Introduction to Large Vision + Language Models

Presnted by: XuDong Wang

09/09, 2024

CS 294-43 Fall 2024

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Berkeley
UNIVERSITY OF CALIFORNIA

Content Writing

Teaching and Education

Analytics

Market Research

Email Marketing

Social Media Management

SEO and Website Optimization

Chatbot Development

Multimodal Intelligence

Text

Image/Video

Audio

Touch

Smell

Taste

Gemini

ChatGPT-4o

Claude 3

Vision Language Model

Object Localization

Is one cat behind another?

Segmentation

Segment: striped cat

Visual QA

What is the breed of these cats?

Learning w/ Instructions

Stripped cats are called tabby cats. What's the breed of the cats in the image?

Vision Language Model

Yes, one cat is behind the other in the image. The cat in the back us facing the camera, while the cat in front it is facing away from the camera.

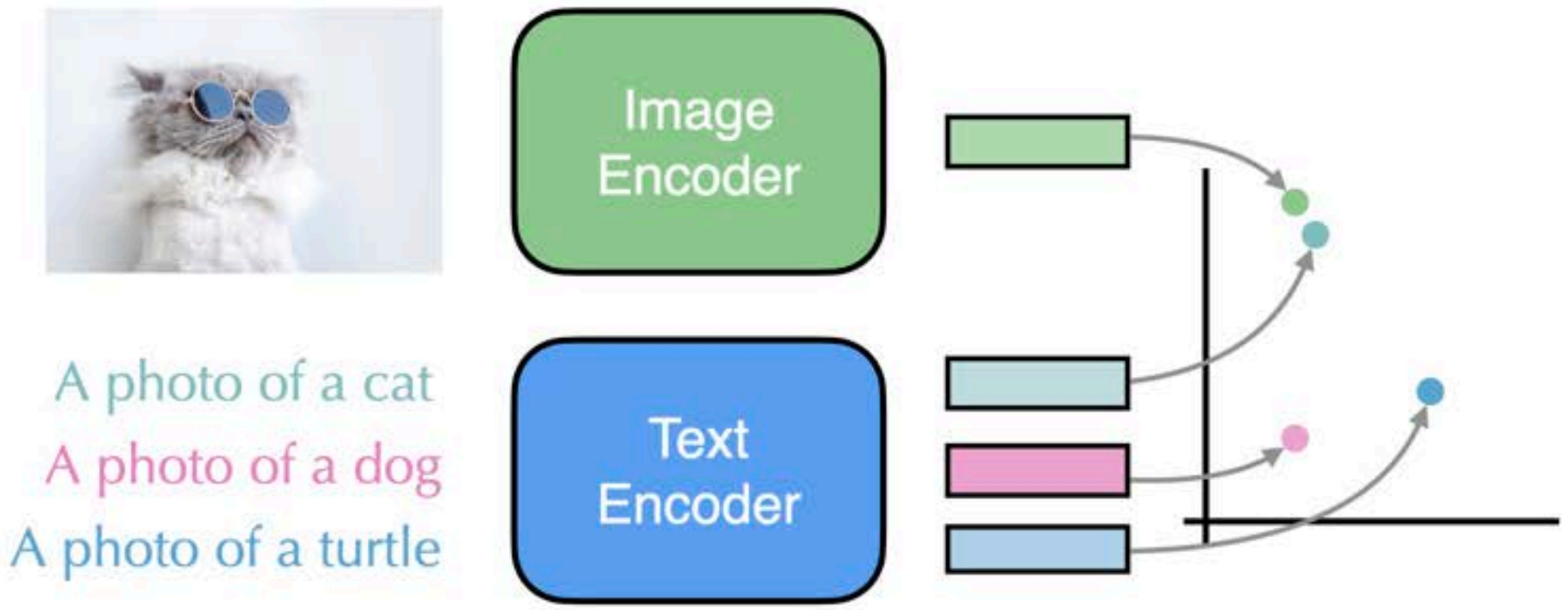
The cats in the image appear to be domestric shorthair

The cats in the image are tabby cats. Tabby cats are a common domestic cat bread and are characterized by their distinctive coat pattern, stripes on the body and a ringed tail.

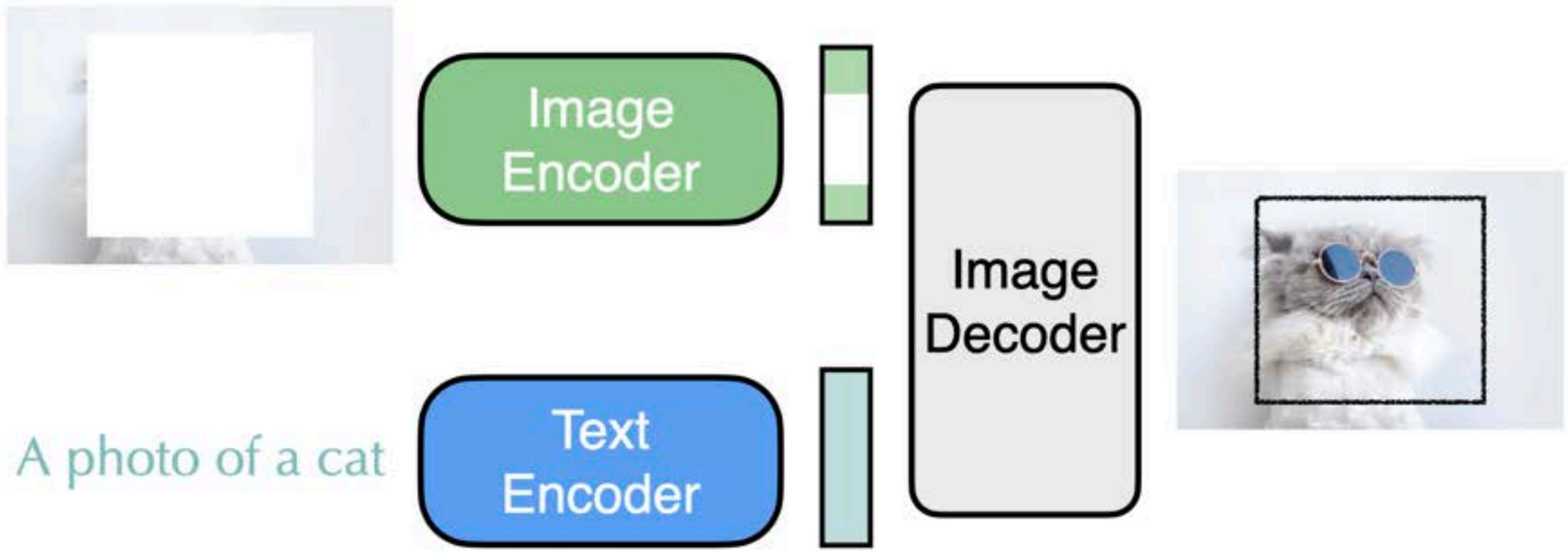
What Will Be Covered Today?

1. Families of VLMs
2. VLM Training
3. VLM Evaluation

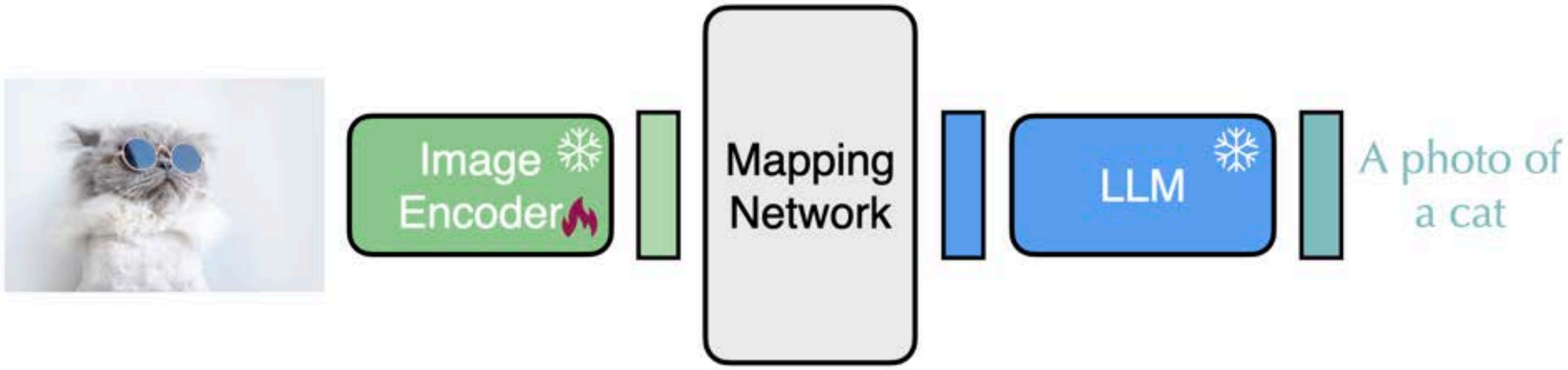
The Families of VLMs



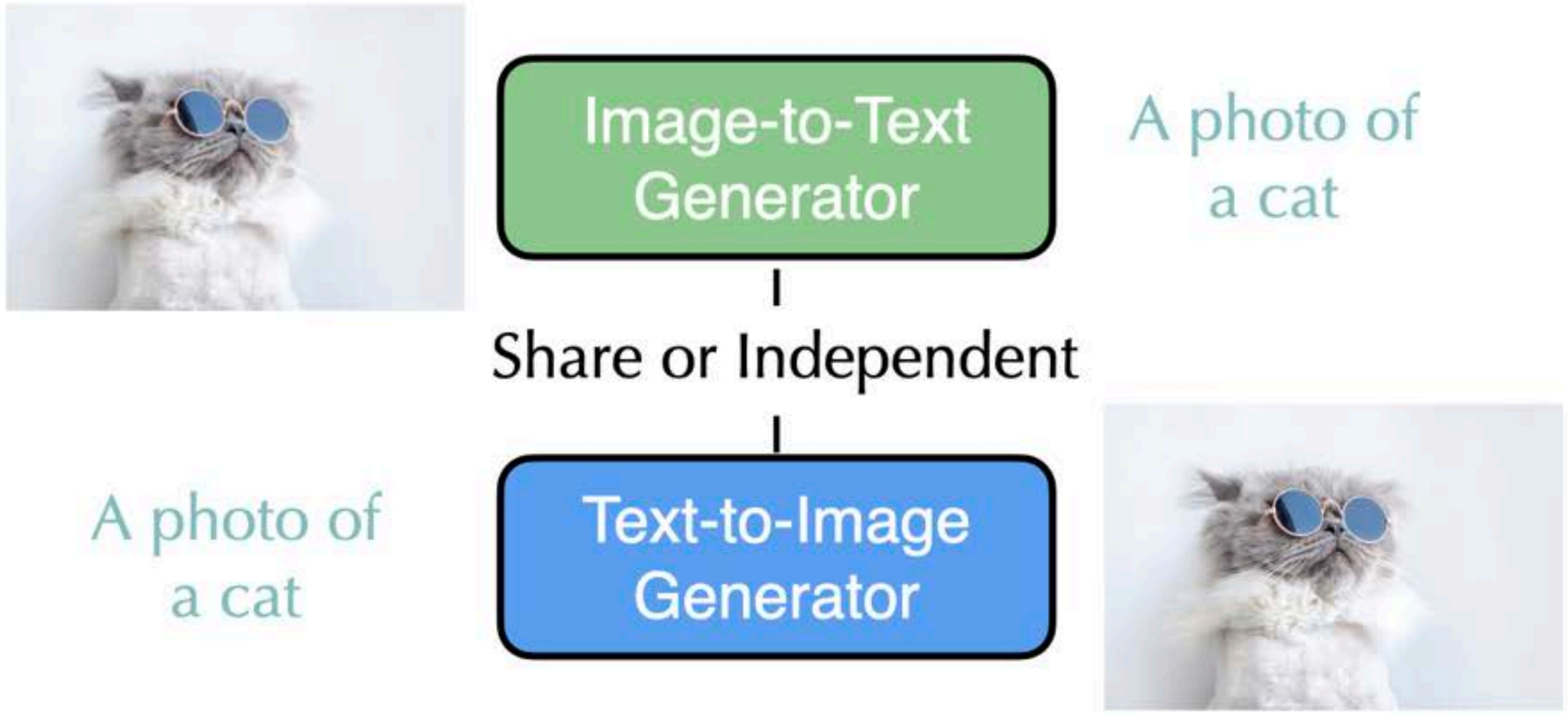
Contrastive-Based



Masking Objective



VLMs from Pretrained Backbones



Generative-Based

The Families of VLMs: 1) Contrastive-Based

CLIP [ICML 2021]: InfoNCE Loss

$$\mathcal{L} = -\frac{1}{2N}\sum_{i=1}^{N}\left(\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{N}e^{t\mathbf{x}_i\cdot\mathbf{y}_j}}}^{\text{image}\rightarrow\text{text softmax}} + \overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{N}e^{t\mathbf{x}_j\cdot\mathbf{y}_i}}}^{\text{text}\rightarrow\text{image softmax}}\right)$$

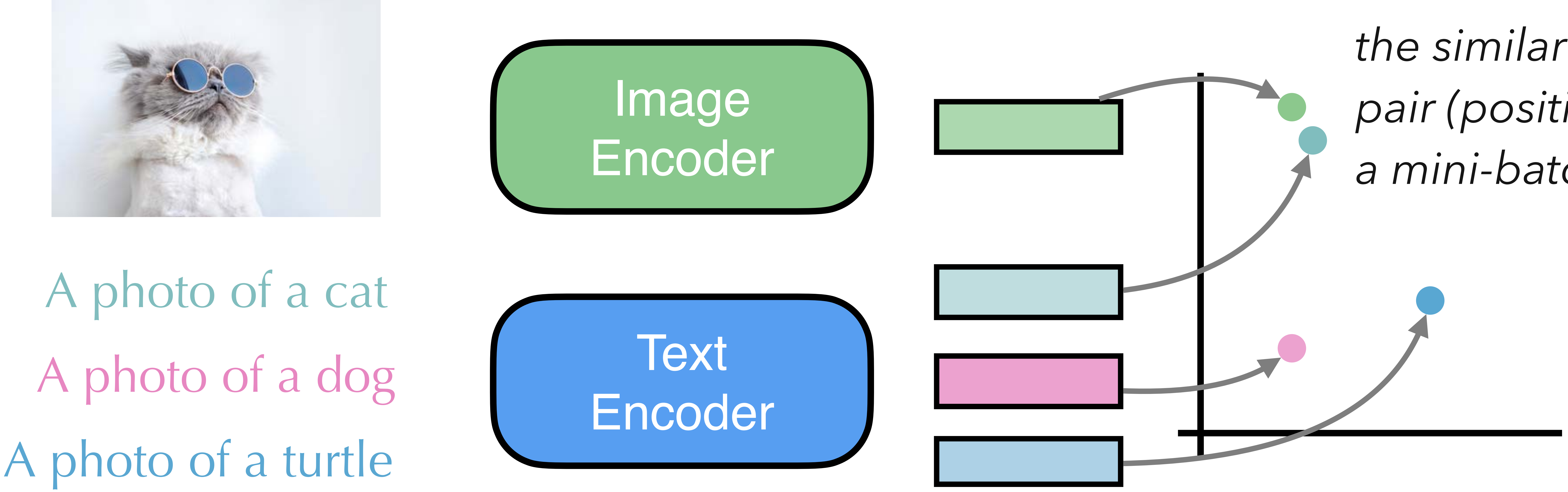
Every positive pair is normalized by **all** negative pairs

SigLIP [ICCV 2023]: Sigmoid Loss

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\log\frac{1}{1+e^{z_{ij}(t\mathbf{x}_i\cdot\mathbf{y}_j+b)}} \quad \text{s.t.} \quad z_{ij} = \begin{cases}1, & \text{for positive pairs.}\\ -1, & \text{for negative pairs.}\end{cases}$$

Every pair (positive/negative) is independent of other pairs

both CLIP and SigLIP compute the similarity between every pair (positives/negatives) within a mini-batch.



A photo of a cat

A photo of a dog

A photo of a turtle

Image Encoder

Text Encoder

The Families of VLMs: 2) Masking Objective - MaskVLM [1]

Fill the masked caption given an unmasked image

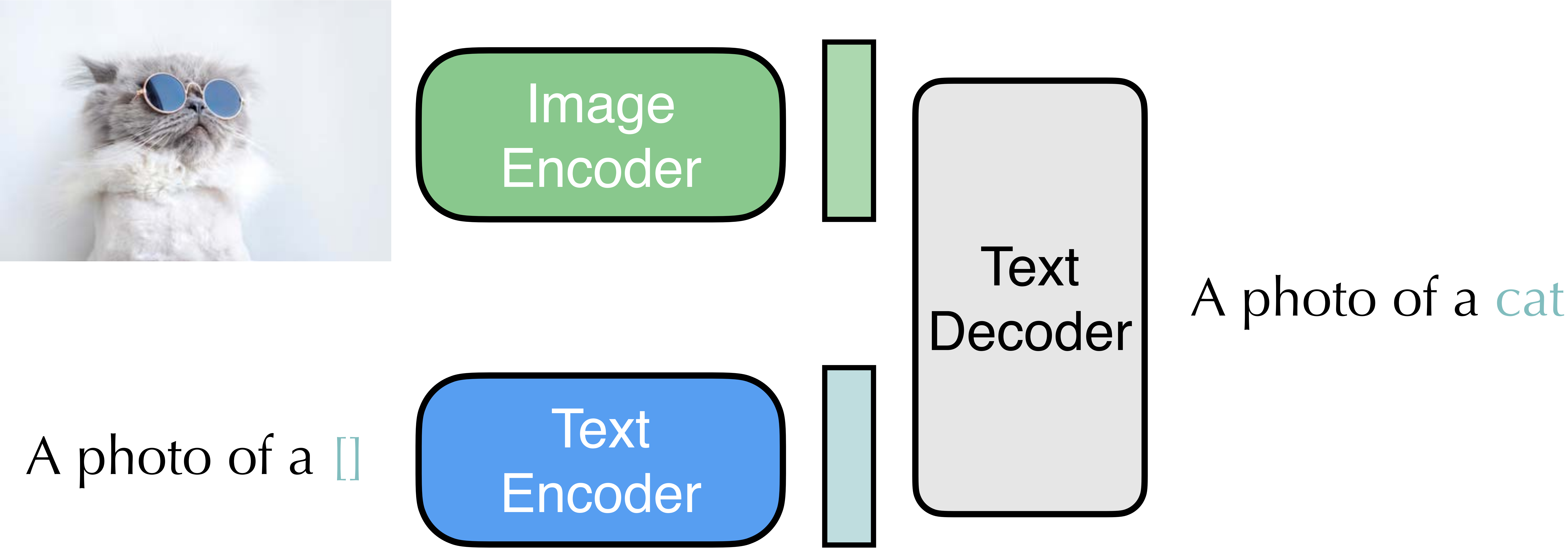


Image Encoder

Text Encoder

Text Decoder

A photo of a []

A photo of a cat

---

[1] Kwon, Gukyeong, et al. "Masked Vision and Language Modeling for Multi-modal Representation Learning." *ICLR 2023*

[2] Singh, Amanpreet, et al. "Flava: A foundational language and vision alignment model." *CVPR 2022*

The Families of VLMs: 2) Masking Objective - MaskVLM [1]

Fill the masked image given an unmasked text caption

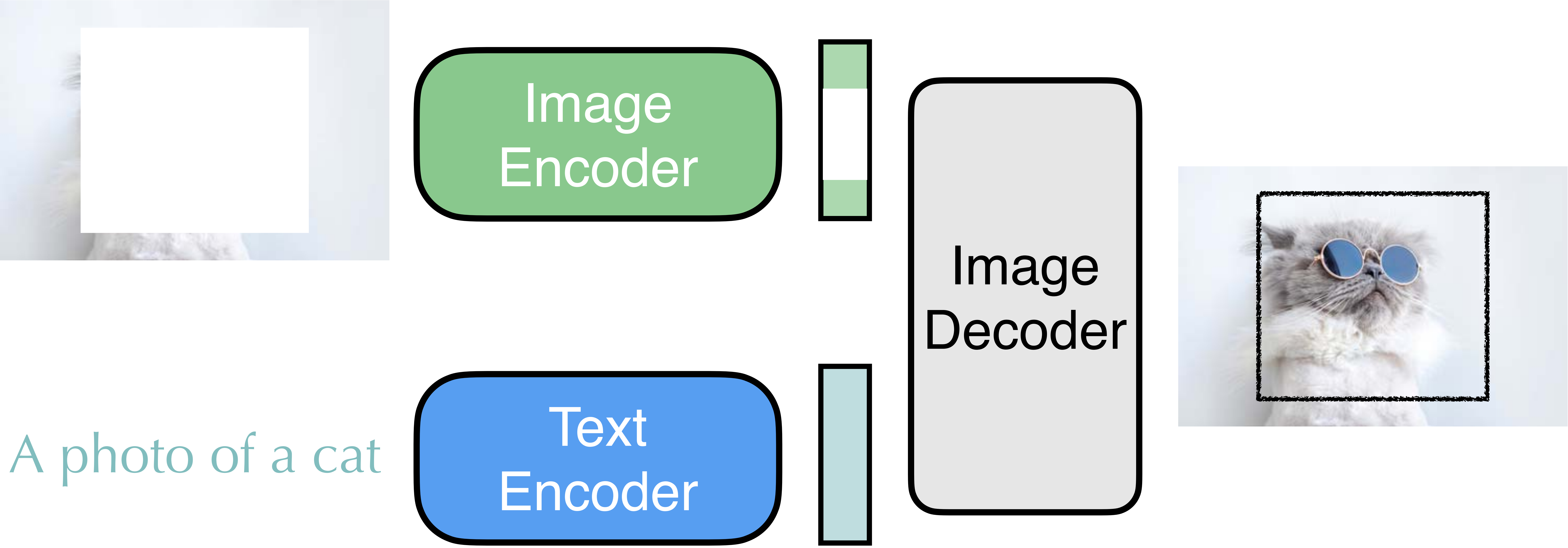


Image Encoder

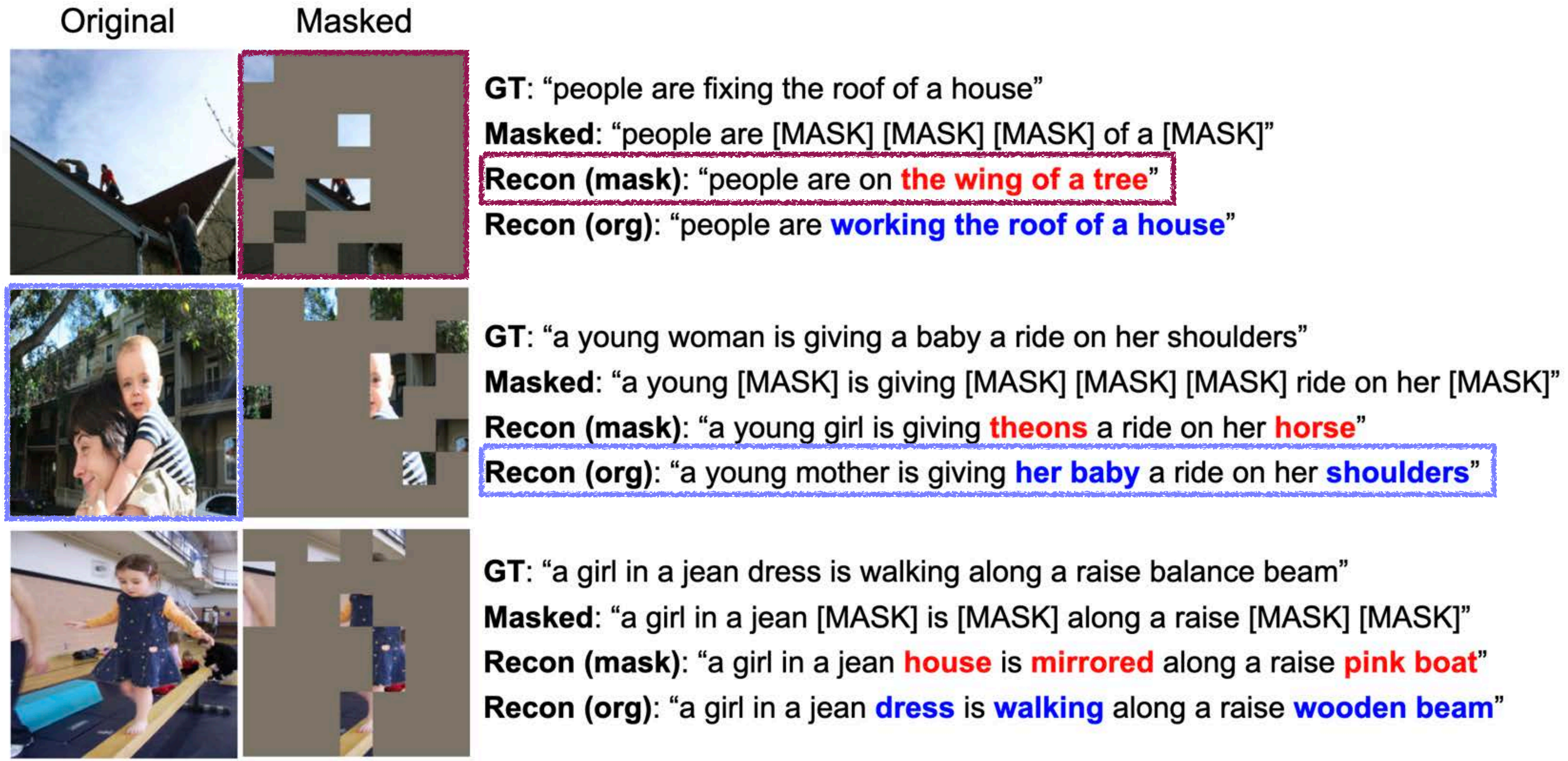
A photo of a cat

Text Encoder

Image Decoder

[1] Kwon, Gukyeong, et al. "Masked Vision and Language Modeling for Multi-modal Representation Learning." *ICLR 2023*
[2] Singh, Amanpreet, et al. "Flava: A foundational language and vision alignment model." *CVPR 2022*

The Families of VLMs: 2) Masking Objective - MaskVLM [1]



GT: “people are fixing the roof of a house”

Masked: “people are [MASK] [MASK] [MASK] of a [MASK]”

Recon (mask): “people are on **the wing of a tree**”

Recon (org): “people are **working the roof of a house**”

GT: “a young woman is giving a baby a ride on her shoulders”

Masked: “a young [MASK] is giving [MASK] [MASK] [MASK] ride on her [MASK]”

Recon (mask): “a young girl is giving **theons** a ride on her **horse**”

Recon (org): “a young mother is giving **her baby** a ride on her **shoulders**”

GT: “a girl in a jean dress is walking along a raise balance beam”

Masked: “a girl in a jean [MASK] is [MASK] along a raise [MASK] [MASK]”

Recon (mask): “a girl in a jean **house** is **mirrored** along a raise **pink boat**”

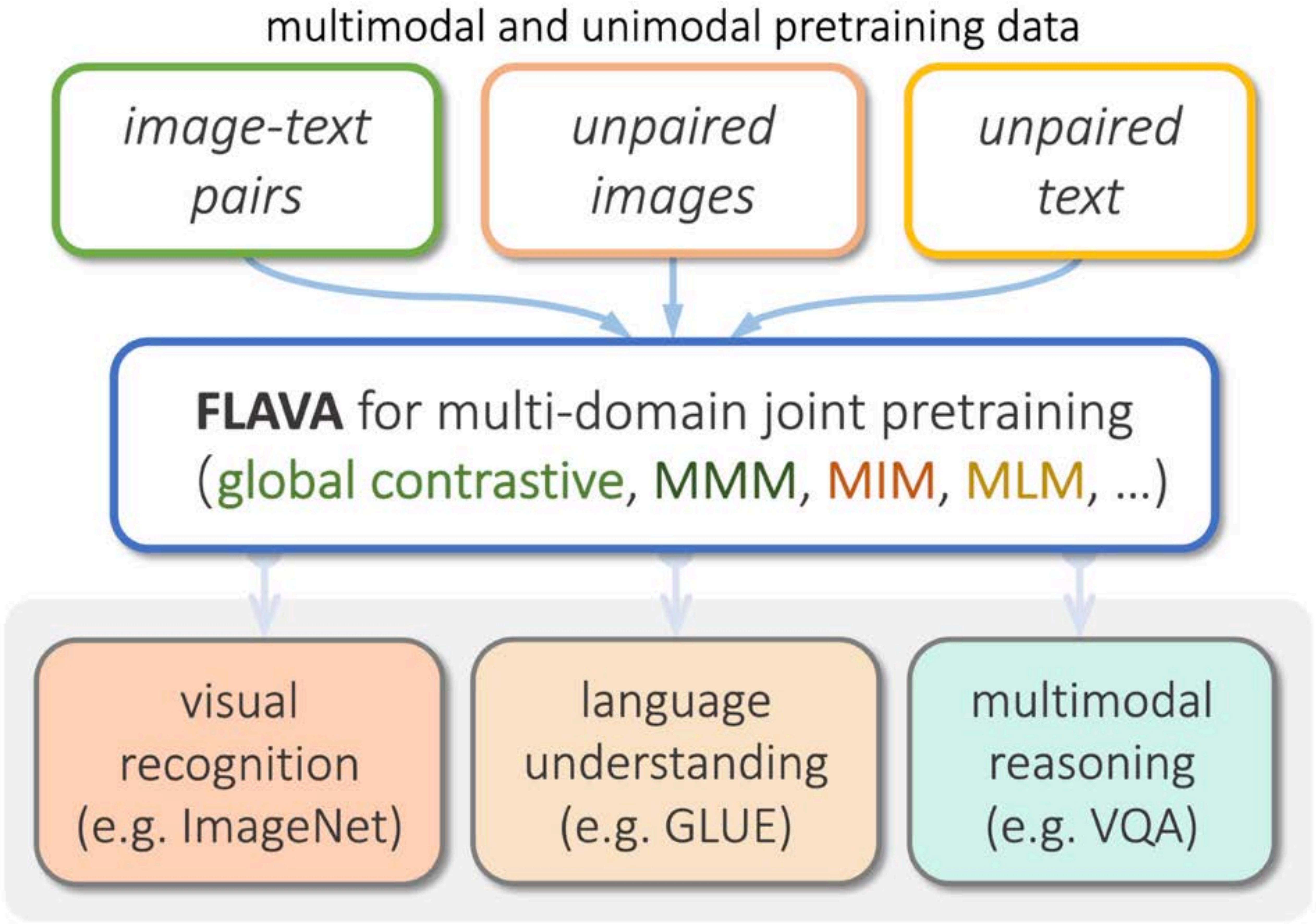
Recon (org): “a girl in a jean **dress** is **walking** along a raise **wooden beam**”

[1] Kwon, Gukyeong, et al. "Masked Vision and Language Modeling for Multi-modal Representation Learning." *ICLR 2023*
[2] Singh, Amanpreet, et al. "Flava: A foundational language and vision alignment model." *CVPR 2022*

The Families of VLMs: 2) Masking Objective - FLAVA [2]

However, there are many unpaired samples! To address it, FLAVA learns strong representations from:

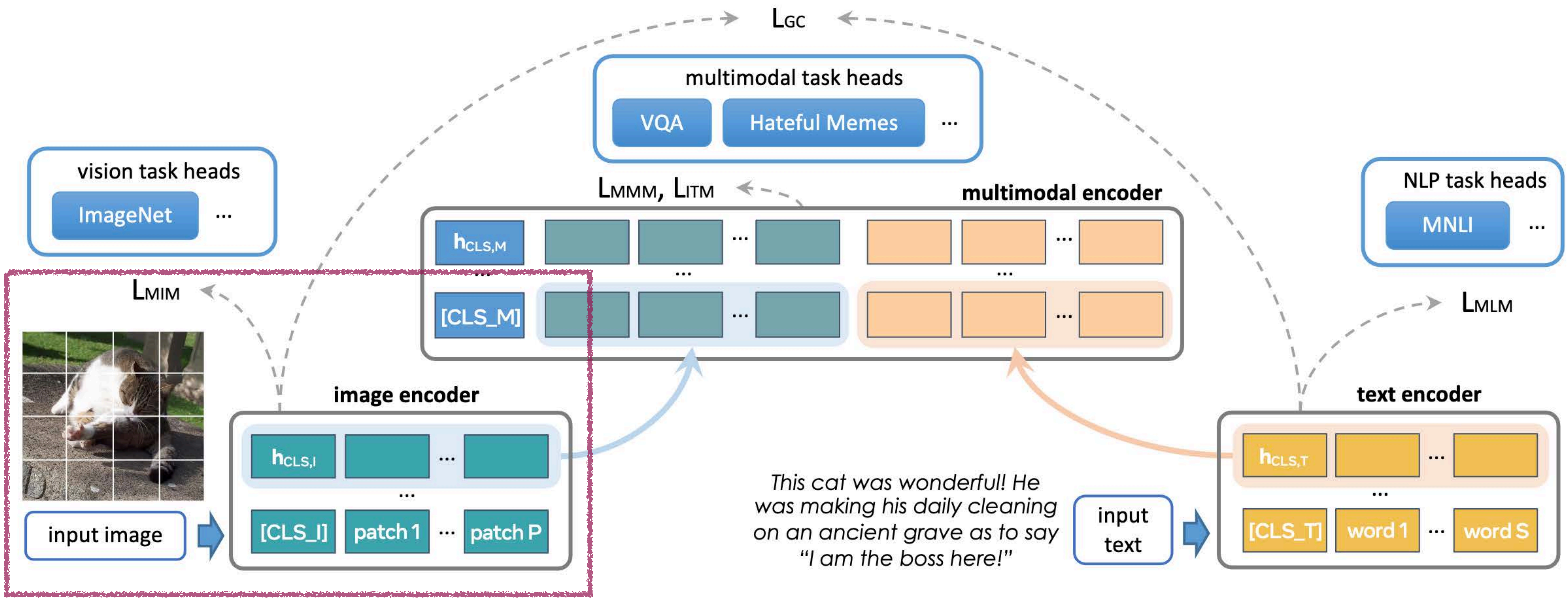
1. Image-text pairs
2. Unpaired images and texts


multimodal and unimodal pretraining data
image-text pairs
unpaired images
unpaired text
FLAVA for multi-domain joint pretraining
(global contrastive, MMM, MIM, MLM, ...)
visual recognition (e.g. ImageNet)
language understanding (e.g. GLUE)
multimodal reasoning (e.g. VQA)

[1] Kwon, Gukyeong, et al. "Masked Vision and Language Modeling for Multi-modal Representation Learning." *ICLR 2023*
[2] Singh, Amanpreet, et al. "Flava: A foundational language and vision alignment model." *CVPR 2022*

The Families of VLMs: 2) Masking Objective - FLAVA [2]

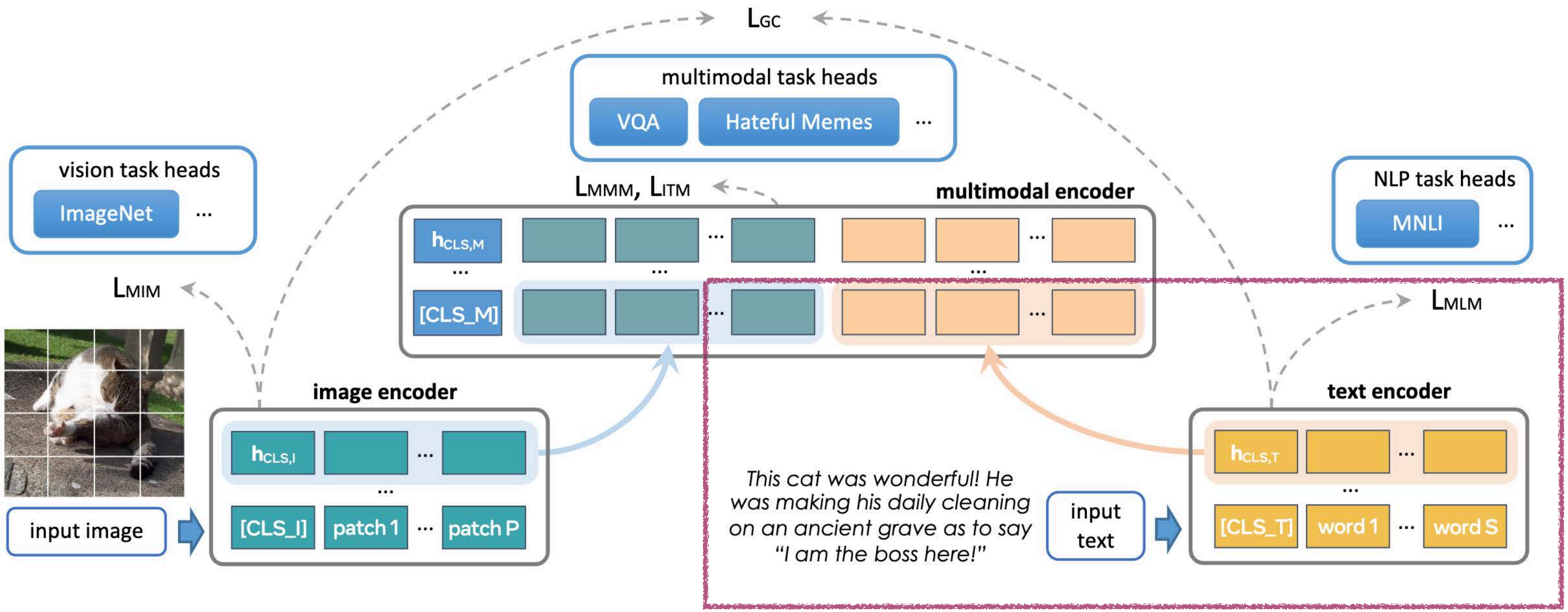
Image encoder (ViT): tokenize image w/ discrete VAE and train w/ with patch masking



[1] Kwon, Gukyeong, et al. "Masked Vision and Language Modeling for Multi-modal Representation Learning." *ICLR 2023*
[2] Singh, Amanpreet, et al. "Flava: A foundational language and vision alignment model." *CVPR 2022*

The Families of VLMs: 2) Masking Objective - FLAVA [2]

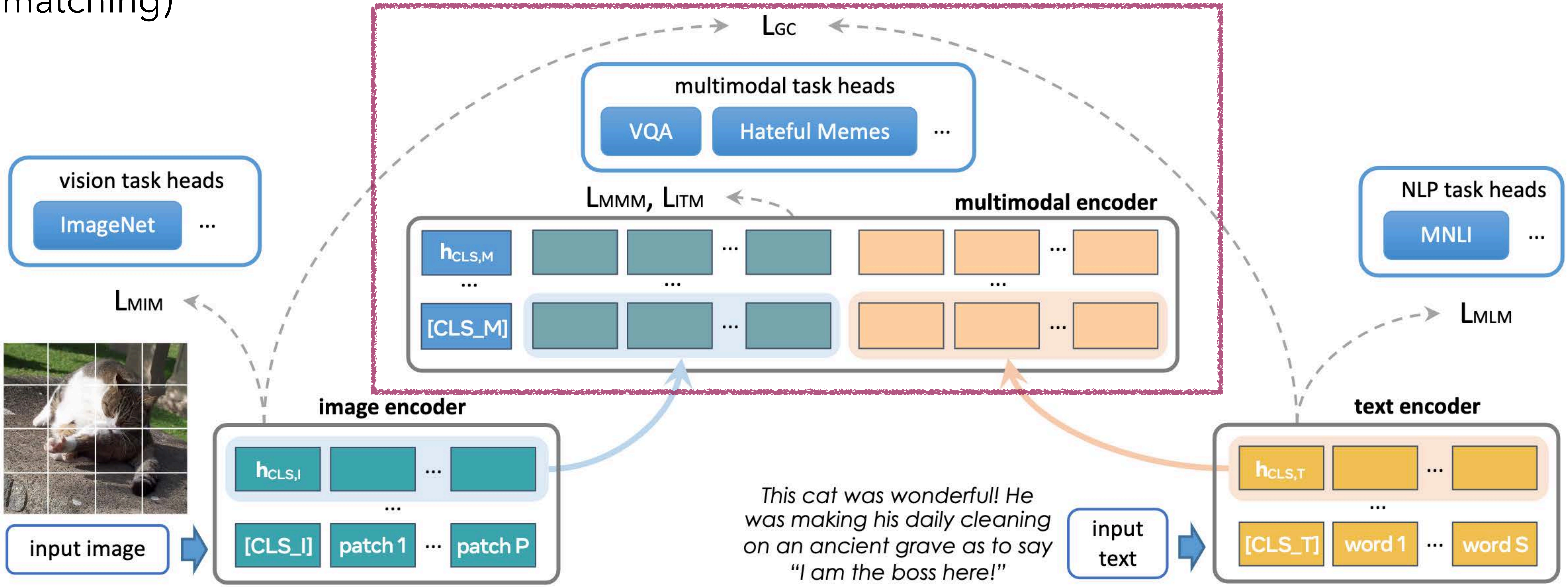
Text encoder (ViT): tokenize w/ BERT tokenizer and train w/ text token masking



[1] Kwon, Gukyeong, et al. "Masked Vision and Language Modeling for Multi-modal Representation Learning." *ICLR 2023*
[2] Singh, Amanpreet, et al. "Flava: A foundational language and vision alignment model." *CVPR 2022*

The Families of VLMs: 2) Masking Objective - FLAVA [2]

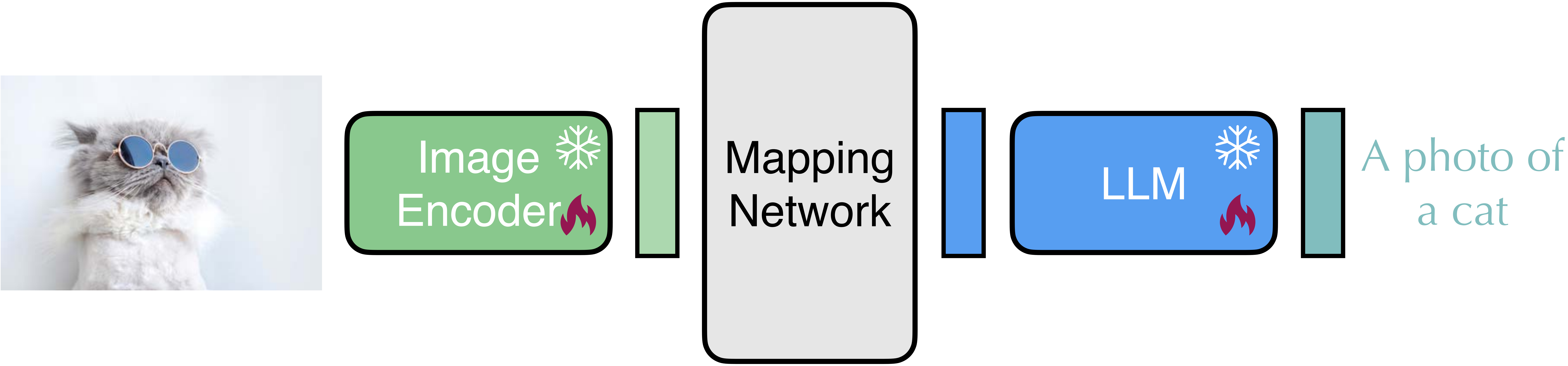
Multimodal encoder: multimodal losses (contrastive, masked multimodal loss, image-text matching)



[1] Kwon, Gukyeong, et al. "Masked Vision and Language Modeling for Multi-modal Representation Learning." *ICLR 2023*
[2] Singh, Amanpreet, et al. "Flava: A foundational language and vision alignment model." *CVPR 2022*

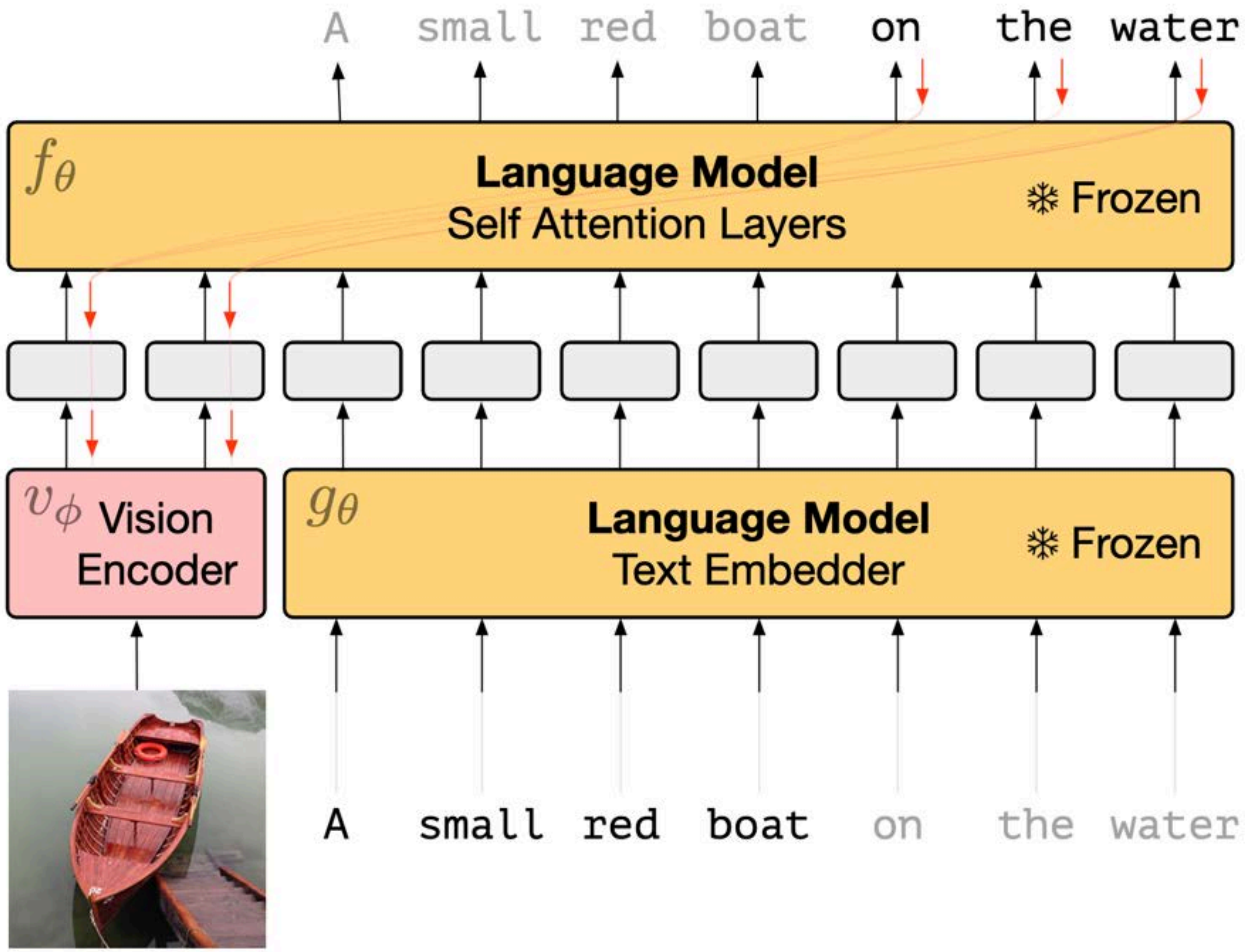
The Families of VLMs: 3) Pretrained Backbones

Learn a mapping between an image encoder and the open-sourced LLM



The Families of VLMs: 3) Pretrained Backbones - Frozen

Frozen is a first example of leveraging a pretrained LLM to training VLMs.



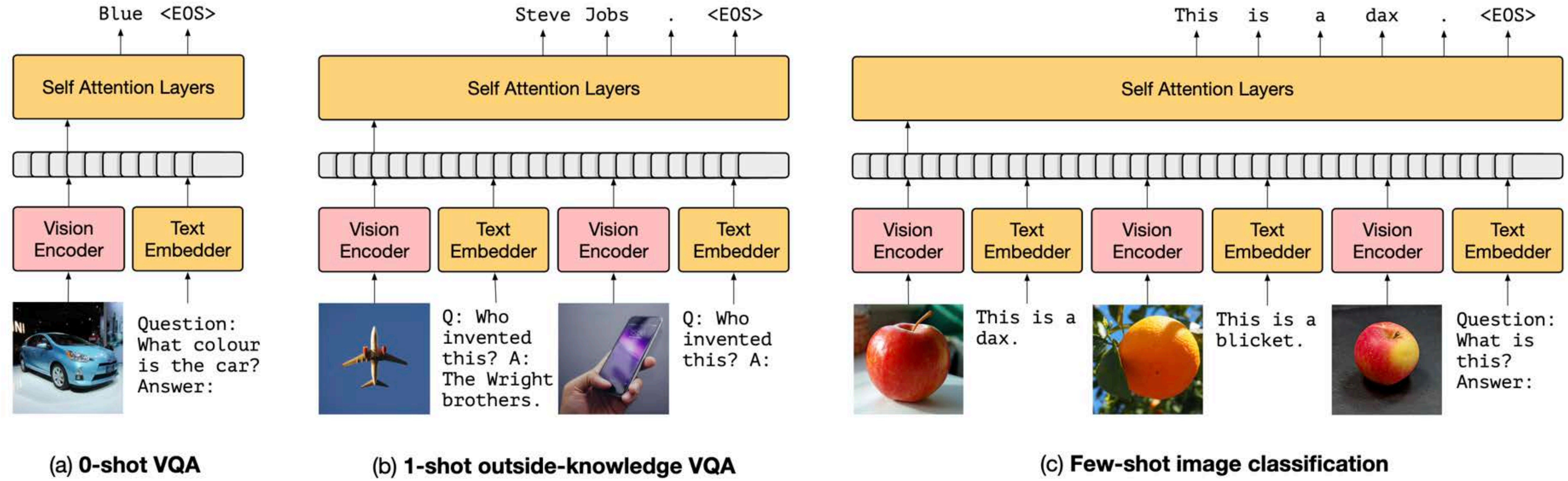
Self-Attention layers: Transformer
Text Eembedder: SentencePiece tokenizer
Vision encoder: NF-ResNet-50 (after pooling)

Frozen (#citation=645) was proposed 2~3 years earlier than LLaVA (#citation=2735)!

Tsimpoukelli, Maria, et al. "Multimodal few-shot learning with frozen language models." *NeurIPS 2021*.

The Families of VLMs: 3) Pretrained Backbones - Frozen

Frozencan supports (a) visual question answering, (b) outside-knowledge question answering and (c) few-shot image classification via **in-context learning**.

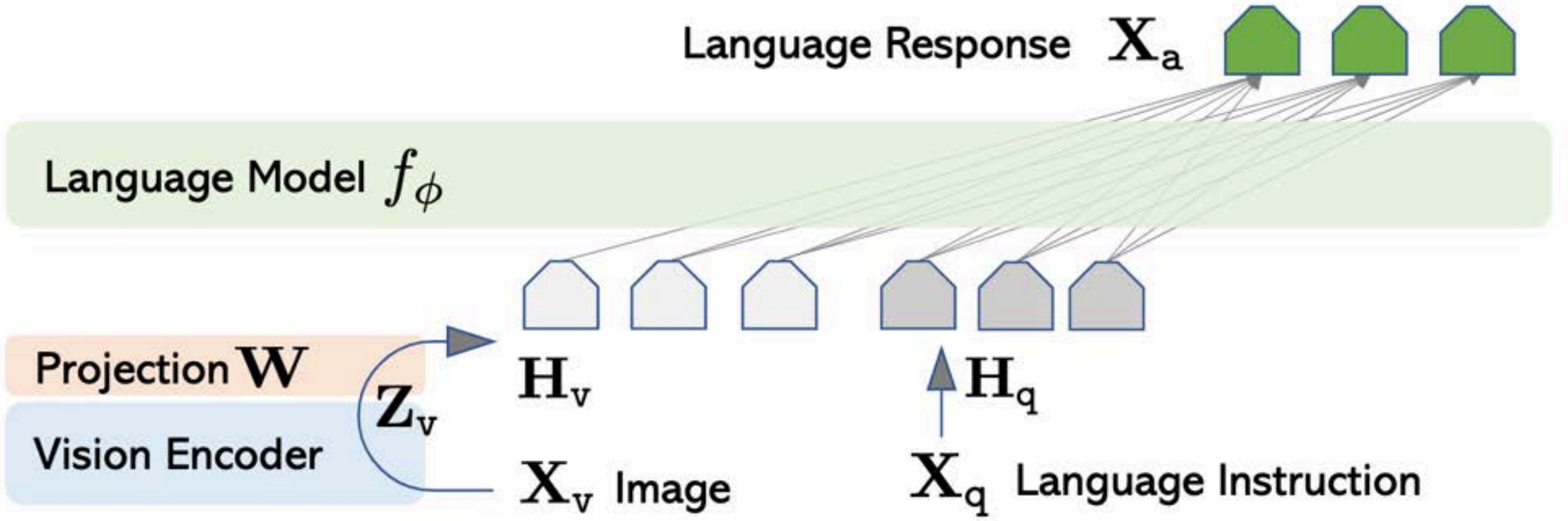


(a) **0-shot VQA** (b) **1-shot outside-knowledge VQA** (c) **Few-shot image classification**

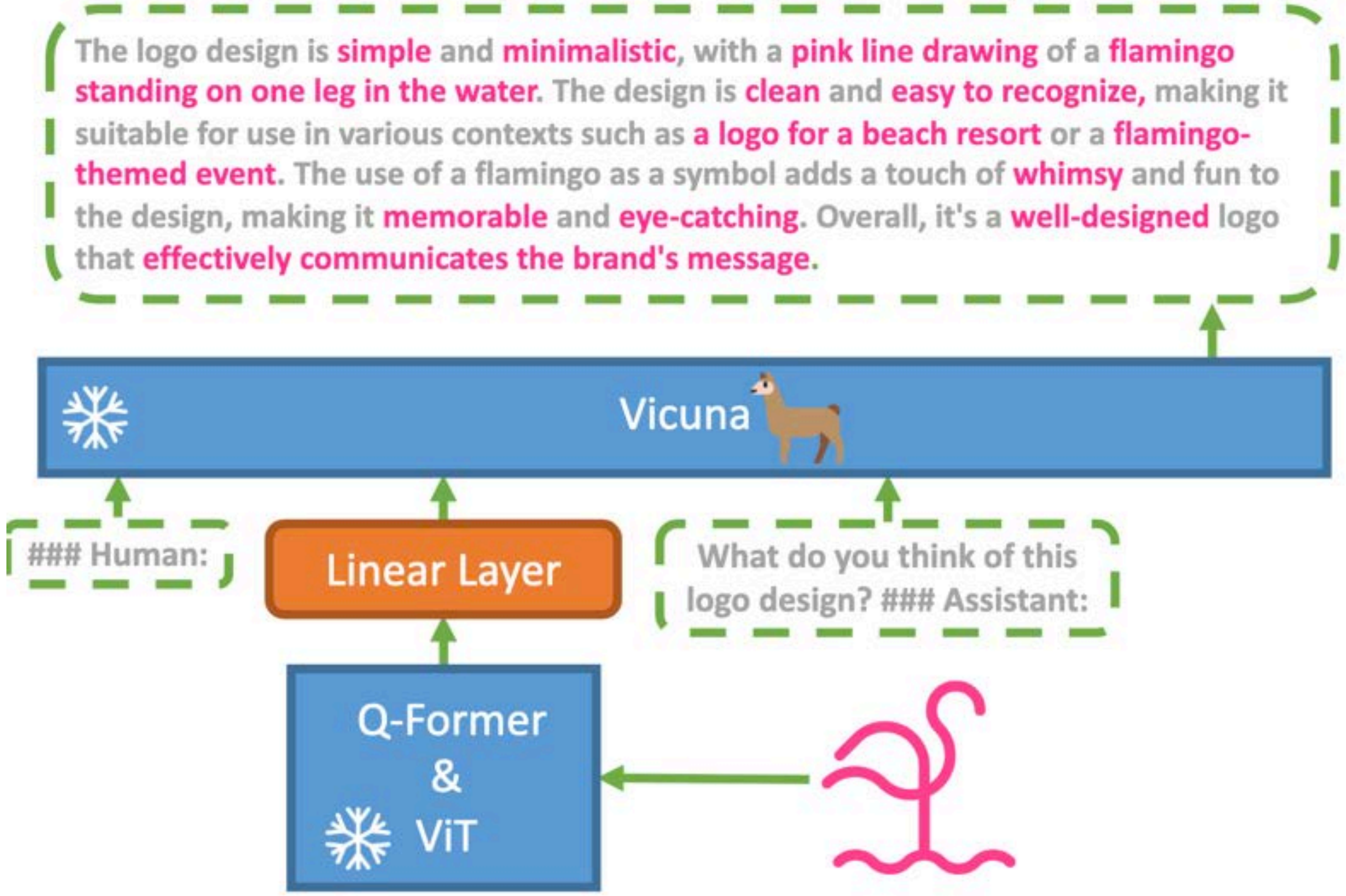
Tsimpoukelli, Maria, et al. "Multimodal few-shot learning with frozen language models." *NeurIPS 2021*.

The Families of VLMs: 3) Pretrained Backbones - LLaVA / MiniGPT-4

MiniGPT-4/LLaVA require only training the project layer: the visual encoder and LLM are already pretrained and used as off-the-shelf from prior work, such as CLIP and Vicuna



LLaVA [1]



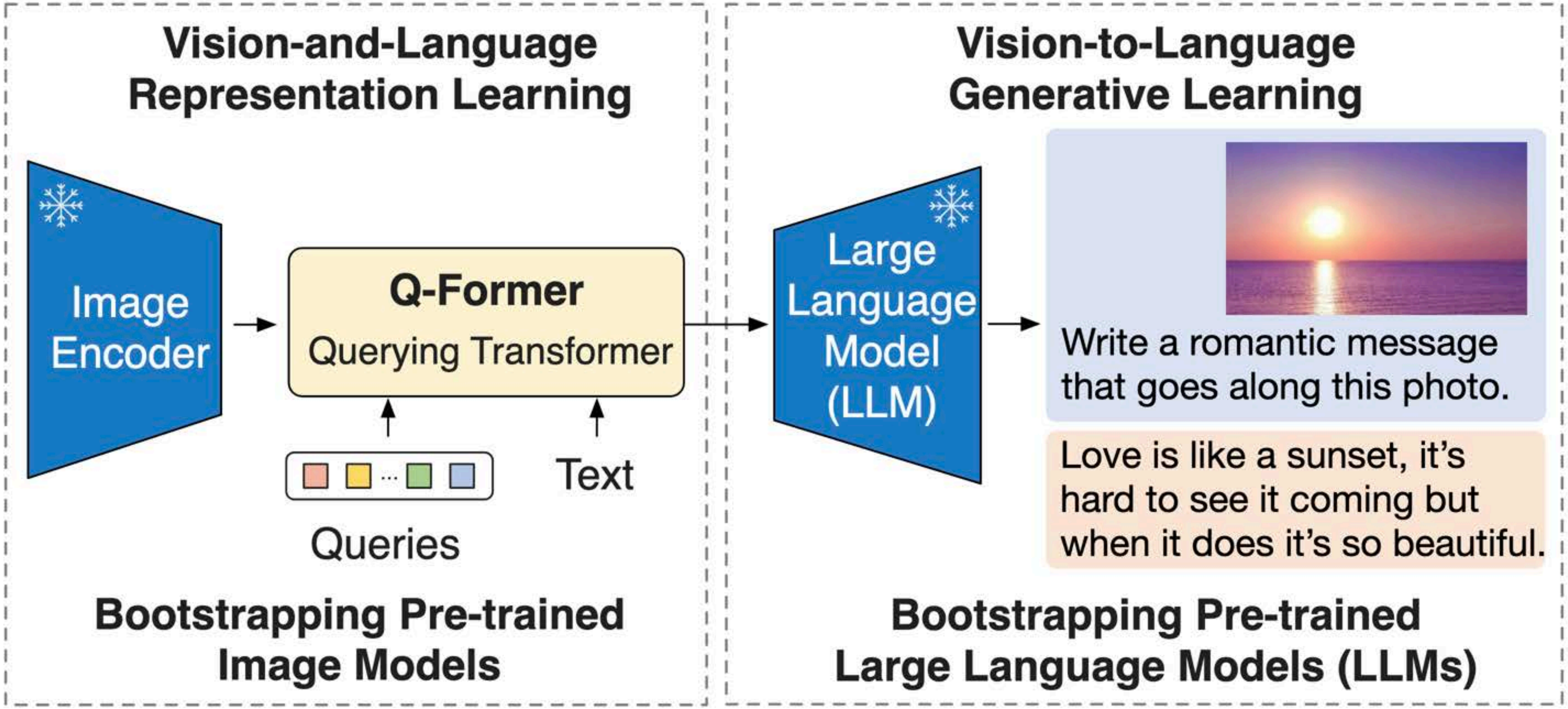
The logo design is **simple** and **minimalistic**, with a **pink line drawing** of a **flamingo standing on one leg in the water**. The design is **clean** and **easy to recognize**, making it suitable for use in various contexts such as **a logo for a beach resort** or a **flamingo-themed event**. The use of a flamingo as a symbol adds a touch of **whimsy** and fun to the design, making it **memorable** and **eye-catching**. Overall, it's a **well-designed** logo that **effectively communicates the brand's message**.

MiniGPT-4 [2]

[1] Liu, Haotian, et al. "Visual instruction tuning." *NeurIPS 2024*.
[2] Zhu, Deyao, et al. "MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models." *ICLR 2023*

The Families of VLMs: 3) Pretrained Backbones - BLIP-2

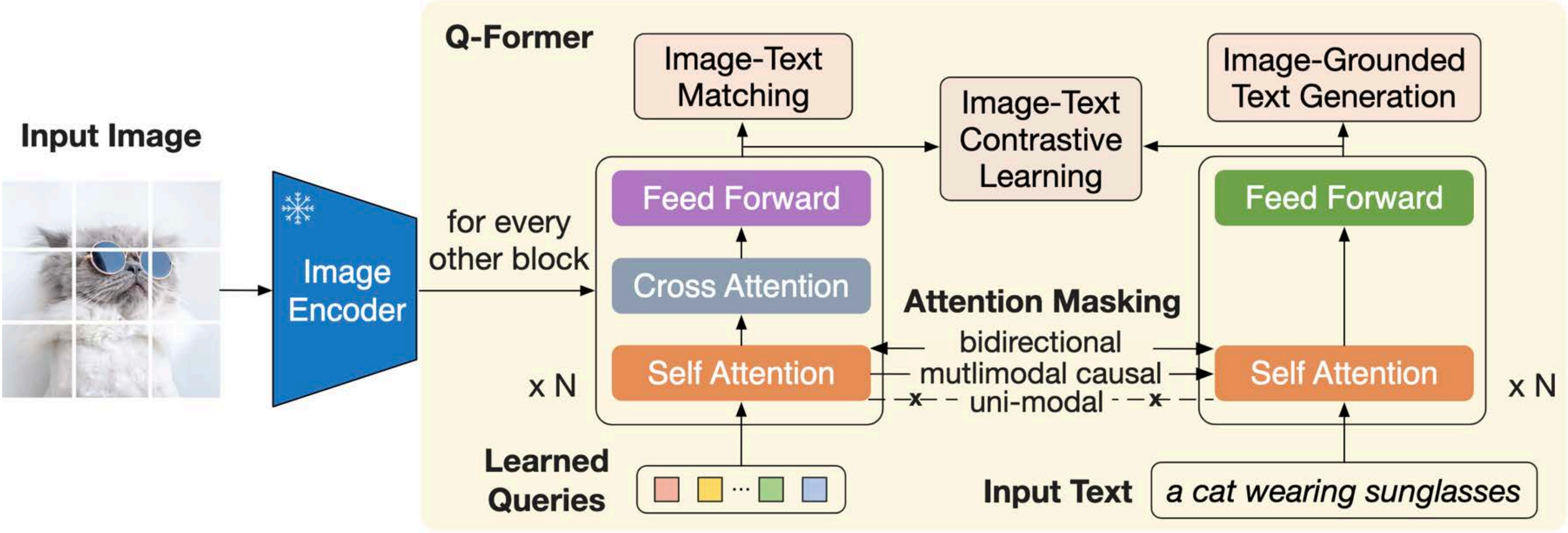
Bootstrapping language-image pre-training with frozen image encoders and LLMs.
Uses CLIP visual encoder and uses OPT and FlanT5 language models



Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." ICML 2023.

The Families of VLMs: 3) Pretrained Backbones - BLIP-2

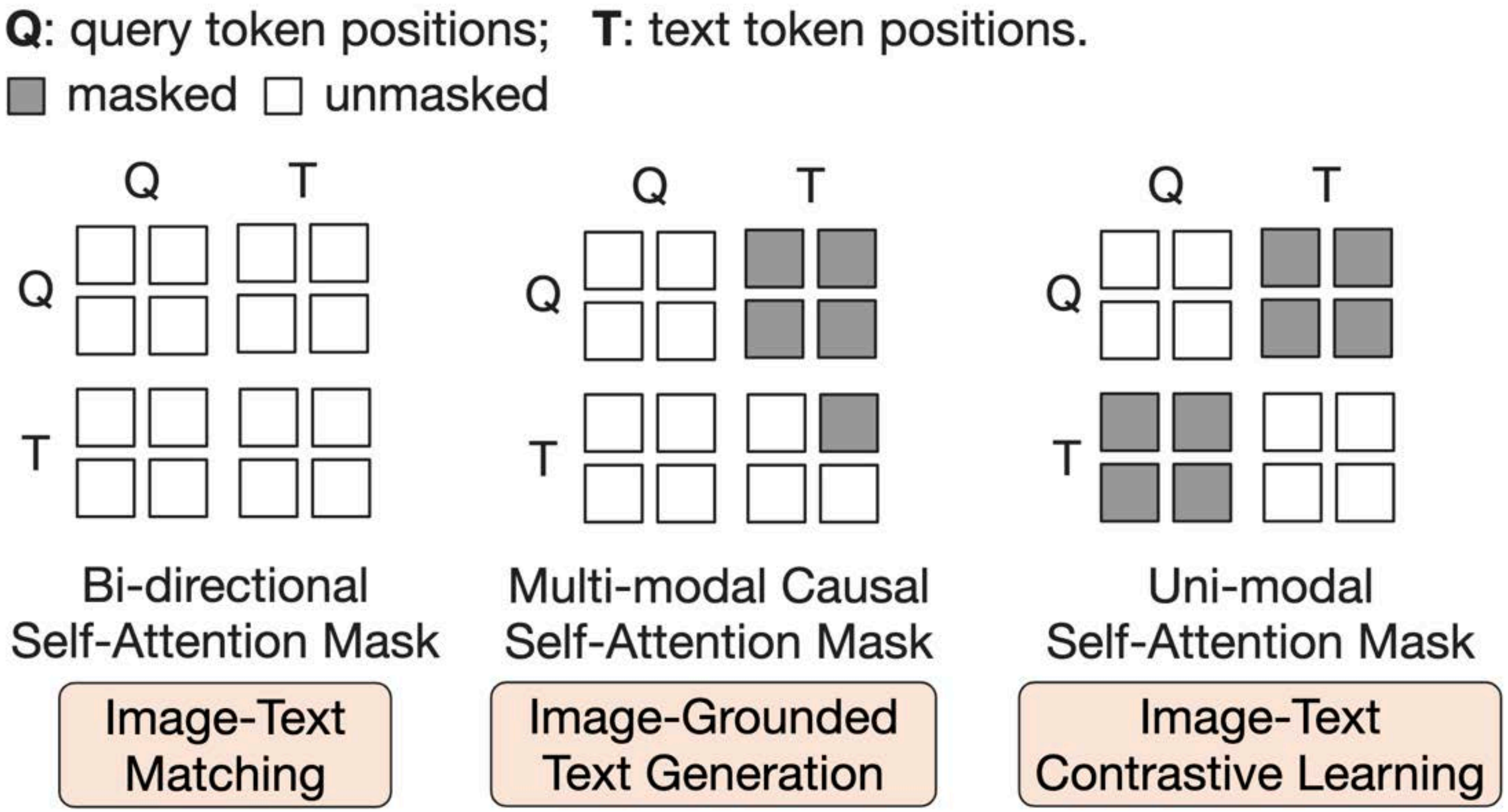
Jointly optimize three objectives which enforce the queries (a set of learnable embeddings) to **extract visual representation most relevant to the text.**



Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." ICML 2023.

The Families of VLMs: 3) Pretrained Backbones - BLIP-2

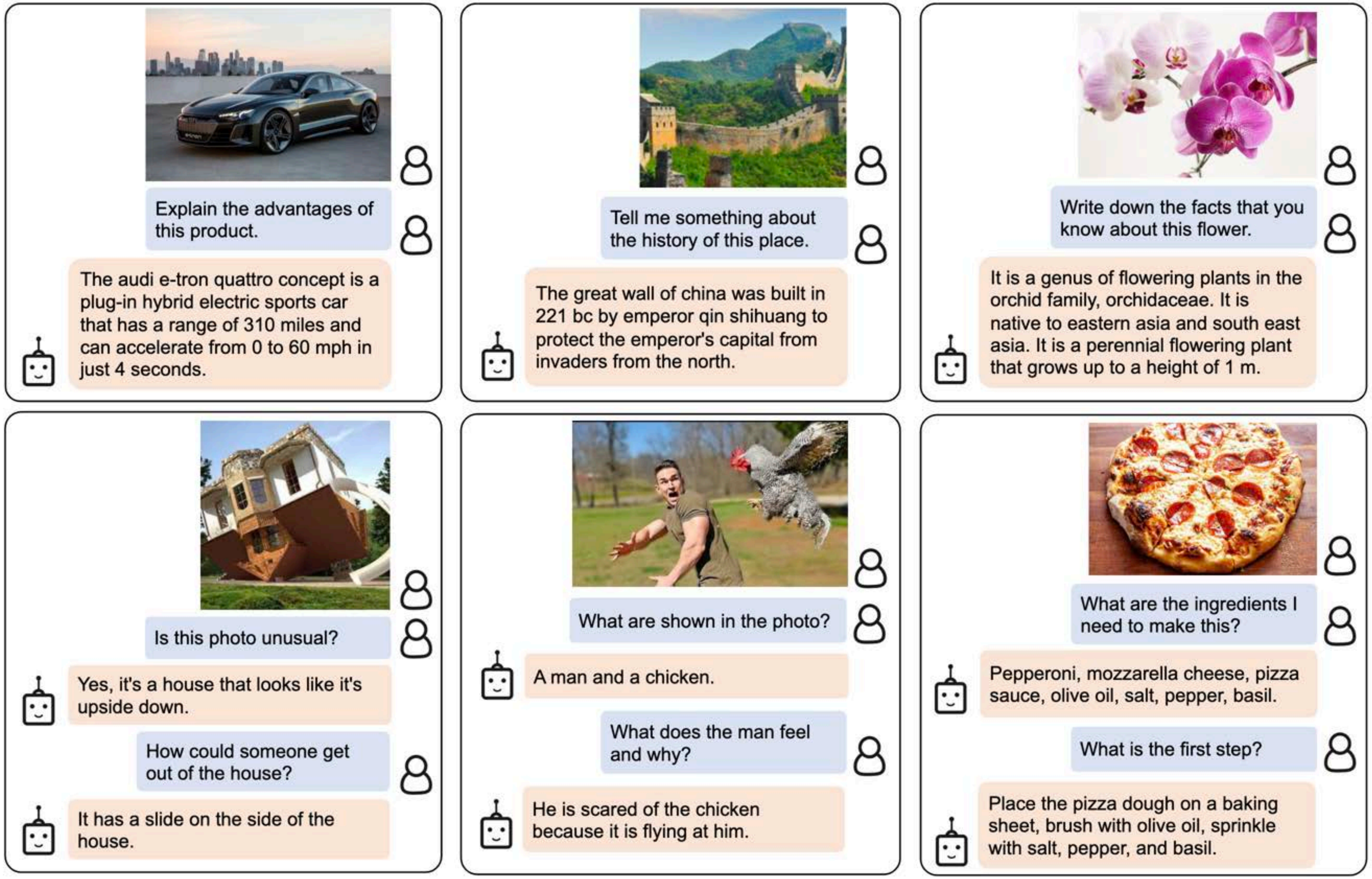
Image-text matching: pick-up similar texts sing similarity matrix and do binary clasification
Image-grounded text generation: given embedding Q, generate the corresponding text
Image-text contrastive loss: compare similarity between text and image embeddings


Q: query token positions; **T**: text token positions.
▨ masked ☐ unmasked
Q
T
Bi-directional Self-Attention Mask
Multi-modal Causal Self-Attention Mask
Uni-modal Self-Attention Mask
Image-Text Matching
Image-Grounded Text Generation
Image-Text Contrastive Learning

Depending on the pre-trained tasks, they apply different masks to control query-text interaction

Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." ICML 2023.

The Families of VLMs: 3) Pretrained Backbones - BLIP-2



Explain the advantages of this product.

The audi e-tron quattro concept is a plug-in hybrid electric sports car that has a range of 310 miles and can accelerate from 0 to 60 mph in just 4 seconds.

Tell me something about the history of this place.

The great wall of china was built in 221 bc by emperor qin shihuang to protect the emperor's capital from invaders from the north.

Write down the facts that you know about this flower.

It is a genus of flowering plants in the orchid family, orchidaceae. It is native to eastern asia and south east asia. It is a perennial flowering plant that grows up to a height of 1 m.

Is this photo unusual?

Yes, it's a house that looks like it's upside down.

How could someone get out of the house?

It has a slide on the side of the house.

What are shown in the photo?

A man and a chicken.

What does the man feel and why?

He is scared of the chicken because it is flying at him.

What are the ingredients I need to make this?

Pepperoni, mozzarella cheese, pizza sauce, olive oil, salt, pepper, basil.

What is the first step?

Place the pizza dough on a baking sheet, brush with olive oil, sprinkle with salt, pepper, and basil.

Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." ICML 2023.

The Families of VLMs: 4) Generative-based

VLMs are trained in such a way they can generate entire images or very long captions

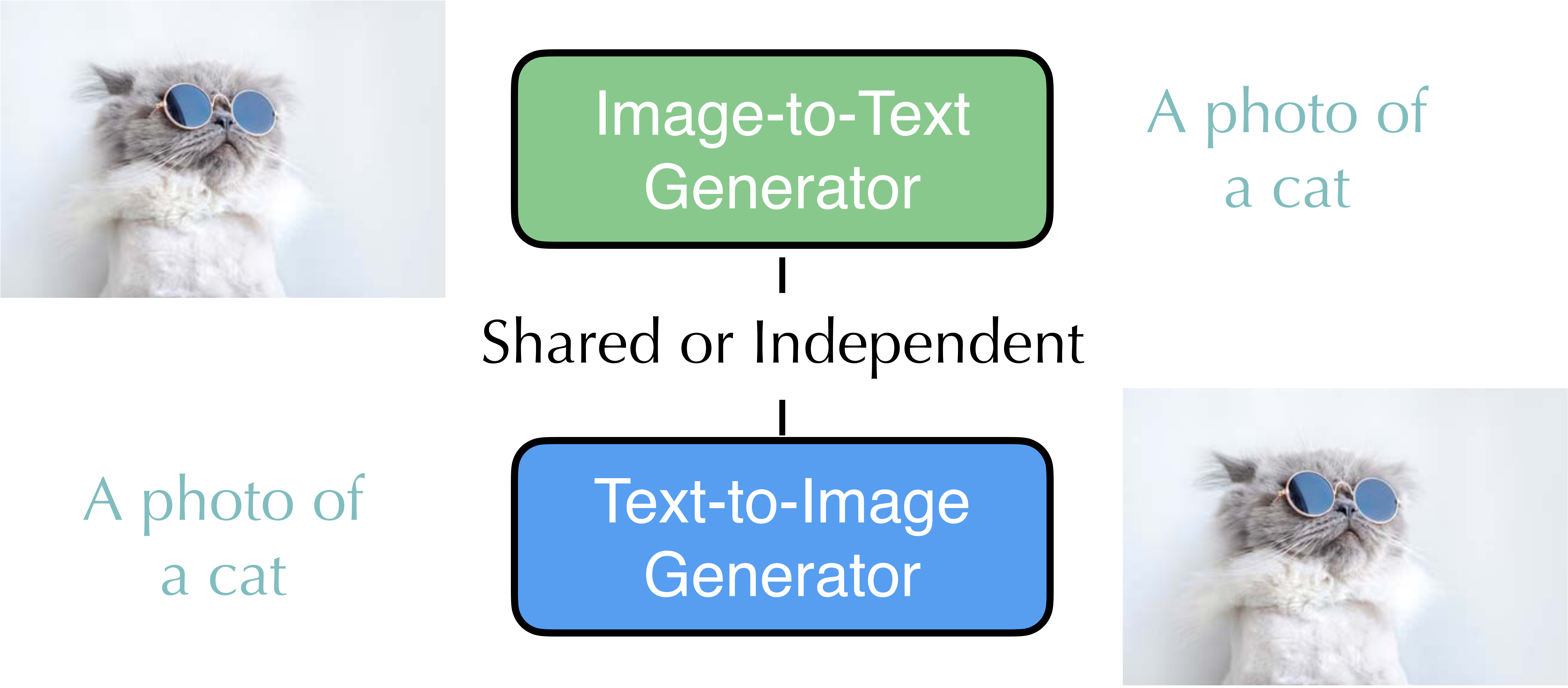


Image-to-Text Generator

A photo of a cat

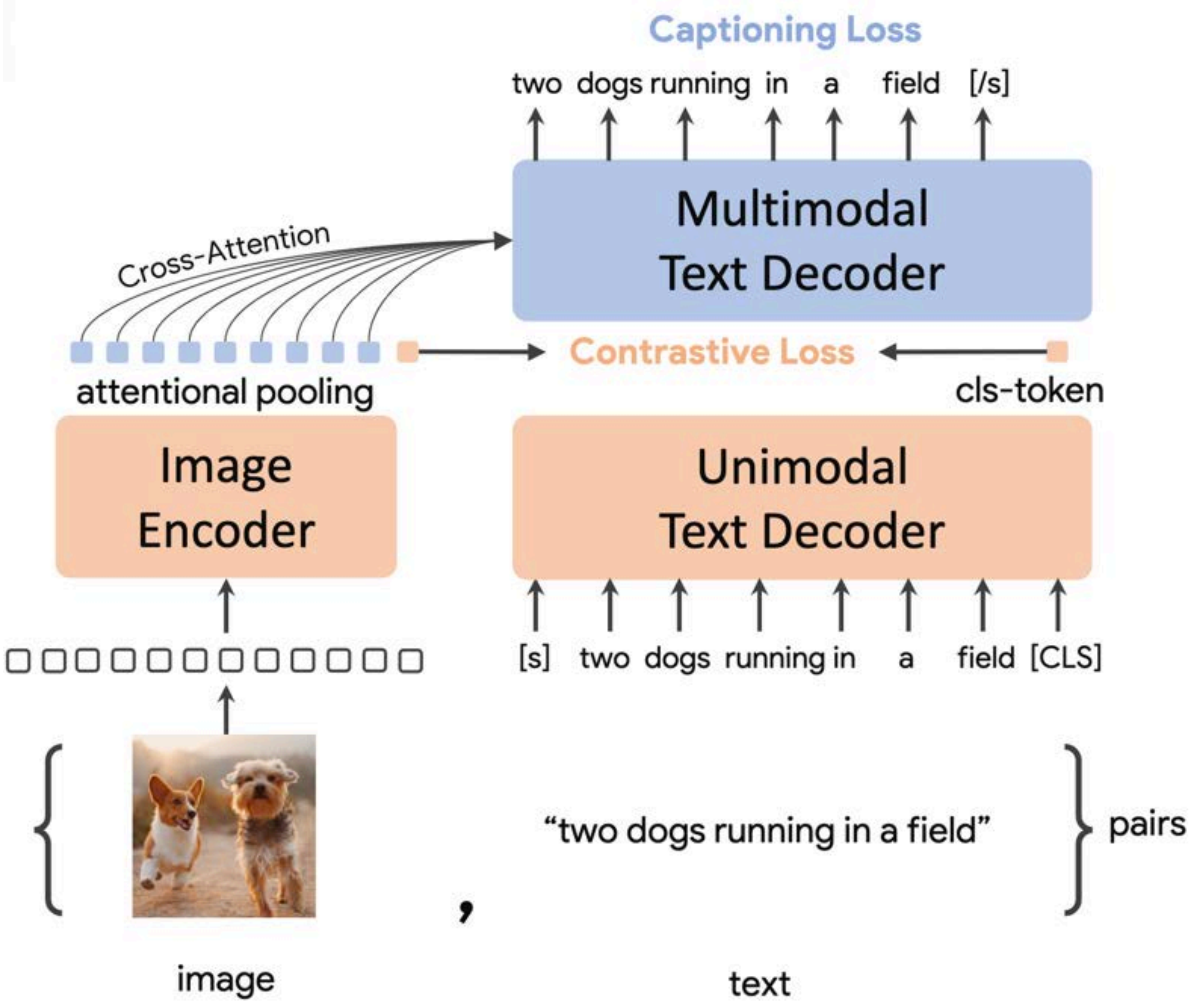
Shared or Independent

A photo of a cat

Text-to-Image Generator

The Families of VLMs: 4) Generative-based - CoCa

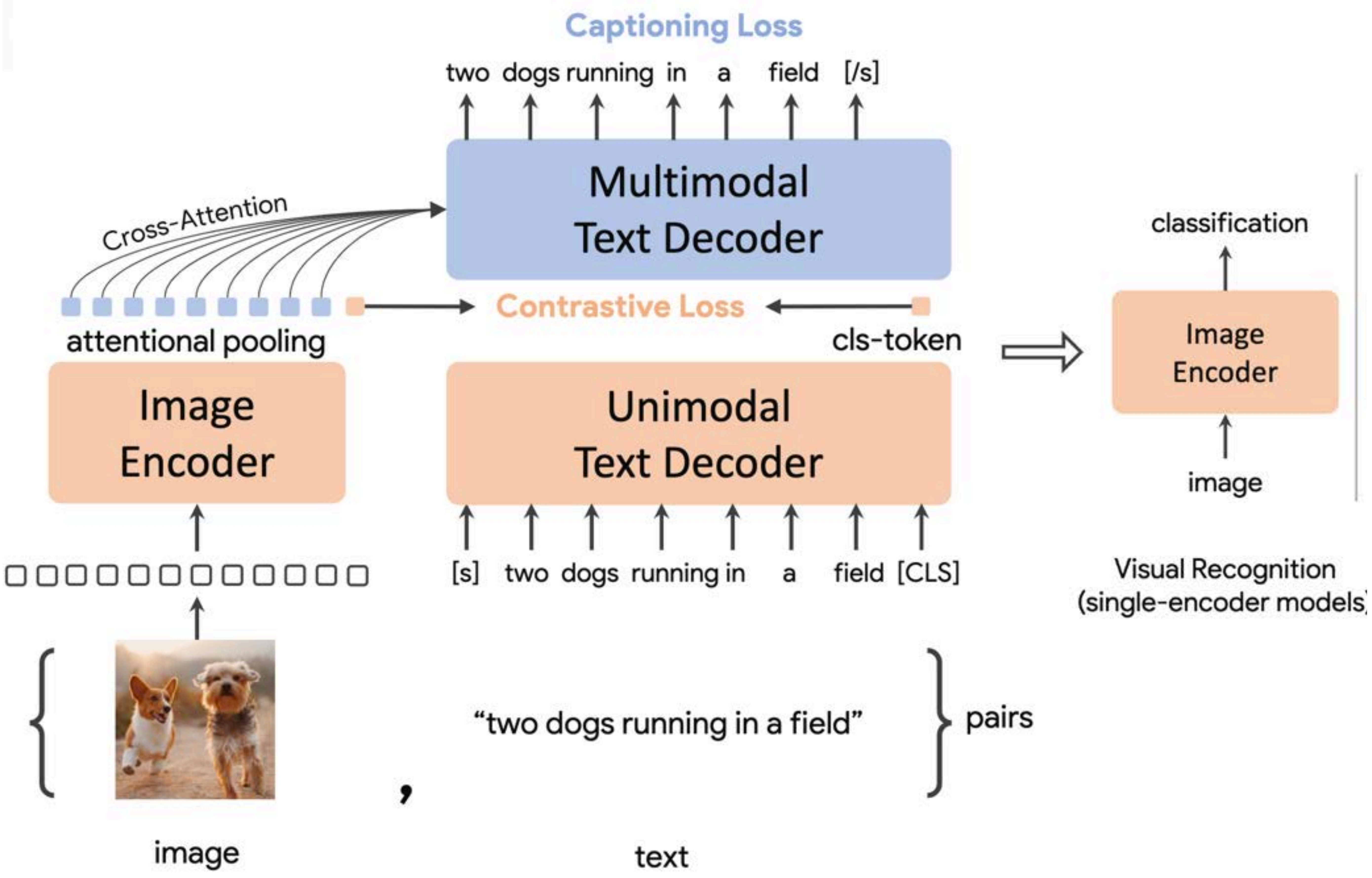
Pretrain an image-text encoder-decoder model with contrastive and captioning loss.



Yu, Jiahui, et al. "CoCa: Contrastive Captioners are Image-Text Foundation Models." *TMLR 2022*

The Families of VLMs: 4) Generative-based - CoCa

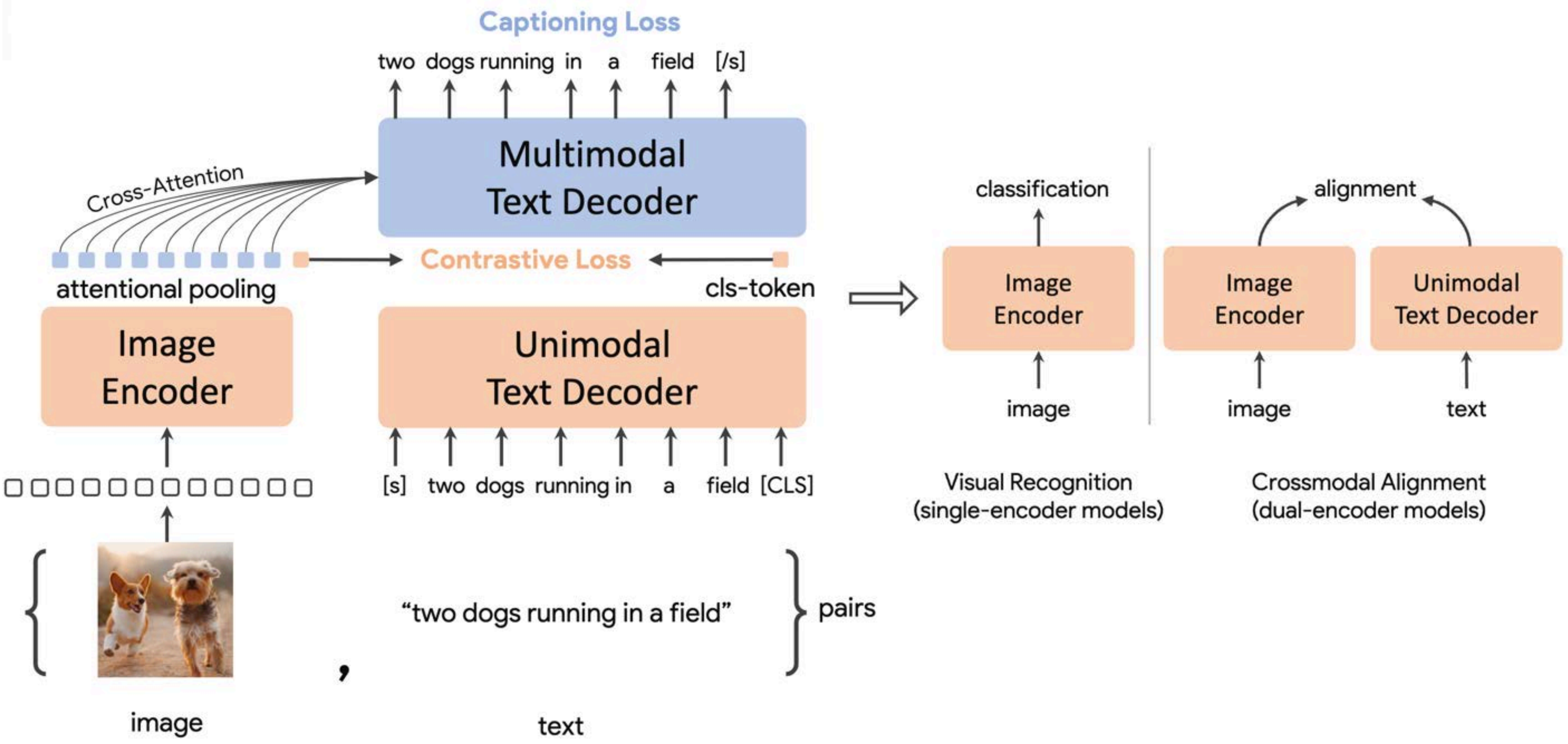
The pretrained CoCa can be used for visual recognition, vision-language alignment, image captioning and multimodal understanding with zero-shot transfer, frozen-feature evaluation or end-to-end finetuning.



Yu, Jiahui, et al. "CoCa: Contrastive Captioners are Image-Text Foundation Models." *TMLR 2022*

The Families of VLMs: 4) Generative-based - CoCa

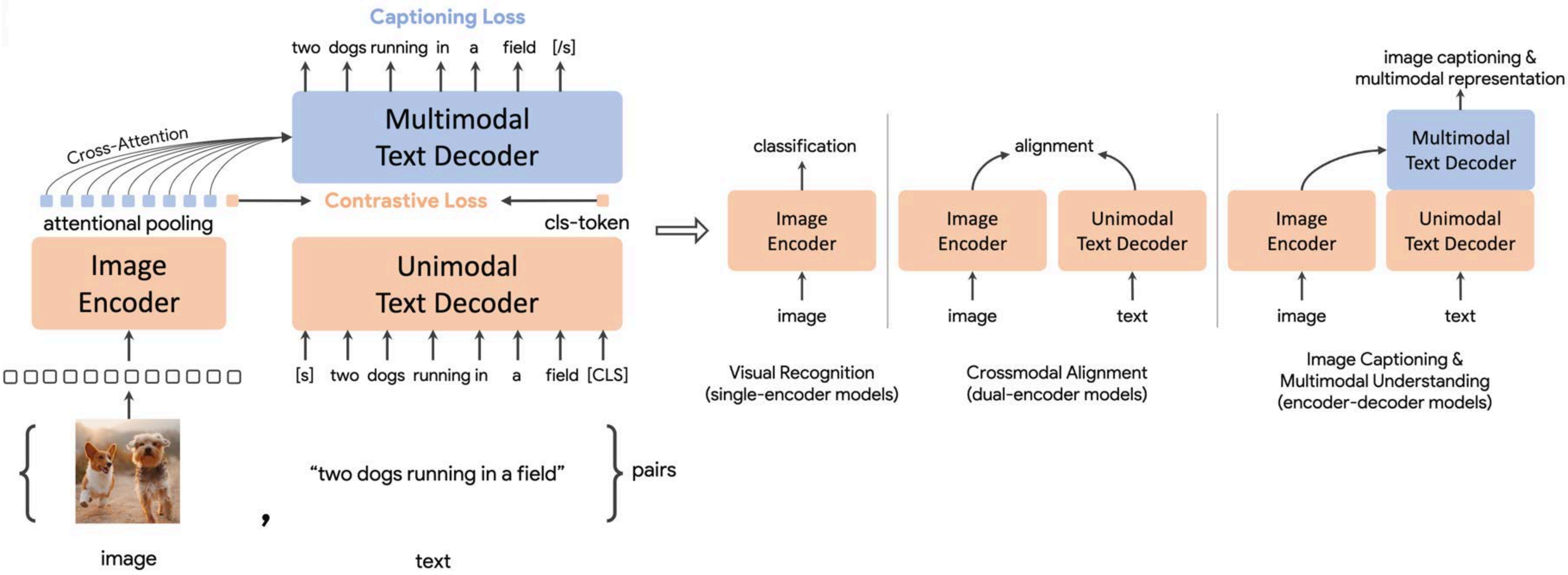
The pretrained CoCa can be used for visual recognition, vision-language alignment, image captioning and multimodal understanding with zero-shot transfer, frozen-feature evaluation or end-to-end finetuning.



Yu, Jiahui, et al. "CoCa: Contrastive Captioners are Image-Text Foundation Models." *TMLR 2022*

The Families of VLMs: 4) Generative-based - CoCa

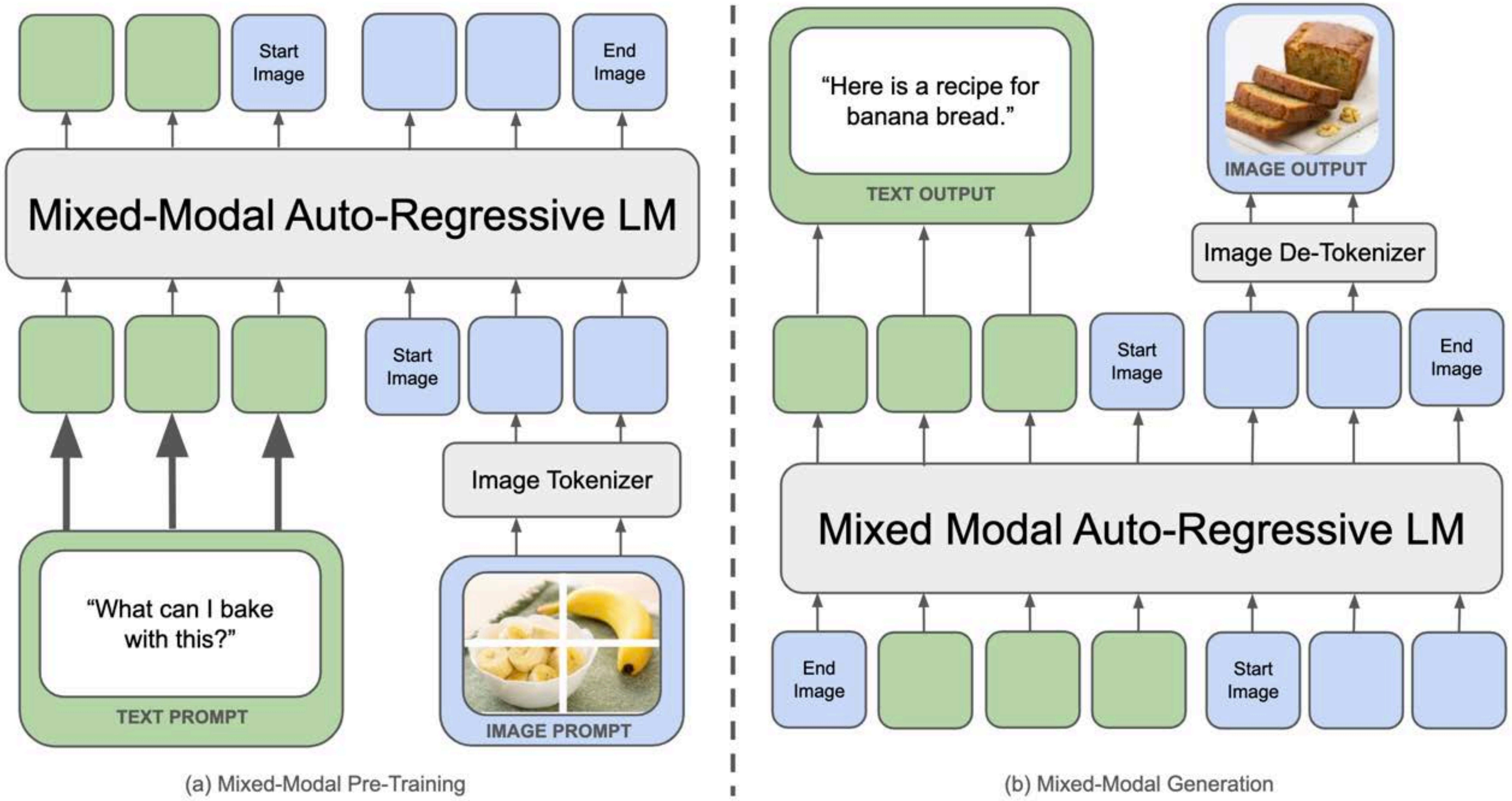
The pretrained CoCa can be used for visual recognition, vision-language alignment, image captioning and multimodal understanding with zero-shot transfer, frozen-feature evaluation or end-to-end finetuning.



Yu, Jiahui, et al. "CoCa: Contrastive Captioners are Image-Text Foundation Models." *TMLR 2022*

The Families of VLMs: 4) Generative-based - Chameleon

Early-fusion token-based mixed-modal models capable of understanding and generating images and text in any arbitrary sequence

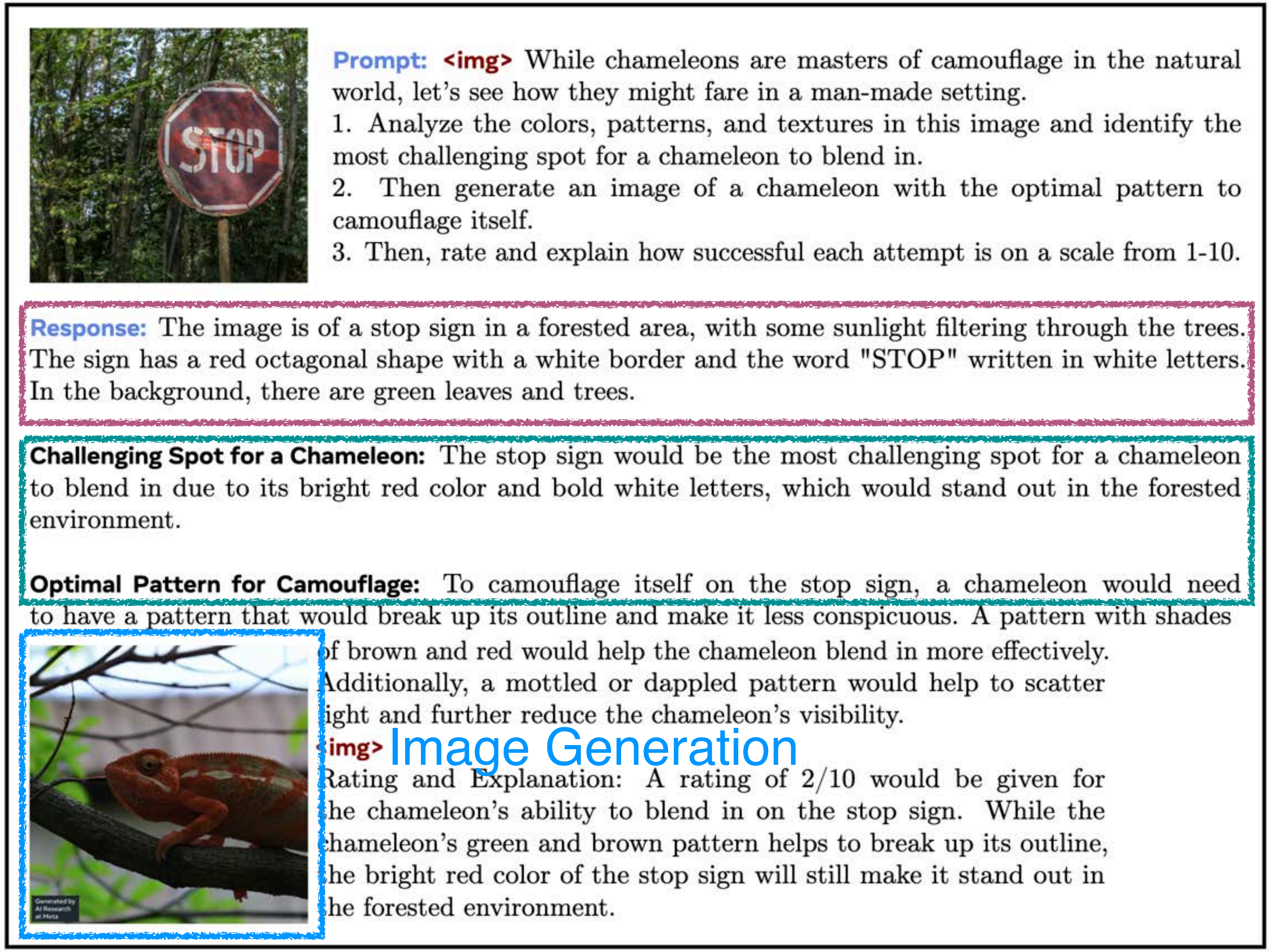


(a) Mixed-Modal Pre-Training

(b) Mixed-Modal Generation

Chameleon Team. "Chameleon: Mixed-modal early-fusion foundation models." *arXiv preprint arXiv:2405.09818* (2024).

The Families of VLMs: 4) Generative-based - Chameleon



Prompt: <img> While chameleons are masters of camouflage in the natural world, let's see how they might fare in a man-made setting.

1. Analyze the colors, patterns, and textures in this image and identify the most challenging spot for a chameleon to blend in.
2. Then generate an image of a chameleon with the optimal pattern to camouflage itself.
3. Then, rate and explain how successful each attempt is on a scale from 1-10.

Response: The image is of a stop sign in a forested area, with some sunlight filtering through the trees. The sign has a red octagonal shape with a white border and the word "STOP" written in white letters. In the background, there are green leaves and trees.

Challenging Spot for a Chameleon: The stop sign would be the most challenging spot for a chameleon to blend in due to its bright red color and bold white letters, which would stand out in the forested environment.

Optimal Pattern for Camouflage: To camouflage itself on the stop sign, a chameleon would need to have a pattern that would break up its outline and make it less conspicuous. A pattern with shades of brown and red would help the chameleon blend in more effectively. Additionally, a mottled or dappled pattern would help to scatter ight and further reduce the chameleon's visibility.

img> Image Generation

Rating and Explanation: A rating of 2/10 would be given for he chameleon's ability to blend in on the stop sign. While the hameleon's green and brown pattern helps to break up its outline, he bright red color of the stop sign will still make it stand out in he forested environment.

Prompt: Text + Image

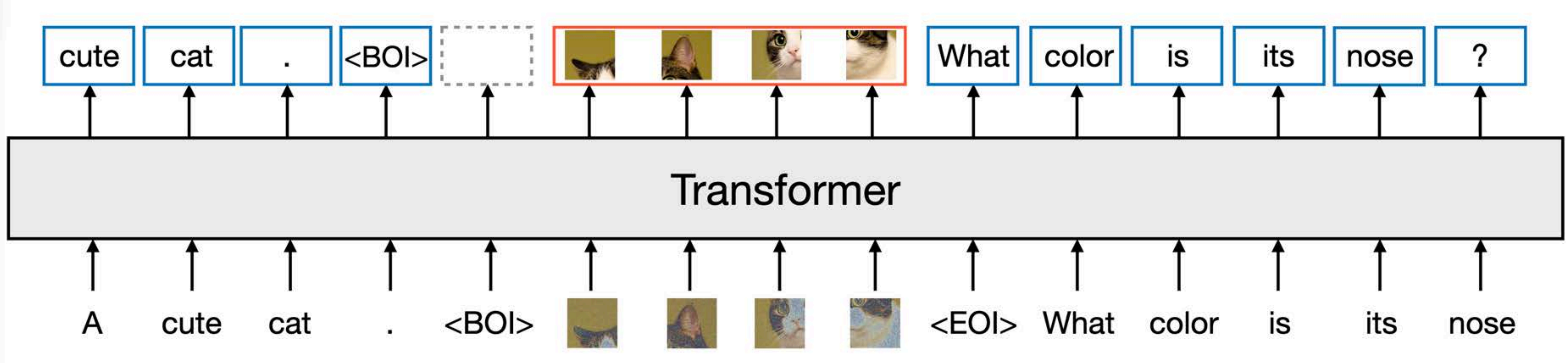
Output: Text + Image

Undertanding

Reasoning

The Families of VLMs: 4) Generative-based - Transfusion

vs. Chameleon: uses continues image vectors and trains on the diffusion objective.
The image generation results can be significantly improved.



Zhou, Chunting, et al. "Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model." arXiv preprint arXiv:2408.11039 (2024).

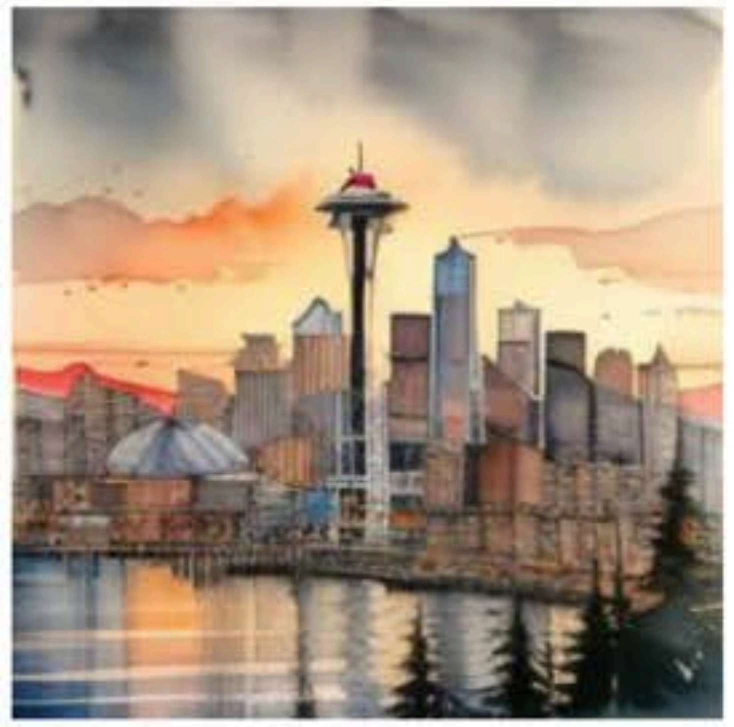
The Families of VLMs: 4) Generative-based - Transfusion

a monarch butterfly.

A rowboat on a lake with a bike on it.

An expressive oil painting of a chocolate chip cookie being dipped in a glass of milk, depicted as an explosion of flavors.

An angry duck doing heavy weightlifting at the gym.



Downtown Seattle at sunrise. detailed ink wash.

A car made out of vegetables.

An emoji of a baby panda wearing a red hat, green gloves, red shirt, and green pants.

A tranquil, anime-style koi pond in a serene Japanese garden, featuring blossoming cherry trees.

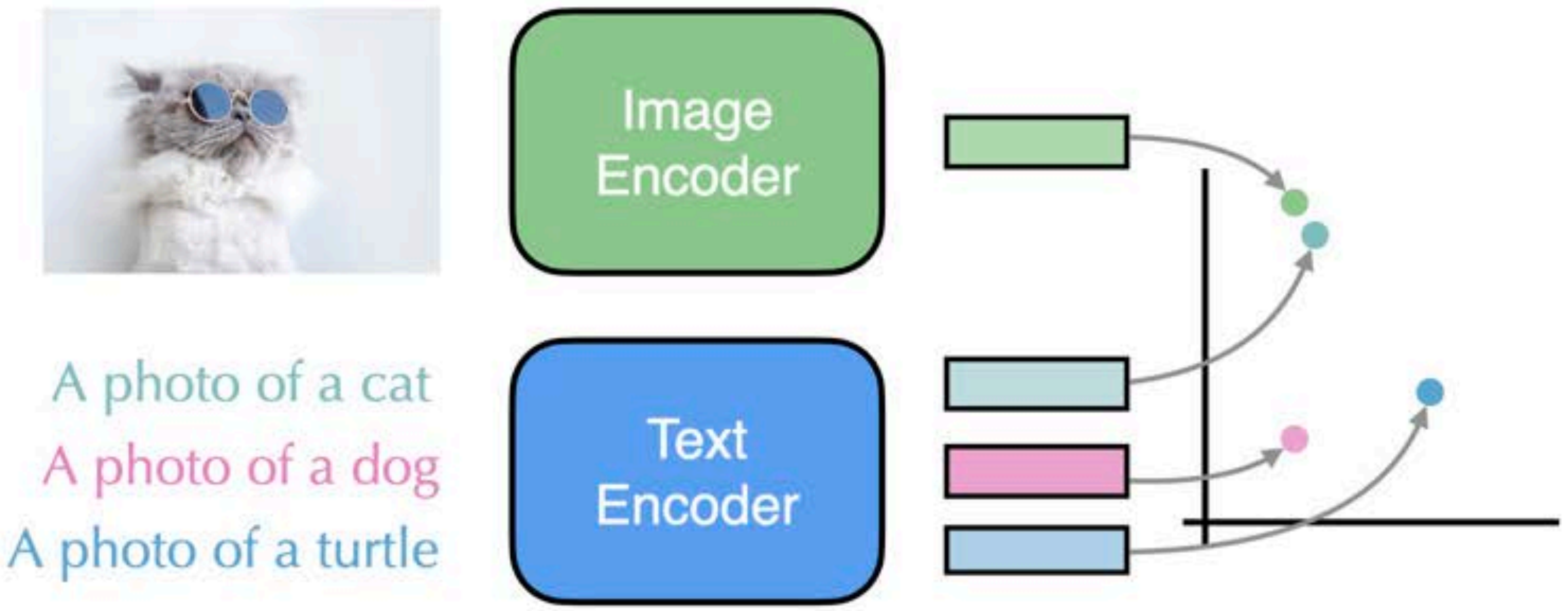
a massive alien space ship that is shaped like a pretzel.

graffiti of a funny dog on a street wall.

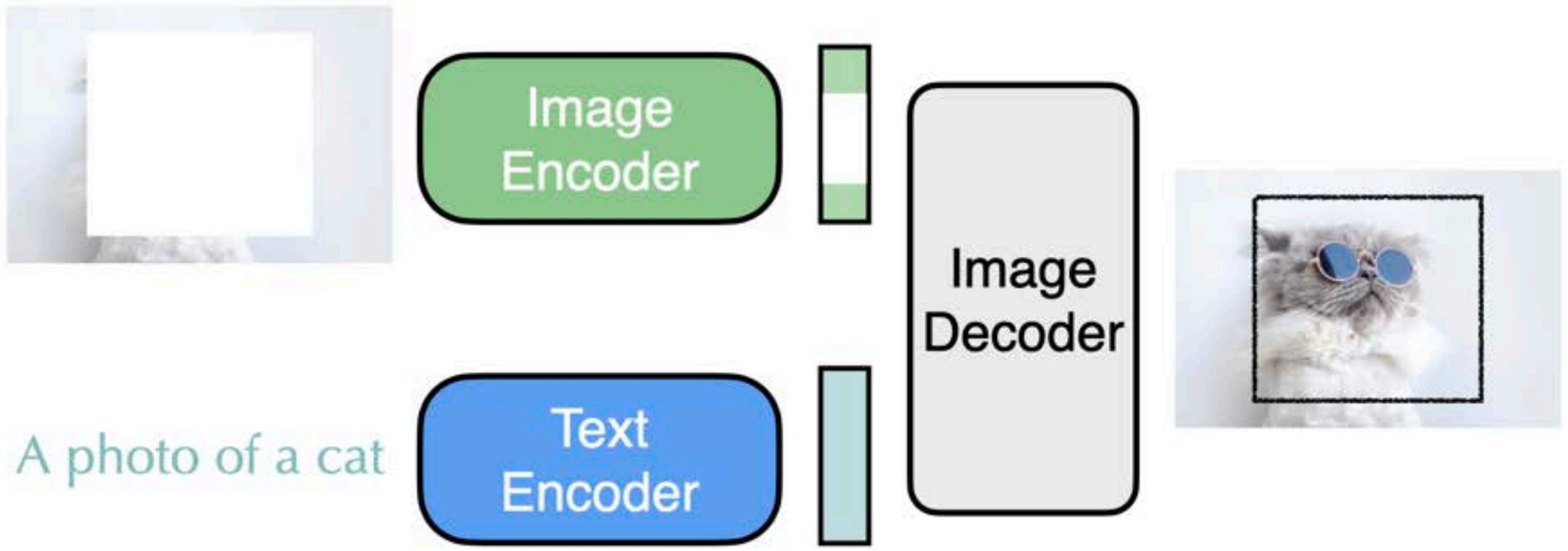
A sign that says "Diffusion".

A black basketball shoe with a lightning bolt on it.

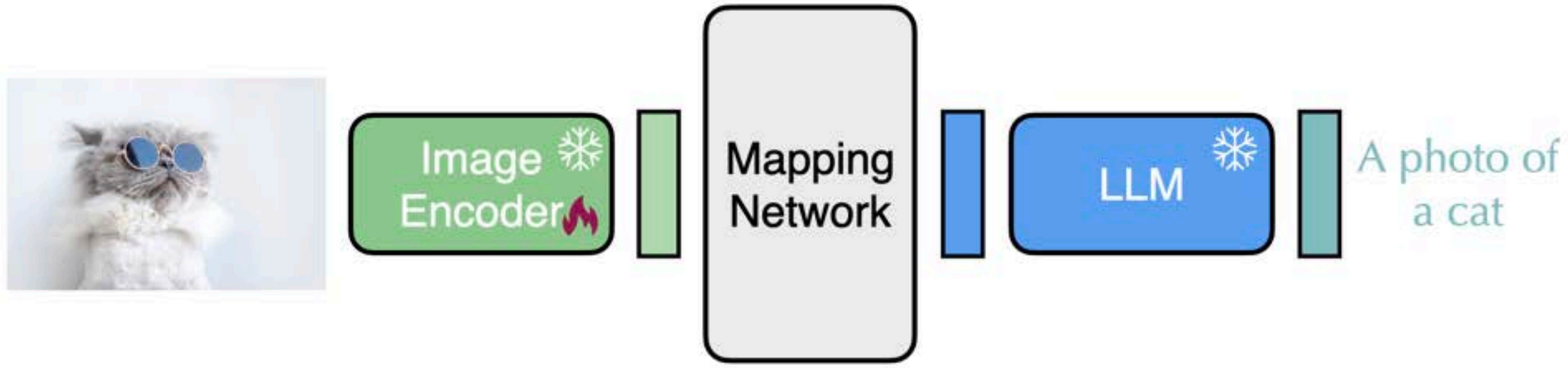
The Families of VLMs



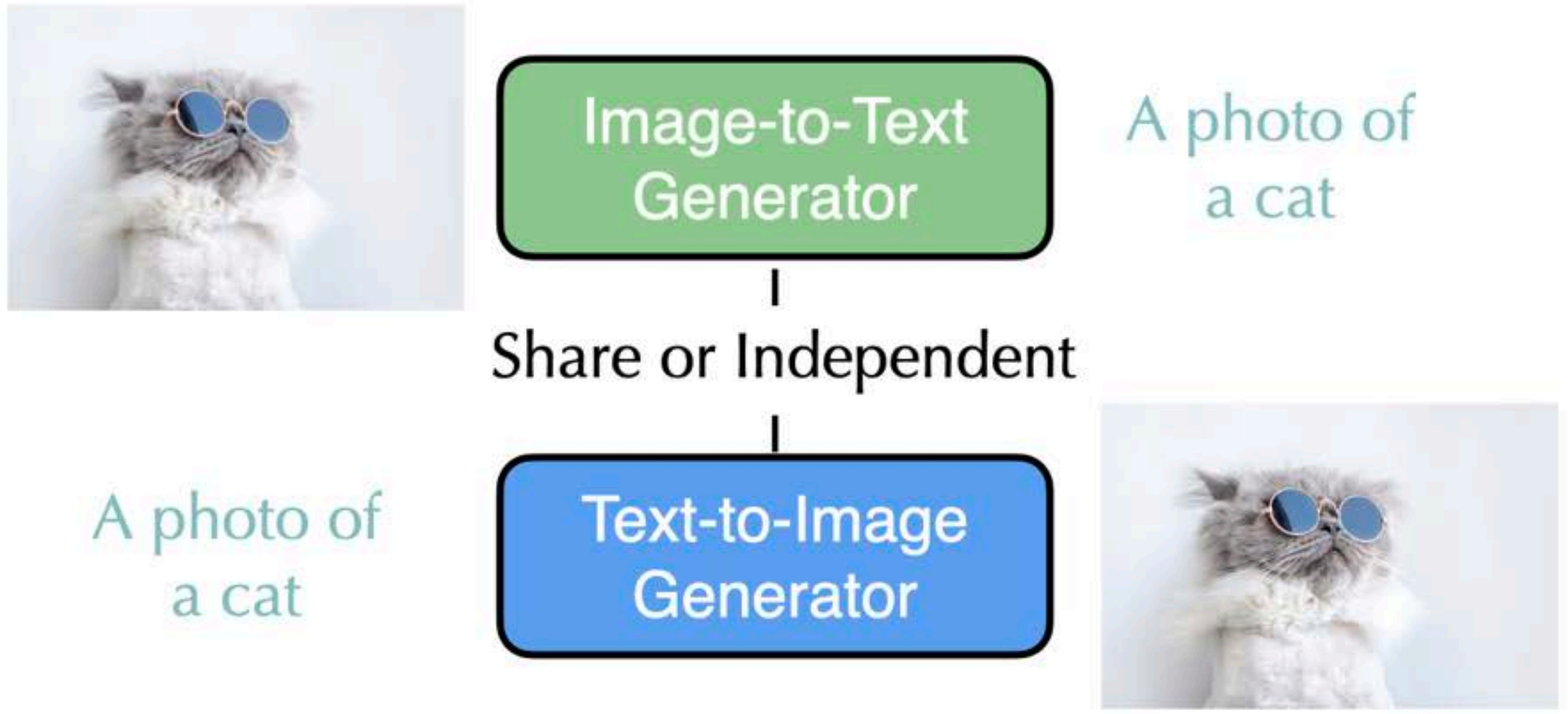
Contrastive-Based



Masking Objective

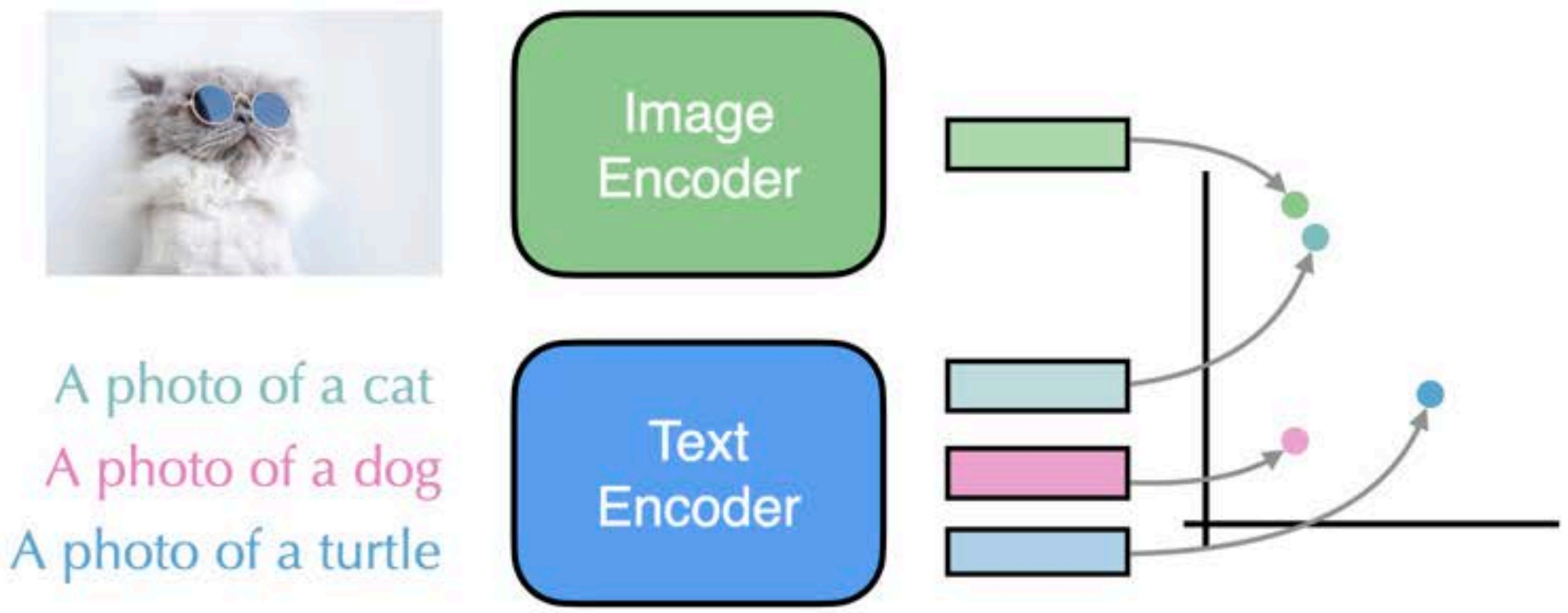


VLMs from Pretrained Backbones



Generative-Based

When to Use Contrastive Models?



Contrastive-Based

Pros:

1. associate text with visual concepts while keeping a **simple training paradigm**
2. **a good base for building more complex model**
3. **retrieve** the images (captions) via prompting the CLIP text (image) encoder with words (images)

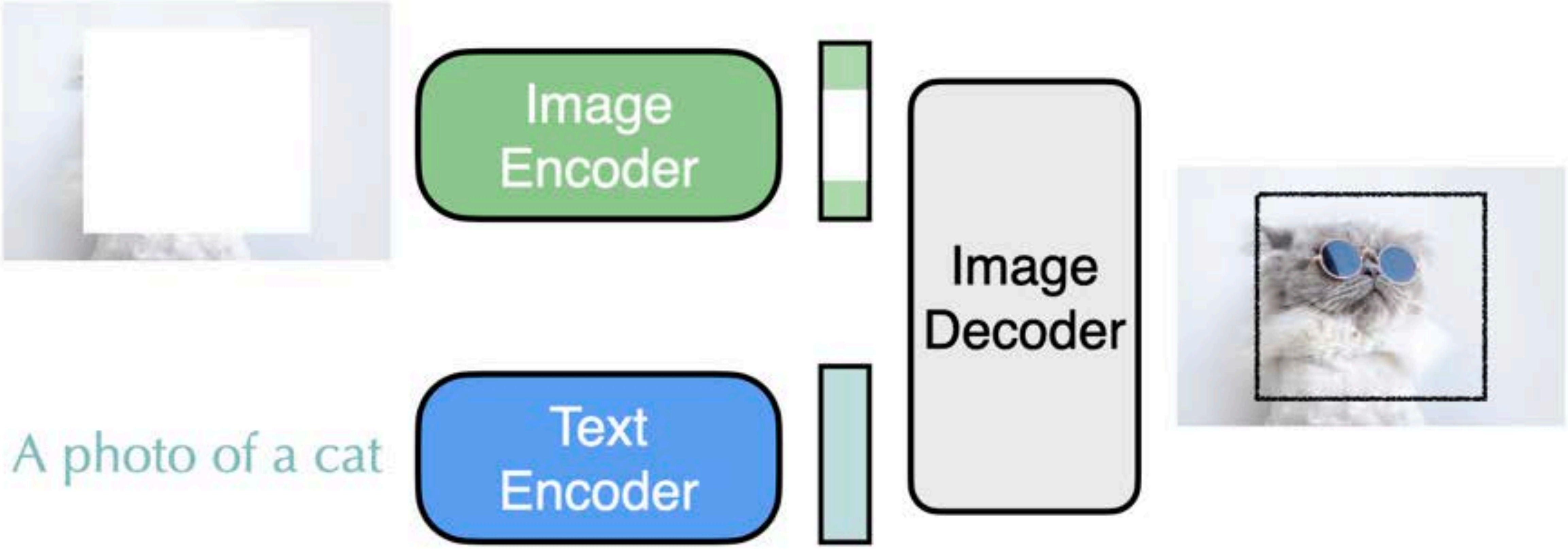
Cons:

1. Is not a generative model, thus it is **not possible to generate a caption**
2. current CLIP models **cannot be used to provide high-level descriptions** of a given image
3. usually needs **a very large dataset as well as large batch sizes** to offer decent performances

When to Use Masking-based Models?

Pros:

1. by learning to reconstruct data from both masked images and text, it **jointly models their distributions**
2. Removing negative examples can enable the use of **smaller mini-batches** without the need to finetune additional hyper-parameters



Masking Objective

Cons:

1. **need to leverage a decoder** to map back the representation to the input space
2. an additional decoder might add an additional bottleneck which might make these methods **less efficient than a purely contrastive one.**

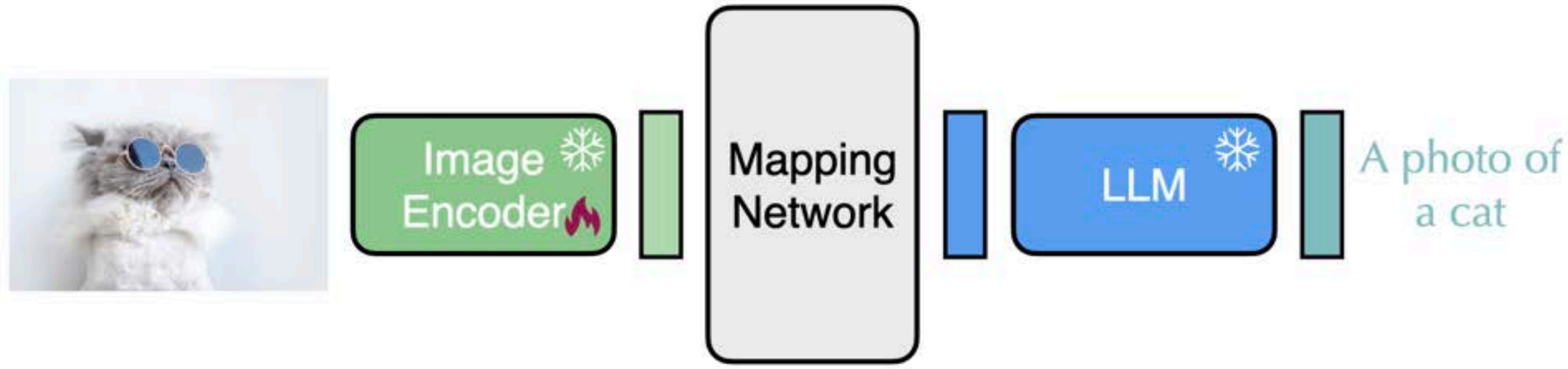
When to Use Pretrained Backbones?

Pros:

1. Can work with limited resource
2. Simple pipeline and framework

Cons:

1. VLMs will be impacted by the potential **hallucination of the LLM**.
2. VLMs could also be impacted by any **bias coming from the pretrained models**.
3. there might be an additional overhead in trying to correct the defect of the vision model or of the LLM.



VLMs from Pretrained Backbones

When to Use Generative Objectives?

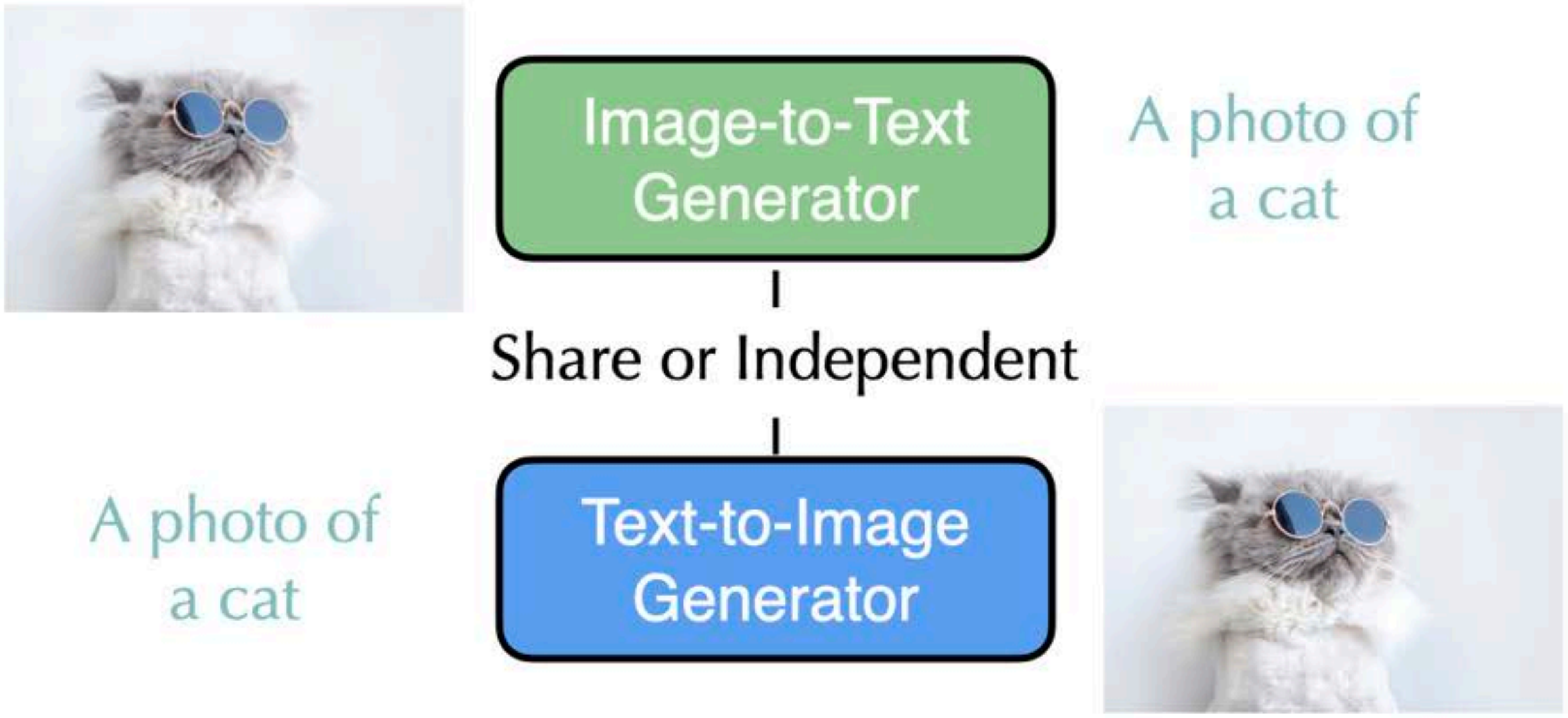
Pros:

1. it might be easier to understand and assess what the model has learned when it is able to decode abstract representations in the input data space
2. can learn an implicit joint distribution between text and images, which might be more suited for learning good representations than leveraging pretrained unimodal encoders.

Cons:

1. they are more computationally expensive to train than their contrastive learning counterpart.
2. Not easy to train, especially when having two generative tasks (T2I and I2T)

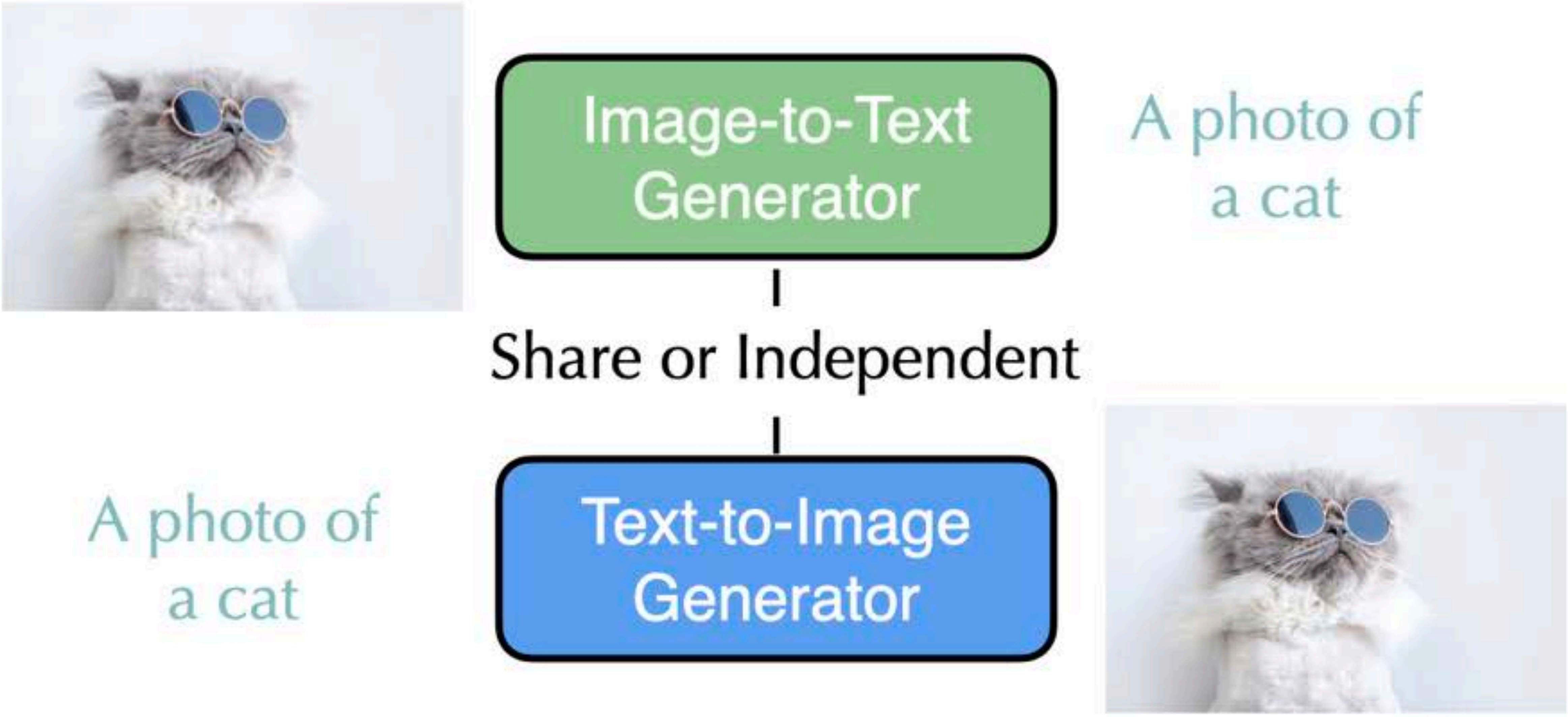
VLMs from Pretrained Backbones



Generative-Based

Is the ability to generate images from words a crucial step towards developing an effective world model?

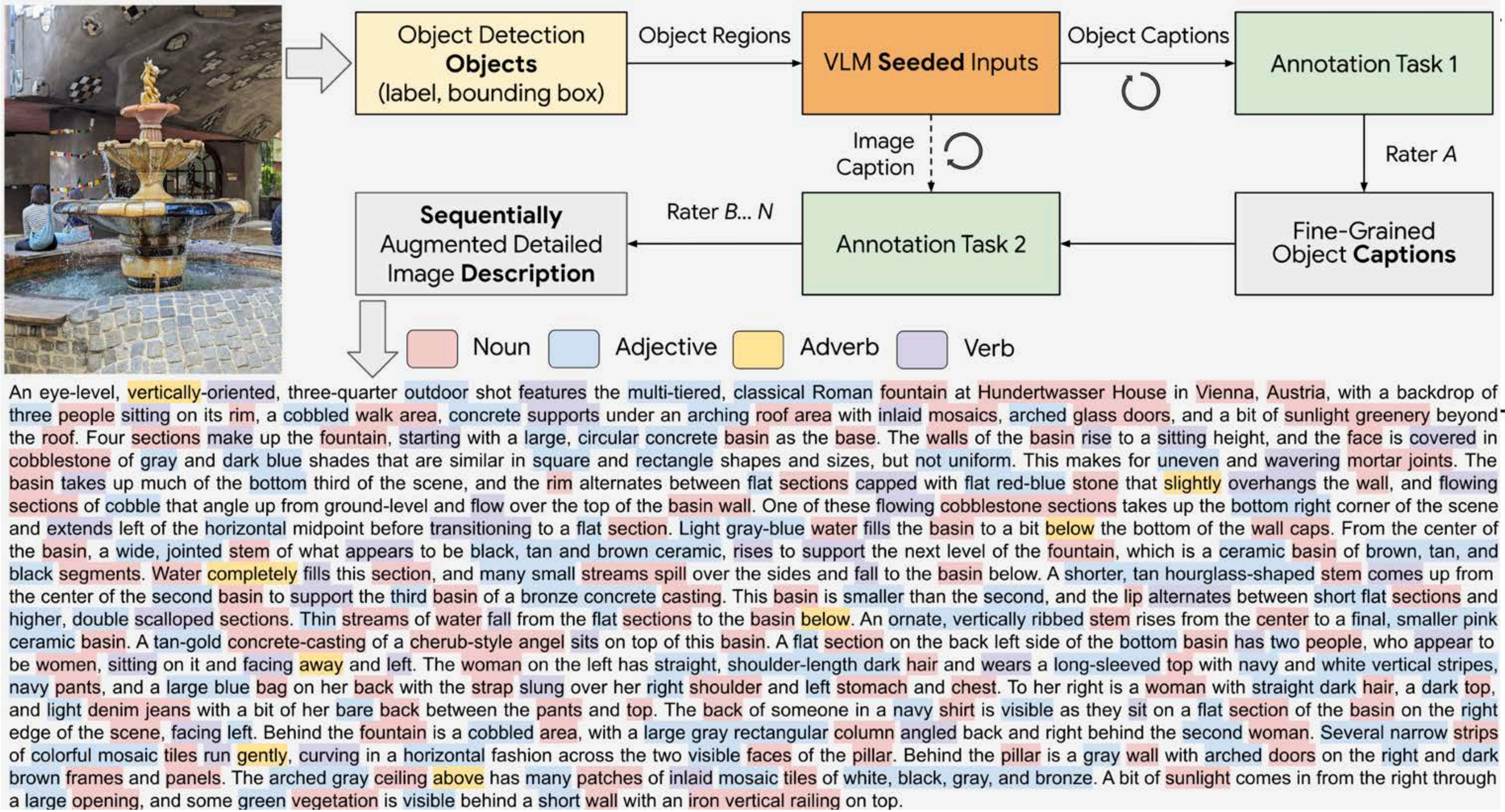
Is such a reconstruction step truly necessary?



Important Considerations When Training VLMs.

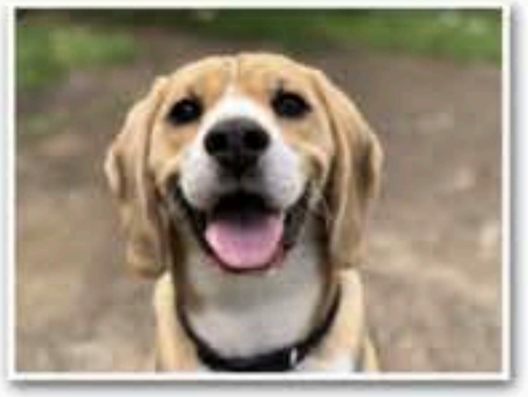
Data
Remove duplicate
Balancing
Pruning
A photo of a dog
A drawing of a bird
Improving
“A beagle dog is smiling and looking at the camera”

Important Considerations When Training VLMs.



An eye-level, vertically-oriented, three-quarter outdoor shot features the multi-tiered, classical Roman fountain at Hundertwasser House in Vienna, Austria, with a backdrop of three people sitting on its rim, a cobbled walk area, concrete supports under an arching roof area with inlaid mosaics, arched glass doors, and a bit of sunlight greenery beyond the roof. Four sections make up the fountain, starting with a large, circular concrete basin as the base. The walls of the basin rise to a sitting height, and the face is covered in cobblestone of gray and dark blue shades that are similar in square and rectangle shapes and sizes, but not uniform. This makes for uneven and wavering mortar joints. The basin takes up much of the bottom third of the scene, and the rim alternates between flat sections capped with flat red-blue stone that slightly overhangs the wall, and flowing sections of cobble that angle up from ground-level and flow over the top of the basin wall. One of these flowing cobblestone sections takes up the bottom right corner of the scene and extends left of the horizontal midpoint before transitioning to a flat section. Light gray-blue water fills the basin to a bit below the bottom of the wall caps. From the center of the basin, a wide, jointed stem of what appears to be black, tan and brown ceramic, rises to support the next level of the fountain, which is a ceramic basin of brown, tan, and black segments. Water completely fills this section, and many small streams spill over the sides and fall to the basin below. A shorter, tan hourglass-shaped stem comes up from the center of the second basin to support the third basin of a bronze concrete casting. This basin is smaller than the second, and the lip alternates between short flat sections and higher, double scalloped sections. Thin streams of water fall from the flat sections to the basin below. An ornate, vertically ribbed stem rises from the center to a final, smaller pink ceramic basin. A tan-gold concrete-casting of a cherub-style angel sits on top of this basin. A flat section on the back left side of the bottom basin has two people, who appear to be women, sitting on it and facing away and left. The woman on the left has straight, shoulder-length dark hair and wears a long-sleeved top with navy and white vertical stripes, navy pants, and a large blue bag on her back with the strap slung over her right shoulder and left stomach and chest. To her right is a woman with straight dark hair, a dark top, and light denim jeans with a bit of her bare back between the pants and top. The back of someone in a navy shirt is visible as they sit on a flat section of the basin on the right edge of the scene, facing left. Behind the fountain is a cobbled area, with a large gray rectangular column angled back and right behind the second woman. Several narrow strips of colorful mosaic tiles run gently, curving in a horizontal fashion across the two visible faces of the pillar. Behind the pillar is a gray wall with arched doors on the right and dark brown frames and panels. The arched gray ceiling above has many patches of inlaid mosaic tiles of white, black, gray, and bronze. A bit of sunlight comes in from the right through a large opening, and some green vegetation is visible behind a short wall with an iron vertical railing on top.

Improving



"A beagle dog is smiling and looking at the camera"

ImageInWords: Unlocking Hyper-Detailed Image Descriptions. ArXiv 2024

Important Considerations When Training VLMs.

Data

Remove duplicate

Balancing

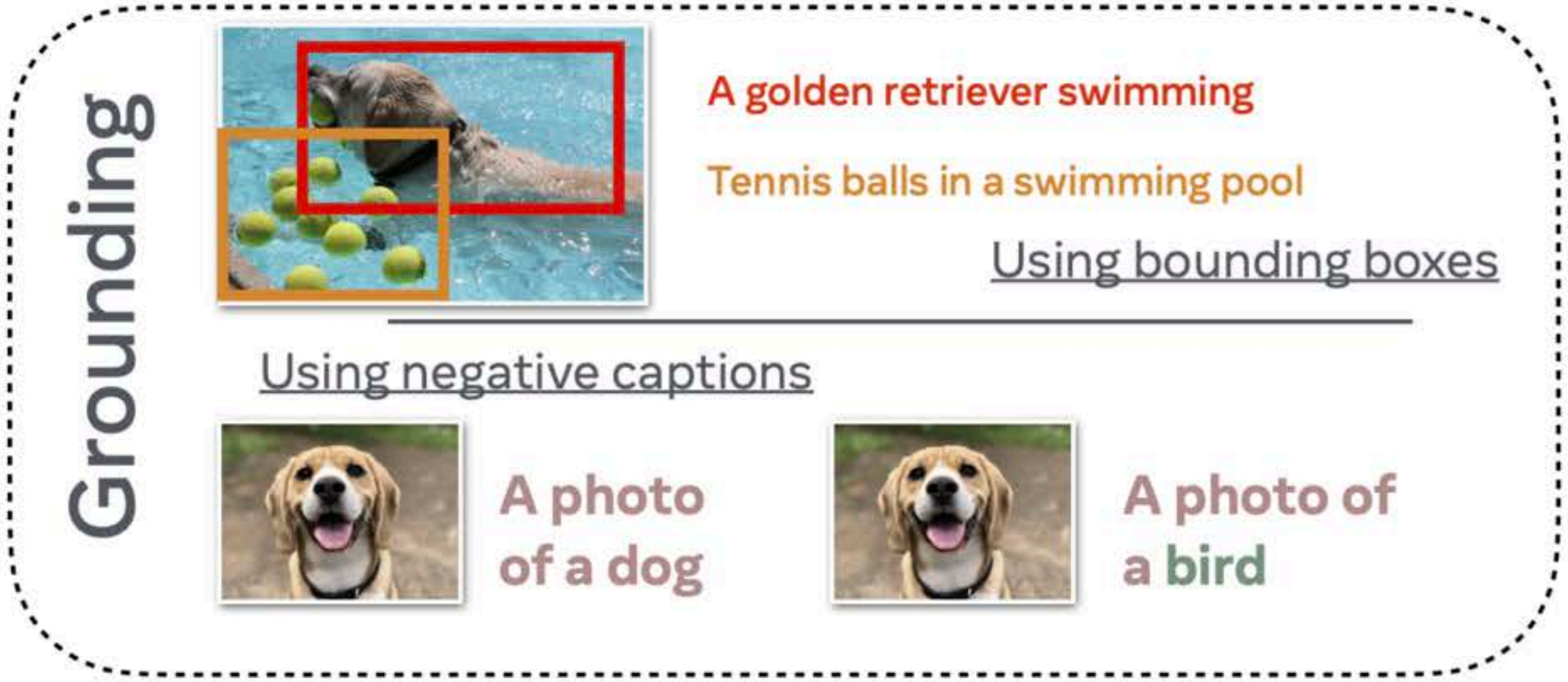
Pruning

A photo of a dog ✓

A drawing of a bird ✗

Improving

“A beagle dog is smiling and looking at the camera”



Grounding

A golden retriever swimming

Tennis balls in a swimming pool

Using bounding boxes

Using negative captions

A photo of a dog

A photo of a bird

Important Considerations When Training VLMs.

Data

Remove duplicate

Balancing

Pruning

A photo of a dog ✓

A drawing of a bird ✗

Improving

"A beagle dog is smiling and looking at the camera"

Grounding

A golden retriever swimming

Tennis balls in a swimming pool

Using bounding boxes

Using negative captions

A photo of a dog

A photo of a bird

Alignement

Instruction: Provide a description of what is presented in the image.

Answer: This is a photo of a smiling dog

Instruction-tuning

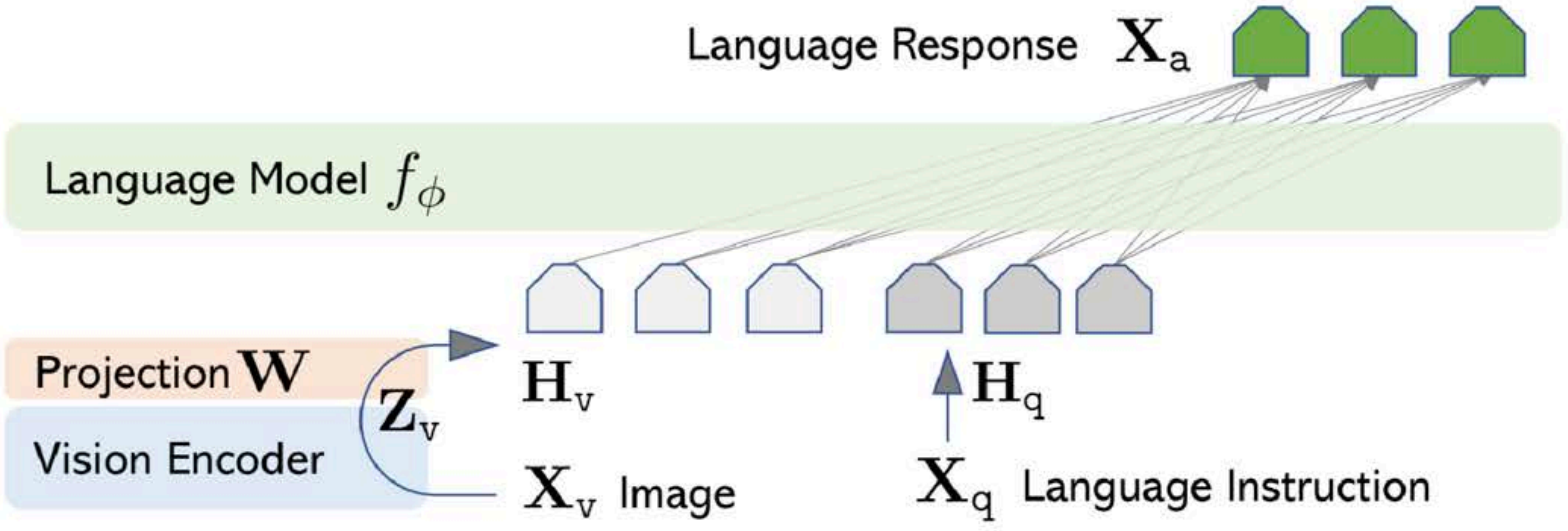
Reinforcement Learning From Human Feedback

VLM: This image is a photo of a smiling dog. 👍

VLM: Hi, I am a robot. It is a photo of a dog. 👎

VLM Training: Alignment / Instruction Tuning - LLaVA

☐ Architecture

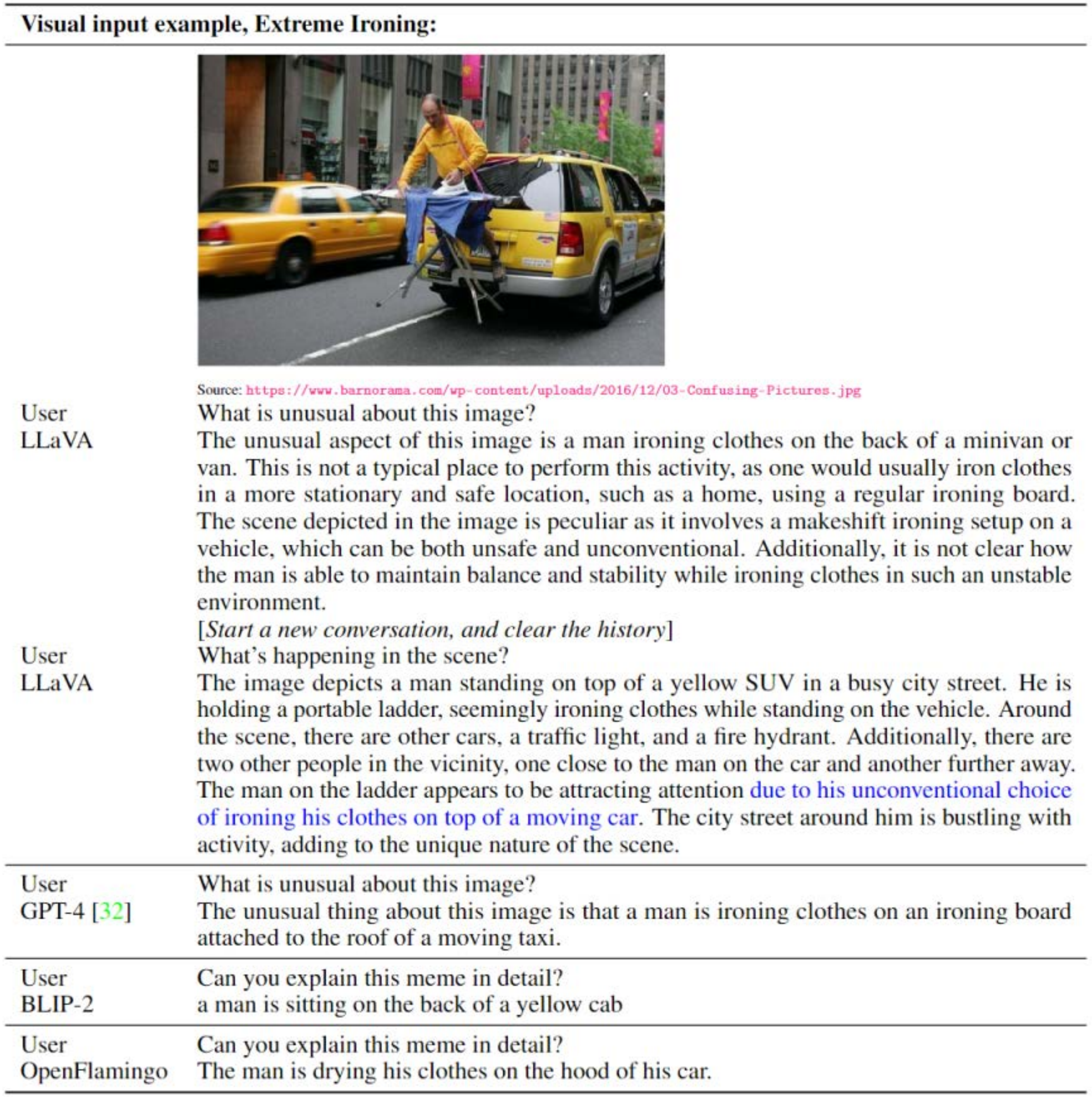


☐ Two-stage Training

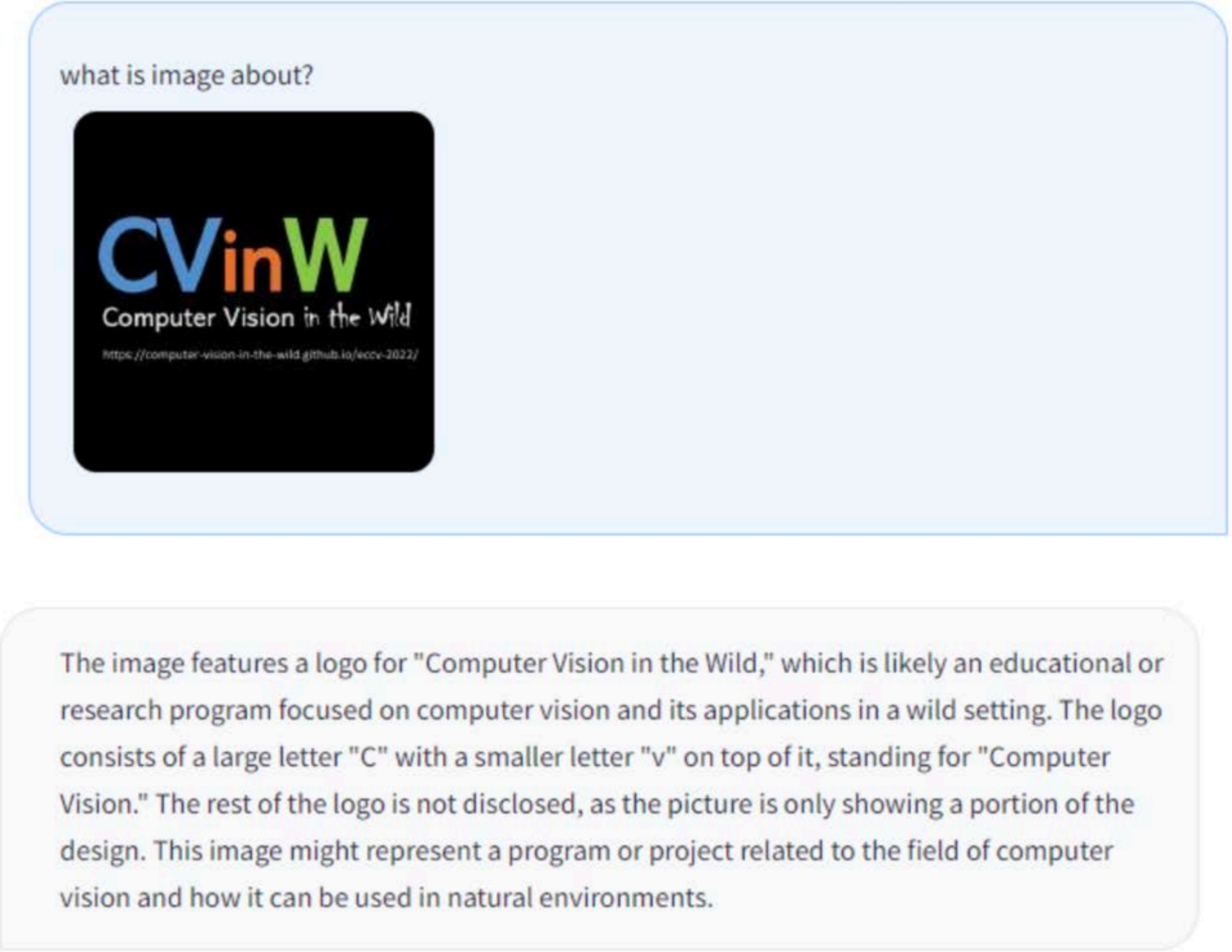
- **Stage 1: Pre-training for Feature Alignment.**
  Only the projection matrix is updated, based on a subset of CC3M.
- **Stage 2: Fine-tuning End-to-End.** Both the projection matrix and LLM are updated
  - **Visual Chat**: Our generated multimodal instruction data for daily user-oriented applications.
  - **Science QA**: Multimodal reasoning dataset for the science domain.

Liu, Haotian, et al. "Visual instruction tuning." *NeurIPS 2023*

VLM Training: Alignment / Instruction Tuning - LLaVA



Visual input example, Extreme Ironing:	
	Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg
User	What is unusual about this image?
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment. [*Start a new conversation, and clear the history*]
User	What's happening in the scene?
LLaVA	The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.
User	What is unusual about this image?
GPT-4 [32]	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
User	Can you explain this meme in detail?
BLIP-2	a man is sitting on the back of a yellow cab
User	Can you explain this meme in detail?
OpenFlamingo	The man is drying his clothes on the hood of his car.

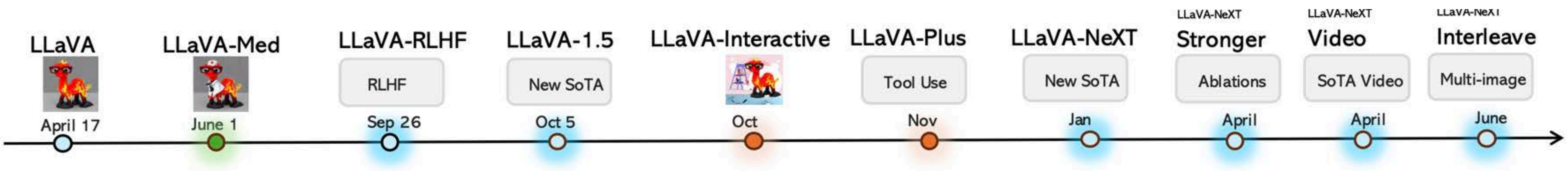


what is image about?

CVinW
Computer Vision in the Wild
https://computer-vision-in-the-wild.github.io/eccv-2022/

The image features a logo for "Computer Vision in the Wild," which is likely an educational or research program focused on computer vision and its applications in a wild setting. The logo consists of a large letter "C" with a smaller letter "v" on top of it, standing for "Computer Vision." The rest of the logo is not disclosed, as the picture is only showing a portion of the design. This image might represent a program or project related to the field of computer vision and how it can be used in natural environments.

VLM Training: Alignment / Instruction Tuning - LLaVA



LLaVA
April 17

LLaVA-Med
June 1

LLaVA-RLHF
RLHF
Sep 26

LLaVA-1.5
New SoTA
Oct 5

LLaVA-Interactive
Oct

LLaVA-Plus
Tool Use
Nov

LLaVA-NeXT
New SoTA
Jan

LLaVA-NeXT
Stronger
Ablations
April

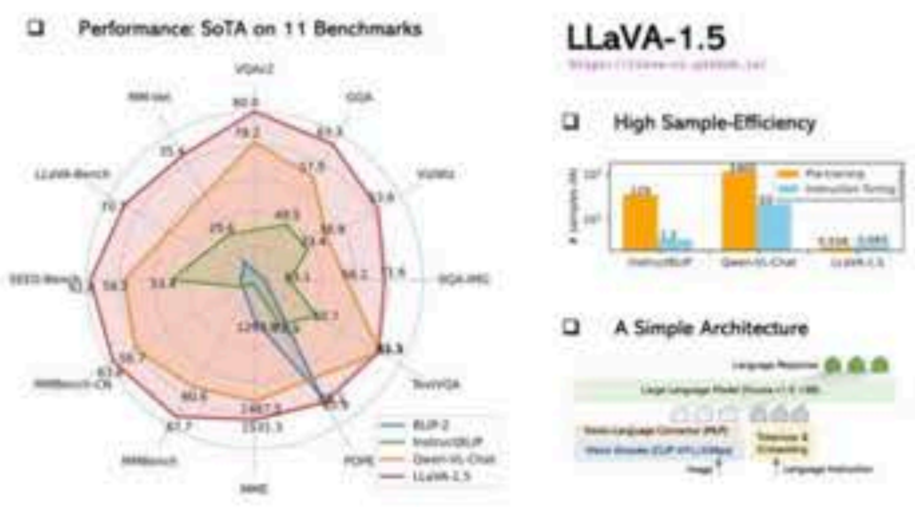
LLaVA-NeXT
Video
SoTA Video
April

LLaVA-NeXT
Interleave
Multi-image
June

Performance

- LLaVA-1.5
- LLaVA-NeXT

New SoTA among open LMMs



Applications

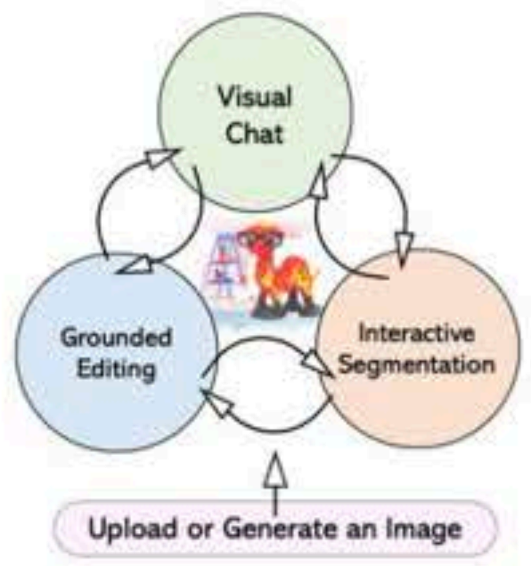
- LLaVA-Med
- LLaVA-Rad

The first open-source healthcare LMM

New Capabilities

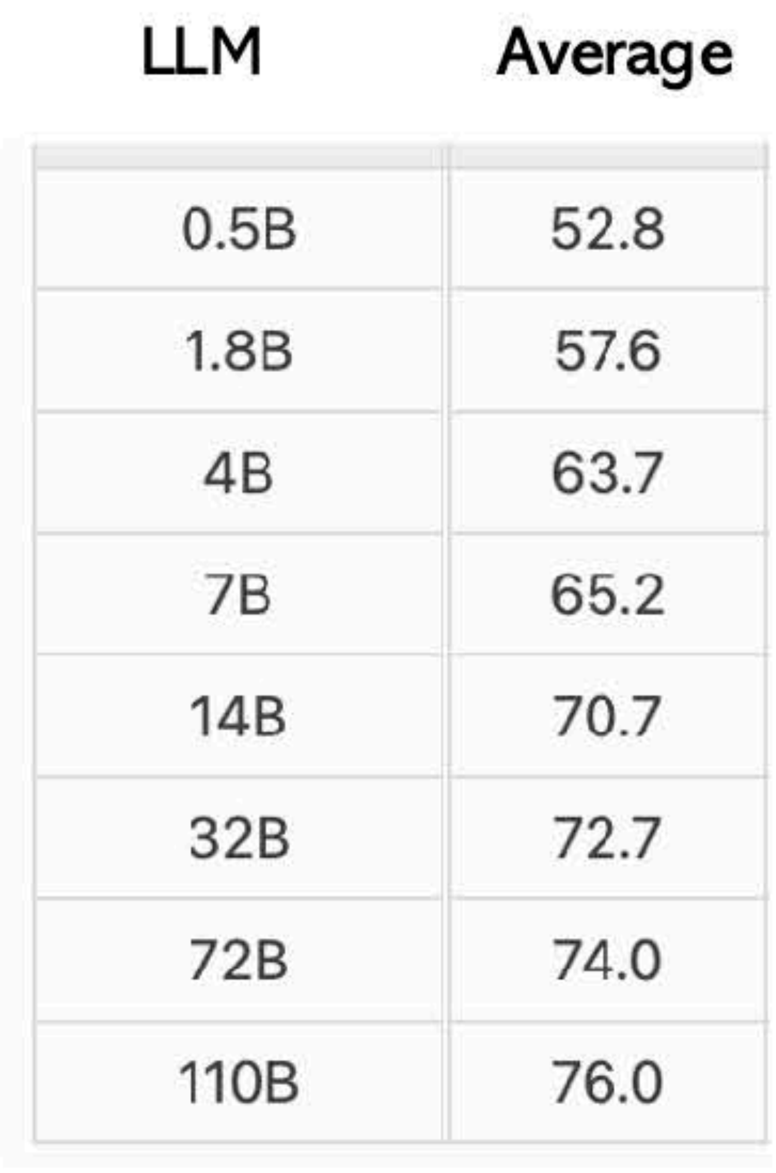
- LLaVA Interactive
- LLaVA-Plus

Tool use to expand capabilities

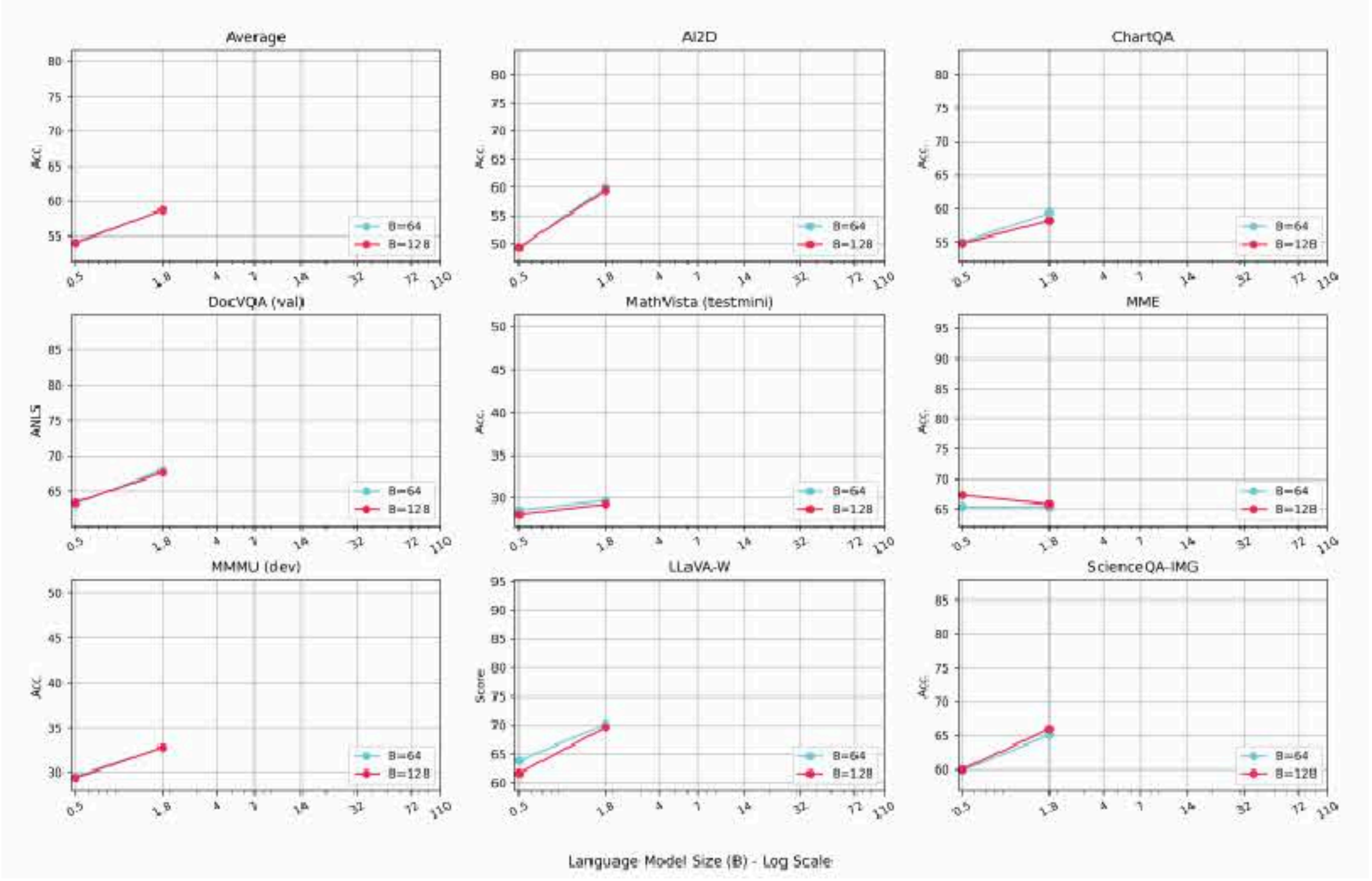


Practical Tricks to Improve VLMs: 1) Architecture

Scaling Language Models
(0.5B - 110B)

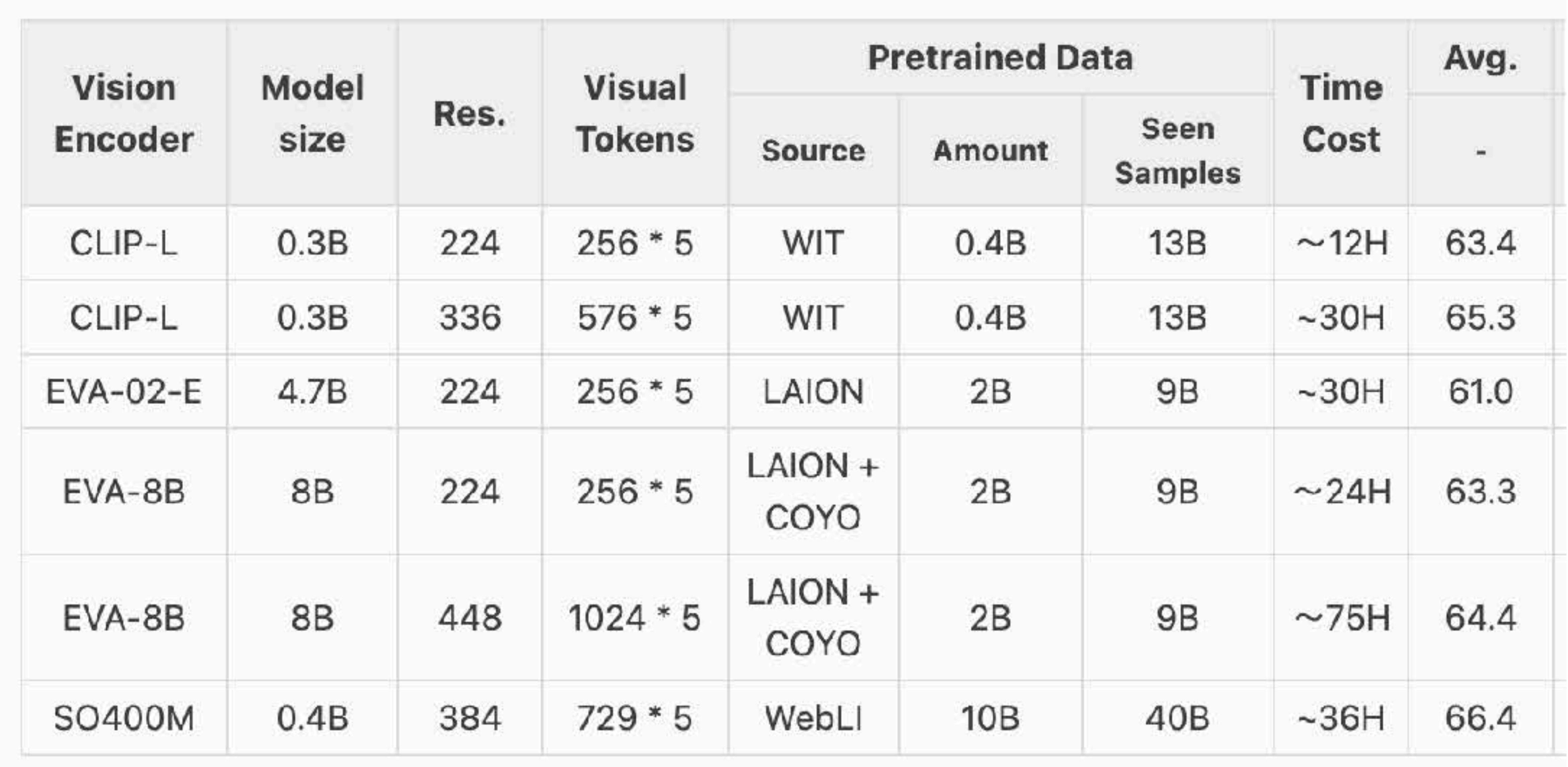


LLM	Average
0.5B	52.8
1.8B	57.6
4B	63.7
7B	65.2
14B	70.7
32B	72.7
72B	74.0
110B	76.0



Model size scaling is effective

Scaling Vision Encoders
(0.3B - 10B)



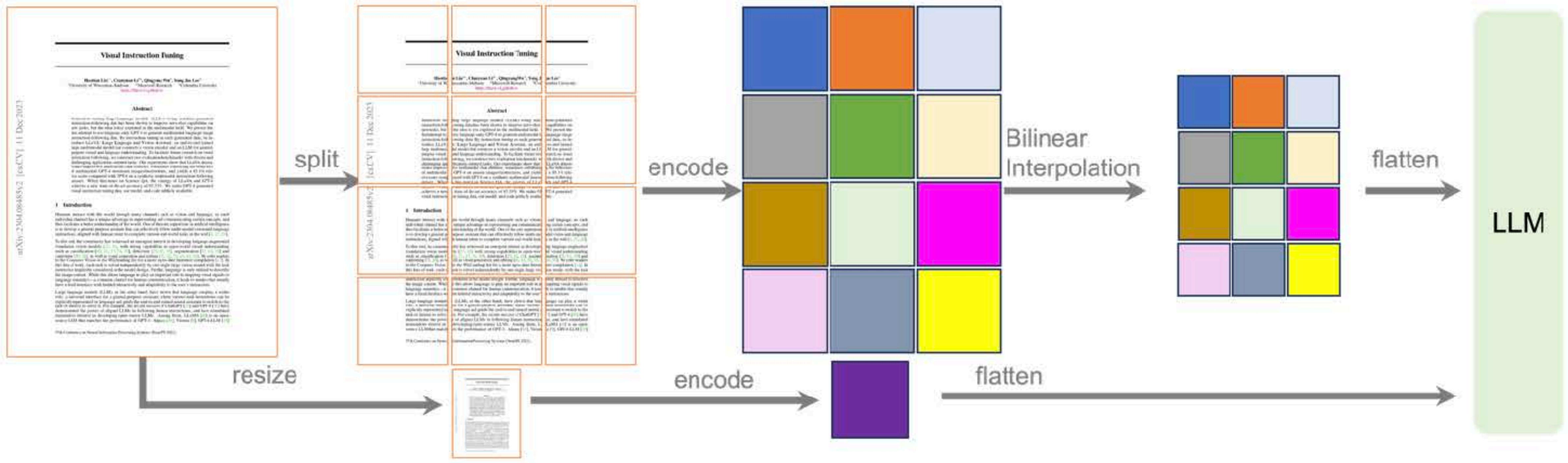
Vision Encoder	Model size	Res.	Visual Tokens	Pretrained Data			Time Cost	Avg.
				Source	Amount	Seen Samples		-
CLIP-L	0.3B	224	256 * 5	WIT	0.4B	13B	~12H	63.4
CLIP-L	0.3B	336	576 * 5	WIT	0.4B	13B	~30H	65.3
EVA-02-E	4.7B	224	256 * 5	LAION	2B	9B	~30H	61.0
EVA-8B	8B	224	256 * 5	LAION + COYO	2B	9B	~24H	63.3
EVA-8B	8B	448	1024 * 5	LAION + COYO	2B	9B	~75H	64.4
SO400M	0.4B	384	729 * 5	WebLI	10B	40B	~36H	66.4

Model size scaling is less effective than training data and visual representations

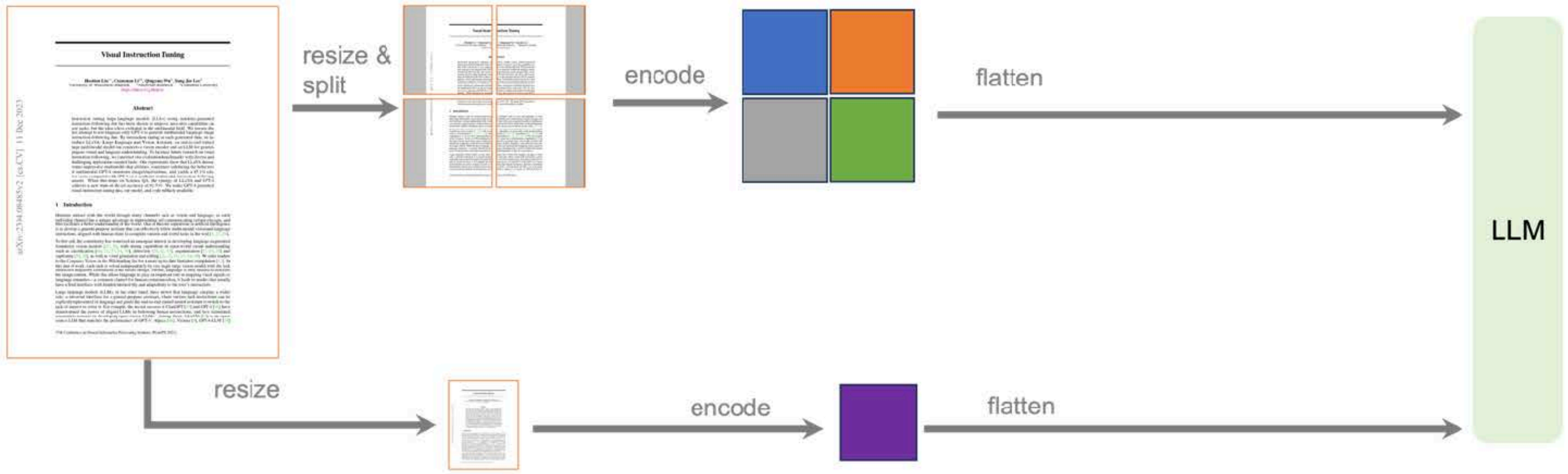
Liu, Haotian, et al. "LLaVA-NeXT: Improved reasoning, OCR, and world knowledge" *2024*

Practical Tricks to Improve VLMs: 2) Visual Representation

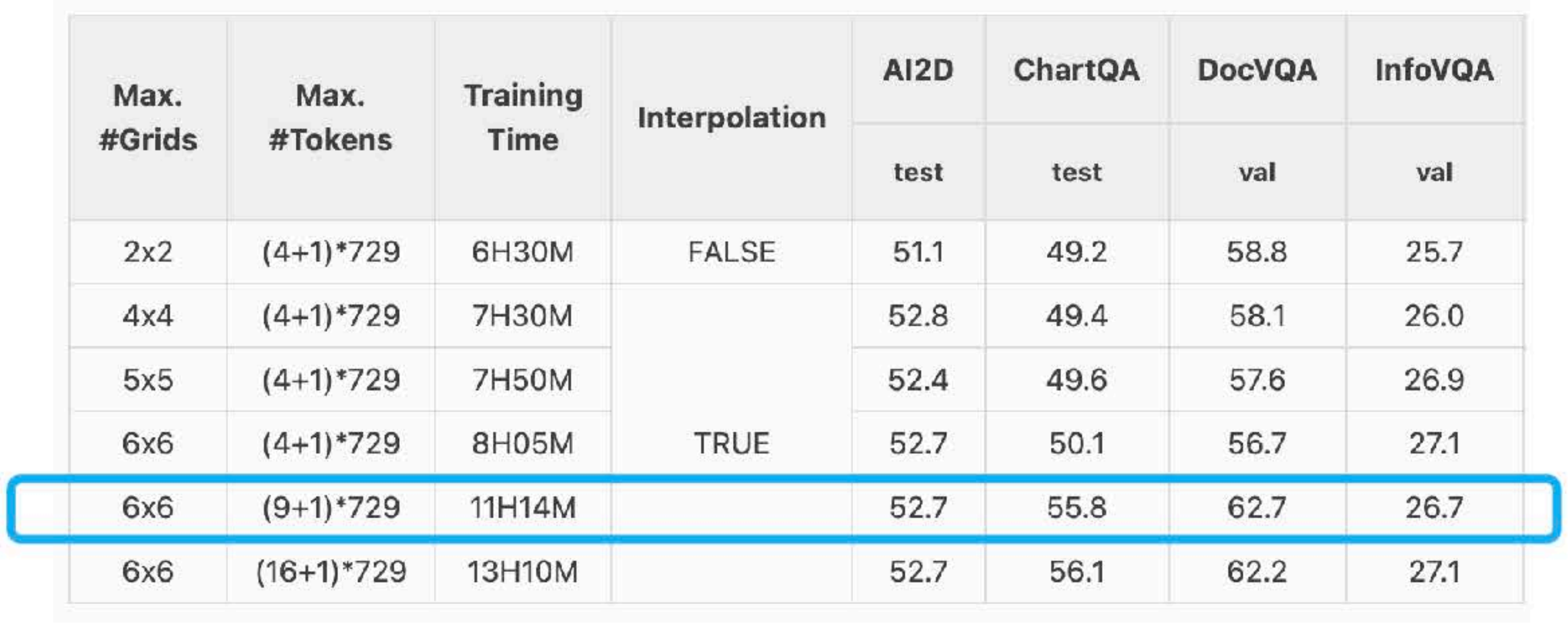
Scaling Up Both Image Resolution and #Tokens
(# Tokens: 729 - (9+1)*729)



(a) Higher-AnyRes with Bilinear Interpolation



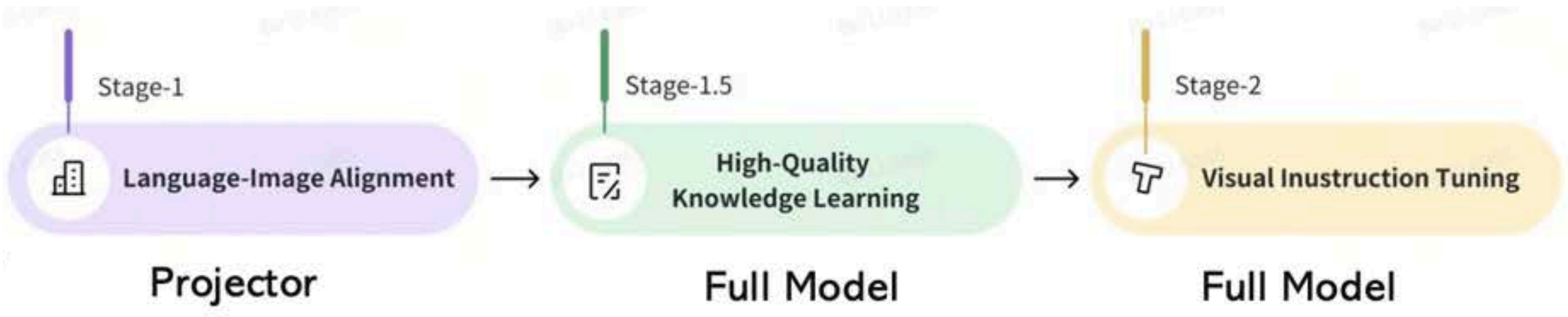
(b) The original AnyRes



Max. #Grids	Max. #Tokens	Training Time	Interpolation	AI2D	ChartQA	DocVQA	InfoVQA
				test	test	val	val
2x2	(4+1)*729	6H30M	FALSE	51.1	49.2	58.8	25.7
4x4	(4+1)*729	7H30M	TRUE	52.8	49.4	58.1	26.0
5x5	(4+1)*729	7H50M		52.4	49.6	57.6	26.9
6x6	(4+1)*729	8H05M		52.7	50.1	56.7	27.1
6x6	(9+1)*729	11H14M		52.7	55.8	62.7	26.7
6x6	(16+1)*729	13H10M		52.7	56.1	62.2	27.1

Liu, Haotian, et al. "LLaVA-NeXT: Improved reasoning, OCR, and world knowledge" *2024*

Practical Tricks to Improve VLMs: 3) Training Strategy

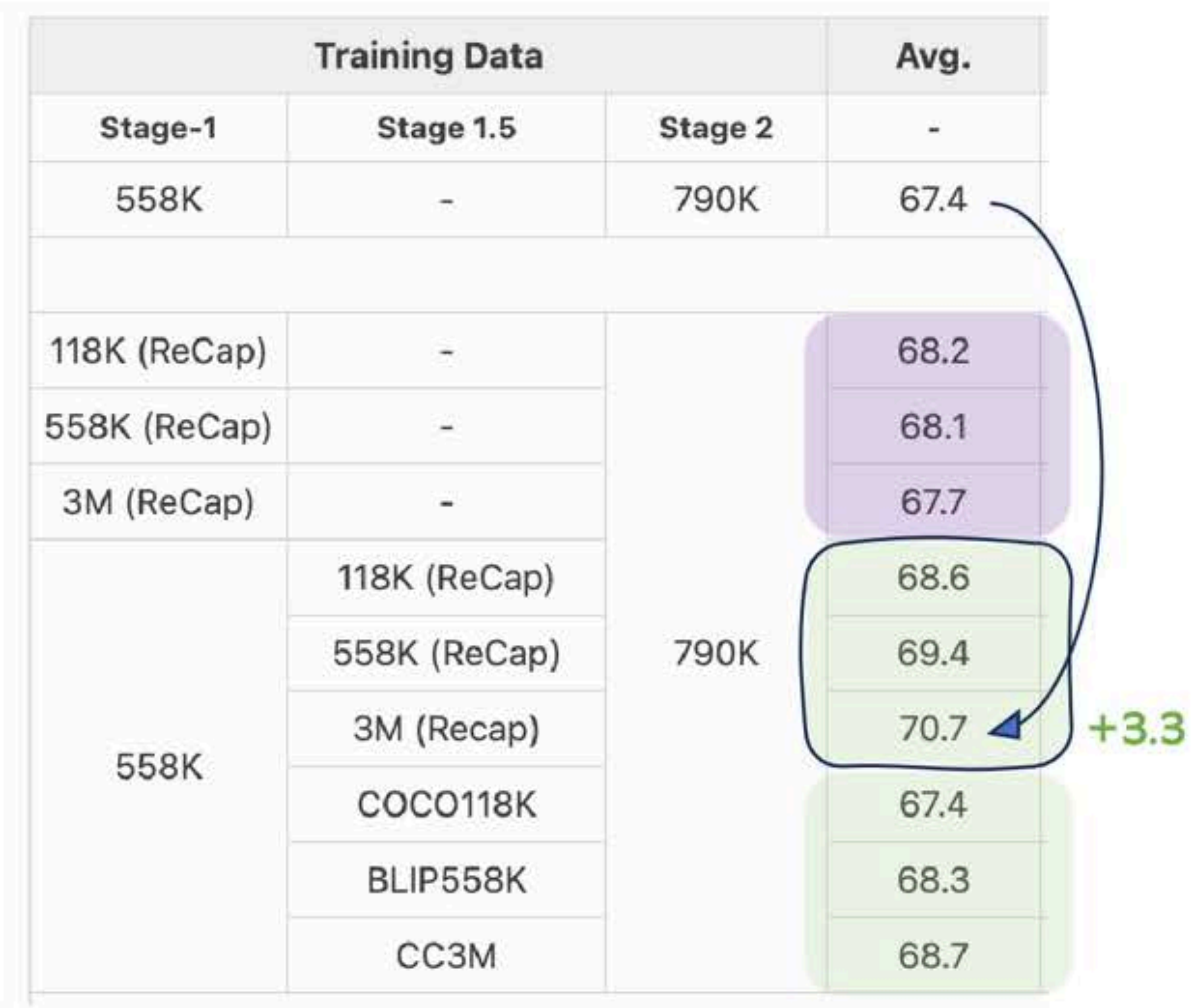


Projector

Full Model

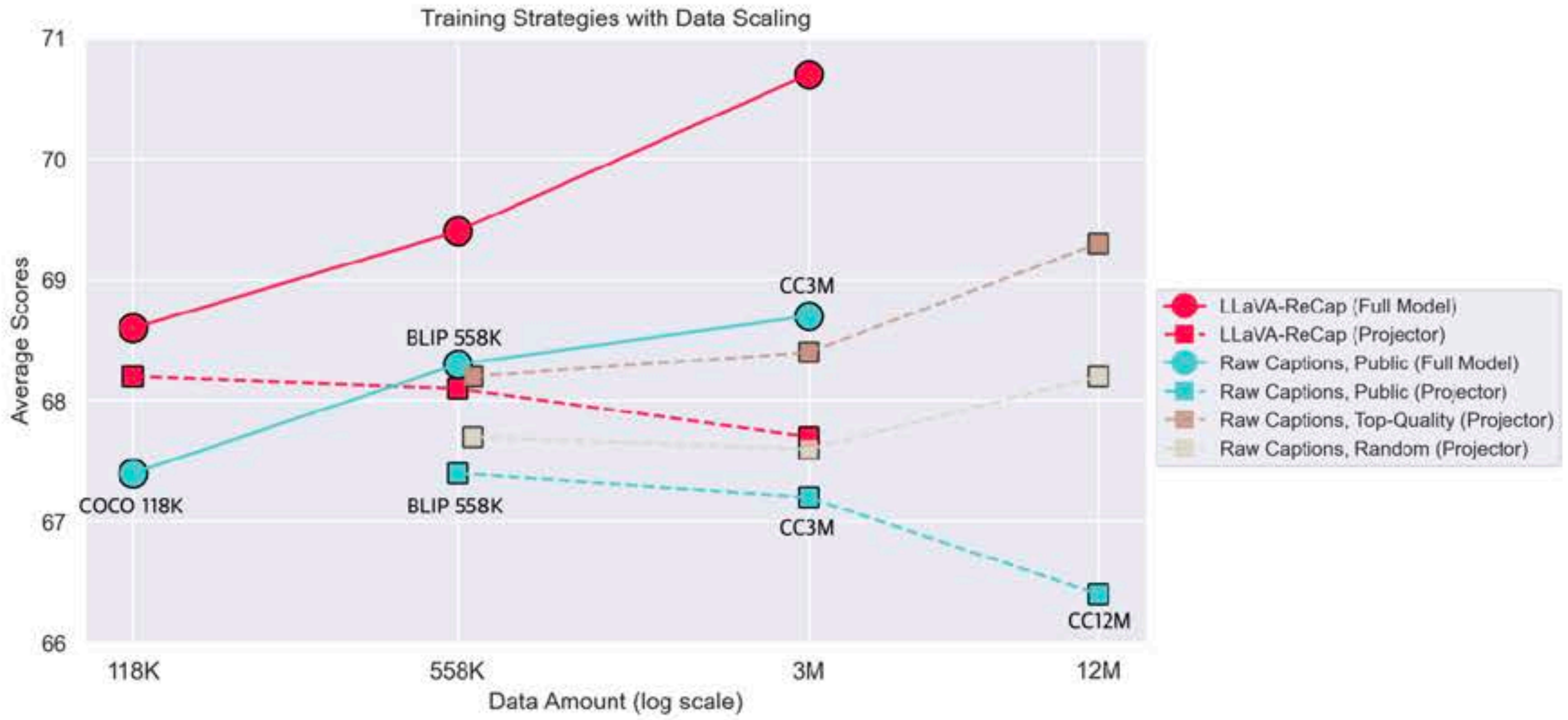
Full Model

The Best Training Strategy:
fully model, high-quality recap data



Training Data			Avg.
Stage-1	Stage 1.5	Stage 2	-
558K	-	790K	67.4
118K (ReCap)	-	790K	68.2
558K (ReCap)	-		68.1
3M (ReCap)	-		67.7
558K	118K (ReCap)		68.6
	558K (ReCap)		69.4
	3M (Recap)		70.7
	COCO118K		67.4
	BLIP558K		68.3
	CC3M		68.7

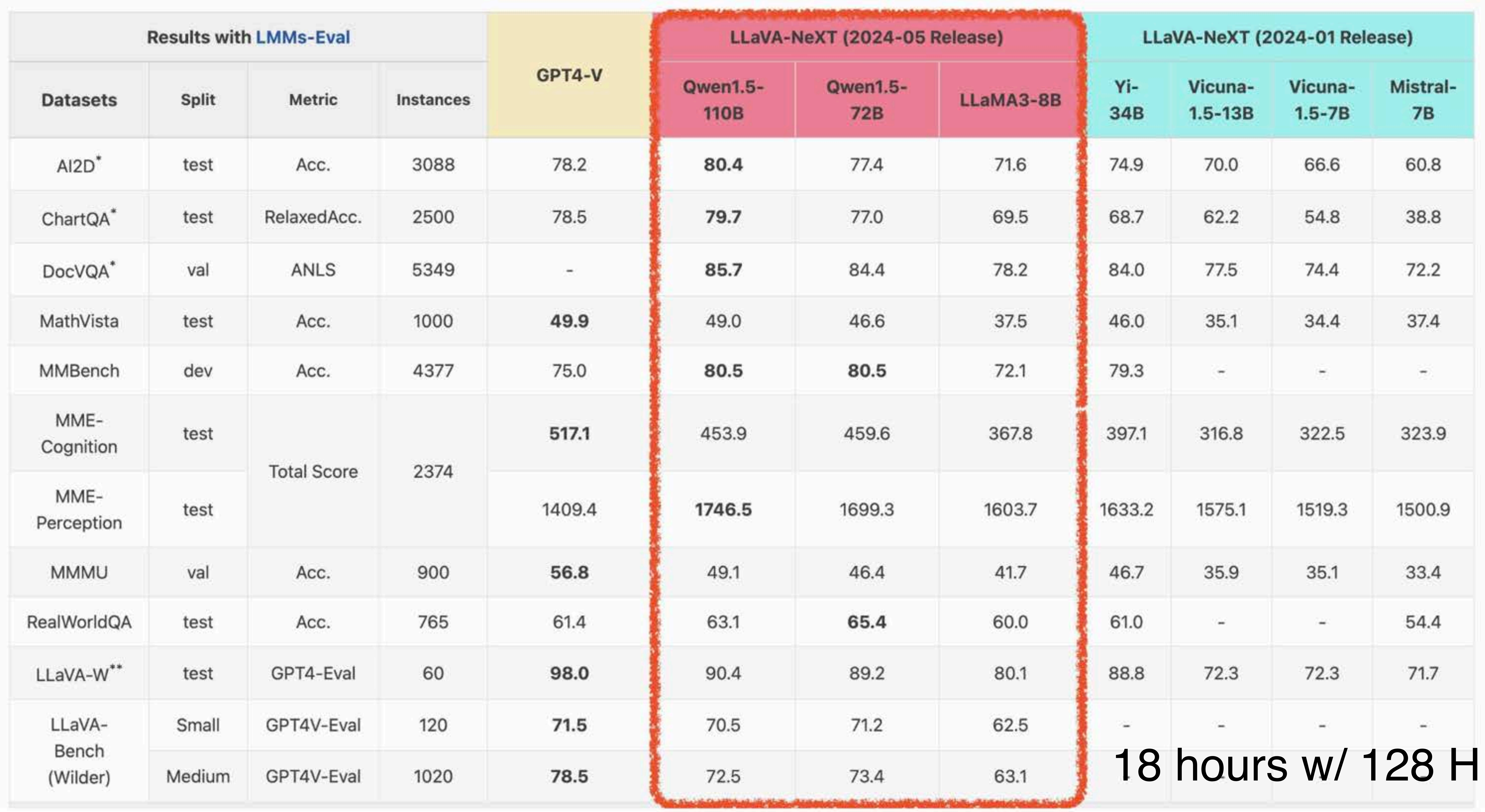
+3.3



Liu, Haotian, et al. "LLaVA-NeXT: Improved reasoning, OCR, and world knowledge" *2024*

VLM Training: Alignment / Instruction Tuning - LLaVA

90% performance on benchmarks!



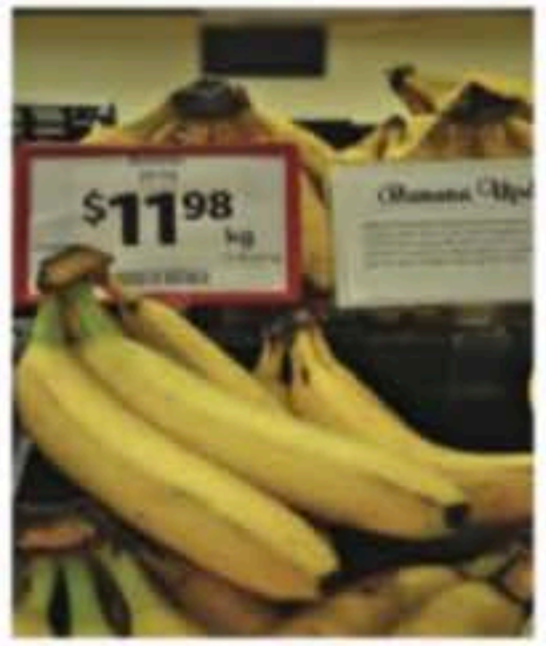
Results with LMMs-Eval				GPT4-V	LLaVA-NeXT (2024-05 Release)			LLaVA-NeXT (2024-01 Release)			
Datasets	Split	Metric	Instances		Qwen1.5-110B	Qwen1.5-72B	LLaMA3-8B	Yi-34B	Vicuna-1.5-13B	Vicuna-1.5-7B	Mistral-7B
AI2D*	test	Acc.	3088	78.2	**80.4**	77.4	71.6	74.9	70.0	66.6	60.8
ChartQA*	test	RelaxedAcc.	2500	78.5	**79.7**	77.0	69.5	68.7	62.2	54.8	38.8
DocVQA*	val	ANLS	5349	-	**85.7**	84.4	78.2	84.0	77.5	74.4	72.2
MathVista	test	Acc.	1000	**49.9**	49.0	46.6	37.5	46.0	35.1	34.4	37.4
MMBench	dev	Acc.	4377	75.0	**80.5**	**80.5**	72.1	79.3	-	-	-
MME-Cognition	test	Total Score	2374	**517.1**	453.9	459.6	367.8	397.1	316.8	322.5	323.9
MME-Perception	test			1409.4	**1746.5**	1699.3	1603.7	1633.2	1575.1	1519.3	1500.9
MMMU	val	Acc.	900	**56.8**	49.1	46.4	41.7	46.7	35.9	35.1	33.4
RealWorldQA	test	Acc.	765	61.4	63.1	**65.4**	60.0	61.0	-	-	54.4
LLaVA-W**	test	GPT4-Eval	60	**98.0**	90.4	89.2	80.1	88.8	72.3	72.3	71.7
LLaVA-Bench (Wilder)	Small	GPT4V-Eval	120	**71.5**	70.5	71.2	62.5	-	-	-	-
	Medium	GPT4V-Eval	1020	**78.5**	72.5	73.4	63.1				

18 hours w/ 128 H100

VLM Evaluation

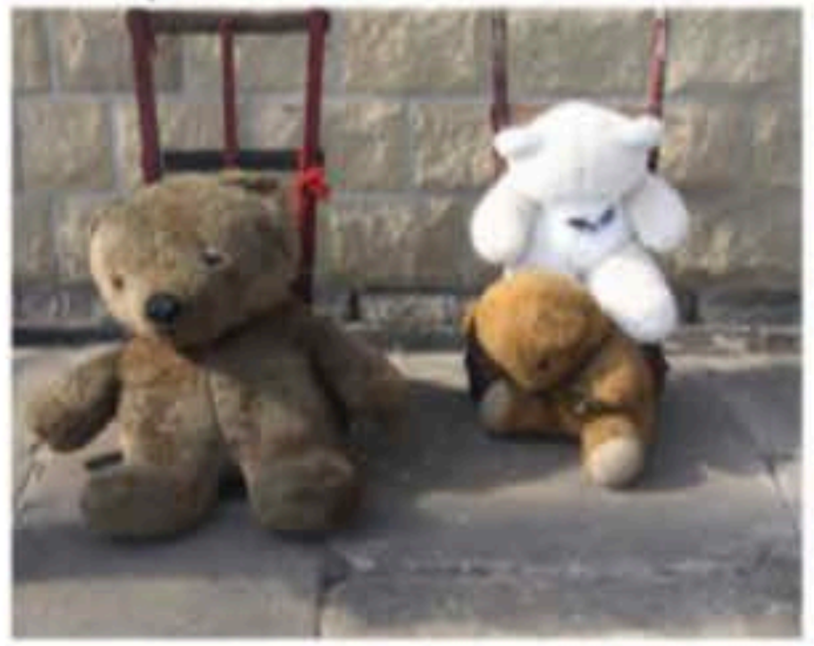
VQA

TextVQA:
Text Reasoning



Q: What is the price of the bananas per kg ? A: $11.98

GQA:
Spatial Reasoning



Q: What animal is sitting on the sidewalk ? A: Bear

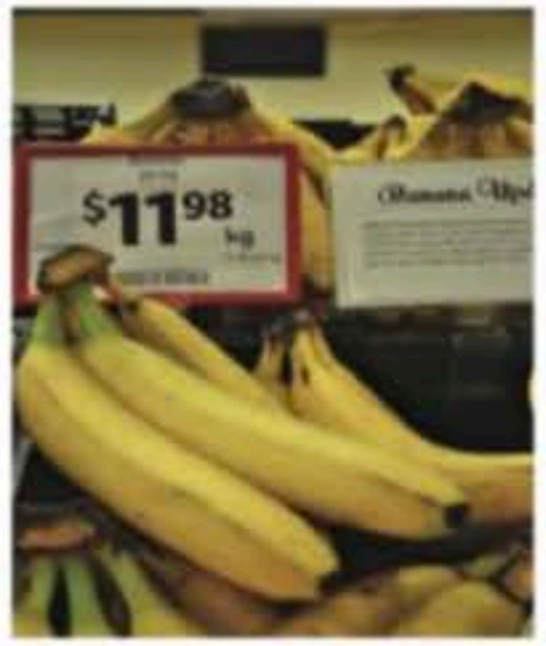
VizWiz:
Unanswerable Questions

Q: What spice is that ?
A: unanswerable

VLM Evaluation

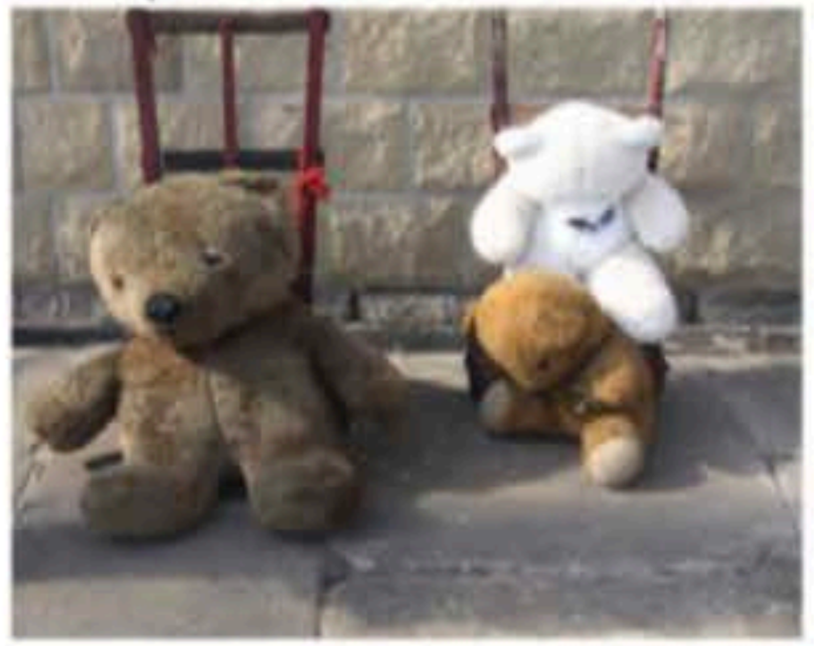
VQA

TextVQA:
Text Reasoning



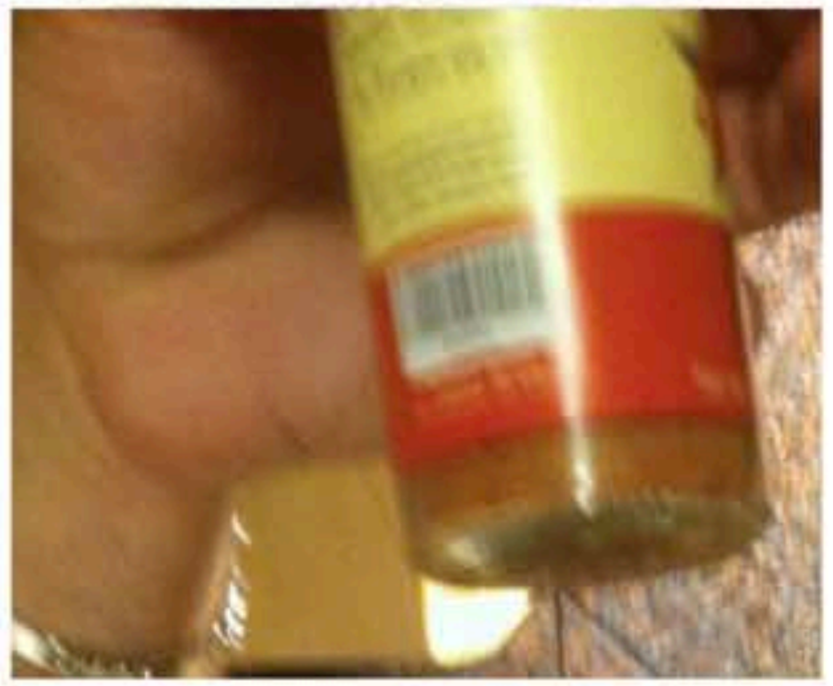
Q: What is the price of the bananas per kg ? A: $11.98

GQA:
Spatial Reasoning



Q: What animal is sitting on the sidewalk ? A: Bear

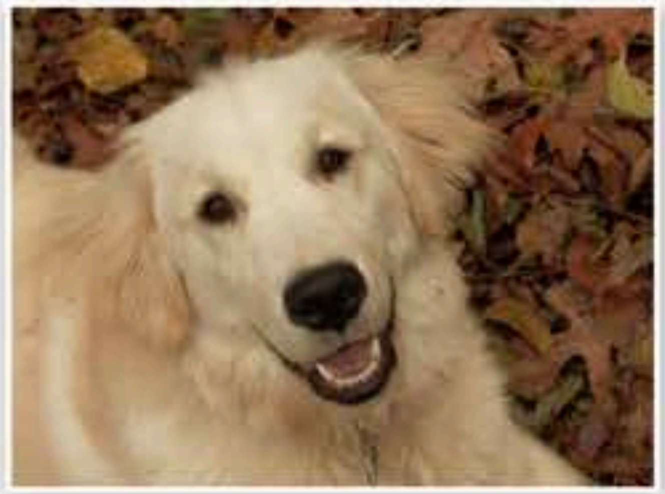
VizWiz:
Unanswerable Questions



Q: What spice is that ?
A: unanswerable

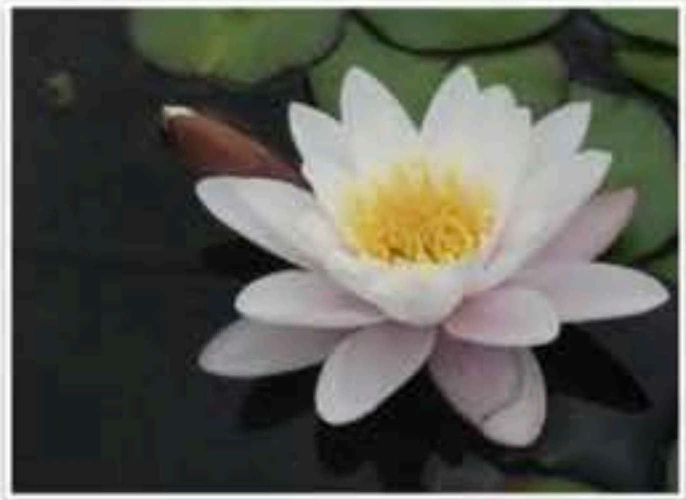
Zero-shot

ImageNet



A photo of a:
- Plane
- Golden Retriever
- Goldfish
- ...

Flower 102



A photo of a:
- Water Lily
- Rose
- Petunia
- ...

Food 101

A photo of a:
- Cake
- Pizza
- Soup
- ...

VLM Evaluation

Reasoning

a) Plant surrounding lightbulb

b) Lightbulb surrounding plant

Winoground

Horse is eating the grass ✓

Grass is eating the horse ✗

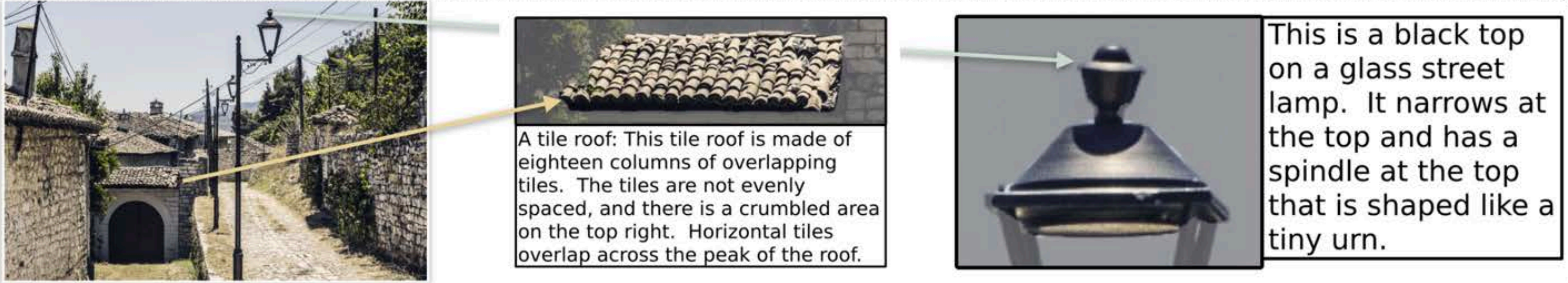
The paved road and the white house ✓

The white road and the paved house ✗

ARO

Dense

DCI dataset

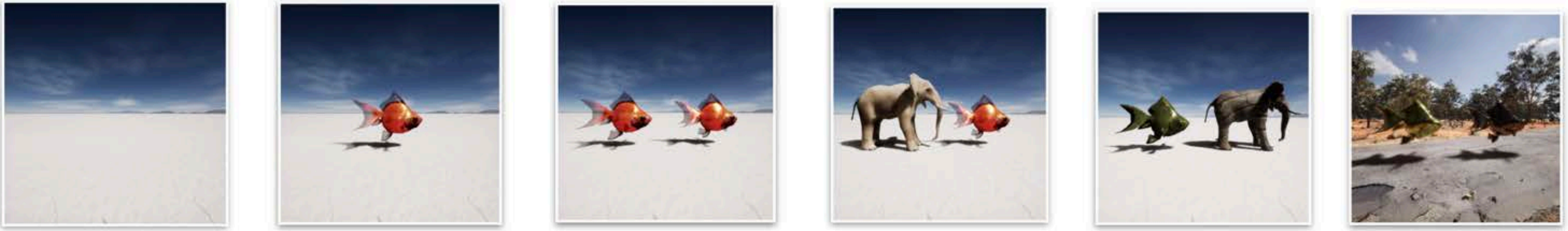


A tile roof: This tile roof is made of eighteen columns of overlapping tiles. The tiles are not evenly spaced, and there is a crumbled area on the top right. Horizontal tiles overlap across the peak of the roof.

This is a black top on a glass street lamp. It narrows at the top and has a spindle at the top that is shaped like a tiny urn.

Synthetic

PUG



VLM Evaluation

MMMU

A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

VLM Evaluation - MMMU Benchmark

Comprehensive Disciplines

- Engineering (26%)
- Art & Design (11%)
- Business (14%)
- Science (23%)
- Humanities & Social Sci. (9%)
- Medicine (17%)

Art & Design

Question: Among the following harmonic intervals, which one is constructed incorrectly?

Options:

(A) Major third <*image 1*>

(B) Diminished fifth <*image 2*>

(C) Minor seventh <*image 3*>

(D) Diminished sixth <*image 4*>

Subject: Music; **Subfield:** Music;
Image Type: Sheet Music;
Difficulty: Medium

Business

Question: ...The graph shown is compiled from data collected by Gallup <*image 1*>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82?

Options:

(A) 0 (B) 0.2142

(C) 0.3571 (D) 0.5

Subject: Marketing; **Subfield:** Market Research; **Image Type:** Plots and Charts;
Difficulty: Medium

Science

Question: <*image 1*> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.

Options:

(A) $\int_0^{1.5}[f(x)-g(x)]dx$

(B) $\int_0^{1.5}[g(x)-f(x)]dx$

(C) $\int_0^{2}[f(x)-g(x)]dx$

(D) $\int_0^{2}[g(x)-x(x)]dx$

Subject: Math; **Subfield:** Calculus;
Image Type: Mathematical Notations;
Difficulty: Easy

Health & Medicine

Question: You are shown subtraction <*image 1*>, T2 weighted <*image 2*> and T1 weighted axial <*image 3*> from a screening breast MRI. What is the etiology of the finding in the left breast?

Options:

(A) Susceptibility artifact

(B) Hematoma

(C) Fat necrosis (D) Silicone granuloma

Subject: Clinical Medicine; **Subfield:** Clinical Radiology; **Image Type:** Body Scans: MRI, CT.;
Difficulty: Hard

Humanities & Social Science

Question: In the political cartoon, the United States is seen as fulfilling which of the following roles? <*image* 1>

Option:

(A) Oppressor

(B) Imperialist

(C) Savior (D) Isolationist

Subject: History; **Subfield:** Modern History; **Image Type:** Comics and Cartoons;
Difficulty: Easy

Tech & Engineering

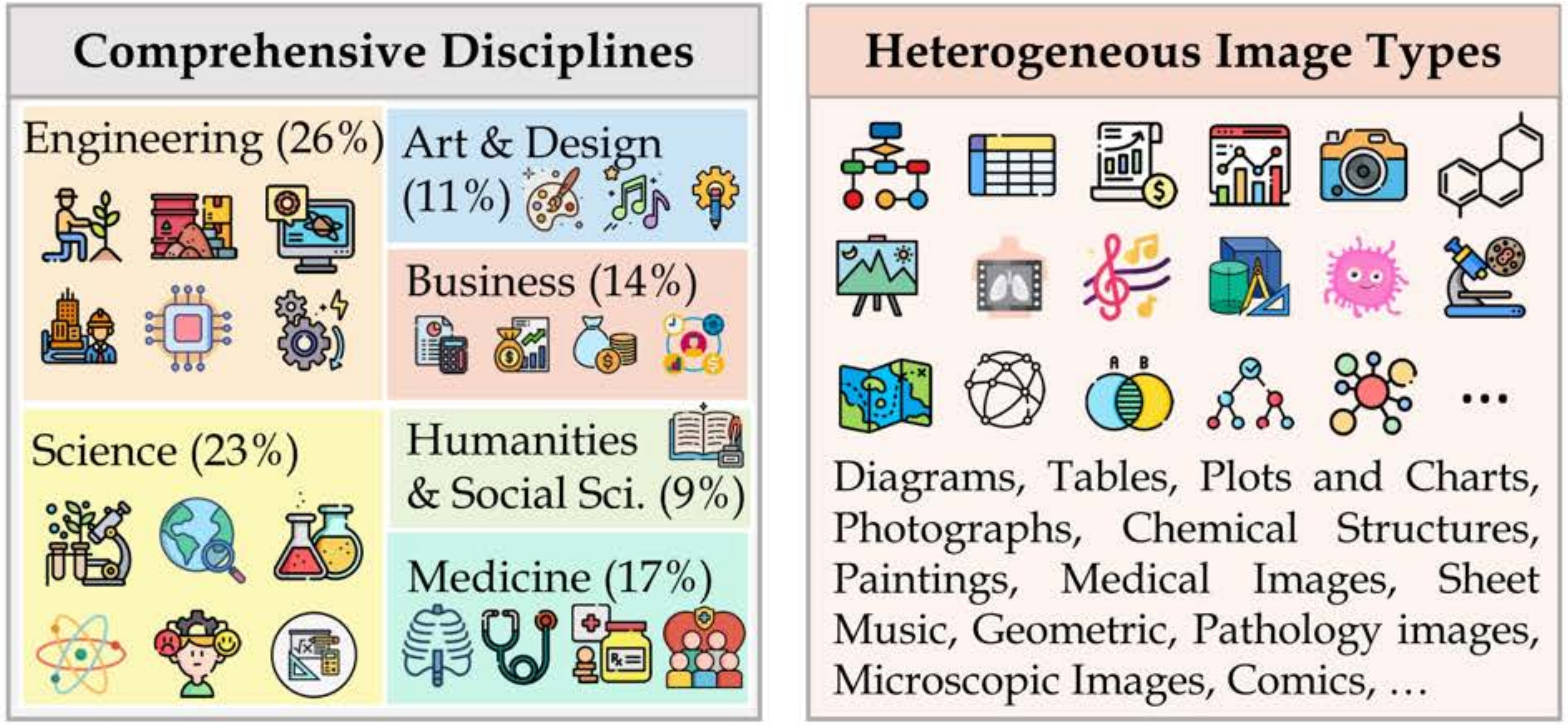
Question: Find the VCE for the circuit shown in <*image 1*>. Neglect VBE

Answer: 3.75

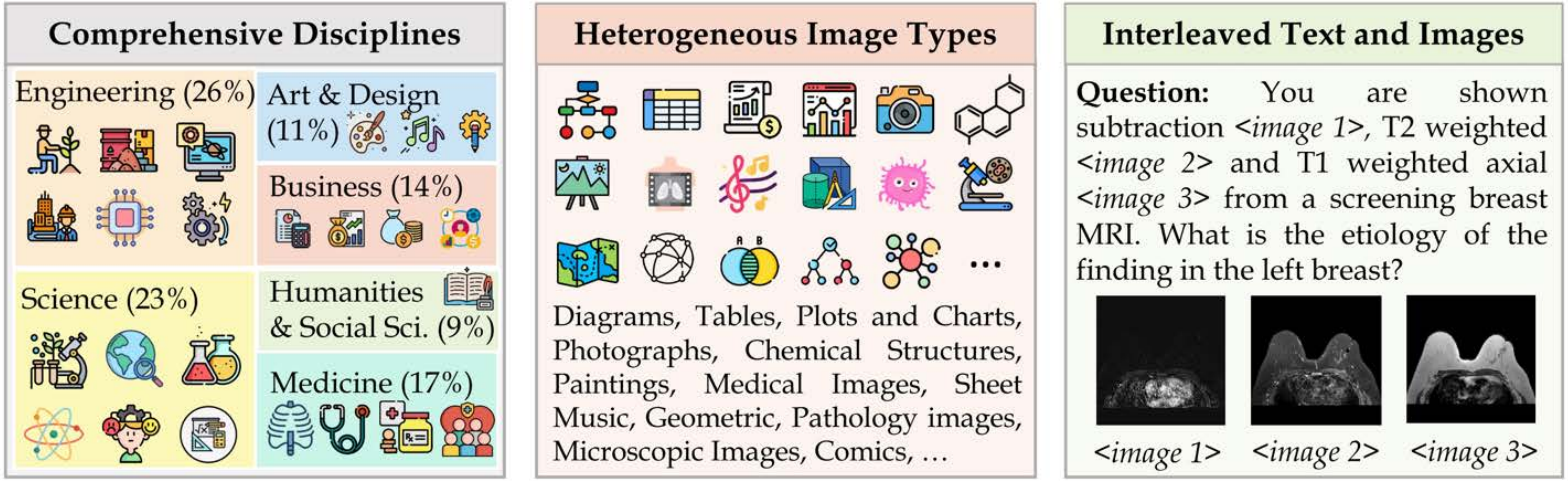
Explanation: ...IE = [(VEE) / (RE)] = [(5 V) / (4 k-ohm)] = 1.25 mA; VCE = VCC - IERL = 10 V - (1.25 mA) 5 k-ohm; VCE = 10 V - 6.25 V = 3.75 V

Subject: Electronics; **Subfield:** Analog electronics; **Image Type:** Diagrams;
Difficulty: Hard

VLM Evaluation - MMMU Benchmark


Comprehensive Disciplines
Engineering (26%)
Art & Design (11%)
Business (14%)
Science (23%)
Humanities & Social Sci. (9%)
Medicine (17%)
Heterogeneous Image Types
...
Diagrams, Tables, Plots and Charts, Photographs, Chemical Structures, Paintings, Medical Images, Sheet Music, Geometric, Pathology images, Microscopic Images, Comics, …

VLM Evaluation - MMMU Benchmark


Comprehensive Disciplines
Engineering (26%)
Art & Design (11%)
Business (14%)
Science (23%)
Humanities & Social Sci. (9%)
Medicine (17%)
Heterogeneous Image Types
Interleaved Text and Images
<image 1>
<image 2>
<image 3>

Diagrams, Tables, Plots and Charts, Photographs, Chemical Structures, Paintings, Medical Images, Sheet Music, Geometric, Pathology images, Microscopic Images, Comics, …

Question: You are shown subtraction *<image 1>*, T2 weighted *<image 2>* and T1 weighted axial *<image 3>* from a screening breast MRI. What is the etiology of the finding in the left breast?

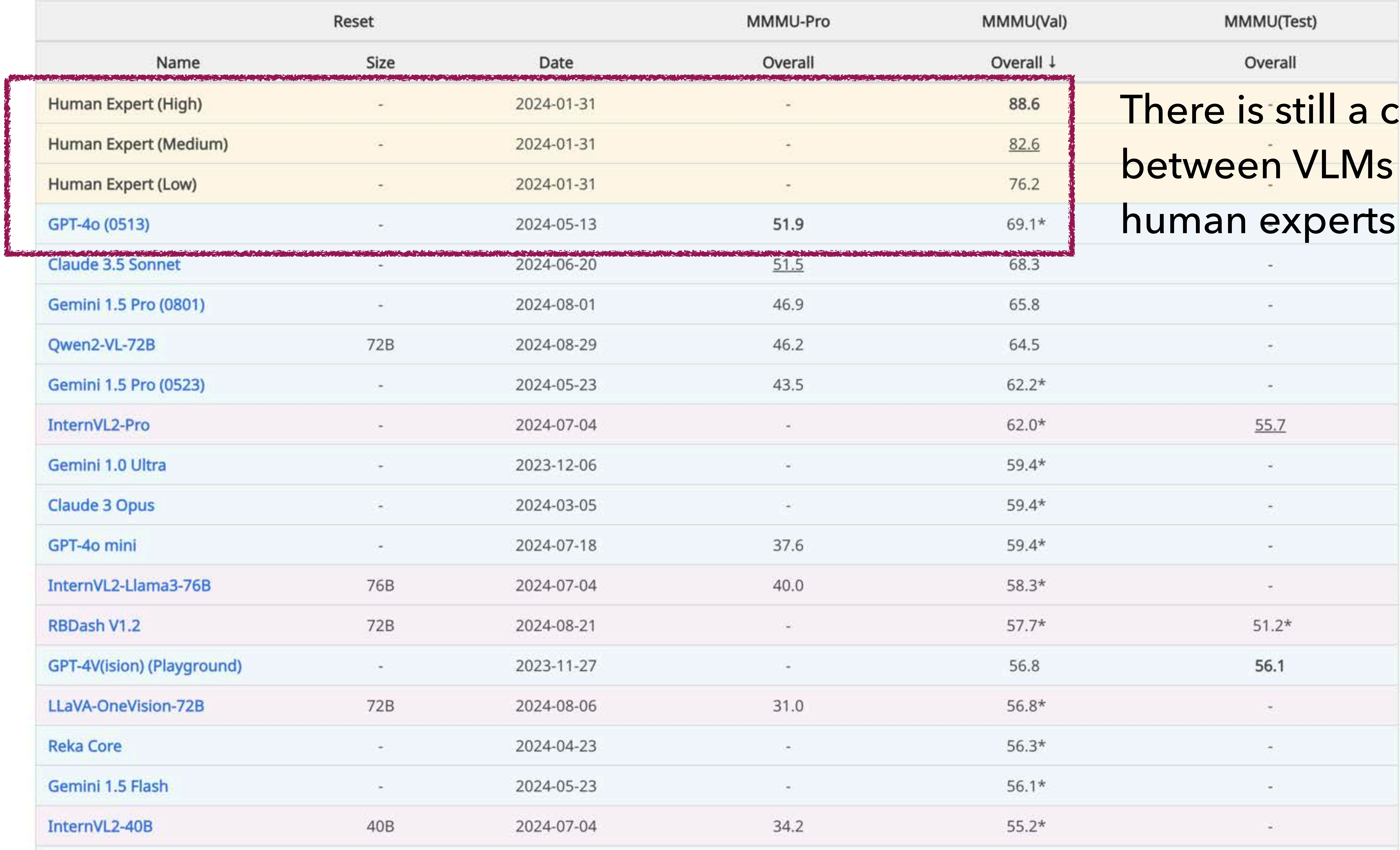
VLM Evaluation - MMMU Benchmark

Comprehensive Disciplines
Engineering (26%)
Art & Design (11%)
Business (14%)
Science (23%)
Humanities & Social Sci. (9%)
Medicine (17%)
Heterogeneous Image Types
...
Interleaved Text and Images
<image 1>
<image 2>
<image 3>
Expert-level Skills Test
Expert-level Visual Perception
Perception
Knowledge
Reasoning
Domain Expertise, World, Linguistic, Visual Knowledge,…
Logical, Spatial Commonsense, Mathematical,…

Diagrams, Tables, Plots and Charts, Photographs, Chemical Structures, Paintings, Medical Images, Sheet Music, Geometric, Pathology images, Microscopic Images, Comics, …

Question: You are shown subtraction <*image* 1>, T2 weighted <*image* 2> and T1 weighted axial <*image* 3> from a screening breast MRI. What is the etiology of the finding in the left breast?

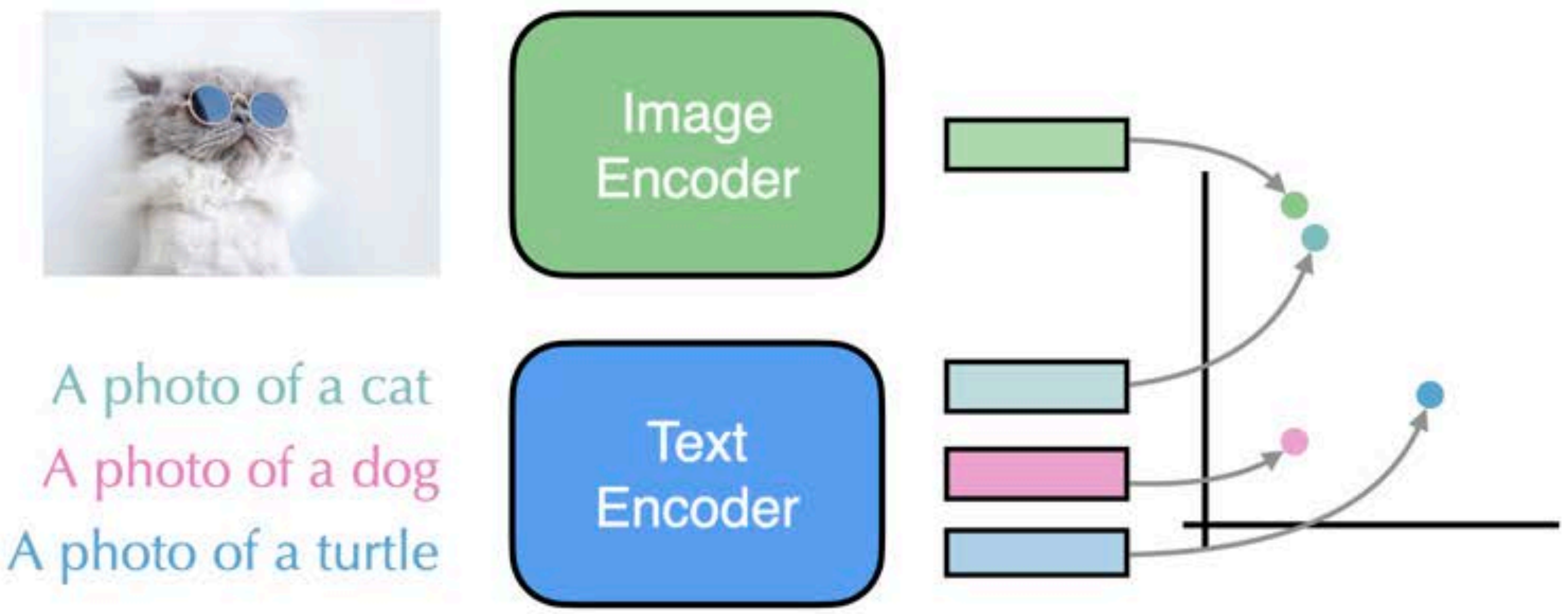
VLM Evaluation - MMMU Benchmark



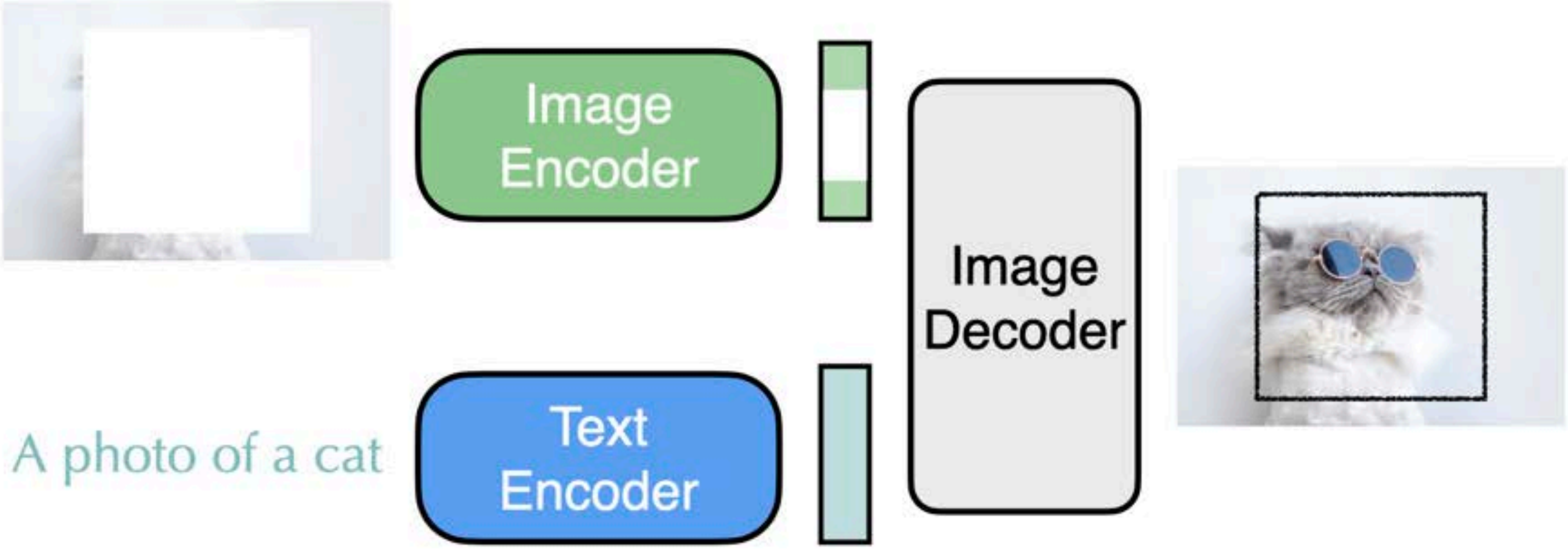
Reset			MMMU-Pro	MMMU(Val)	MMMU(Test)
Name	Size	Date	Overall	Overall ↓	Overall
Human Expert (High)	-	2024-01-31	-	**88.6**	-
Human Expert (Medium)	-	2024-01-31	-	82.6	-
Human Expert (Low)	-	2024-01-31	-	76.2	-
GPT-4o (0513)	-	2024-05-13	**51.9**	69.1*	-
Claude 3.5 Sonnet	-	2024-06-20	51.5	68.3	-
Gemini 1.5 Pro (0801)	-	2024-08-01	46.9	65.8	-
Qwen2-VL-72B	72B	2024-08-29	46.2	64.5	-
Gemini 1.5 Pro (0523)	-	2024-05-23	43.5	62.2*	-
InternVL2-Pro	-	2024-07-04	-	62.0*	55.7
Gemini 1.0 Ultra	-	2023-12-06	-	59.4*	-
Claude 3 Opus	-	2024-03-05	-	59.4*	-
GPT-4o mini	-	2024-07-18	37.6	59.4*	-
InternVL2-Llama3-76B	76B	2024-07-04	40.0	58.3*	-
RBDash V1.2	72B	2024-08-21	-	57.7*	51.2*
GPT-4V(ision) (Playground)	-	2023-11-27	-	56.8	**56.1**
LLaVA-OneVision-72B	72B	2024-08-06	31.0	56.8*	-
Reka Core	-	2024-04-23	-	56.3*	-
Gemini 1.5 Flash	-	2024-05-23	-	56.1*	-
InternVL2-40B	40B	2024-07-04	34.2	55.2*	-

There is still a clear gap between VLMs and human experts!

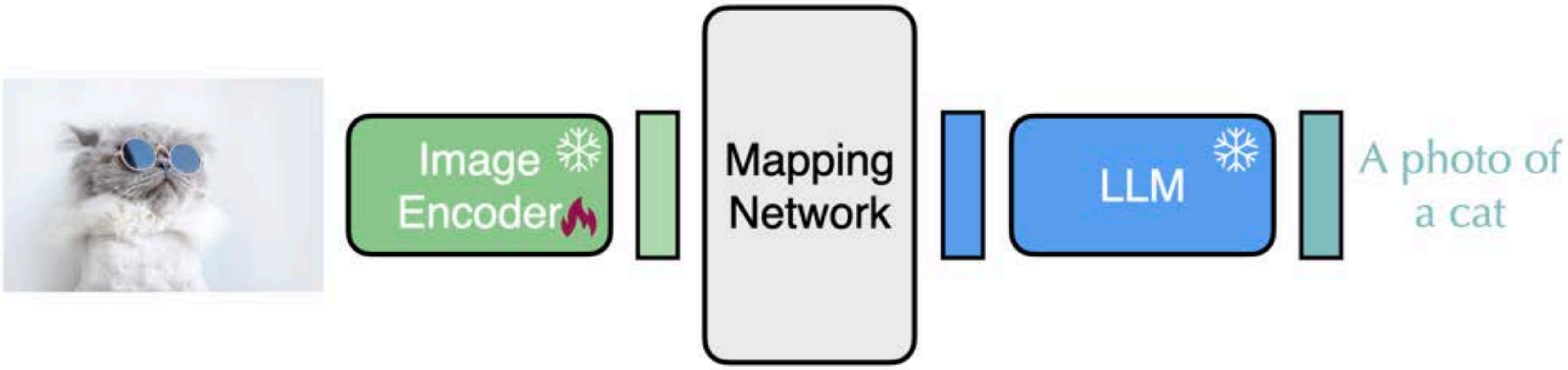
The Families of VLMs



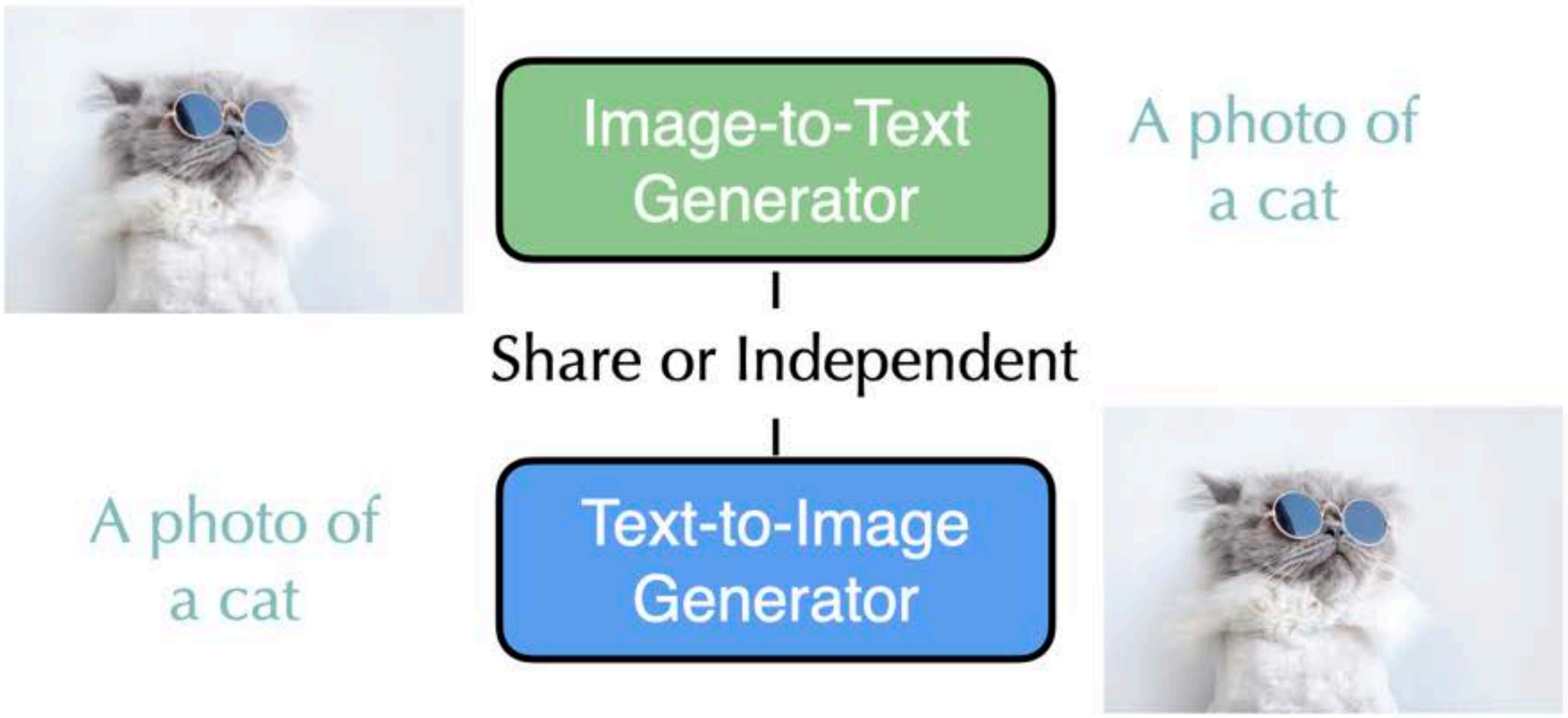
Contrastive-Based



Masking Objective



VLMs from Pretrained Backbones



Generative-Based

Important Considerations When Training VLMs.

Data

Remove duplicate

Balancing

Pruning

A photo of a dog ✓

A drawing of a bird ✗

Improving

"A beagle dog is smiling and looking at the camera"

Grounding

A golden retriever swimming

Tennis balls in a swimming pool

Using bounding boxes

Using negative captions

A photo of a dog

A photo of a bird

Alignement

Instruction: Provide a description of what is presented in the image.

Answer: This is a photo of a smiling dog

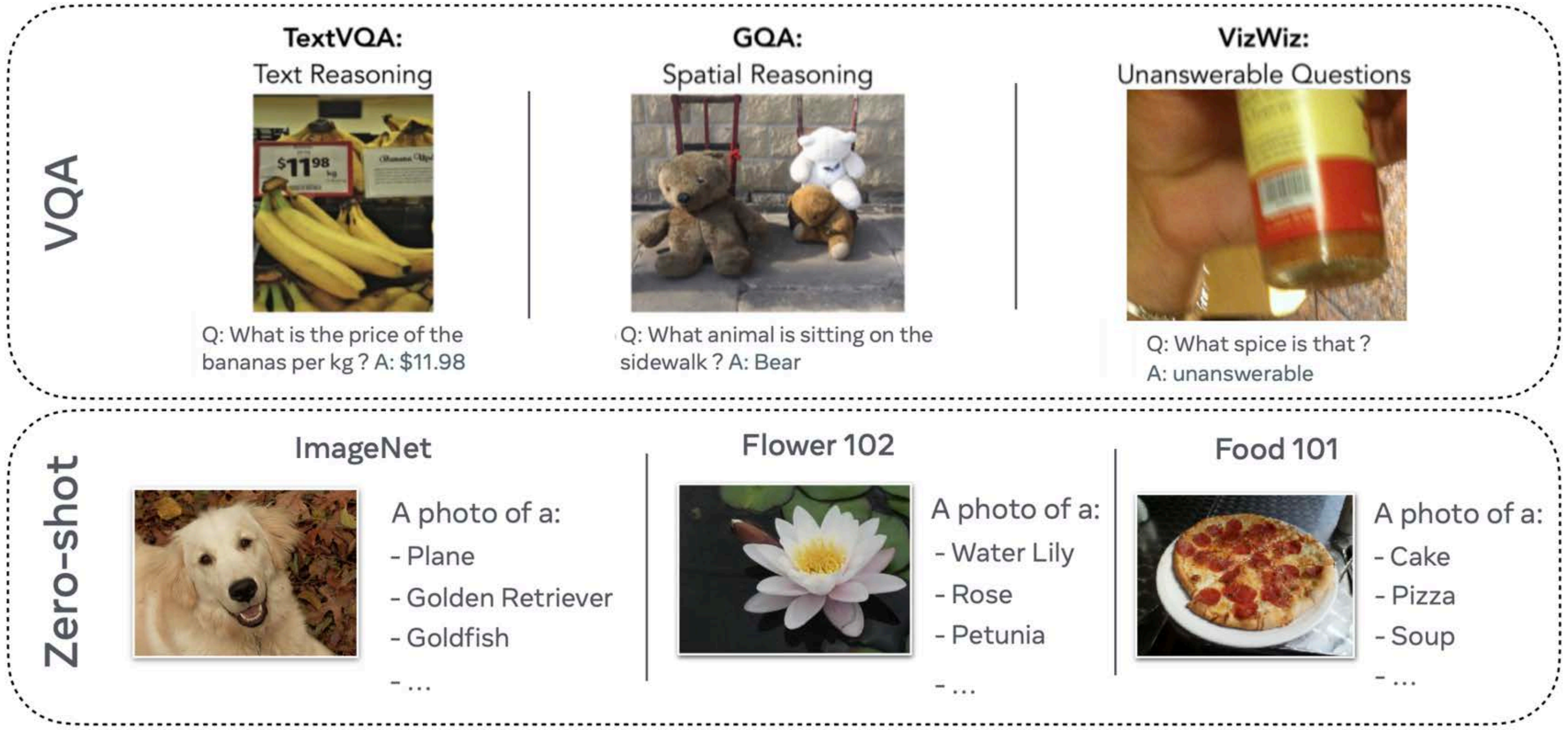
Instruction-tuning

Reinforcement Learning From Human Feedback

VLM: This image is a photo of a smiling dog.

VLM: Hi, I am a robot. It is a photo of a dog.

VLM Evaluation



VQA

TextVQA:
Text Reasoning
Q: What is the price of the bananas per kg ? A: $11.98

GQA:
Spatial Reasoning
Q: What animal is sitting on the sidewalk ? A: Bear

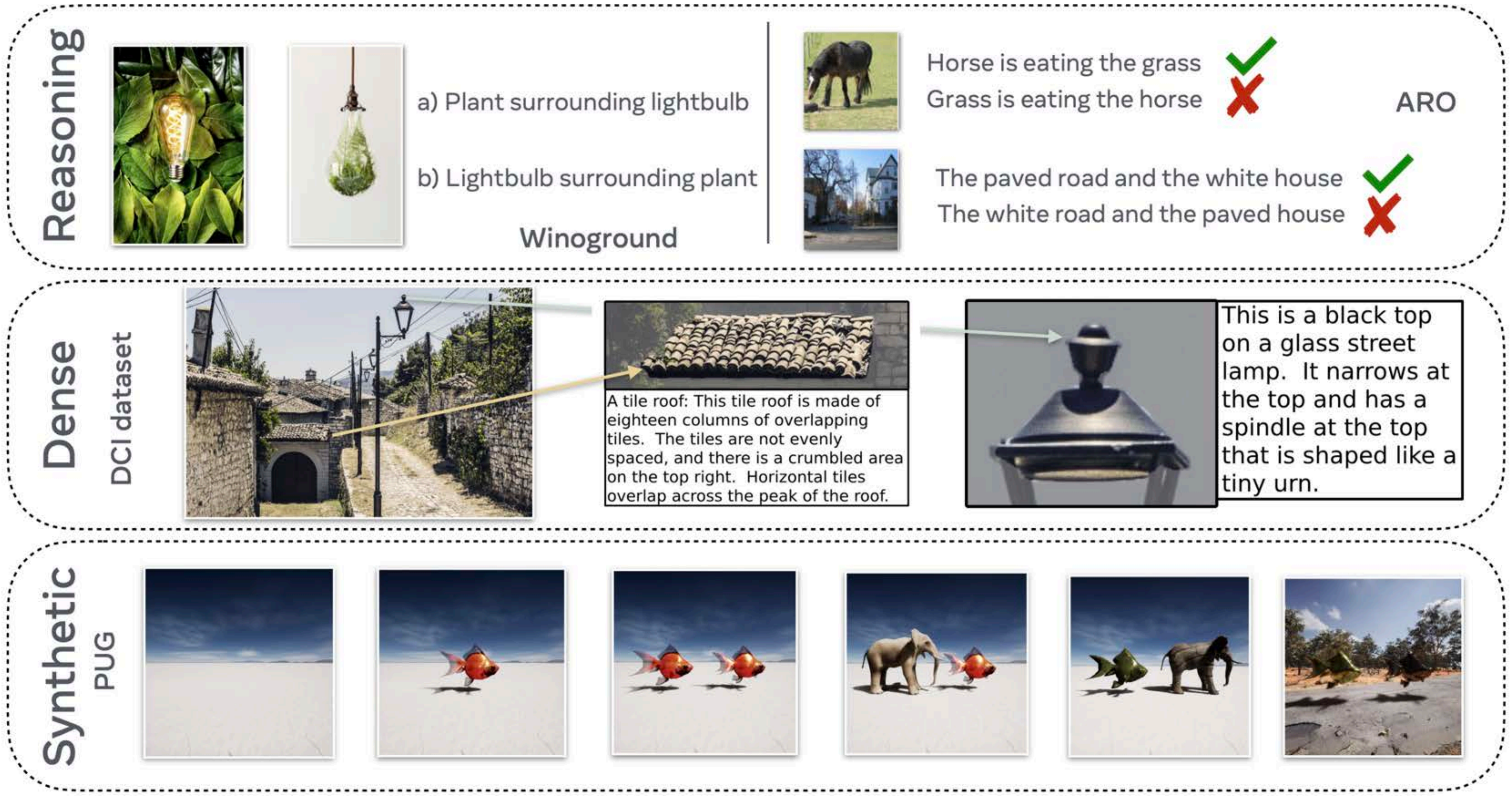
VizWiz:
Unanswerable Questions
Q: What spice is that ?
A: unanswerable

Zero-shot

ImageNet
A photo of a:
- Plane
- Golden Retriever
- Goldfish
- ...

Flower 102
A photo of a:
- Water Lily
- Rose
- Petunia
- ...

Food 101
A photo of a:
- Cake
- Pizza
- Soup
- ...



Reasoning

a) Plant surrounding lightbulb
b) Lightbulb surrounding plant
Winoground

Horse is eating the grass ✓
Grass is eating the horse ✗
The paved road and the white house ✓
The white road and the paved house ✗
ARO

Dense
DCI dataset

A tile roof: This tile roof is made of eighteen columns of overlapping tiles. The tiles are not evenly spaced, and there is a crumbled area on the top right. Horizontal tiles overlap across the peak of the roof.

This is a black top on a glass street lamp. It narrows at the top and has a spindle at the top that is shaped like a tiny urn.

Synthetic
PUG

Thank You!!!