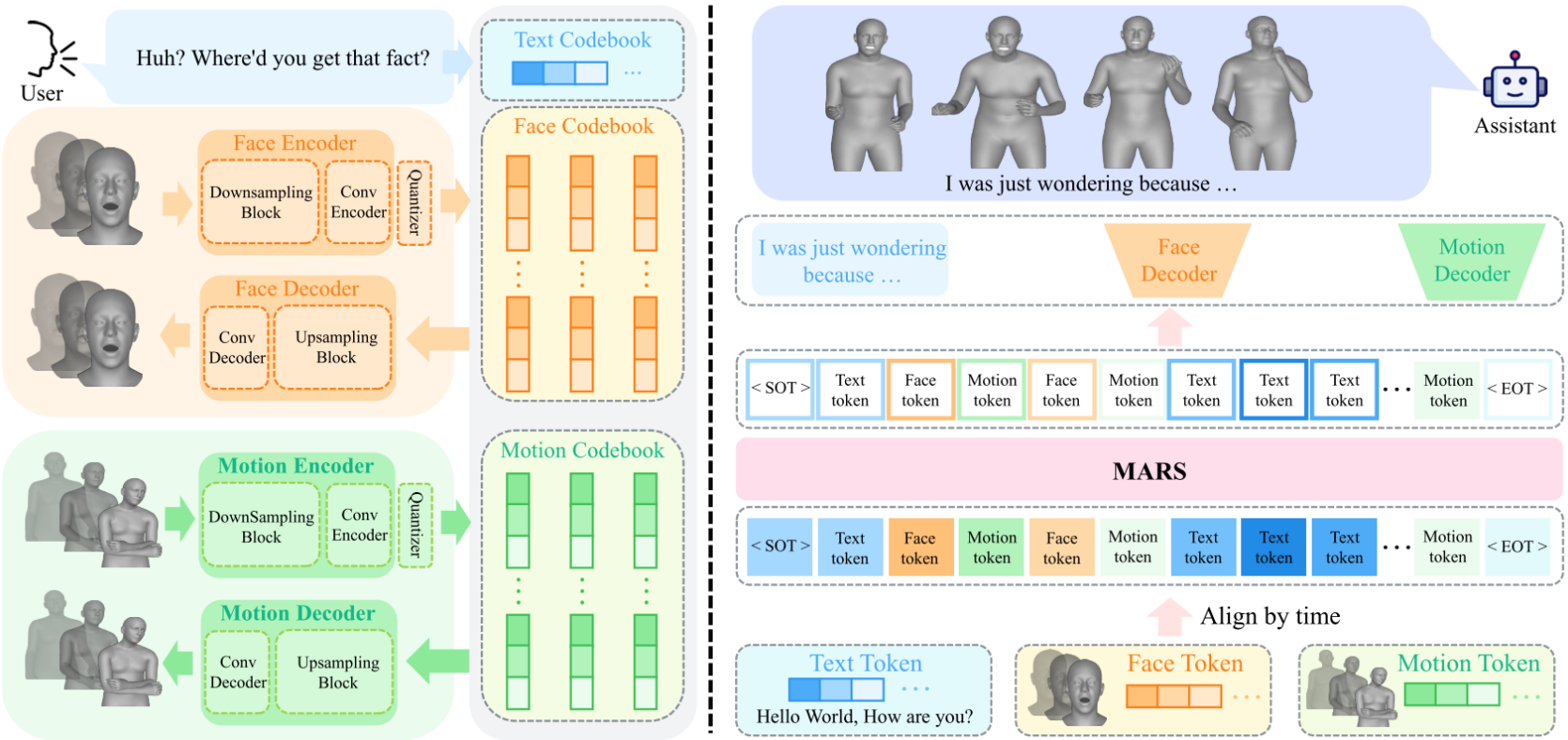
AnyGPT [29]

Permite de usarlos en un LLM para que sea multimodal!



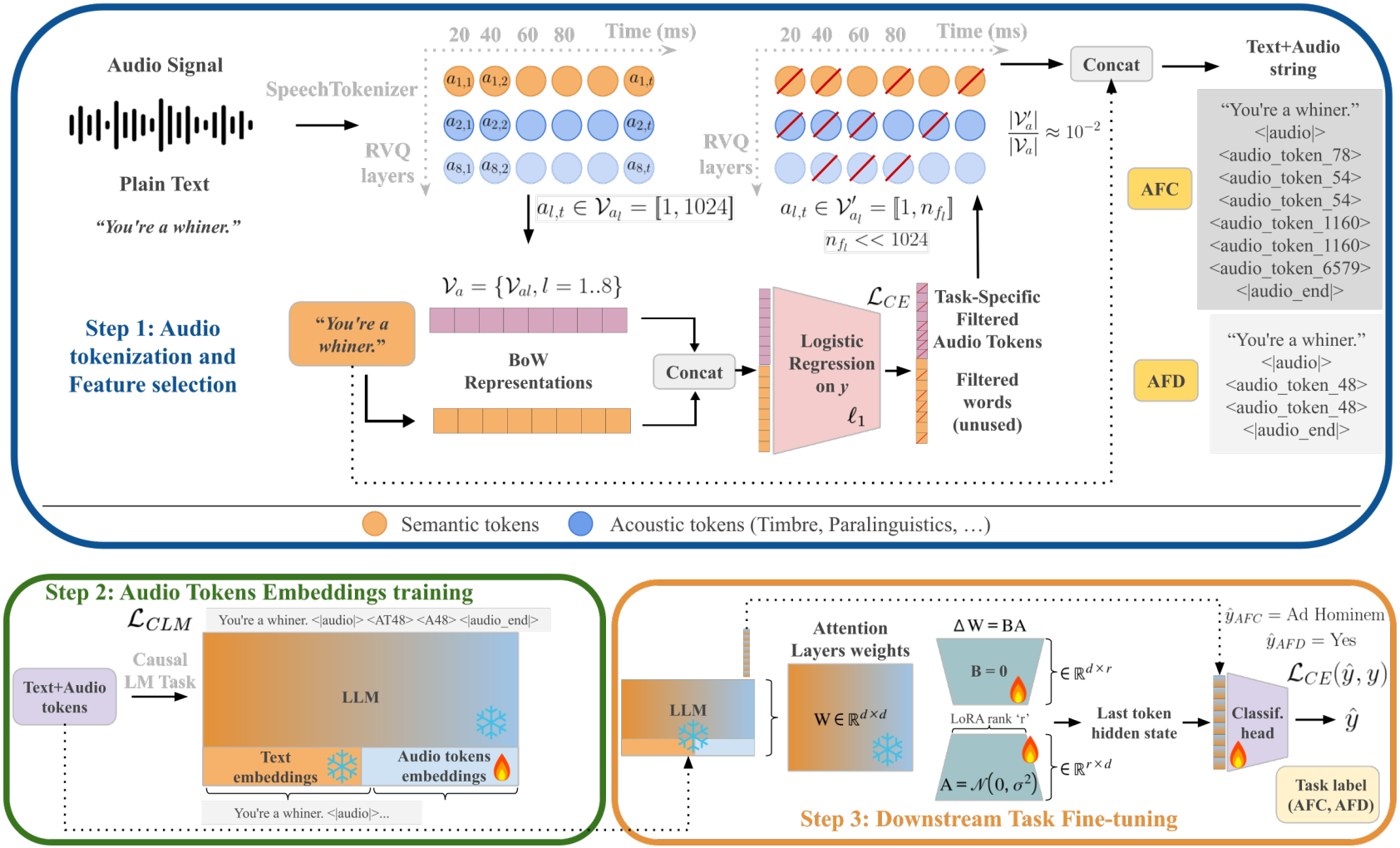
Ayuda a mejorar los agentes! [11]

Usando VQ-VAE, es posible de encontrar un codebook de expresiones faciales y gestos. Hicieron una tokenizacion!



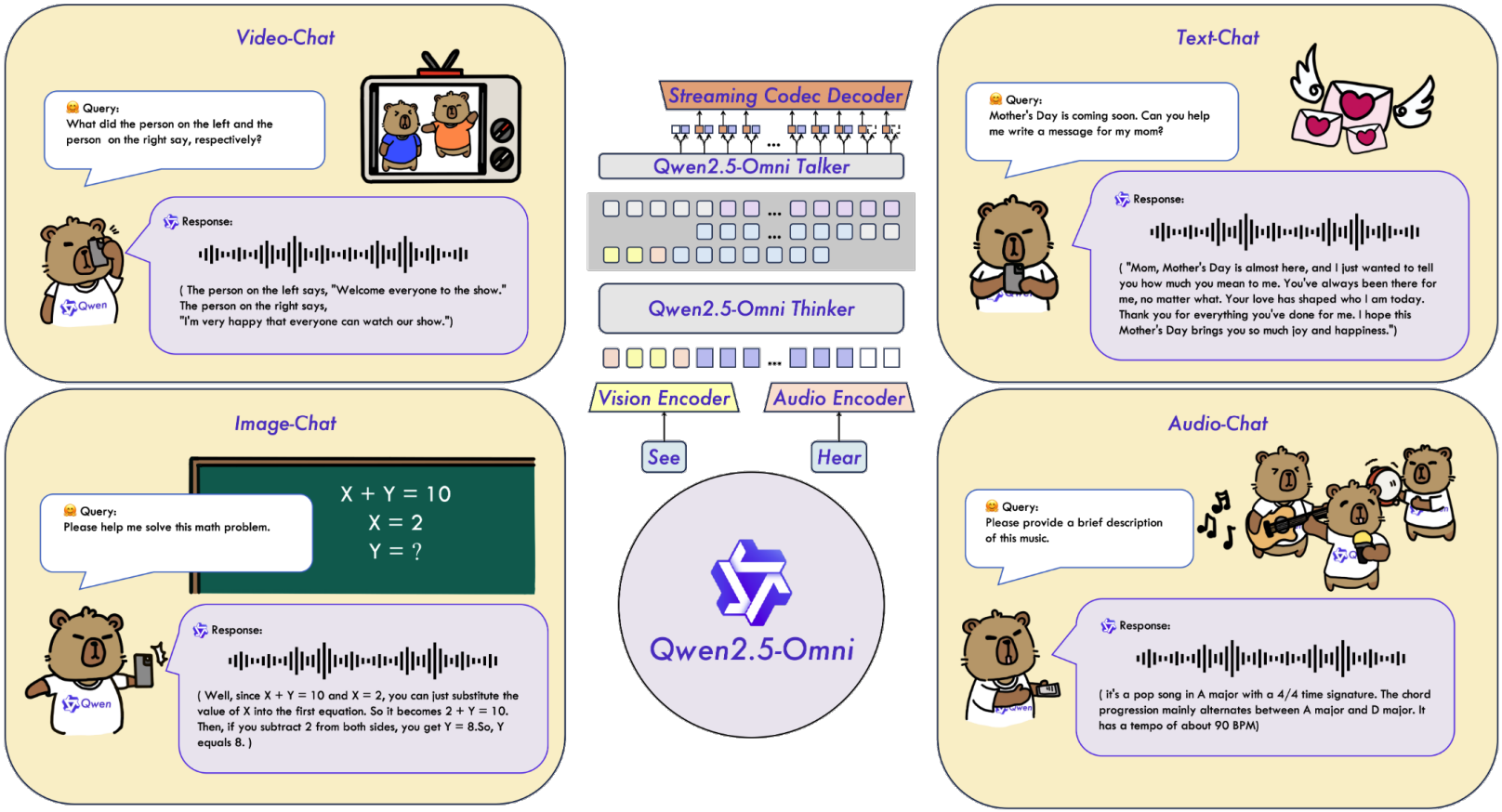
Vieron que el agente tiene un comportamiento social mas adaptados!

Como aumentar un LLM sin muchos datos? [3]



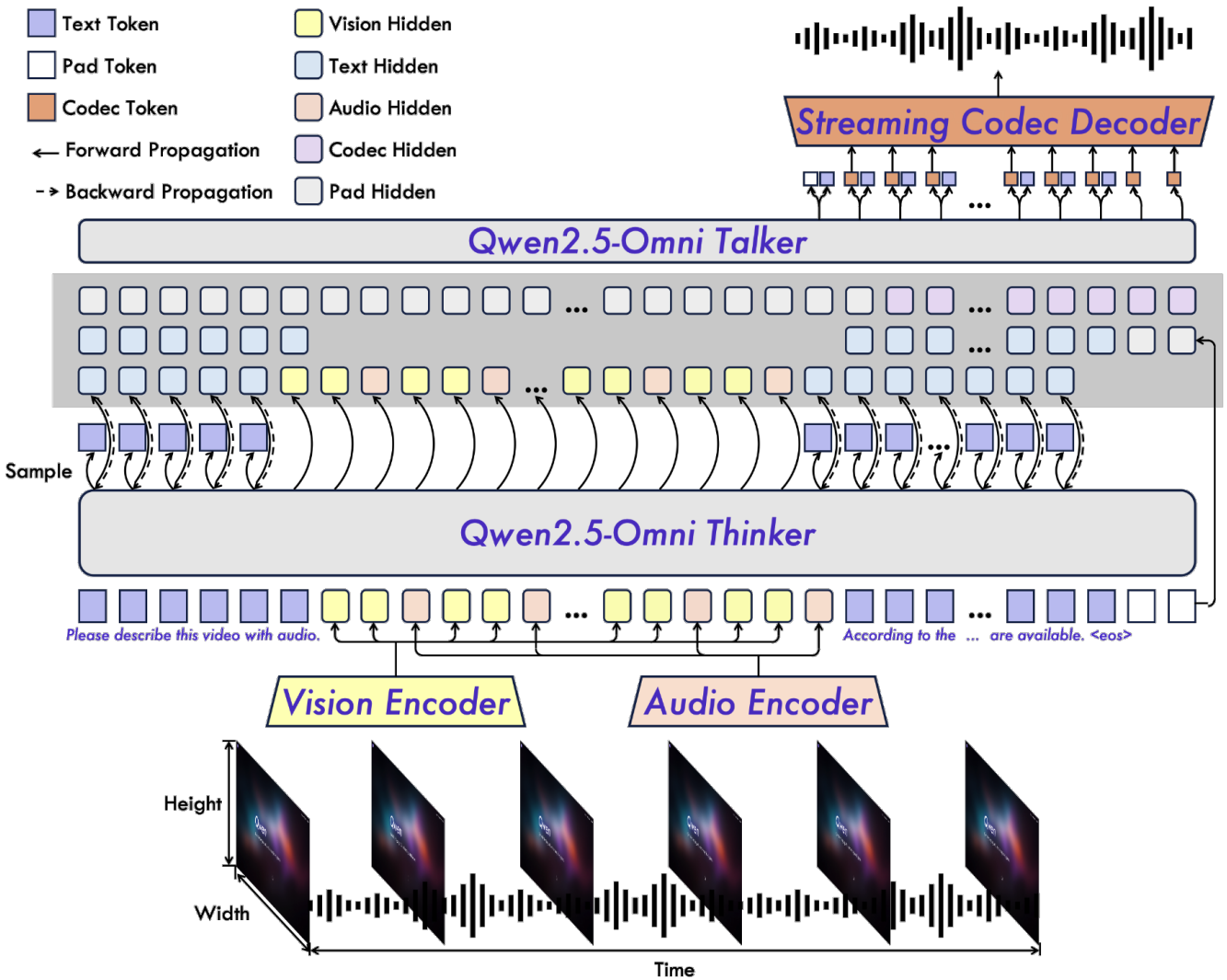
Elijemos los tokenes de audio los mas importante por una tarea especifica, y entrenamos un LLM a usarlos. Mejora en tareas de detección de falacias, de emociones y sentimiento.

Modelos Omni [27]



Basado en Qwen2.5 [25], el modelo se afina (fine-tuning) sobre datos secuenciales multimodales.

Modelos Omni [27]

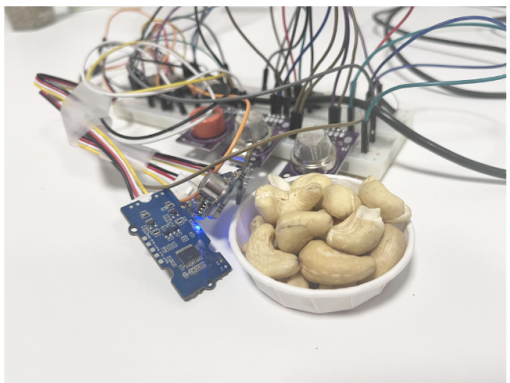


Basado en Qwen2.5 [25], el modelo se afina (fine-tuning) sobre datos secuenciales multimodales.

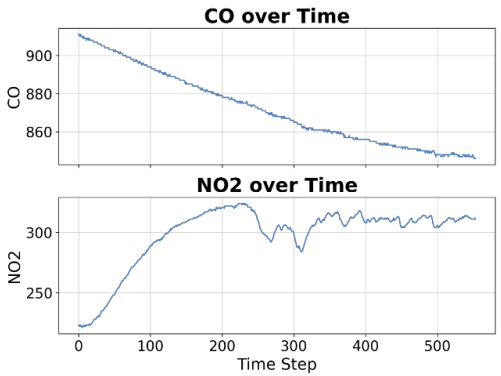
Ampliando la Frontera de las Modalidades

Modalidad #1: el olfato

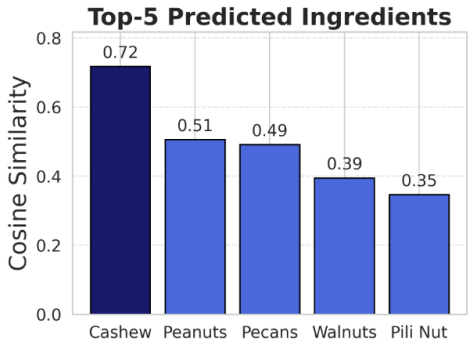
Algunas modalidades son más fáciles de digitalizar, como el texto, la imagen o el audio. El olfato no cuenta con la infraestructura madura de sensores/datasets que sí tienen la visión o el habla.



(a) Sensor setup detecting cashew.



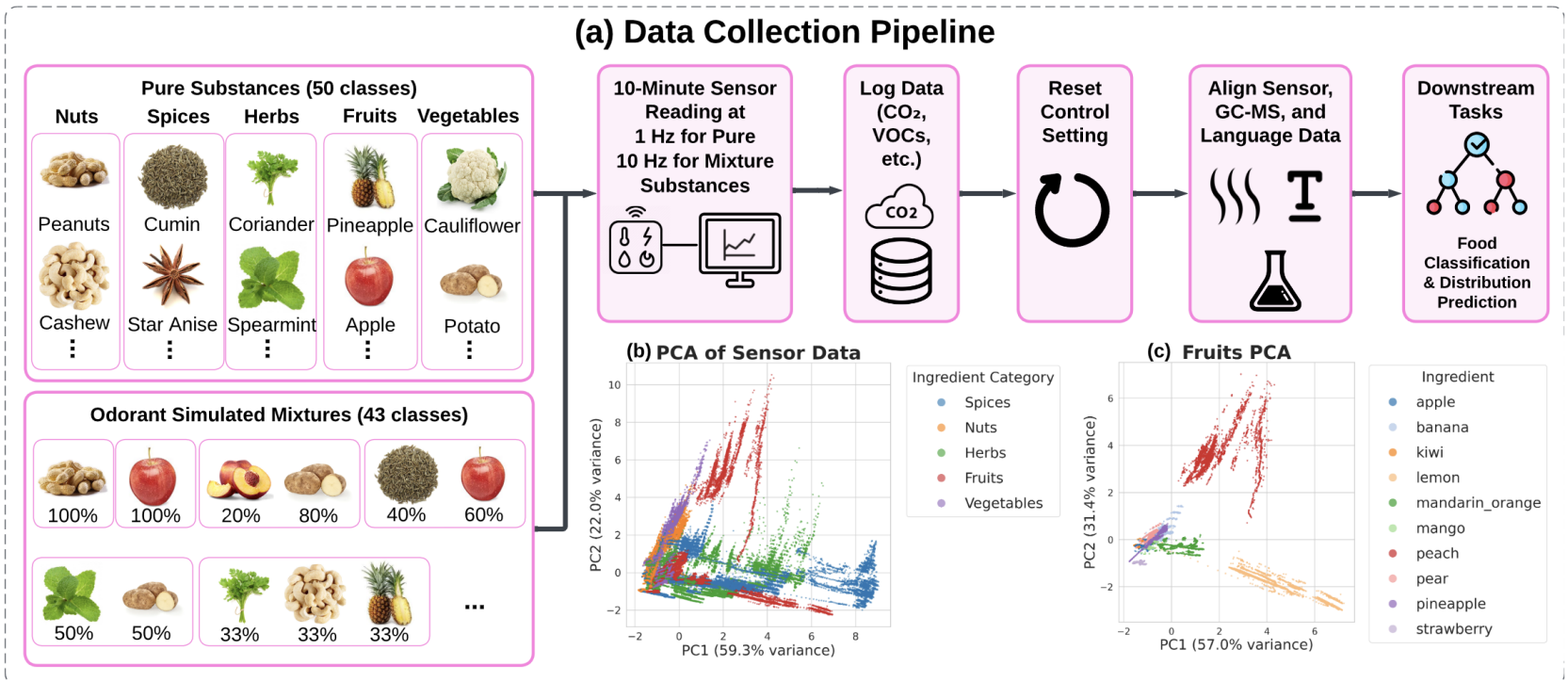
(b) Time-series signals from CO and NO_2 sensors.



(c) Top-5 model predictions using cosine similarity.

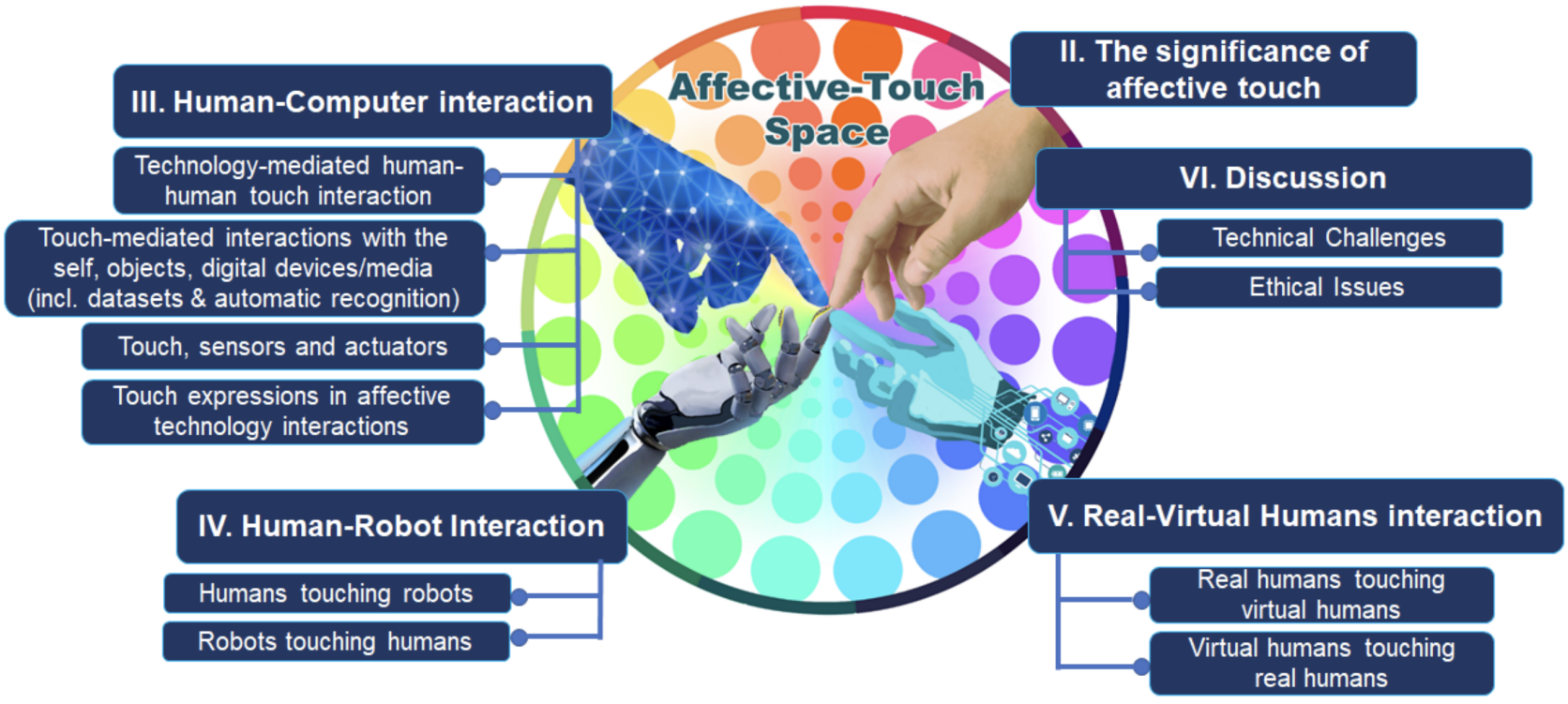
Modalidad #1: el olfato

Algunas modalidades son más fáciles de digitalizar, como el texto, la imagen o el audio. El olfato no cuenta con la infraestructura madura de sensores/datasets que sí tienen la visión o el habla.



SmellNet contiene alrededor de 828.000 puntos de datos de series temporales a través de 50 sustancias, abarcando frutos secos, especias, hierbas, frutas y verduras [4].

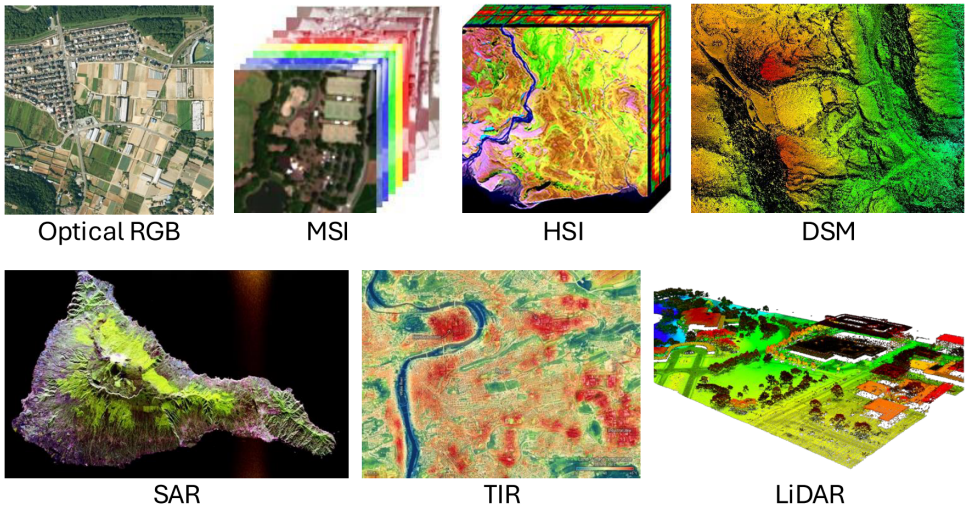
Modalidad #2: el tacto social [19]



- El tacto social ayuda a difuminar los límites entre agentes virtuales controlados por computadora y por humanos [7]
- Datasets egocéntricos con sensores de tacto de grano fino [23]

Modalidades #3: ¿son los datos satelitales realmente "una imagen"?

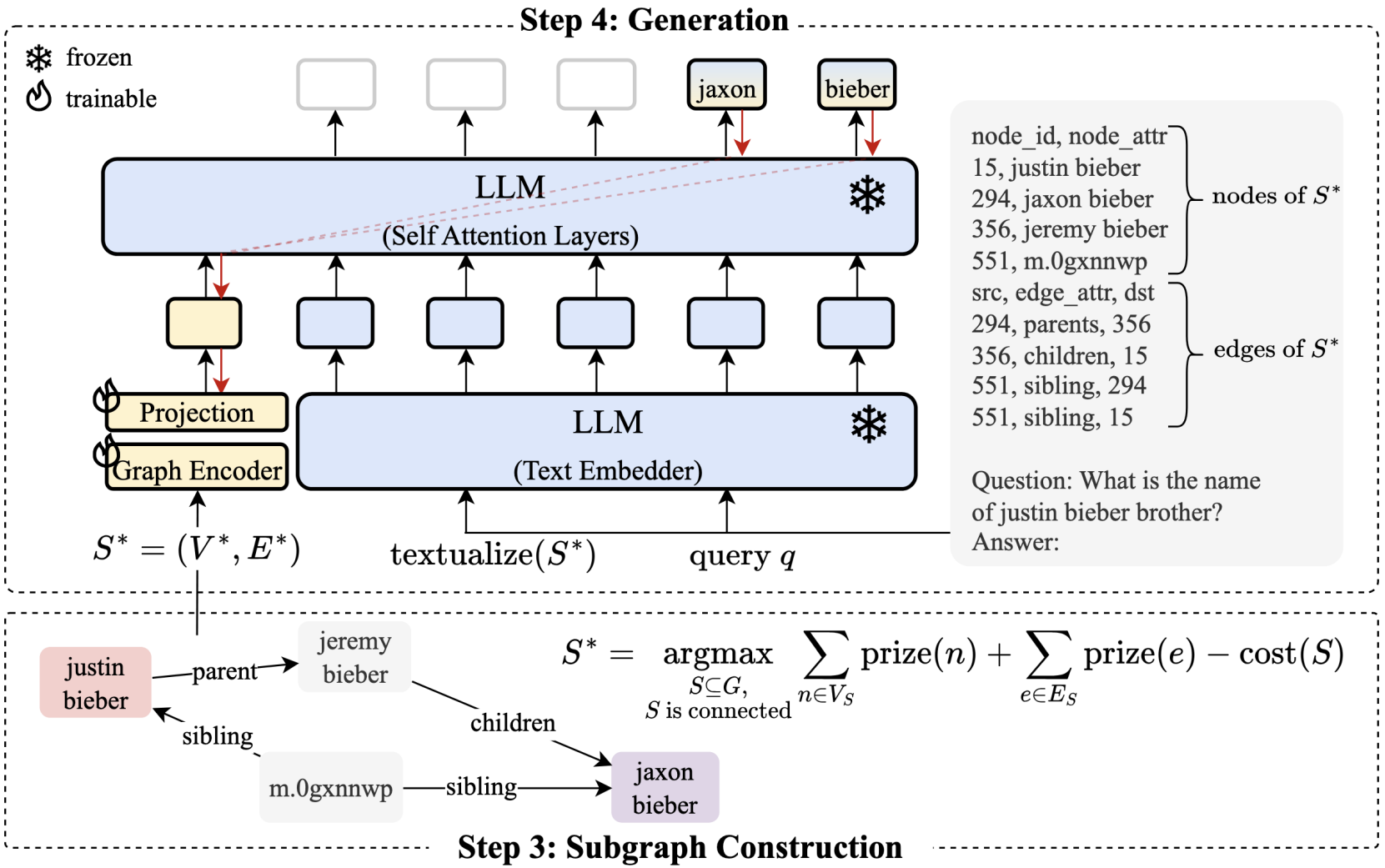
Postura: los datos satelitales / de teledetección deberían tratarse como su propia modalidad distinta, no solo como otro dominio de imagen [22]



- Canales más allá del RGB: multiespectral, SAR, elevación, clima
- Invarianzas diferentes: Rango de escala enorme (un bote vs. un glaciar), estructura temporal de revisita

- Series temporales multimodales: si lo óptico es visión, entonces el SAR es tacto
- El preentrenamiento imagen-texto a escala web no se transfiere gratis

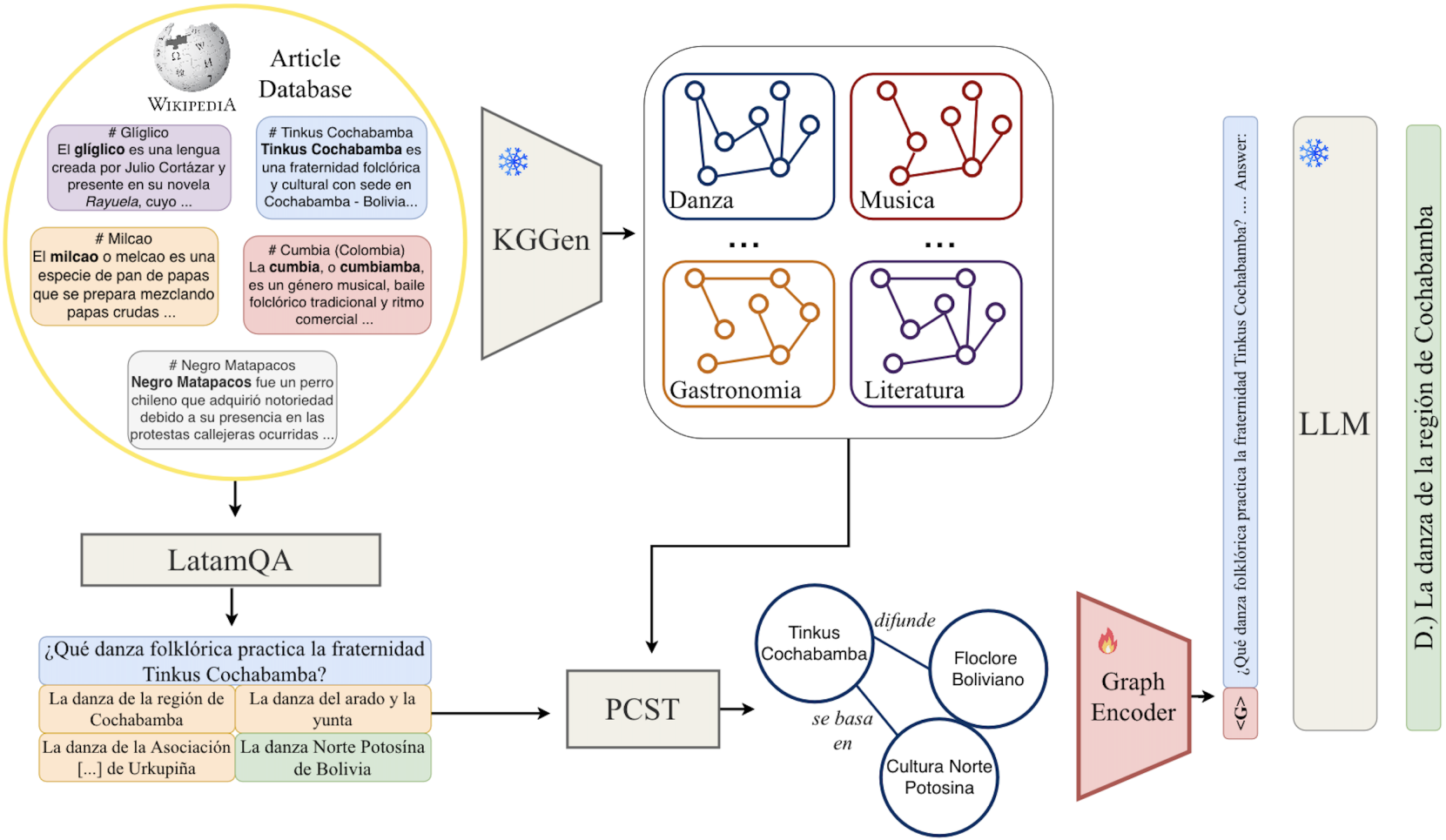
Modalidad #4: Integrar datos de Grafos [6]



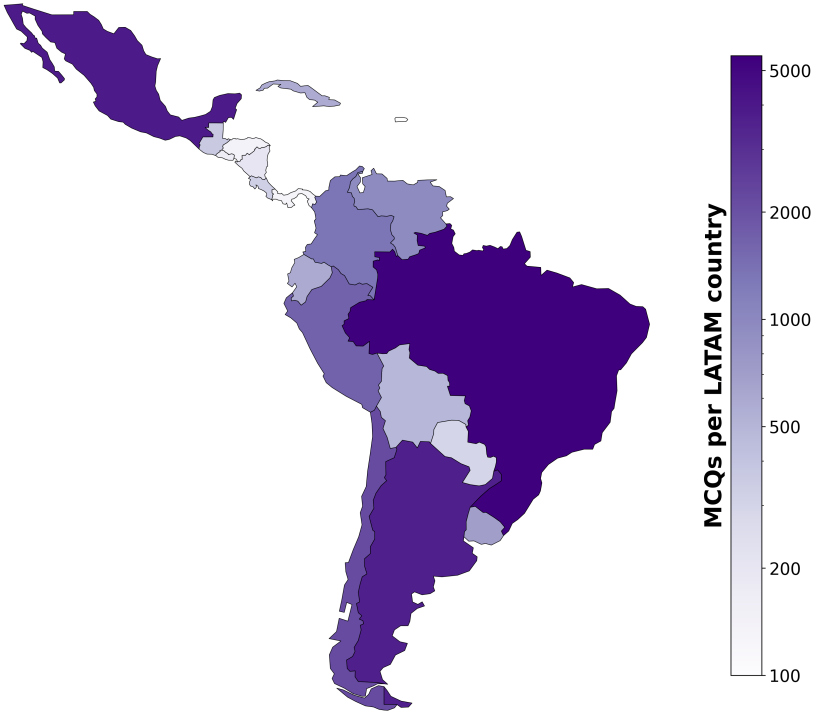
Un GNN encodea el Grafo de Conocimiento (Knowledge Graph: KG) en el espacio de los embeddings del LLM.

Caso de uso: Conocimiento Cultural Latino Americano [20]

Enseñarle a un LLM conocimiento nuevo y específico de una región mediante datos de entrenamiento es difícil: cola larga, catastrophic forgetting,... Los **KG** dan control fino sobre qué entra al contexto, con más explicabilidad y facilidad de actualización.



Caso de uso: Conocimiento Cultural Latino Americano [20]



- `Qwen2.5-3B-Instruct`
- Embedding de `jinaai/jina-embeddings-v3`
- GRAG y RAG superan el modelo Vanilla
- GRAG menos costoso y KG mas simple a manipular

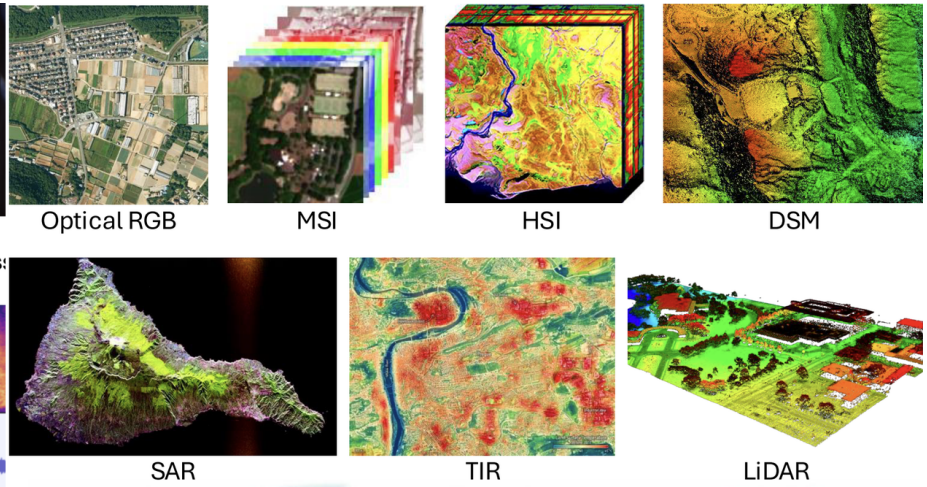
Figure 8: Dataset LatamQA [9]

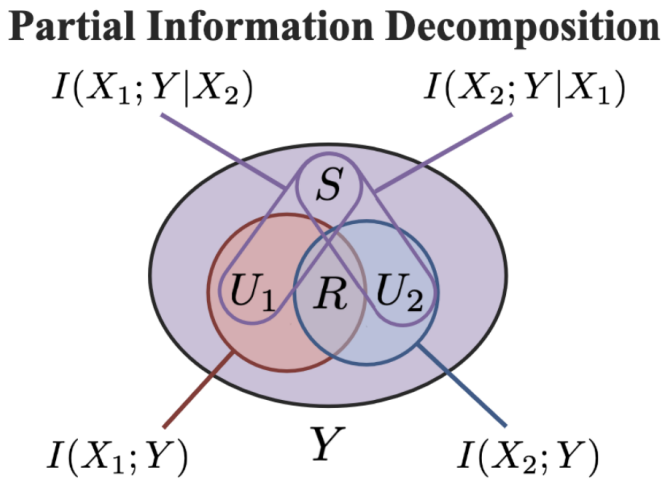
CaZero-shot	RAG	KGGen+G-Retriever
0.6017	0.9338	0.9141

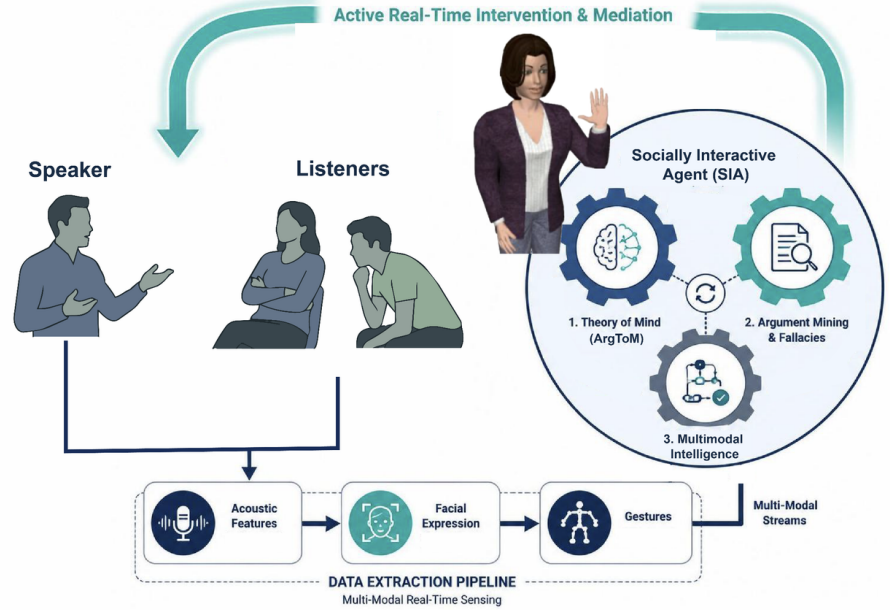
Nuestros trabajos en curso

Datos Humanos y Satelitales hacia un Modelado General



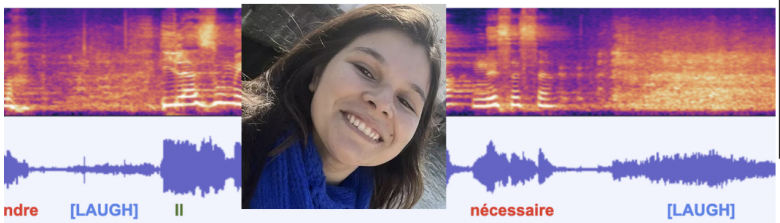




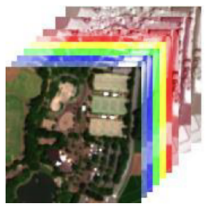


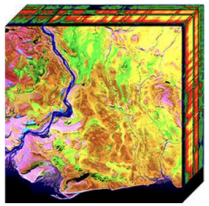
Datos Humanos y Satelitales hacia un Modelado General



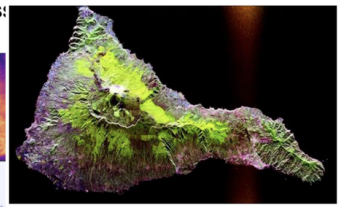


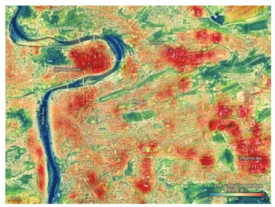

Optical RGB


MSI

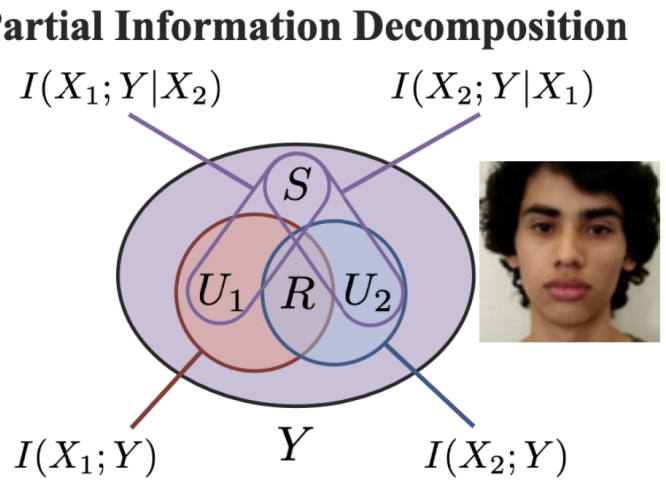

HSI

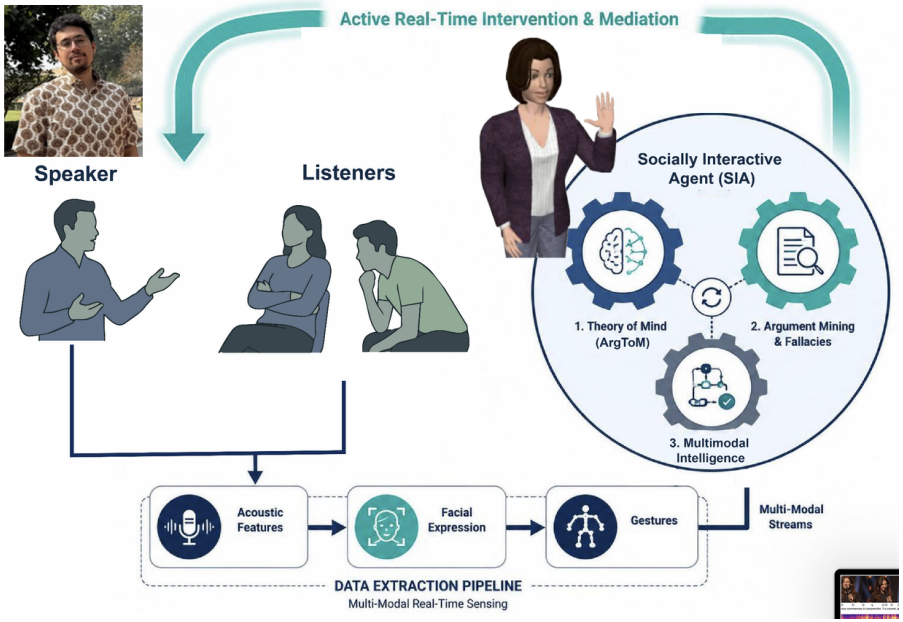



SAR


TIR


LiDAR

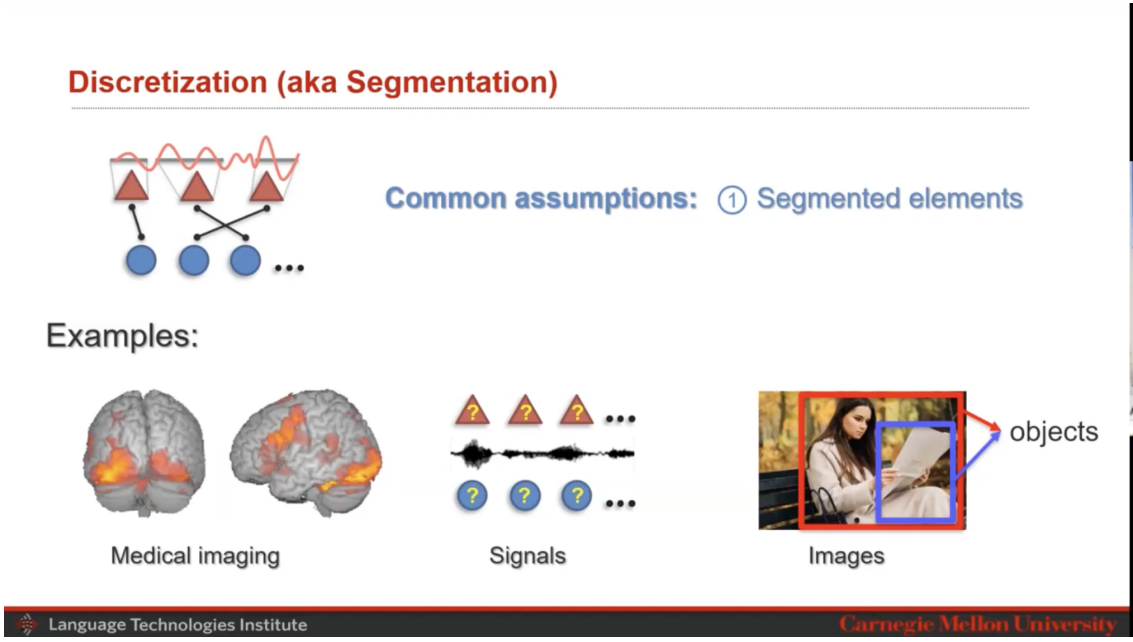




Cierre

¿Quieres aprender más sobre ML Multimodal?

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





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

También puedes revisar estos artículos increíbles [15, 13]. Y estas charlas sobre Fundamentos de la representación multimodal y Fundamentos de AI multisensorial de alta modalidad

Preguntas?

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


N. Calbucura, J. Guillen, and V. Barriere.
A Simple Method to Enhance Pre-trained Language Models with Speech Tokens for Classification.
4 2026.


D. Feng, W. Dai, C. Li, A. Pernigo, Y. Wen, and P. P. Liang.
SmellNet: A Large-scale Dataset for Real-world Smell Recognition.
4 2026.


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.

References iii


X. He, Y. Tian, Y. Sun, N. V. Chawla, T. Laurent, Y. LeCun, X. Bresson, and B. Hooi.
G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering.
5 2024.


M. Hoppe, B. Rossmy, D. P. Neumann, S. Streuber, A. Schmidt, and T.-K. Machulla.
A human touch: Social touch increases the perceived human-likeness of agents in virtual reality.
In Proceedings of the 2020 CHI conference on human factors in computing systems, pages 1–11, 2020.

References iv



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.



Y. Karmim, R. Pino, H. Contreras, H. Lira, S. Cifuentes, S. Escoffier, L. Martí, D. Seddah, and V. Barriere.
Leveraging wikidata for geographically informed sociocultural bias dataset creation: Application to latin america.
In Proceedings of the First Workshop on Multilingual Multicultural Evaluation, pages 177–188, 2026.


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.


Y. Kim, J. Chung, J. Kim, S. Lee, S. Lee, J. Kim, C. Yang, and Y. Yu.
Speaking Beyond Language: A Large-Scale Multimodal Dataset for Learning Nonverbal Cues from Video-Grounded Dialogues.
pages 2247–2265, 2025.


P. P. Liang, Y. Cheng, X. Fan, C. K. Ling, S. Nie, R. Chen, Z. Deng, N. Allen, R. Auerbach, F. Mahmood, R. Salakhutdinov, and L.-P. Morency.
Quantifying & modeling multimodal interactions: An information decomposition framework.
In Advances in Neural Information Processing Systems (NeurIPS), 2023.


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, 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.

References viii


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.


T. Olugbade, L. He, P. Maiolino, D. Heylen, and N. Bianchi-Berthouze.
Touch technology in affective human–, robot–, and virtual–human interactions: A survey.
Proceedings of the IEEE, 111(10):1333–1354, 2023.


P. Poulenard, Y. Karmim, and V. Barriere.
Knowledge-graph based augmentation versus retrieval augmented generation for cultural-related question answering.
Proceedings of ORACLE Workshop at EMNLP26, 2026.


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.


E. Rolf, K. Klemmer, C. Robinson, and H. Kerner.
Position: Mission critical – satellite data is a distinct modality in machine learning.
In International Conference on Machine Learning (ICML), 2024.

References x



Y. R. Song, J. Li, R. Fu, D. Murphy, K. Zhou, R. Shiv, Y. Li, H. Xiong, C. E. Owens, Y. Du, et al.
Opentouch: Bringing full-hand touch to real-world interaction.
arXiv preprint arXiv:2512.16842, 2025.



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.



Q. Team.
Qwen2.5 technical report.
arXiv preprint arXiv:2412.15115v2, 2024.


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


J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, B. Zhang, X. Wang, Y. Chu, and J. Lin.
Qwen2.5-Omni Technical Report.
3 2025.


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.