CS60092: Information Retrieval

Distant Supervision & Multimodal Information Retrieval

Debaditya Roy

The Core Idea

> *"If two entities participate in a relation, any sentence mentioning both likely expresses that relation."*
>
> — Mintz et al., 2009

The hand-labeling bottleneck

- Relation extraction needs labeled (e_1, r, e_2) triples in text
- Expert annotation is slow, expensive, and domain-specific
- Modern KBs (Freebase, Wikidata) already contain millions of triples

Distant supervision solution

- Use existing KB triples as heuristic labels
- Find all sentences co-mentioning (e_1, e_2) in a large corpus
- Label them automatically — no human annotators needed

The Noise Problem

The at-least-one assumption breaks in practice

Both sentences mention (Obama, Hawaii) but mean very different things:
"Obama was born in Hawaii." expresses born_in
"Obama visited Hawaii." does NOT express born_in

False positives

Sentence mentions the entity pair but expresses a different or no relation. The most common noise type in practice.

False negatives

True relation is expressed in the text but the KB does not know it yet, so the sentence gets no label.

KB incompleteness

The KB itself is incomplete; even correct heuristic labeling misses real relation instances in the wild.

Each noise type has its own mitigation strategy — this structure organizes the entire Distance Supervision literature.

Multi-Instance Learning

Bag $B = \{x_1, x_2, \ldots, x_n\}$ all sentences mentioning entity pair (e_1, e_2)

At-least-one assumption: if (e_1, r, e_2) in KB, then there exists x_i in B that expresses r

Max aggregation (Riedel 2010)

$$p(y|B) = \max_{x_i \in B} p(y \mid x_i)$$

Train on the single most expressive sentence in the bag.

Selective attention (Lin 2016)

$$s(B) = \sum_i a_i * x_i$$

$$a_i = \text{softmax}\left(\frac{x_i^\top A r}{\tau}\right)$$

A = weight matrix
r = relation vector
τ = temperature

Training loss (cross-entropy over bags): $L = -\sum \log p(y_B \mid s(B))$

Riedel, S., Yao, L., & McCallum, A. (2010, September). Modeling relations and their mentions without labeled text. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 148-163).
Lin, Y., Shen, S., Liu, Z., Luan, H., & Sun, M. (2016, August). Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 2124-2133).

The PCNN Sentence Encoder

Piecewise CNN encodes inter-entity context explicitly

Divide every sentence into three segments relative to (e_1, e_2)

L

Left [L]
tokens before e_1

M

Middle [M]
between e_1 and e_2

R

Right [R]
tokens after e_2

CNN + max-pool each segment independently, then concatenate:

$$x = [\,pool(L) \;||\; pool(M) \;||\; pool(R)\,]$$

M captures inter-entity context independently – PCNN's key advantage over a flat CNN.

Zeng, D., Liu, K., Chen, Y., & Zhao, J. (2015, September). Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1753-1762).

Temperature in Selective Attention

$$\boldsymbol{a}_i = \text{softmax}\left(\frac{\boldsymbol{x}_i^{\top} A\, \boldsymbol{r}}{\tau}\right)$$

$\tau \to 0$	Collapses to hard-max — only the highest-scoring sentence gets weight 1, all others get 0. *Equivalent to Riedel max-aggregation.*
$\tau \to \infty$	Softmax becomes uniform — every sentence in the bag gets equal weight $1/n$ *Model ignores all attention signal; bags treated as simple averages.*
$\tau = 0.07$	PCNN+ATT uses small τ near this range: sharp but not degenerate *Strong gradient signal; model is sensitive to angular differences between representations.*

Multimodal Information Retrieval

From shared embedding spaces

to cross-modal nearest-neighbor search

Retrieval Problem

Similarity across modalities requires a shared space

Given a query in modality A, return relevant documents in modality B (or A)
You cannot compute similarity between raw pixels and token sequences — both must project into a common geometry

$T \rightarrow I$

Text query
→ Image results

"a dog on a beach"

$I \rightarrow T$

Image query
→ Text results

[photo] -> captions

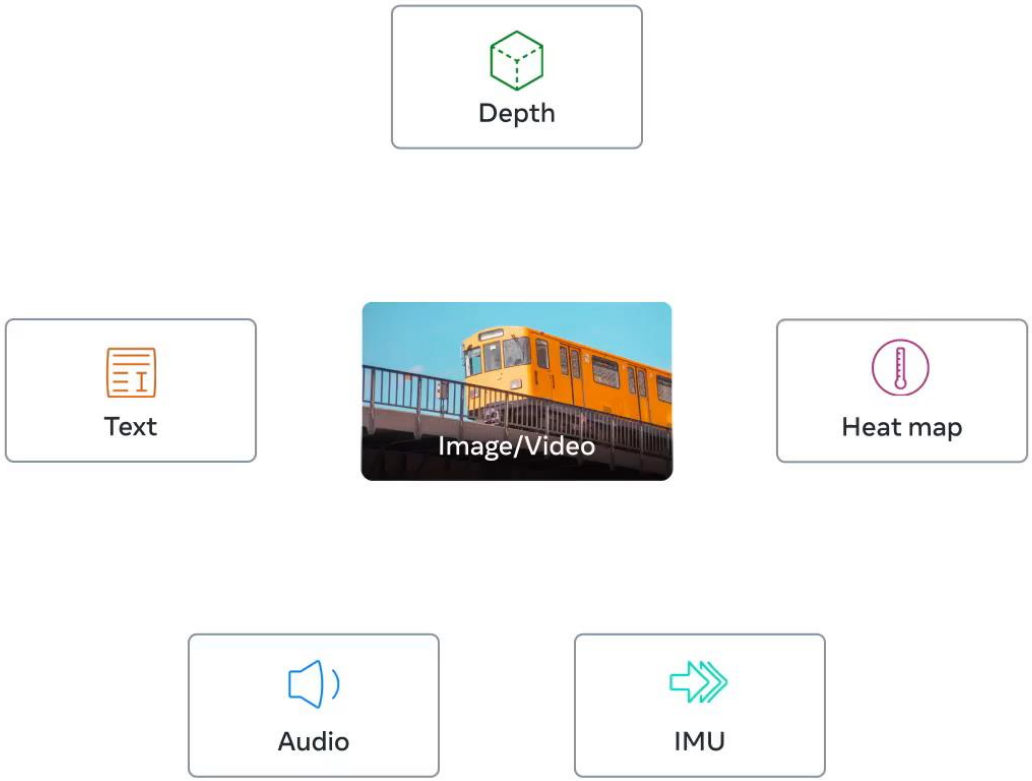
$* \rightarrow *$

Any query
→ Any modality

ImageBind: 6 modalities

Key insight: cross-modal retrieval is a nearest-neighbor problem once you have the shared embedding space.

ImageBind



https://imagebind.metademolab.com/demo?modality=A2GI

Encoders & L_2 Normalization

Visual encoder

$$v_i = \frac{f_v(image_i)}{\|f_v(image_i)\|_2}$$

Text encoder

$$t_i = \frac{f_t(caption_i)}{\|f_t(caption_i)\|_2}$$

Why L_2 normalization is required

- Without normalization: the model can minimize loss by scaling vector magnitudes instead of aligning directions.
- After normalization: dot product equals cosine similarity, bounded in $[-1, 1]$. Geometry becomes interpretable.

CLIP

Image + Text

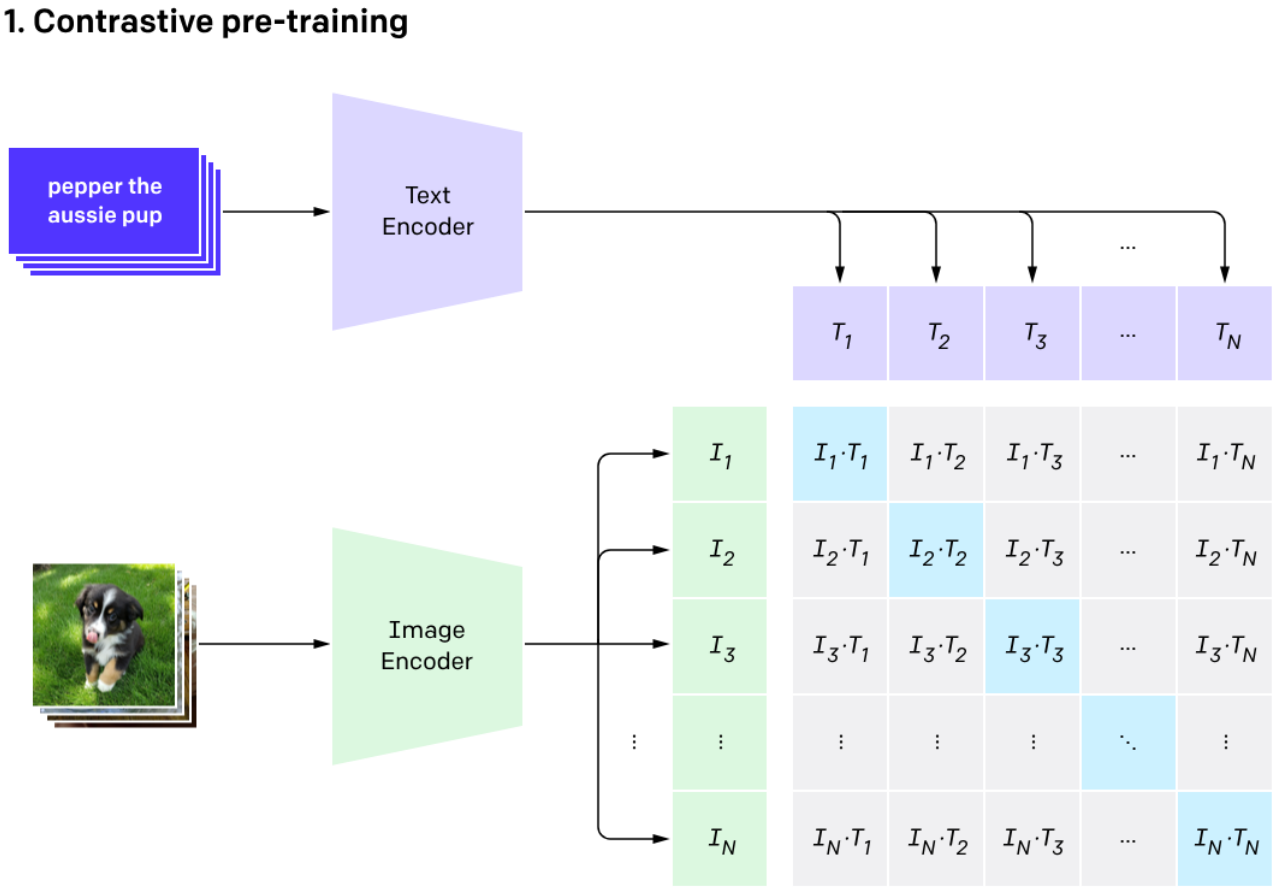
Whisper -> embed

Audio -> Text space

ImageBind

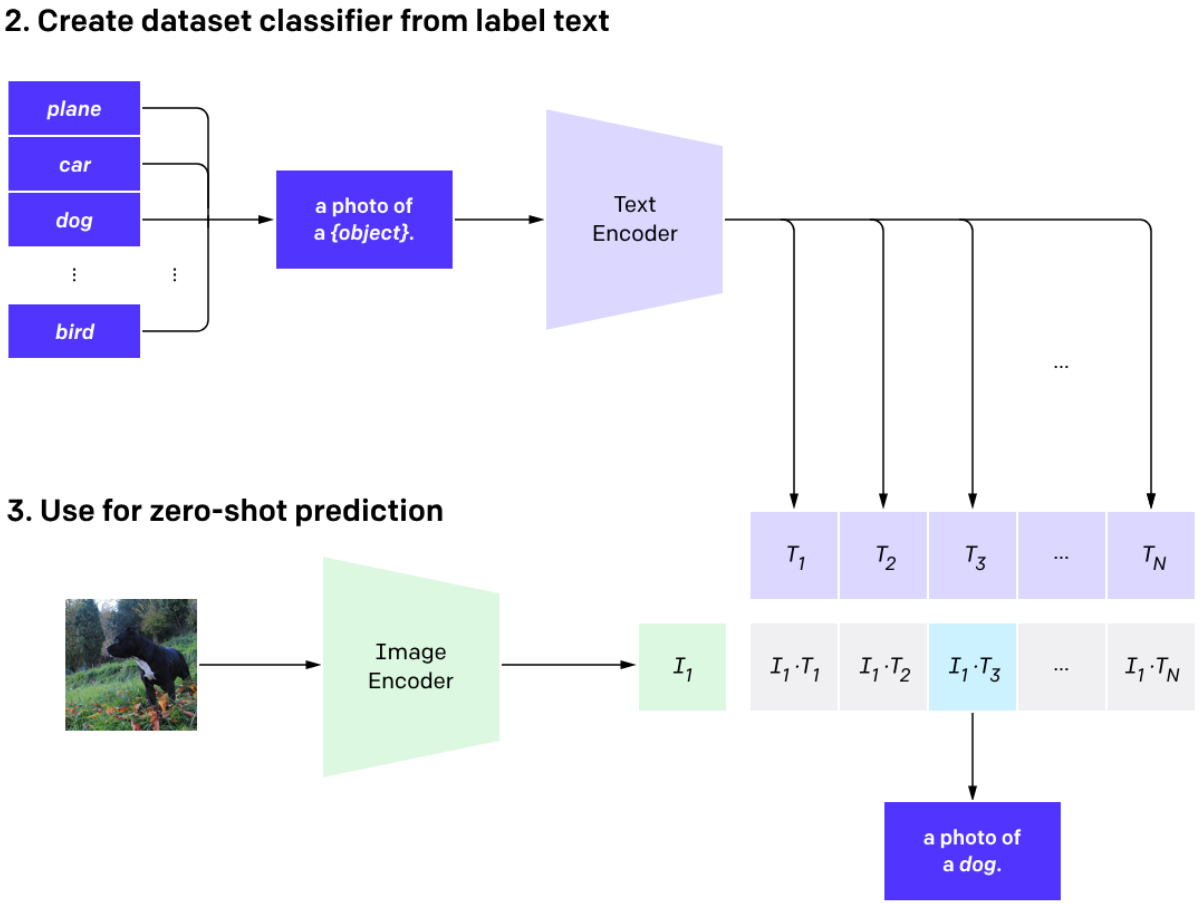
6 modalities (image, text, audio, depth, thermal, IMU)

CLIP



Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.

CLIP



Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.

InfoNCE Loss (CLIP)

A batch of N pairs generates N squared minus N hard negatives

Similarity matrix $S_{ij} = v_i^\top t_j / \tau$

Diagonal = positive pairs
Off-diagonal = $N - 1$ negatives per row

Image → text loss

$$L_{i\to t} = -\left(\frac{1}{N}\right)\sum_i \log \frac{\exp S_{ii}}{\sum_j S_{ij}}$$

Each row: N-class softmax; diagonal is the target.

Text → image loss

$$L_{t\to i} = -\left(\frac{1}{N}\right)\sum_j \log \frac{\exp S_{jj}}{\sum_i S_{ij}}$$

Each column: N-class softmax; diagonal is the target.

Symmetric total loss: $L = \frac{(L_{i\to t} + L_{t\to i})}{2}$

Larger batch N means more negatives per sample, which gives a stronger discriminative signal. CLIP used N = 32,768.

Temperature in CLIP

Learned, initialized at 0.07 — same intuition as Distant Supervision

$S_{ij} = v_i^\top t_j / \tau$ (τ is a learned scalar, initialized to 0.07, clipped to prevent instability)

Low τ	Sharpens softmax — model is sensitive to small angular differences between embeddings. *Risk: gradient instability; minor misalignment causes large loss spikes.*
High τ	Softens distribution — reduced gradient magnitude, slower learning of fine distinctions. *Risk: embeddings collapse into indistinct clusters, negatives are ignored.*
τ learned	CLIP clips $\log(\tau)$ to $[-4, 4]$ during training; prevents pathological regimes.

Same temperature intuition appears in Distant Supervision selective attention

Connection to Distant Supervision: same temperature-softmax mechanism governs both PCNN+ATT and CLIP. One formula, two domains.

Retrieval at Inference & Evaluation

Nearest-neighbor search and cross-modal metrics

Retrieval: rank all indexed items by $score(q, v_i) = q^\top v_i$ (cosine similarity on unit sphere)

Recall @ K

$$= \frac{|correct\ in\ topK|}{|correct\ total|}$$

Fraction of queries where the ground truth appears in the top K results.

Mean Reciprocal Rank

$$= \left(\frac{1}{|Q|}\right) \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

Rewards higher placement of the first correct result; useful when position matters.

Cross-modal R@K

$$= Image \rightarrow Text\ R@K$$
$$+ Text \rightarrow Image\ R@K$$

Always report both directions — they can differ dramatically in difficulty.

Indexing: FAISS HNSW gives $O(\log N)$ amortized retrieval. Product quantization compresses embeddings on the unit sphere.

InfoNCE as Mutual Information

CLIP maximizes a lower bound on $I(V;T)$

$$I(V;T) >= \log(N) - L_{InfoNCE}$$

What this means:

I CLIP is not just a retrieval system — it is a mutual information estimator between vision and language.

N As batch size N grows, the bound tightens (log N increases). This is why large batches matter theoretically, not just empirically.

L Minimizing $L_{InfoNCE}$ simultaneously maximizes the MI lower bound — alignment and information maximization are the same objective.

This MI framing connects CLIP to variational methods and representation learning broadly — a rich thread for advanced students.

Unifying Thread

Supervision without hand-annotation

DS: DS: KB-derived heuristic labels | **Multimodal:** CLIP: web image-caption pairs as labels

Temperature-scaled dot-product attention

DS: attention weight $a_i = softmax(x_i^T A\, r/\tau)$

Multimodal: CLIP: similarity $S_{ij} = v_i^\tau\, t_j/\tau$

Cross-entropy as the training objective

DS: bag-level cross-entropy loss L | **Multimodal:** CLIP: symmetric InfoNCE loss L

Dense representations enabling similarity search

DS: entity pair embeddings for RE | **Multimodal:** CLIP: shared embedding space for retrieval

ViLBERT

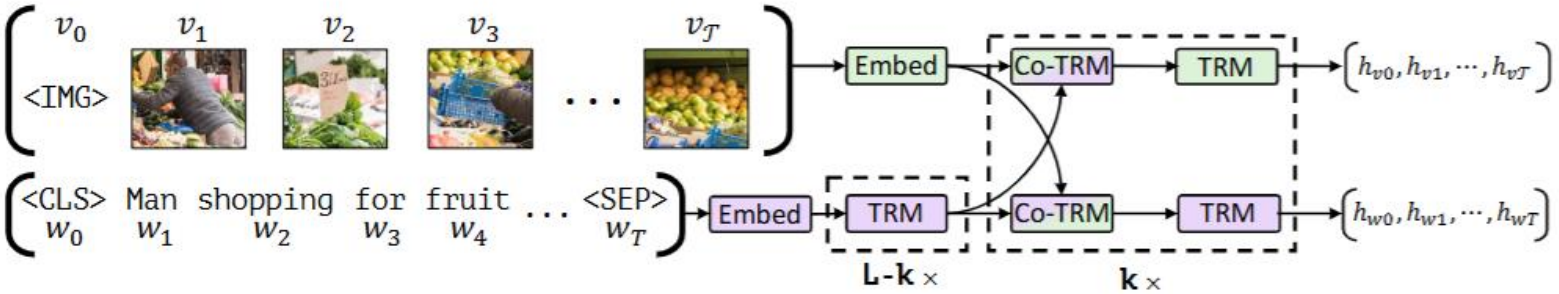
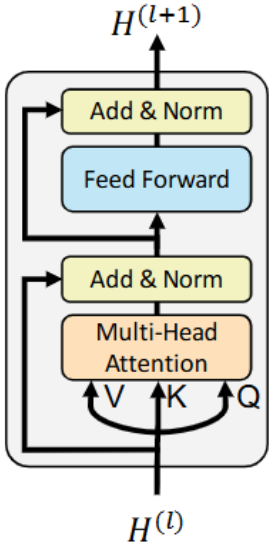


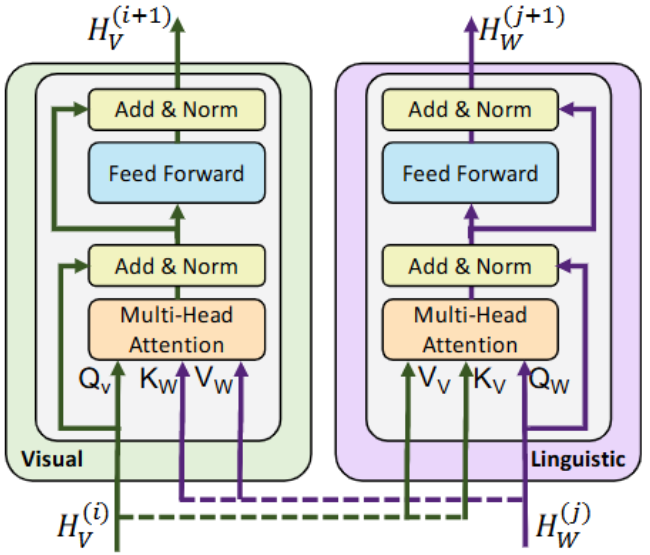
Figure 1: Our ViLBERT model consists of two parallel streams for visual (green) and linguistic (purple) processing that interact through novel co-attentional transformer layers. This structure allows for variable depths for each modality and enables sparse interaction through co-attention. Dashed boxes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

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, *32*.

VilBERT



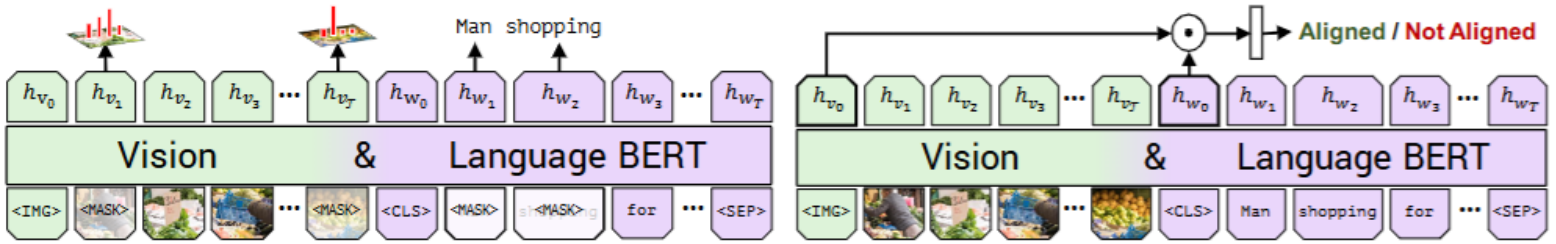
(a) Standard encoder transformer block



(b) Our co-attention transformer layer

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, *32*.

ViLBERT

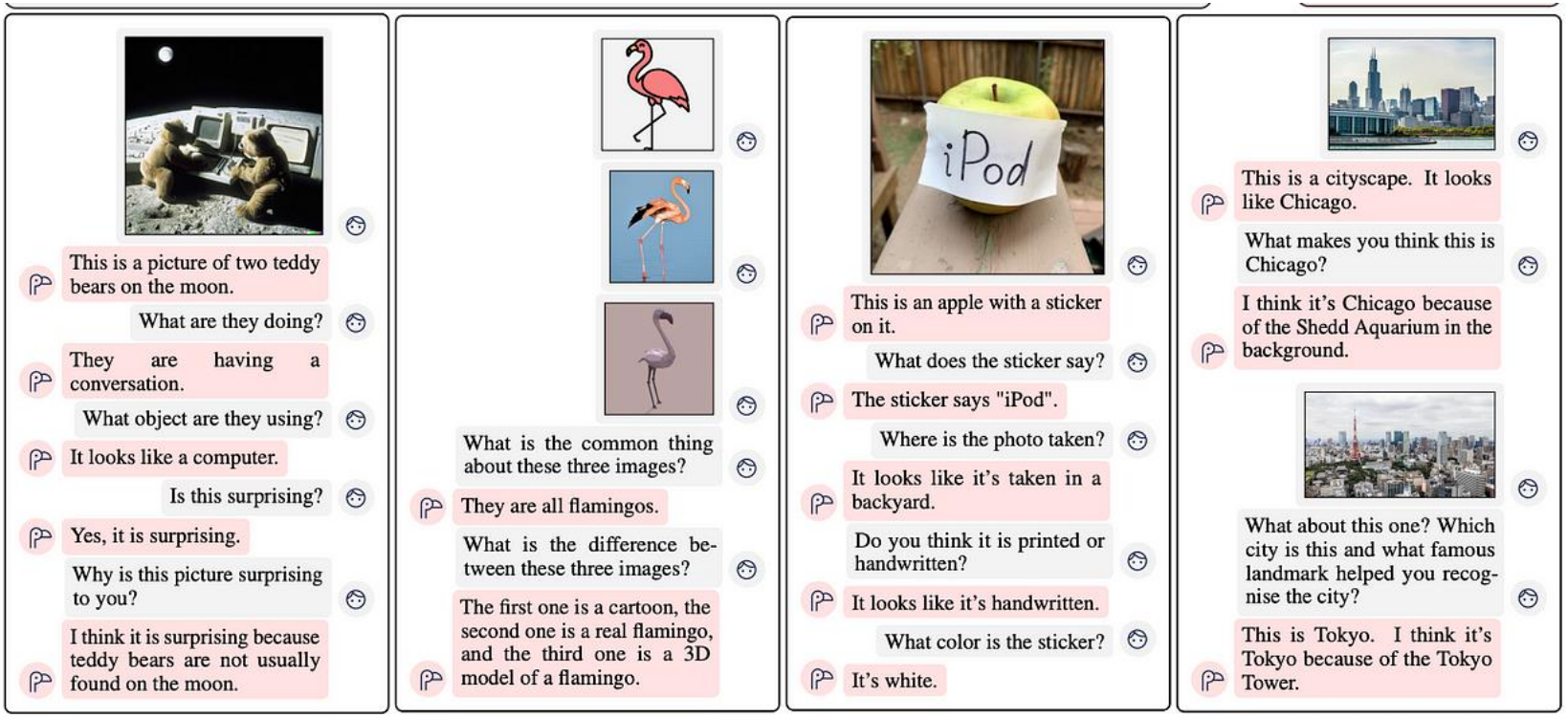


(a) Masked multi-modal learning (b) Multi-modal alignment prediction

Figure 3: We train ViLBERT on the Conceptual Captions [24] dataset under two training tasks to learn visual grounding. In masked multi-modal learning, the model must reconstruct image region categories or words for masked inputs given the observed inputs. In multi-modal alignment prediction, the model must predict whether or not the caption describes the image content.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, *32*.

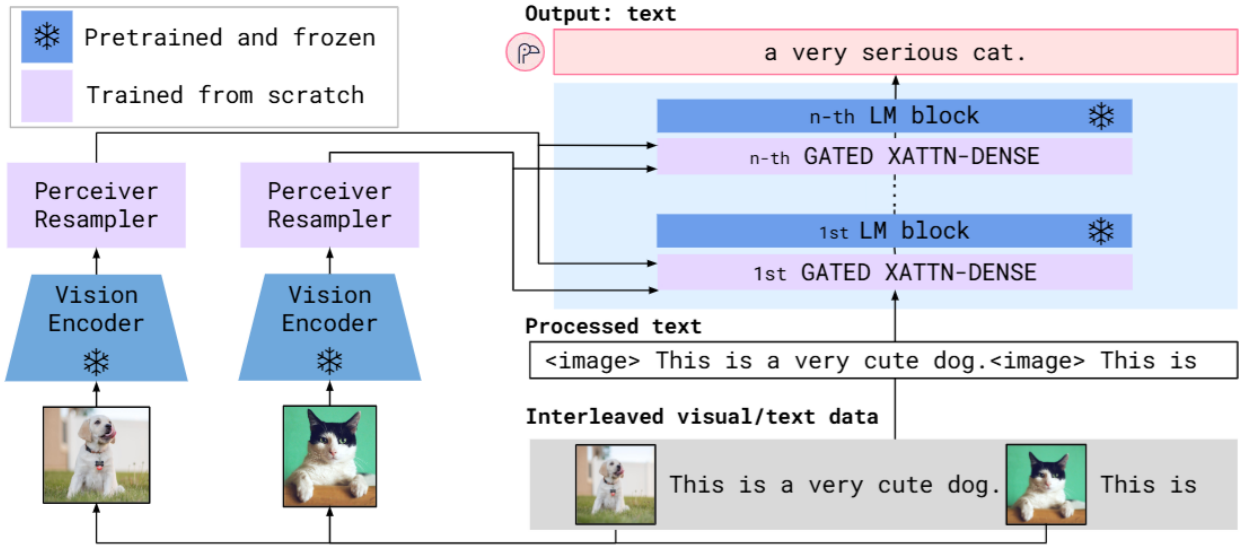
Flamingo



Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, *35*, 23716-23736.

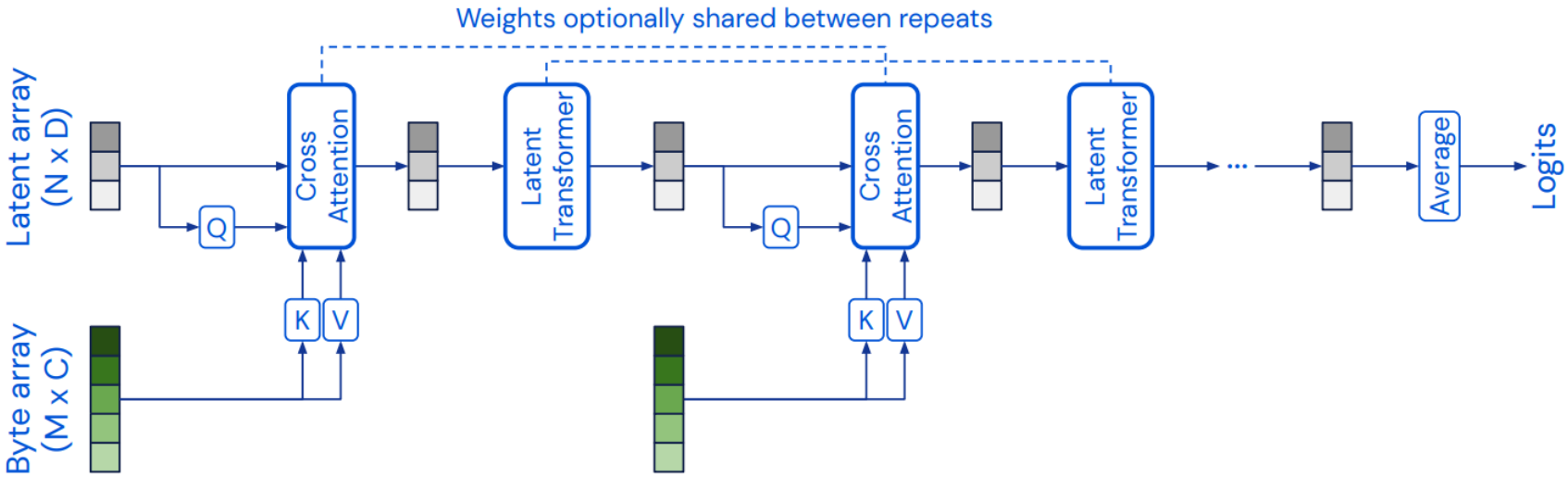
Flamingo

- Vision-language model using few-shot learning
- Inputs: Interleaved image and text tokens

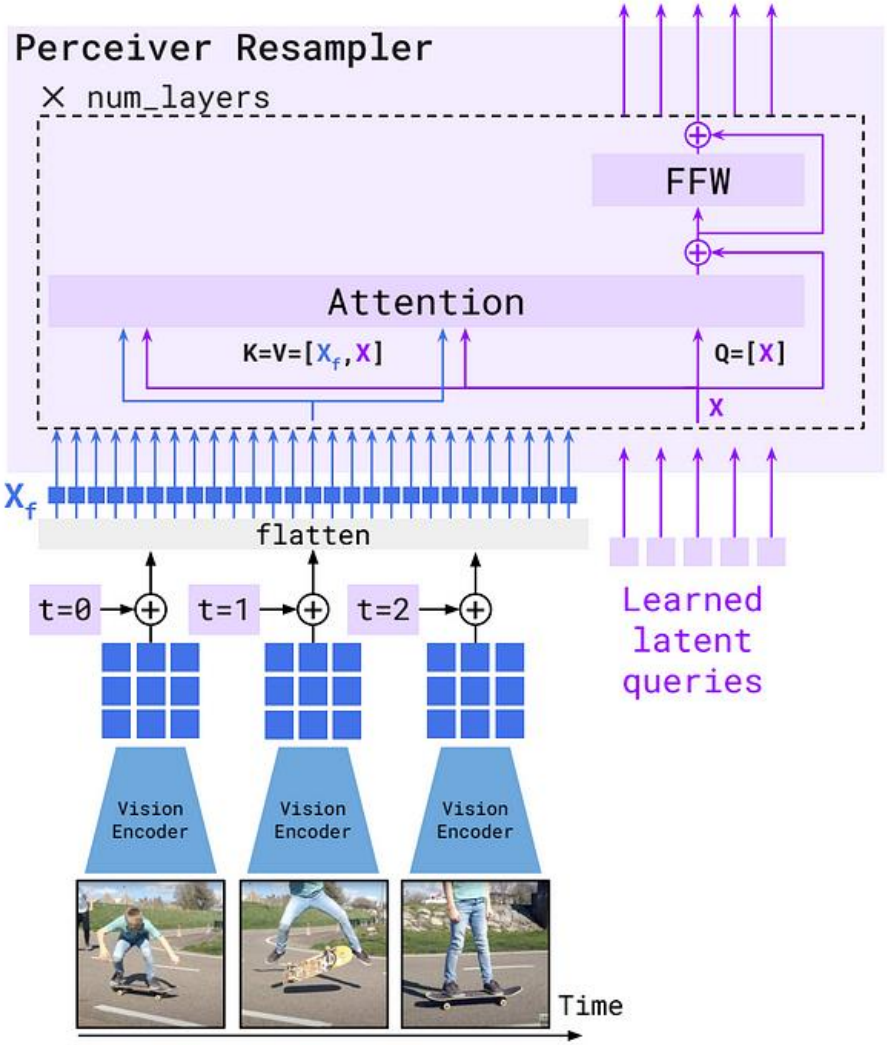


Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, *35*, 23716-23736.

Perceiver Model



Perceiver Resampler



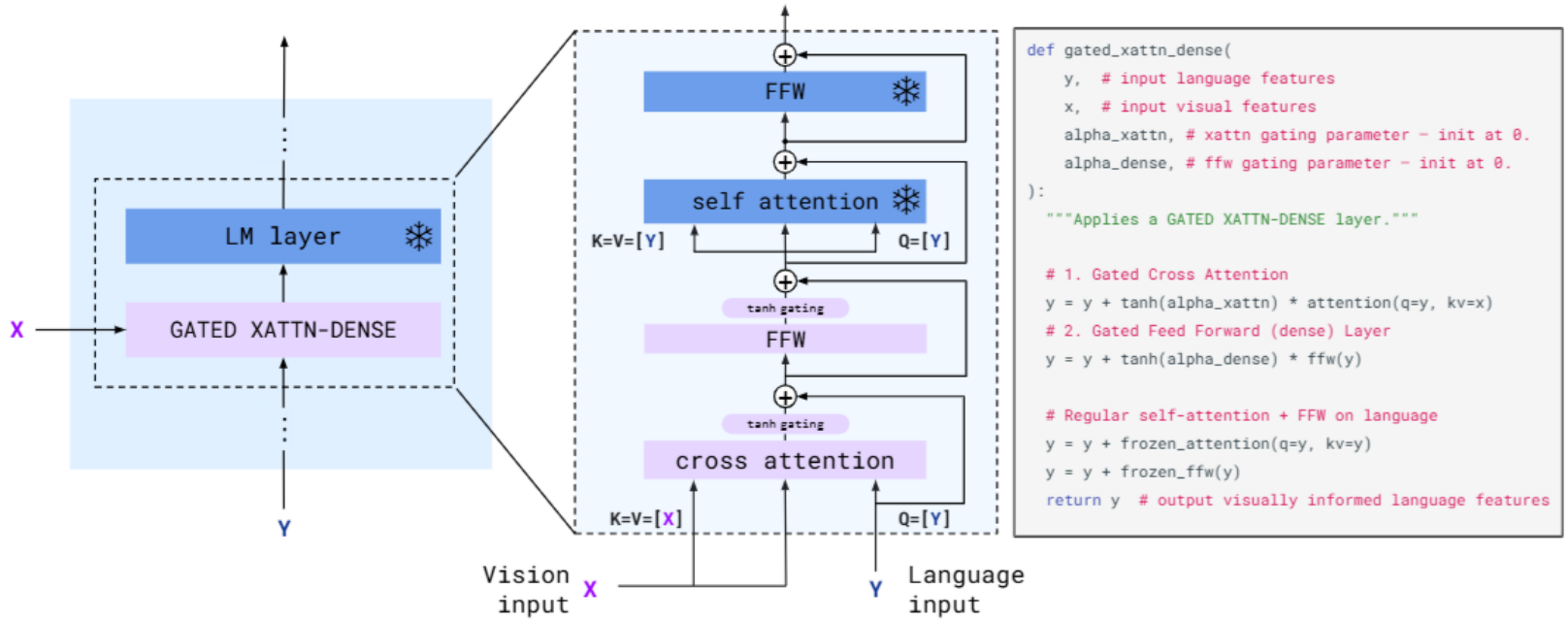

```
def perceiver_resampler(
    x_f,  # The [T, S, d] visual features (T=time, S=space)
    time_embeddings,  # The [T, 1, d] time pos embeddings.
    x,  # R learned latents of shape [R, d]
    num_layers,  # Number of layers
):
  """The Perceiver Resampler model."""

  # Add the time position embeddings and flatten.
  x_f = x_f + time_embeddings
  x_f = flatten(x_f)  # [T, S, d] -> [T * S, d]
  # Apply the Perceiver Resampler layers.
  for i in range(num_layers):
    # Attention.
    x = x + attention_i(q=x, kv=concat([x_f, x]))
    # Feed forward.
    x = x + ffw_i(x)
  return x
```

Flamingo

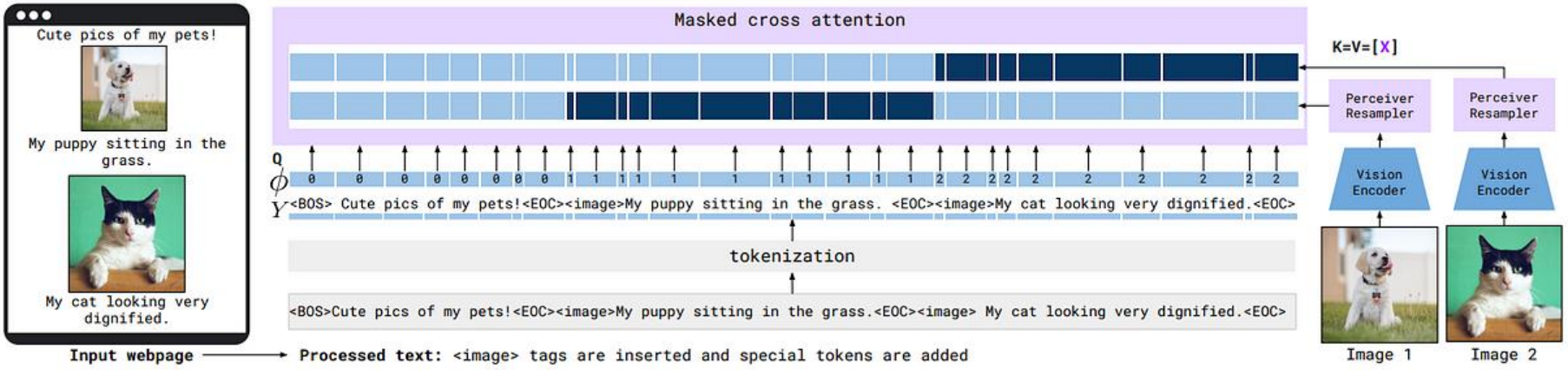
Gated Cross-Modal Attention

a. Keys and values obtained from vision features
b. Queries are derived from the language inputs



Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, *35*, 23716-23736.

Cross-attention Block Visualized

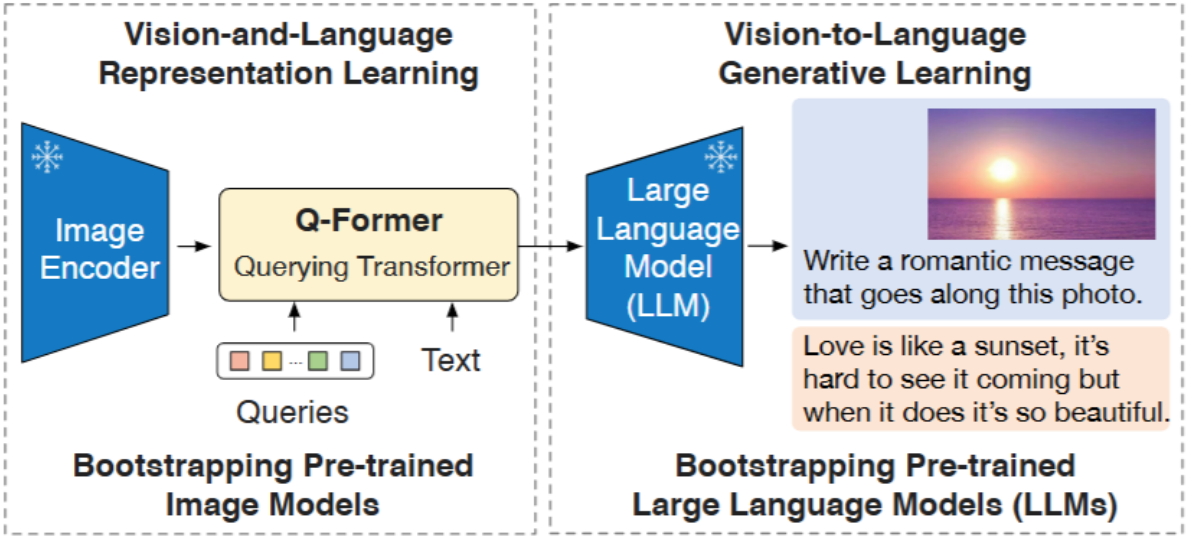


Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, *35*, 23716-23736.

BLIP (Bootstrapping Language-Image Pretraining)

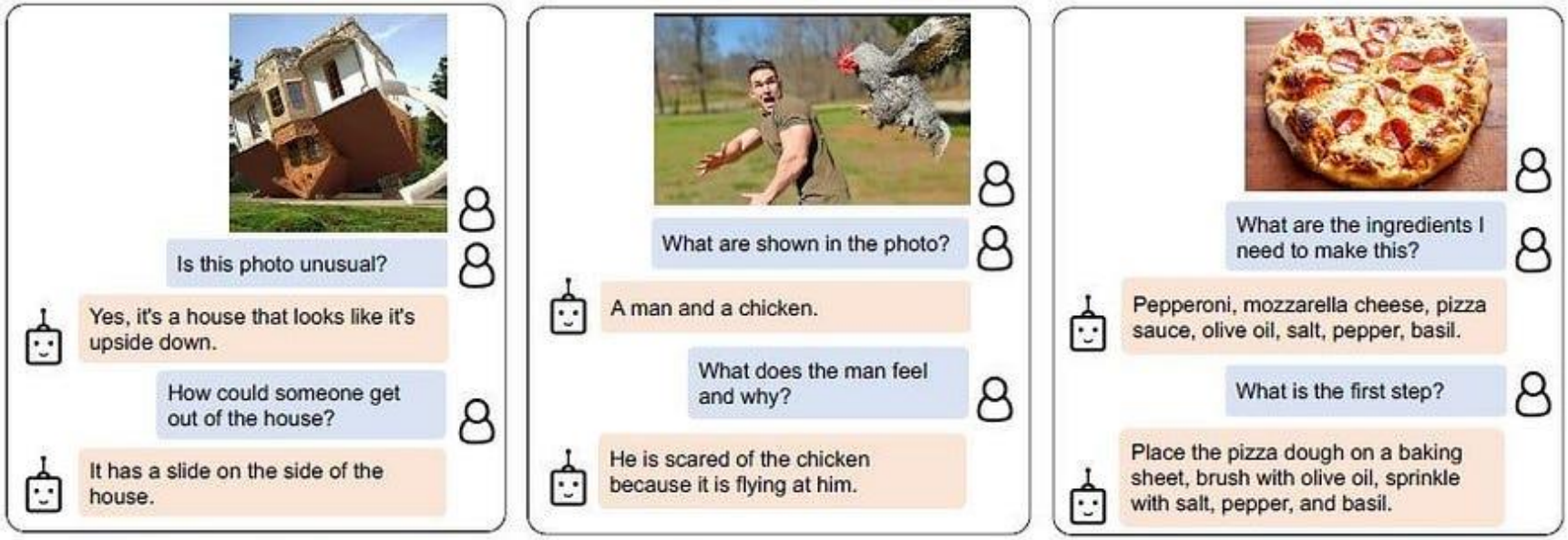
Modular architecture:

- Image encoder (ViT) - **Frozen**
- **Q-Former to query image embeddings**
- Text decoder (e.g., T5, OPT) - **Frozen**



Li, J., Li, D., Savarese, S., & Hoi, S. (2023, July). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (pp. 19730-19742). PMLR.

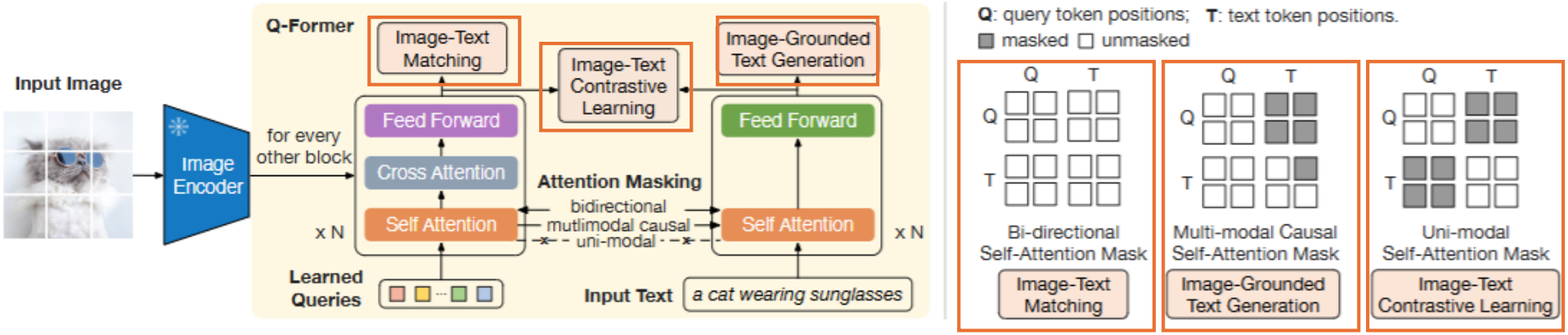
BLIP (Bootstrapping Language-Image Pretraining)



Li, J., Li, D., Savarese, S., & Hoi, S. (2023, July). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (pp. 19730-19742). PMLR.

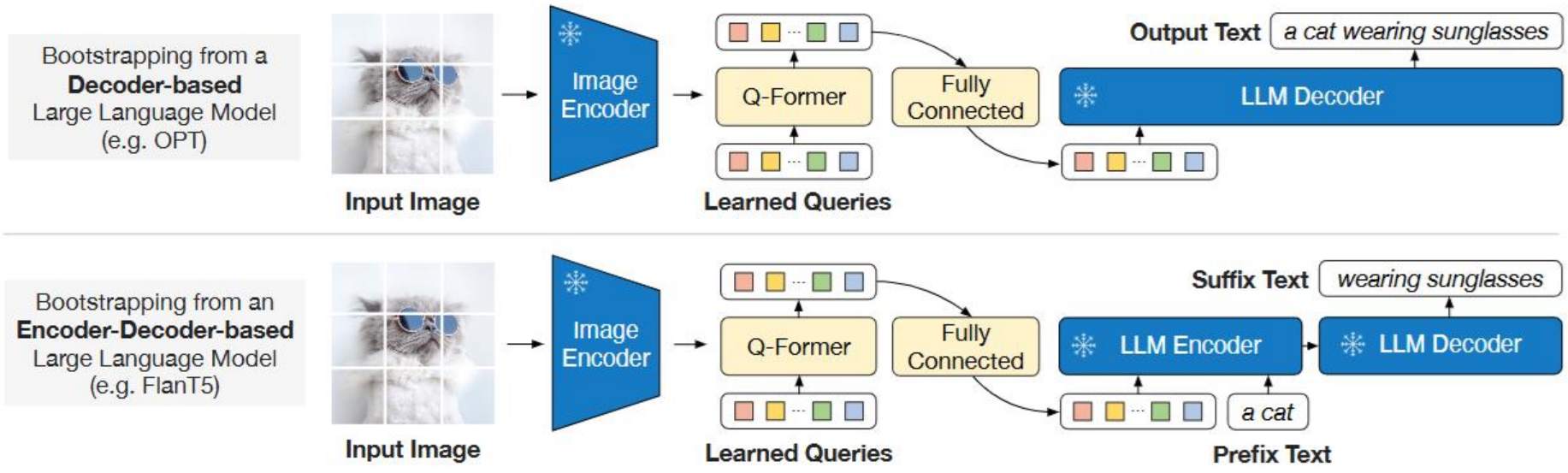
Stage 1 - Vision-Language Representation Learning

- Image-Text Contrastive Learning
- Image-grounded Text Generation
- Image-Text Matching

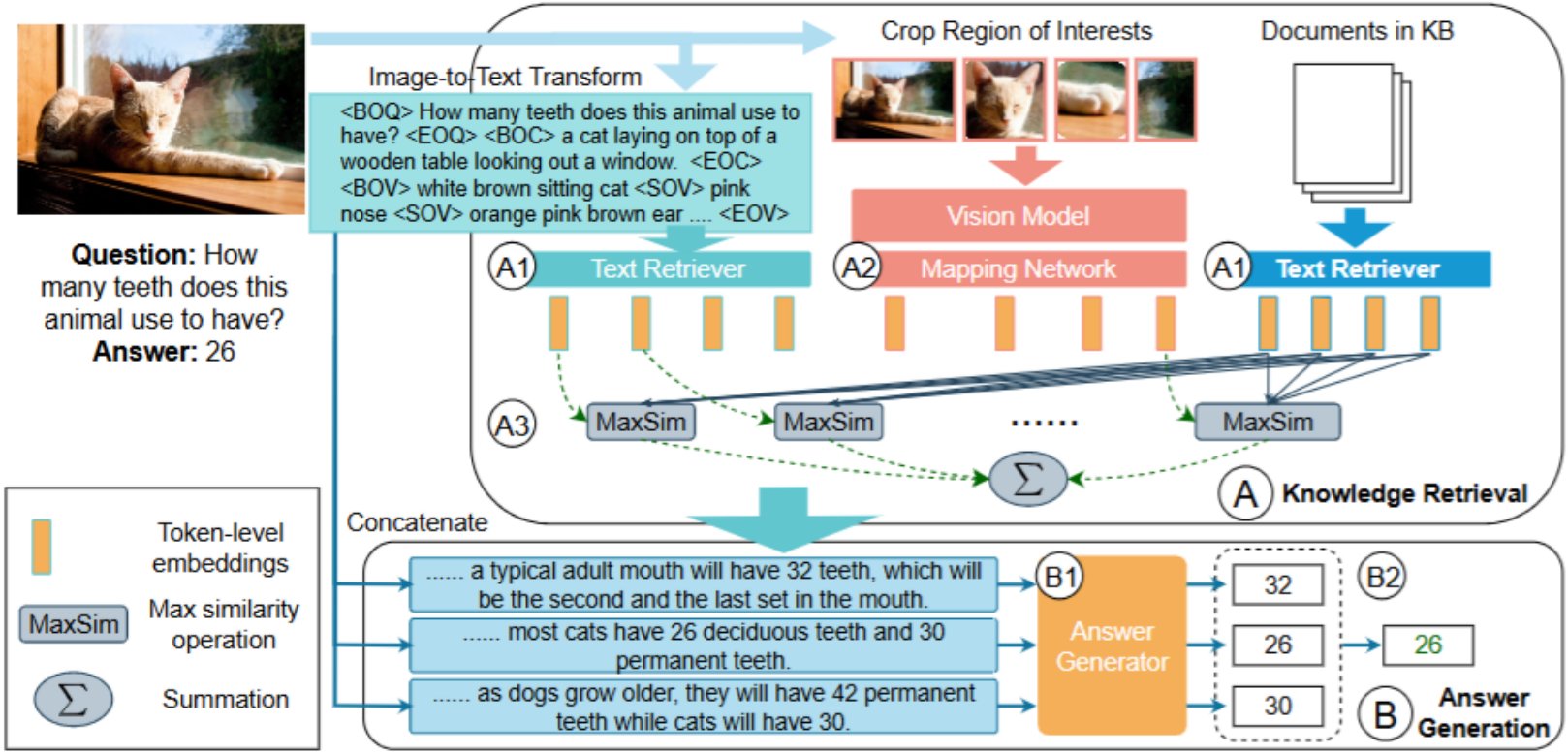


Li, J., Li, D., Savarese, S., & Hoi, S. (2023, July). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (pp. 19730-19742). PMLR.

Stage 2 - Vision-Language Generative Learning



Fine-grained Late-interaction Multi-modal Retrieval (FLMR)



Lin, W., Chen, J., Mei, J., Coca, A., & Byrne, B. (2023). Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, *36*, 22820-22840.

Fine-grained Late-interaction Multi-modal Retrieval (FLMR)

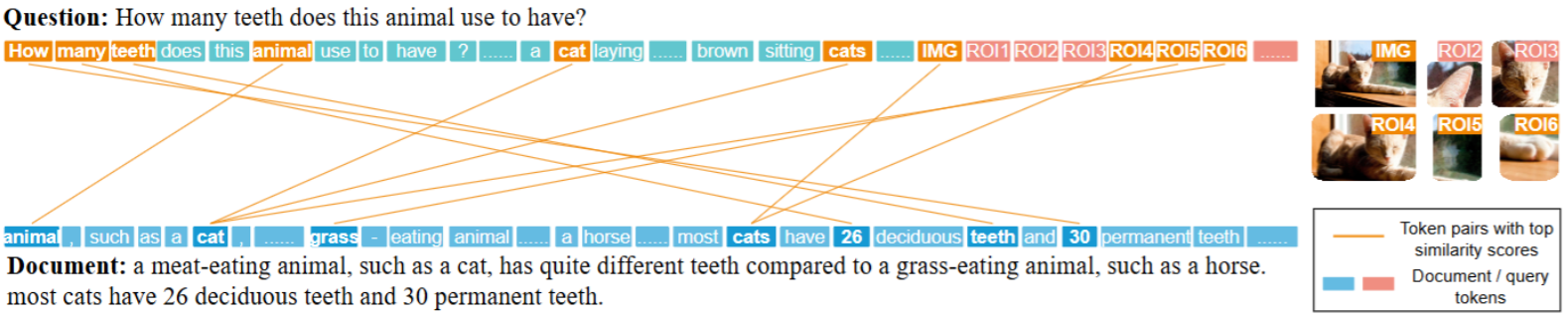
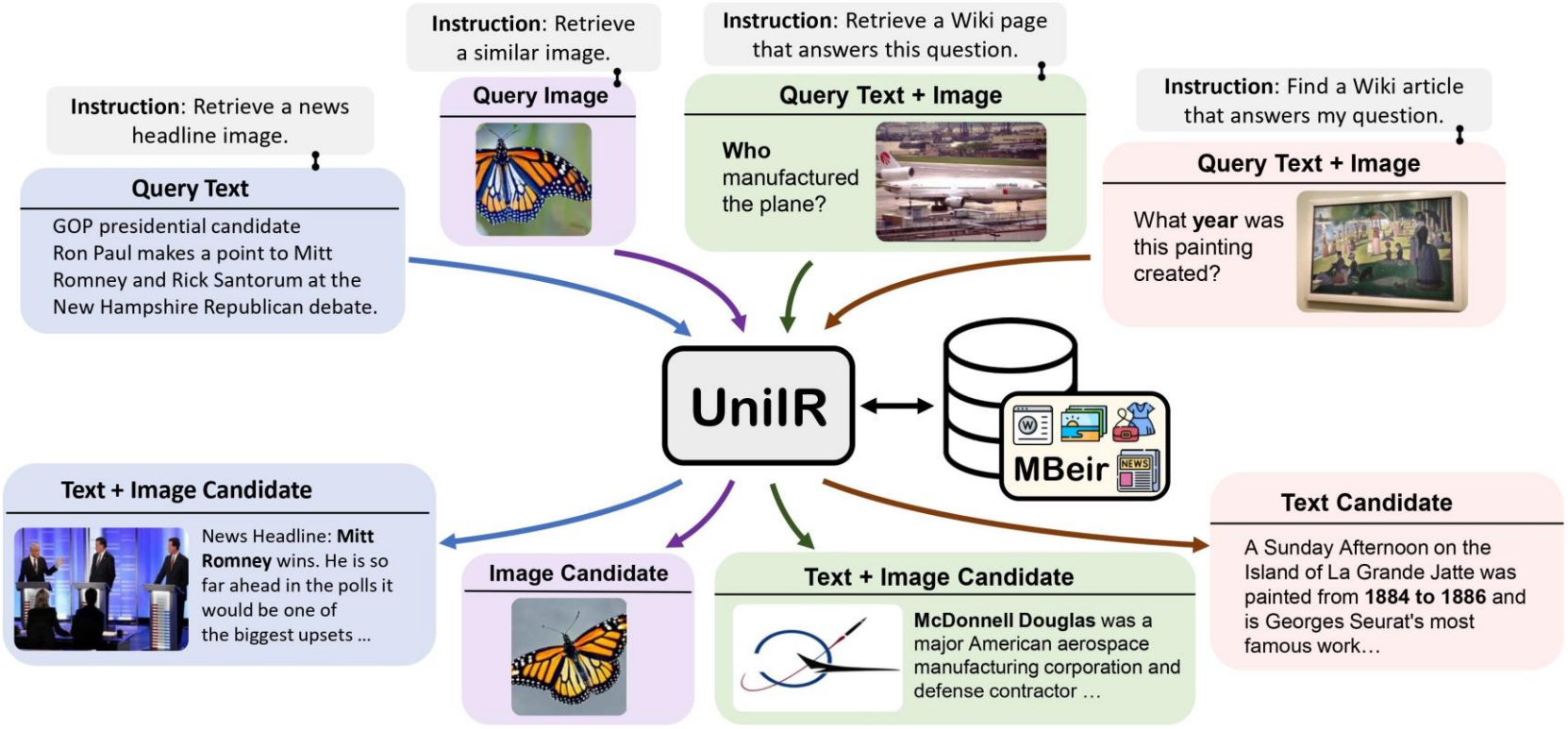


Figure 3: Selected query tokens connected by document tokens that have the highest token-level relevance with them, as computed by FLMR. For example, amongst all document tokens, '26' and '30' have the highest relevance with the query token 'how' and 'many', respectively. This shows that FLMR can capture fine-grained document relevance. Zoom in for better visualization.

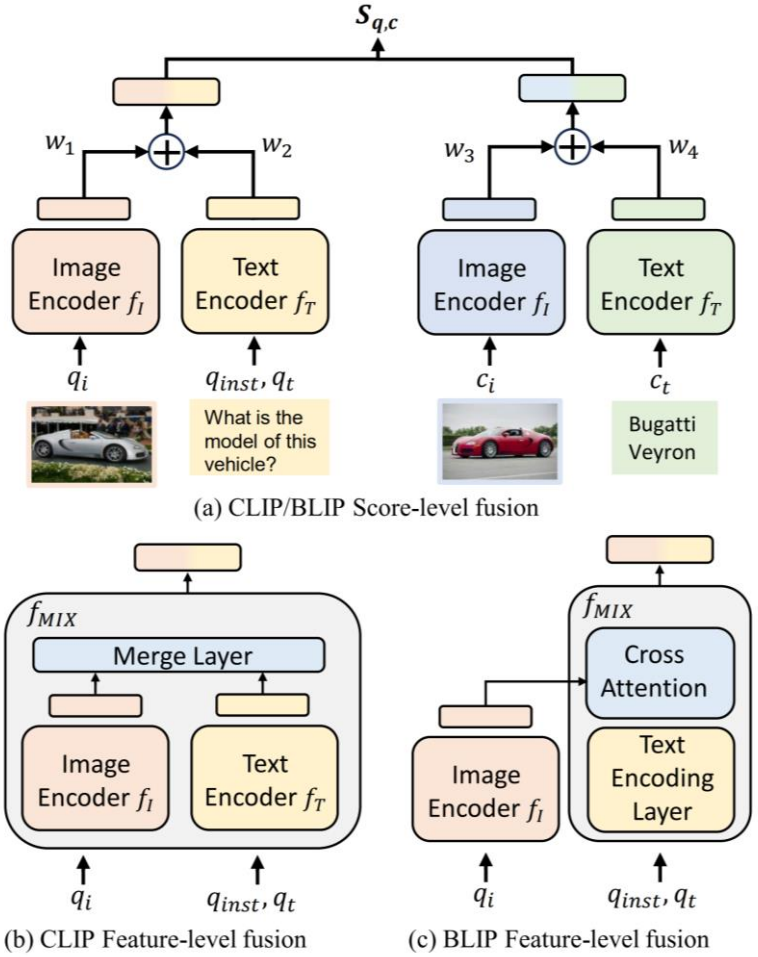
Lin, W., Chen, J., Mei, J., Coca, A., & Byrne, B. (2023). Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, *36*, 22820-22840.

Universal Multimodal Information Retrievers (UniIR)



Wei, C., Chen, Y., Chen, H., Hu, H., Zhang, G., Fu, J., ... & Chen, W. (2024, September). Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision* (pp. 387-404). Cham: Springer Nature Switzerland.

Universal Multimodal Information Retrievers (UniIR)



Fusion

- Score-level
- Feature-level

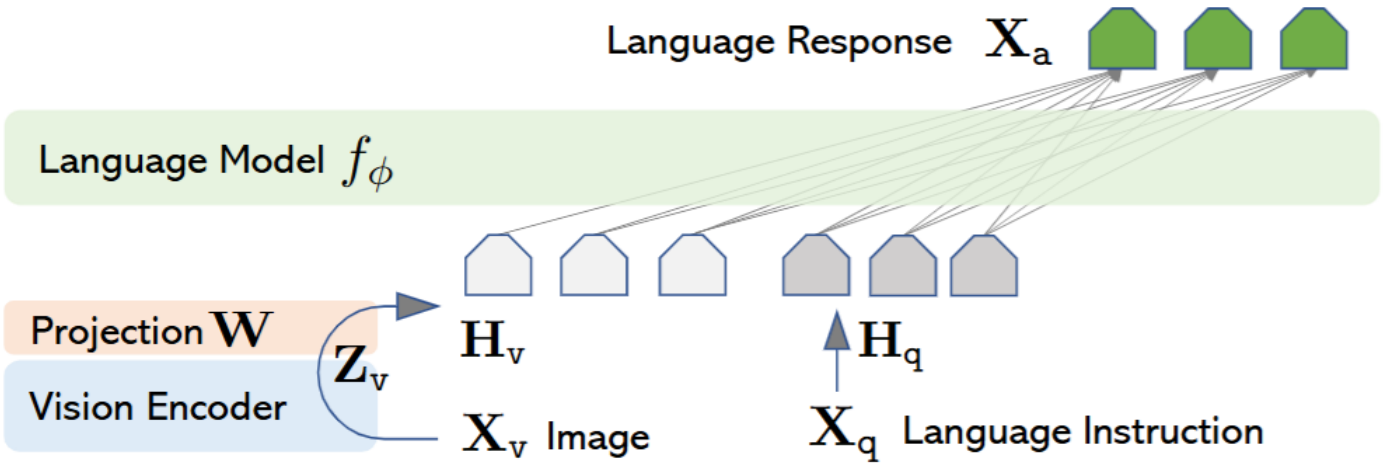
Backbone

- BLIP
- CLIP

Wei, C., Chen, Y., Chen, H., Hu, H., Zhang, G., Fu, J., ... & Chen, W. (2024, September). Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision* (pp. 387-404). Cham: Springer Nature Switzerland.

LLaVA (Large Language and Vision Assistant)

- Combines CLIP-ViT image encoder + Vicuna LLM
- Only trains lightweight projection from vision to language
- **Uses instruction tuning with image-prompt-response pairs**

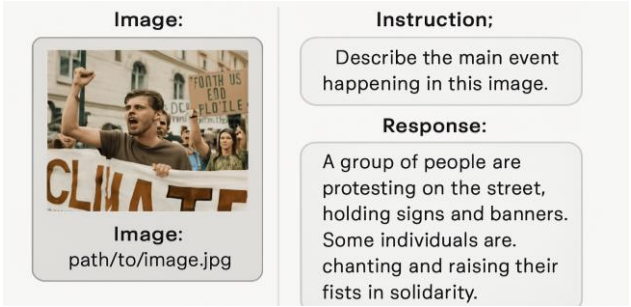


Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *Advances in neural information processing systems*, *36*, 34892-34916.

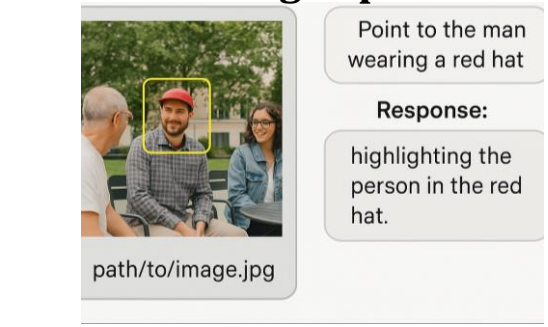
Instruction Tuning

- Fine-tune on natural language instructions
- Input: [Image] + instruction → Output: desired response
- Enables general-purpose assistant-like behavior

Task: Situation Description



Task: Referring Expression Ground



Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.

Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

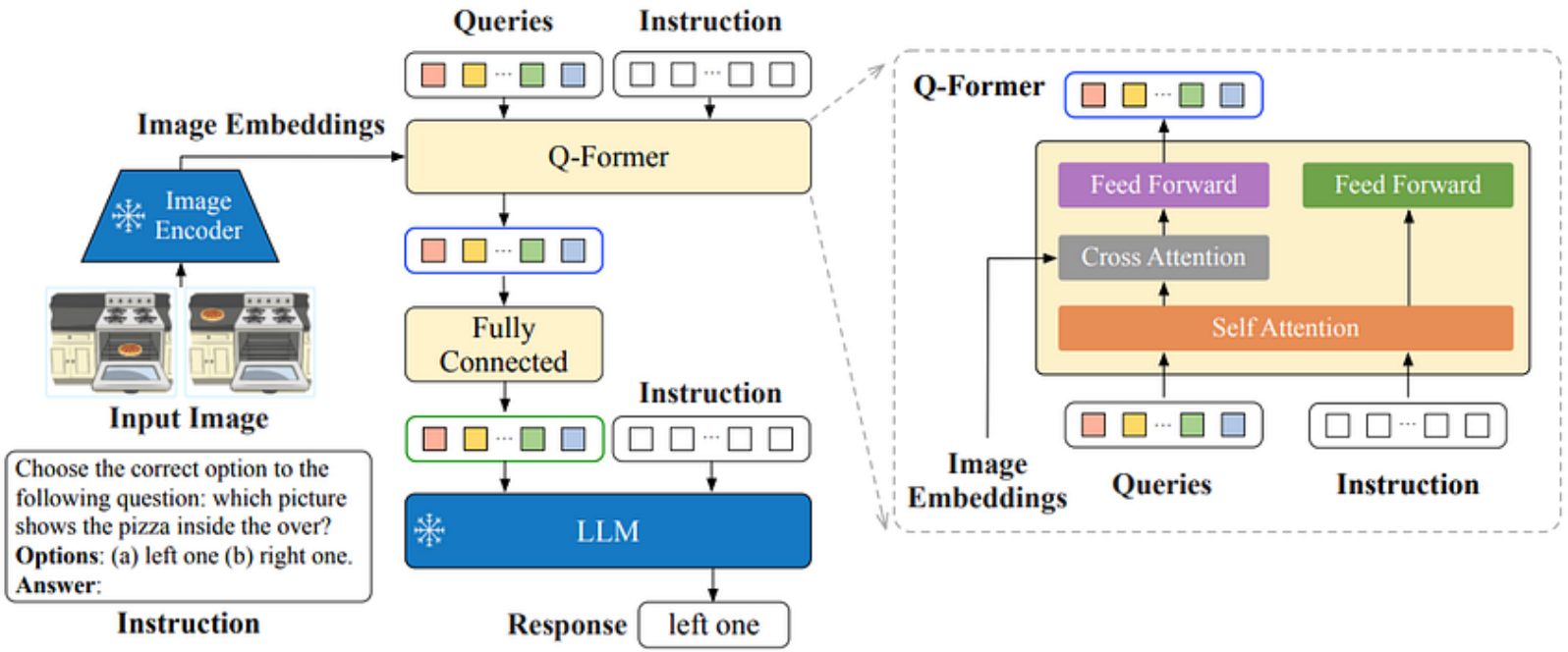
Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>
Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *Advances in neural information processing systems*, *36*, 34892-34916.

How Instruction Tuning Changed Training

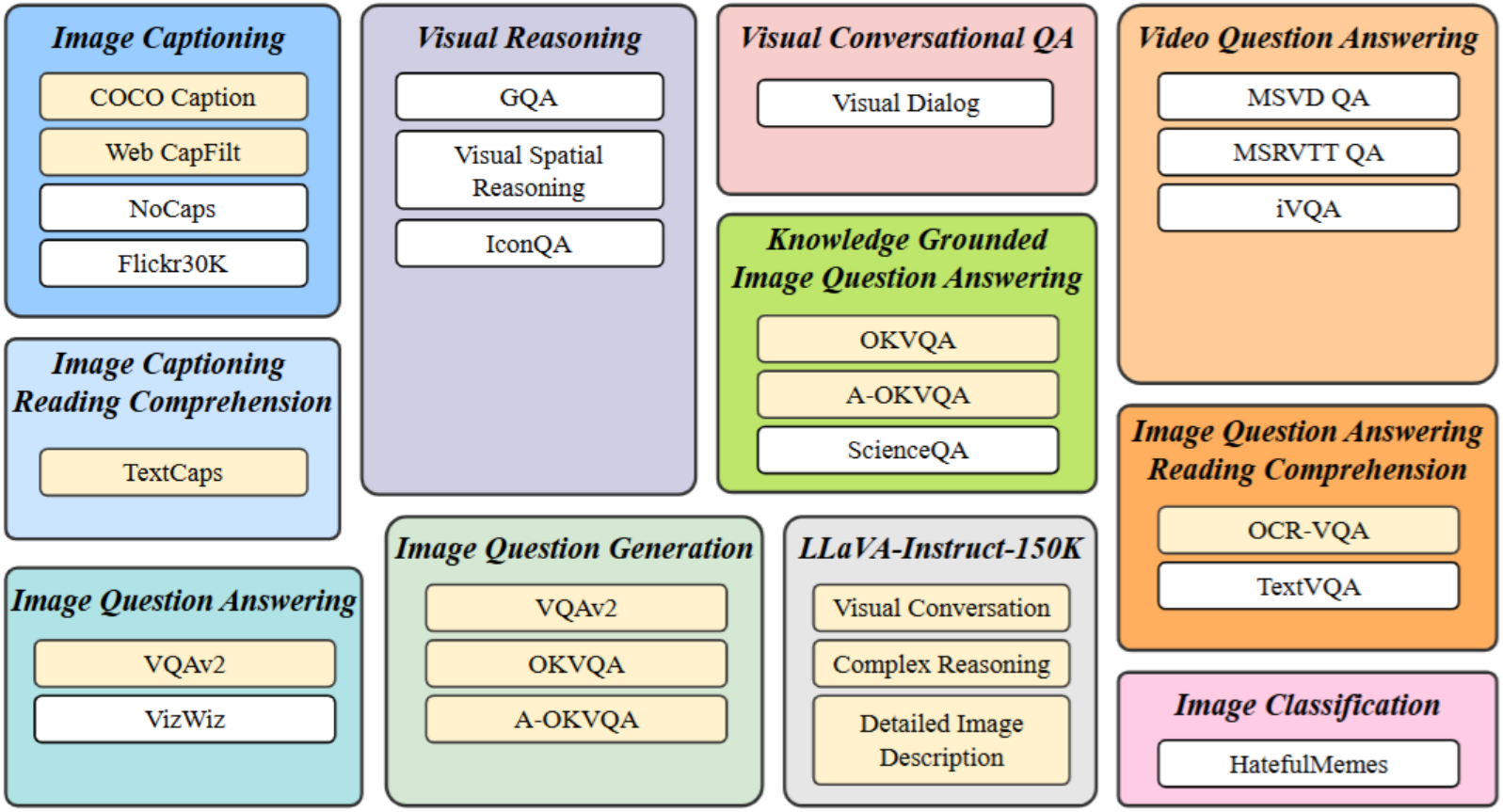
- Shift from task-specific training to general-purpose
- Training data: instruction-response pairs
- Architecture: frozen vision encoder + pretrained LLM
- Encourages reusable models with prompt-based control

InstructBLIP



Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., ... & Hoi, S. (2023). Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, *36*, 49250-49267.

InstructBLIP



Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., ... & Hoi, S. (2023). Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, *36*, 49250-49267.

Qwen3-VL

Shift in Paradigm

- Traditional IR: $retrieve \rightarrow return$
- Qwen3-VL: $retrieve + reason + generate$

Instruct vs Thinking editions

- **Instruct:** Aligned for general use, faster, and cheaper; ideal for most practical applications.
- **Thinking:** Adds internal reasoning traces and better compositional skills, helpful for STEM problems, complex diagrams, and multi-step analysis.

Qwen3-VL is a generative multimodal model that treats images as language-conditioned inputs, enabling unified reasoning, retrieval, and interaction.

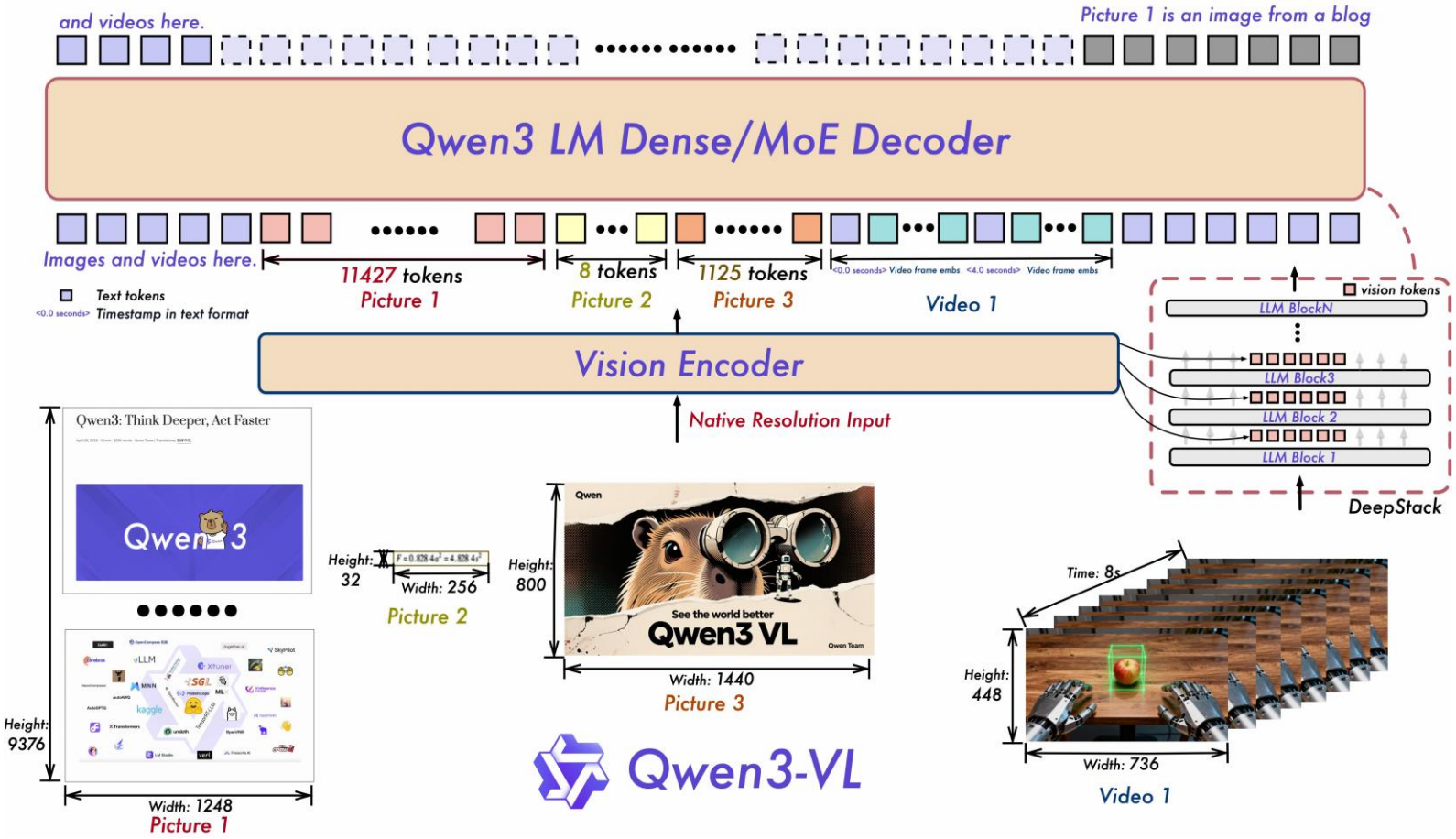
Qwen3-VL

- Existing systems
  - CLIP → retrieval only
  - LLaVA → reasoning but limited grounding
  - OCR systems → text-only

- Build a unified model that can
  - understand images
  - read text inside images (OCR)
  - reason over multimodal inputs
  - follow instructions

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., ... & Zhu, K. (2025). Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Qwen3-VL


and videos here.
Picture 1 is an image from a blog
Qwen3 LM Dense/MoE Decoder
Images and videos here.
11427 tokens
Picture 1
8 tokens
Picture 2
1125 tokens
Picture 3
Video 1
Text tokens
Timestamp in text format
Vision Encoder
Native Resolution Input
vision tokens
LLM BlockN
LLM Block3
LLM Block 2
LLM Block 1
DeepStack
Qwen3: Think Deeper, Act Faster
Qwen 3
Height: 9376
Width: 1248
Picture 1
Height: 32
Width: 256
Picture 2
Height: 800
See the world better
Qwen3 VL
Width: 1440
Picture 3
Qwen3-VL
Time: 8s
Height: 448
Width: 736
Video 1

Qwen3-VL

- **Long-context multimodality:** Processes interleaved text, images, and long videos with memory-efficient attention and improved positional handling across time and space.
- **Strong OCR and document intelligence:** Reads dense pages, noisy scans, and complex layouts with better robustness to blur, tilt, and low light.
- **Spatial and temporal reasoning:** Locates entities precisely, reasons about occlusion and viewpoint, and grounds video answers to timestamps.
- **Agentic UI interaction:** Perceives GUI elements, understands their functions, and can plan steps to complete tasks across desktop or mobile interfaces.

WebQA

Figure 1. Example WEBQA dataset pipeline in which the question requires finding and reasoning about two relevant sources and discarding distractors to produce the correct natural language answer.

Chang, Y., Narang, M., Suzuki, H., Cao, G., Gao, J., & Bisk, Y. (2022). Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16495-16504).

Multimodal QA (MMQA)

Multimodal Context

Steal This Movie!
The film follows Hoffman's (D'Onofrio) relationship with his second wife Anita (Garofalo) and their "awakening" and subsequent conversion to an activist life. The title of the film is a play on Hoffman's 1970 counter-culture guidebook titled "Steal This Book".

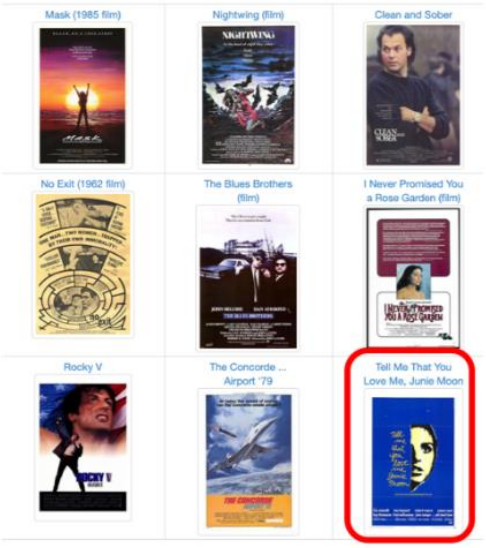
La liceale
La liceale (internationally released as The Teasers, "Under-graduate Girls", "Sophomore Swingers" and "Teasers") is a 1975 commedia sexy all'italiana directed by Michele Massimo Tarantini.
...
Guida. It was followed by "La liceale nella classe dei ripetenti".

Sage Stallone
Stallone made his acting debut alongside his father in Rocky V (1990), the fifth installment of the Rocky franchise, playing Robert Balboa Jr., the onscreen son of his father's title character. He did not, however,
...
After that, he acted in lesser profile films.

Pierino contro tutti
Pierino contro tutti (also known as "Desirable Teacher") is a 1981 comedy film directed by Marino Girolami. The main character of the film is Pierino, an
...
l as a short lived subgenre of joke-films in which the plot basically consists of a series of jokes placed side by side.

Ben Piazza - Filmography

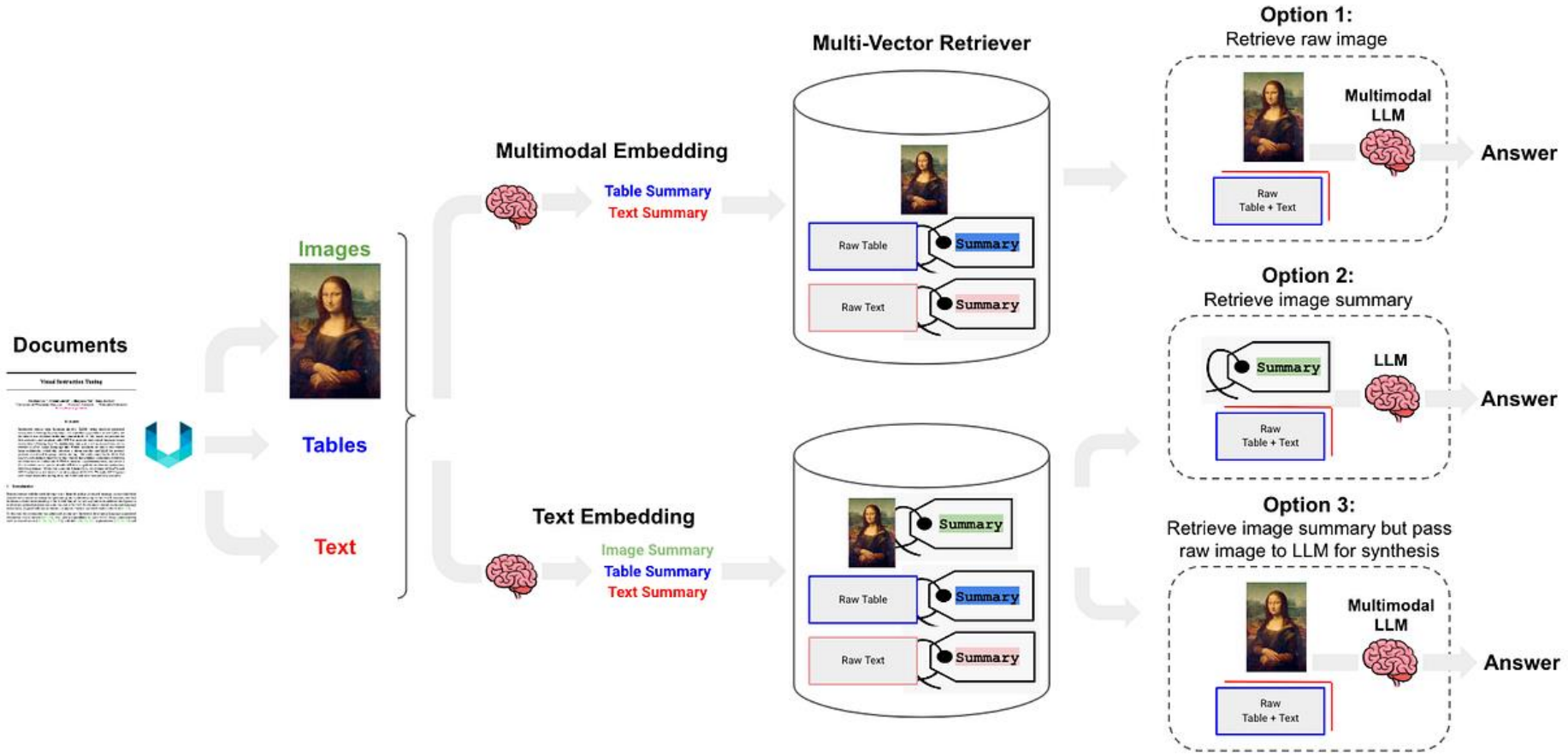
Year	Title	Role
1957	A Dangerous Age	David
1959	The Hanging Tree	Rune
1962	No Exit	Camarero
1970	**Tell Me That You Love Me, Junie Moon**	Jesse
1972	The Outside Man	Desk Clerk
...	...	...
1985	Mask	Mr. Simms
1988	Clean and Sober	Kramer
1990	**Rocky V**	Doctor
1991	Guilty by Suspicion	Darryl Zanuck



Q: Which B. Piazza title came earlier: the movie S. Stallone's son starred in or the movie with half of a lady's face on the poster?
A: Tell Me That You Love Me, Junie Moon

Talmor, A., Yoran, O., Catav, A., Lahav, D., Wang, Y., Asai, A., ... & Berant, J. (2021). Multimodalqa: Complex question answering over text, tables and images. *ICLR 2021*

Multimodal RAG



https://blog.langchain.com/semi-structured-multi-modal-rag/

Multimodal RAG

Joint Embedding and Retrieval

- Utilize models like **CLIP** (Contrastive Language-Image Pre-training) to create unified embeddings for both text and images.
- Implement approximate nearest neighbor search using libraries like **FAISS** for efficient retrieval.
- Feed retrieved multimodal content (raw images and text chunks) to a multimodal LLM such as LLaVa, Qwen-VL for answer generation.

Multimodal RAG

Image-to-Text Conversion

- Generate Summaries from images using models like **LLaVA**
- Use text-based embedding models like **Sentence-BERT** to create embeddings for both original text and image captions.
- Pass the text chunks to an LLM for final answer synthesis.

Multimodal RAG

Hybrid Retrieval with Raw Image Access

- Employ a multimodal LLM to produce text summaries from images.
- Embed and retrieve these summaries with references to raw images, alongside other textual chunks.
- Multi-Vector Retriever with vector databases like Chroma, Milvus to store raw text and images along with their summaries for retrieval.
- For final answer generation, use multimodal models like **Pixtral 12B**, **LLaVa**, **GPT-4V**, **Qwen-VL** that can process both text and raw image inputs simultaneously.