Multimodal Foundation Models 2

Multimodal Pre-traning

Sangdoo Yun and Jin-Hwa Kim

NAVER AI LAB

Today's lecture

- Contents
  - Multimodal pre-training
  - Large-scale multimodal pre-training

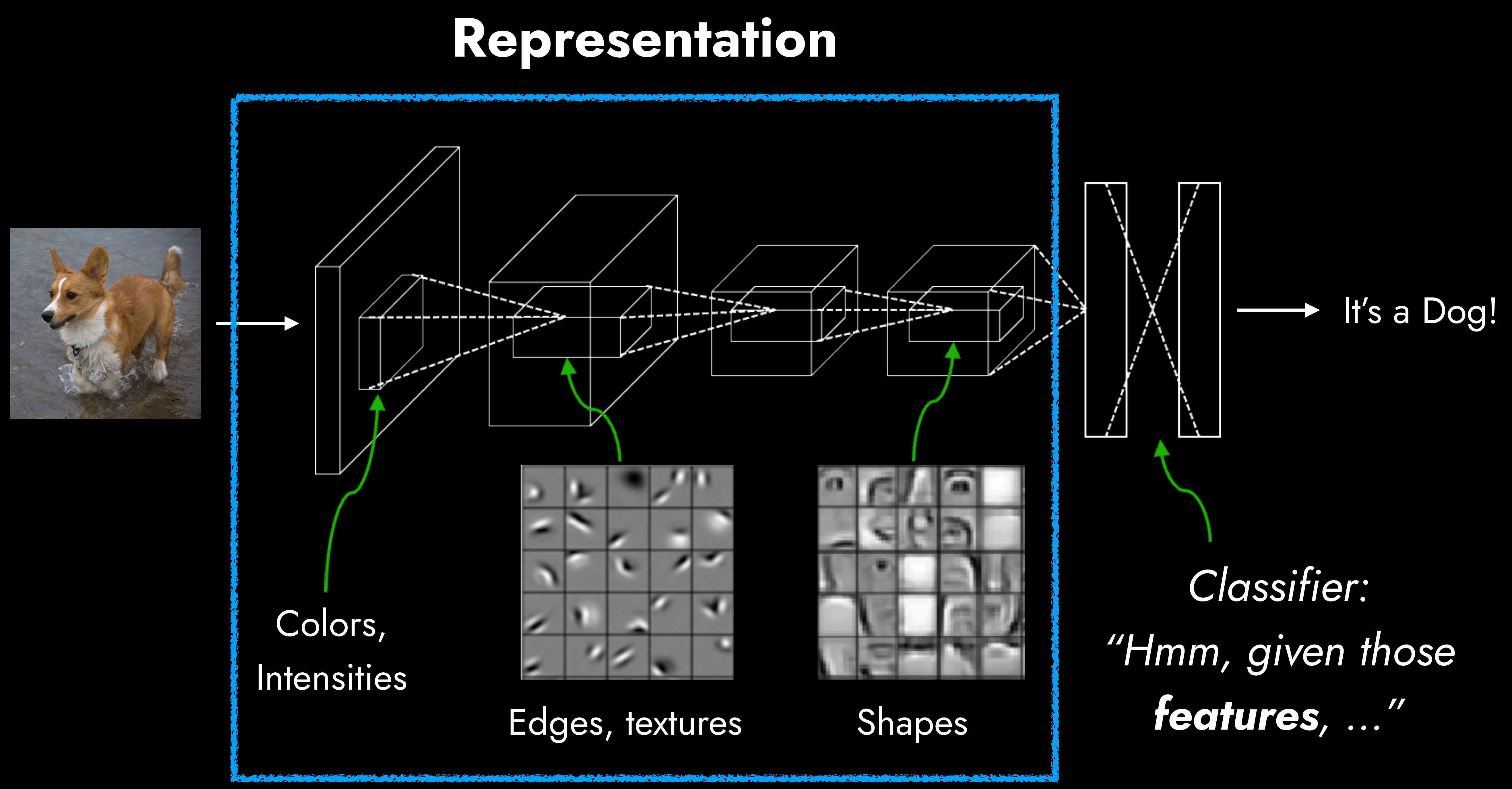
Please don't hesitate to ask questions!

Your questions help everyone (including me) learn better.

Multimodal Pre-training

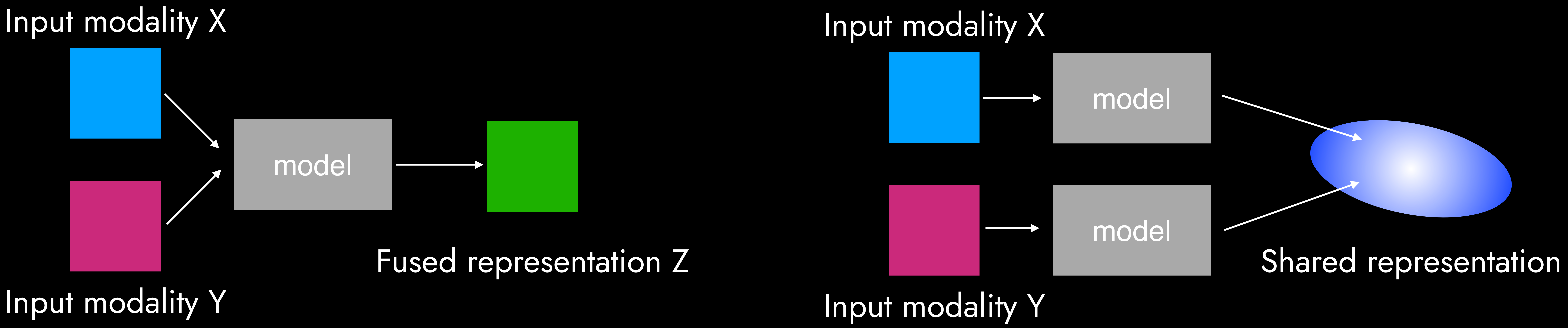
Multimodal representation

- The goal of representation learning (recap. Lecture 3)
  - Image space: $256^{3\times300\times300} \rightarrow 1{,}024$ (representation space)
  - Compact vector
  - Represents input contents
  - Can transfer to other tasks



Multimodal representation

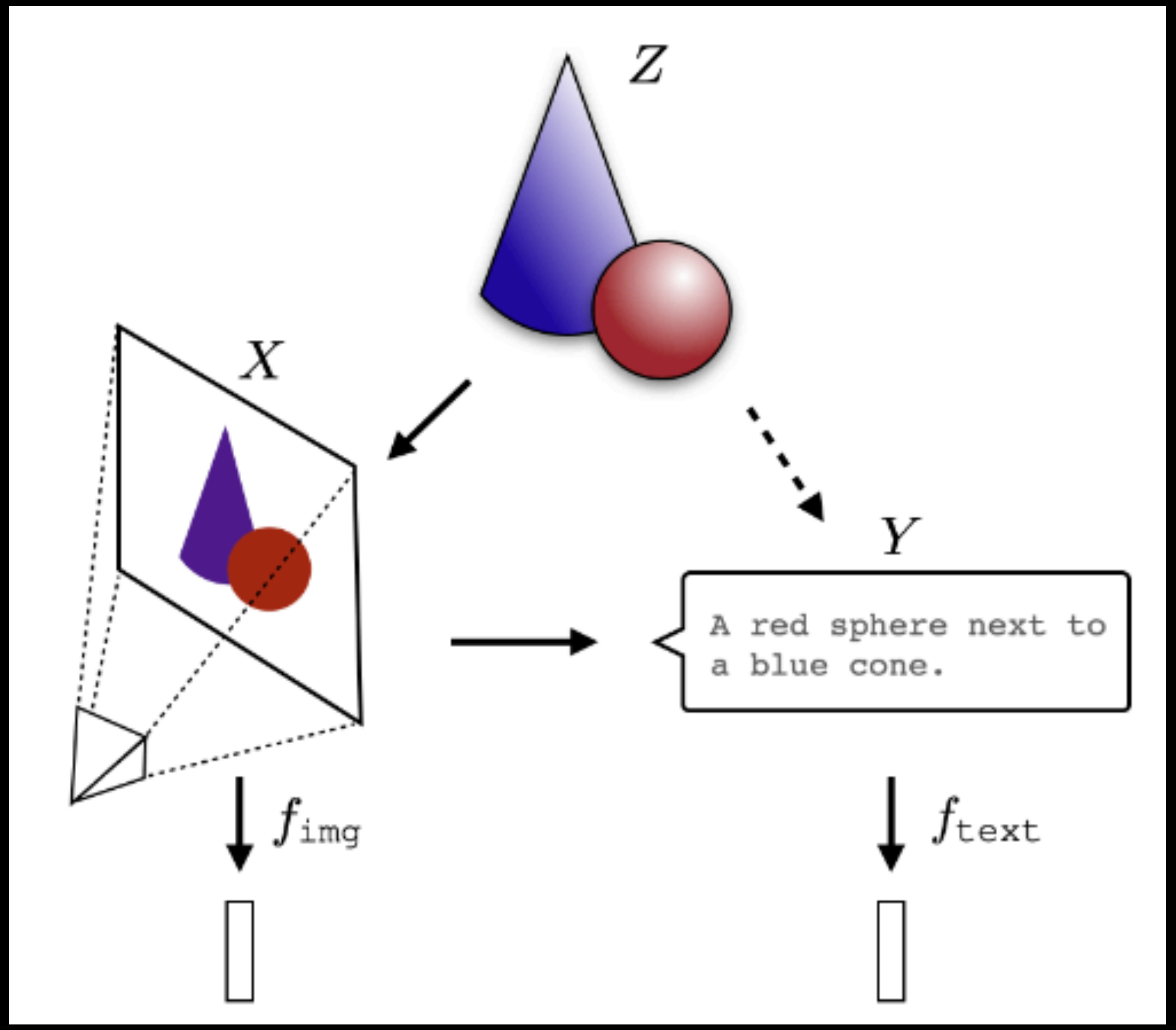
- Multimodal representation (recap. Lecture 2)

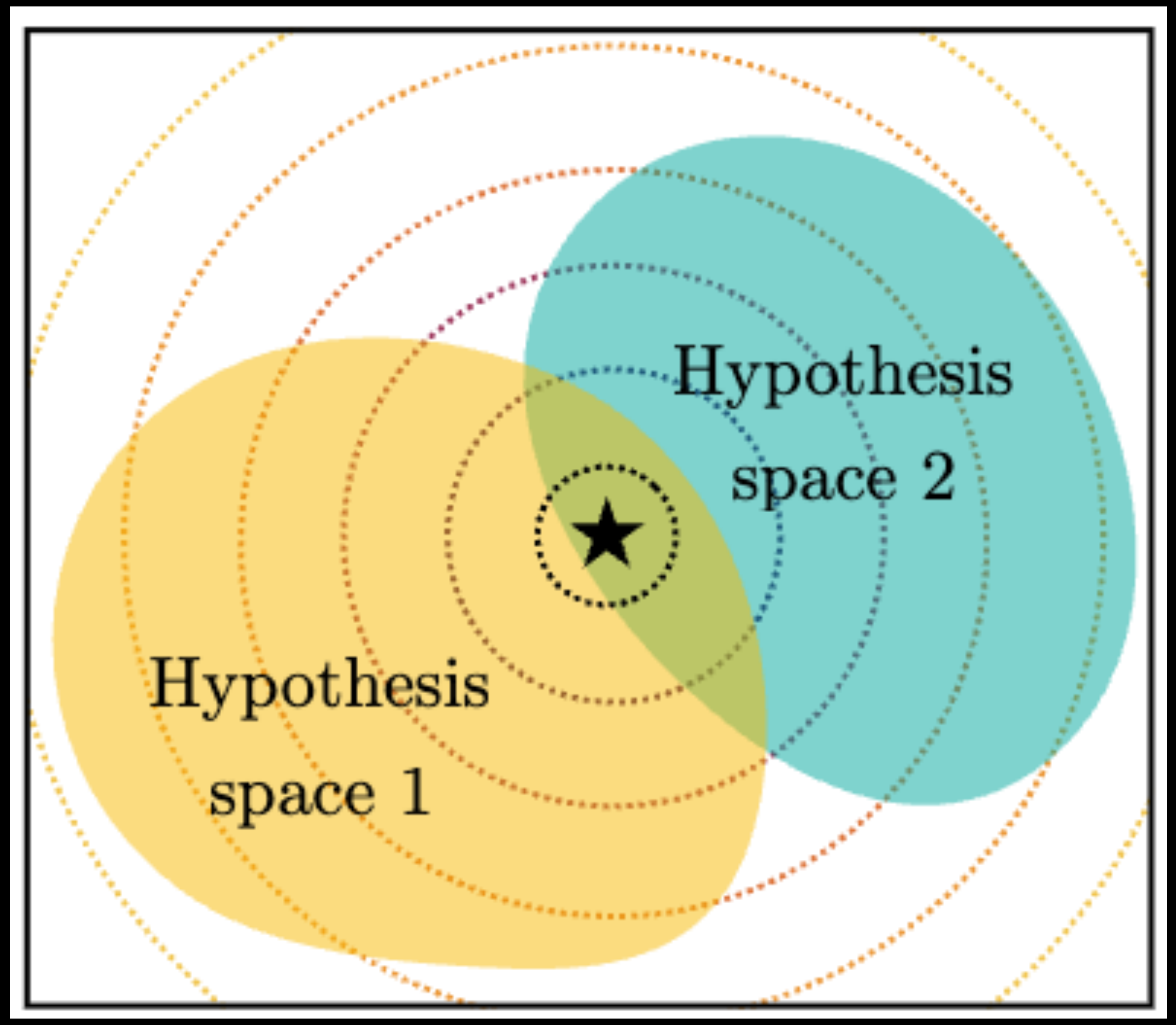


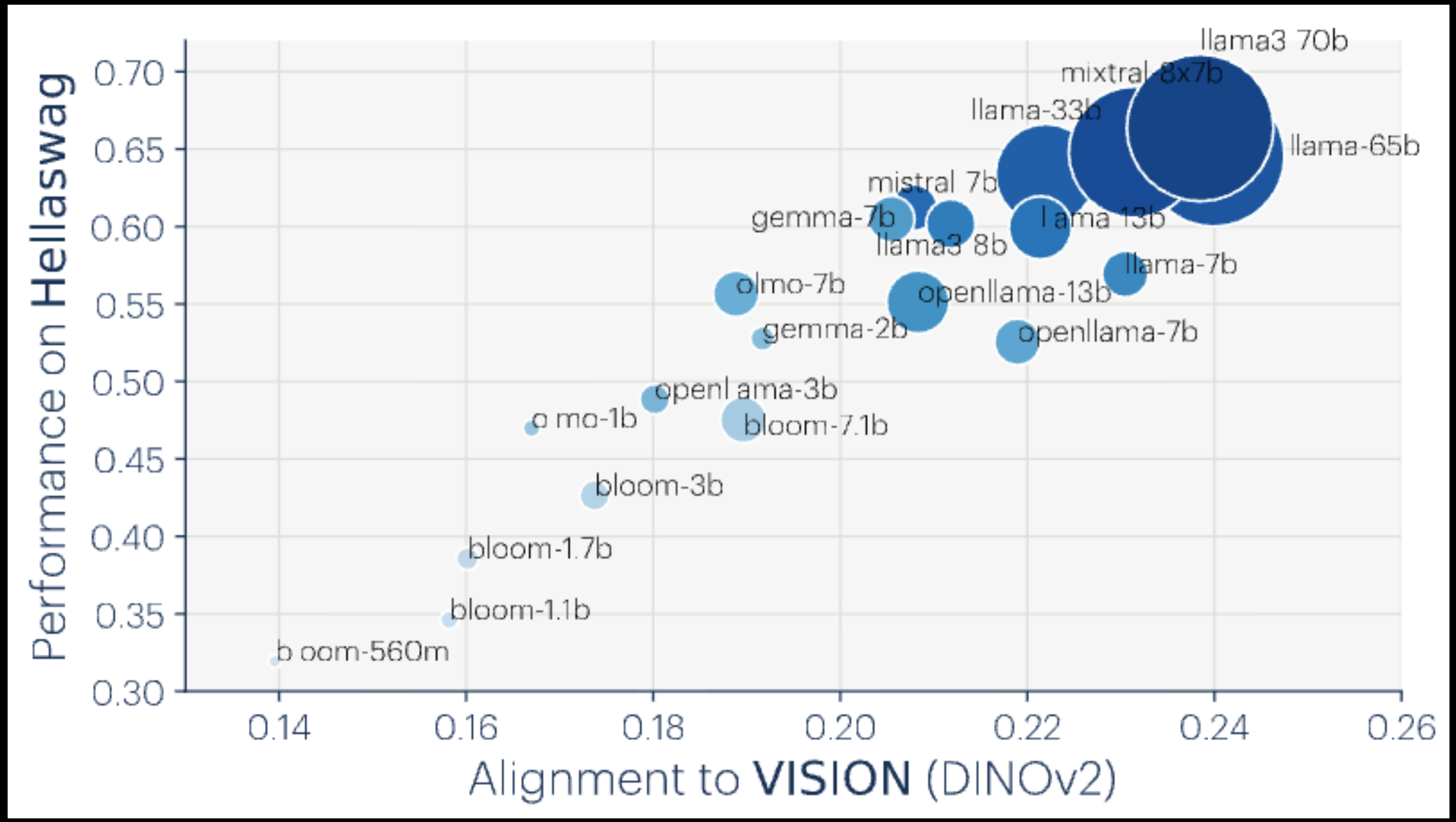
- Goal of multimodal pre-training: obtain a good *multimodal* representation

Platonic Representation Hypothesis

- Neural networks, trained with *different objectives* on *different data* and *modalities*, are *converging* to a shared statistical model of reality in their *representation spaces*.





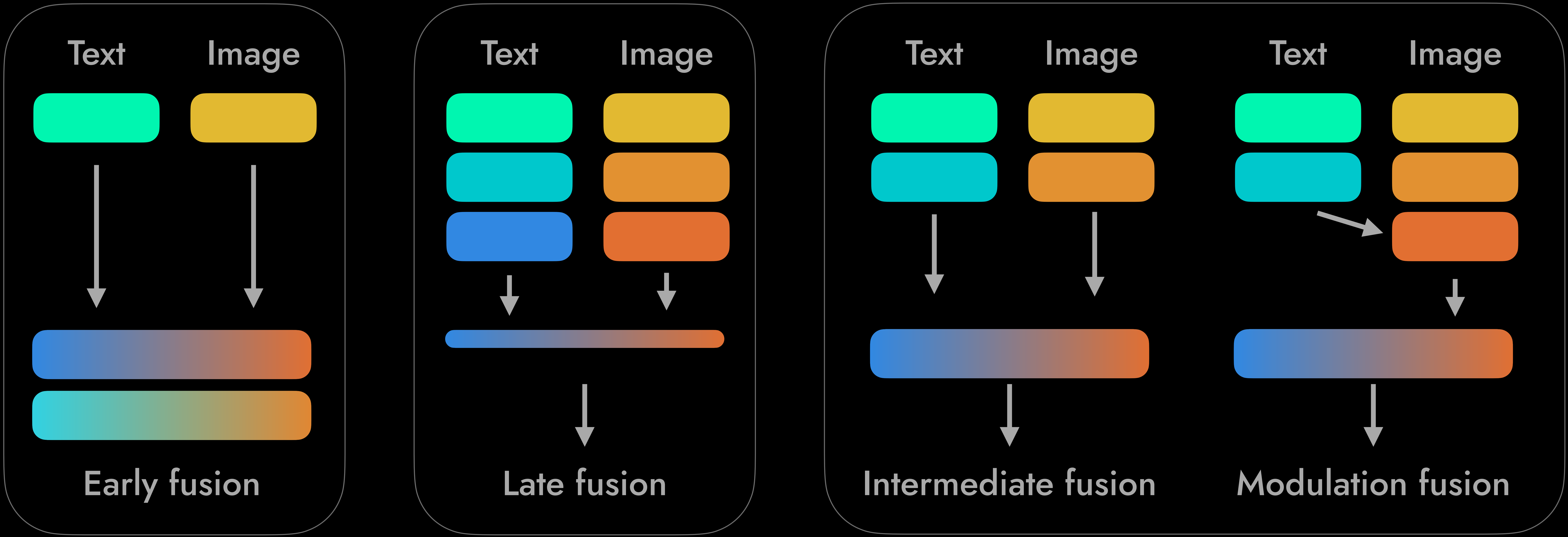


Multimodal pre-training

- Recent multimodal pre-training has been developed based on "Vision-and-Language"
- Why Vision-and-language?
  - Training data: MS-COCO, CC3M, CC12M, LAION-5B, …
  - Evaluation tasks: Image captioning, cross-modal retrieval, VQA, …
- Vision-Language Pre-training (VLP)

Vision-and-language pre-training

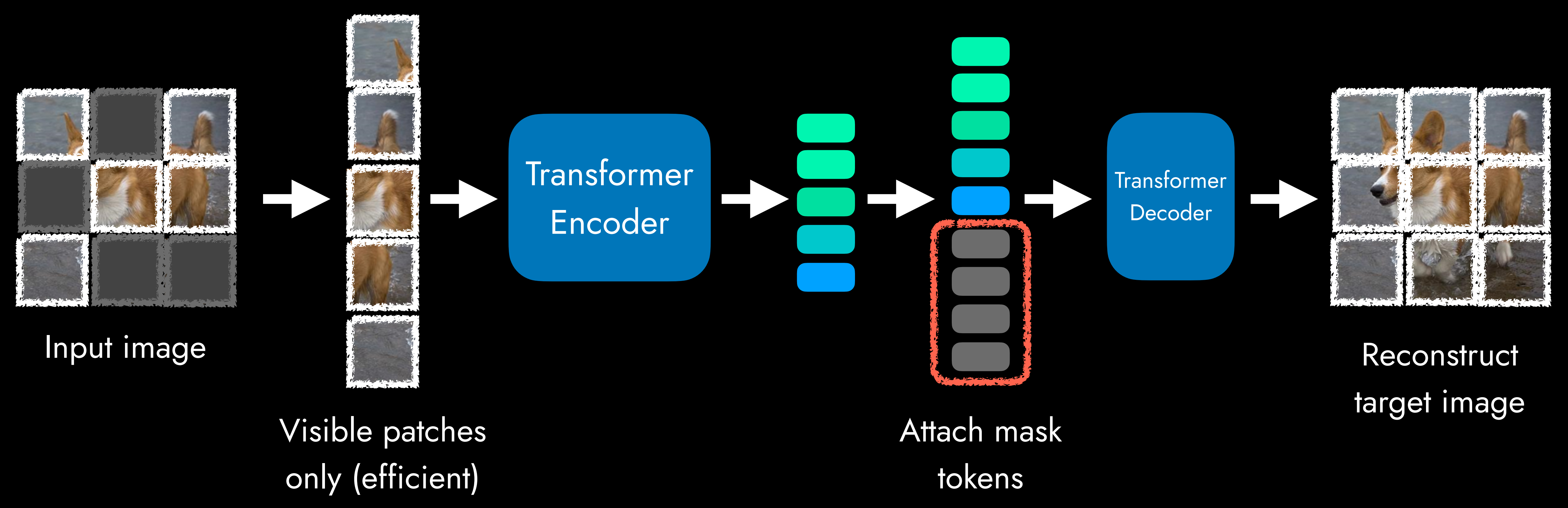
- How to fuse or link different modalities? (recap. Lecture 2)



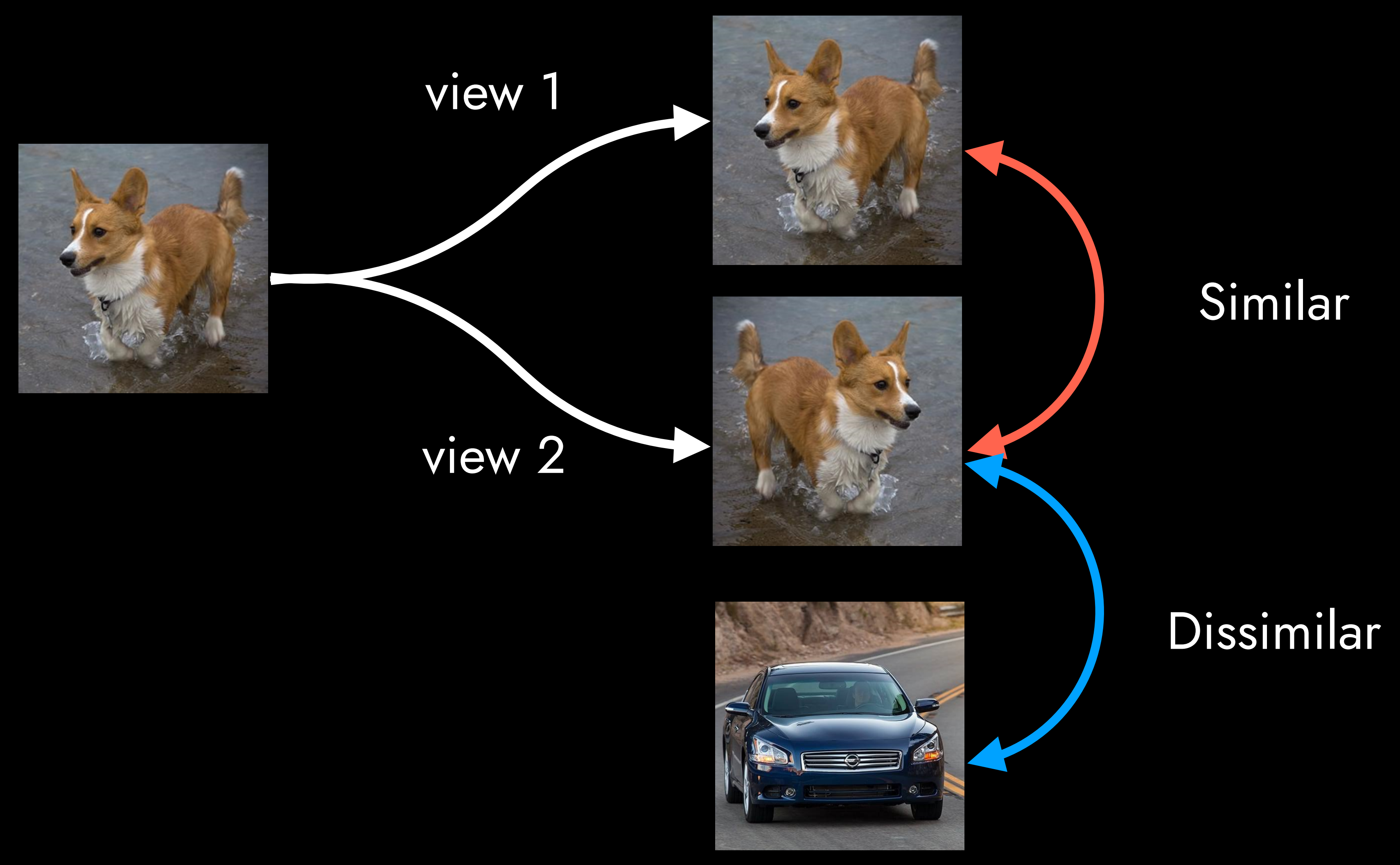
Vision-and-language pre-training

- Objective — Self-supervised learning

Masked Auto-Encoder (MAE) (He et al., 2022)



Contrastive learning



Vision-and-language pre-training

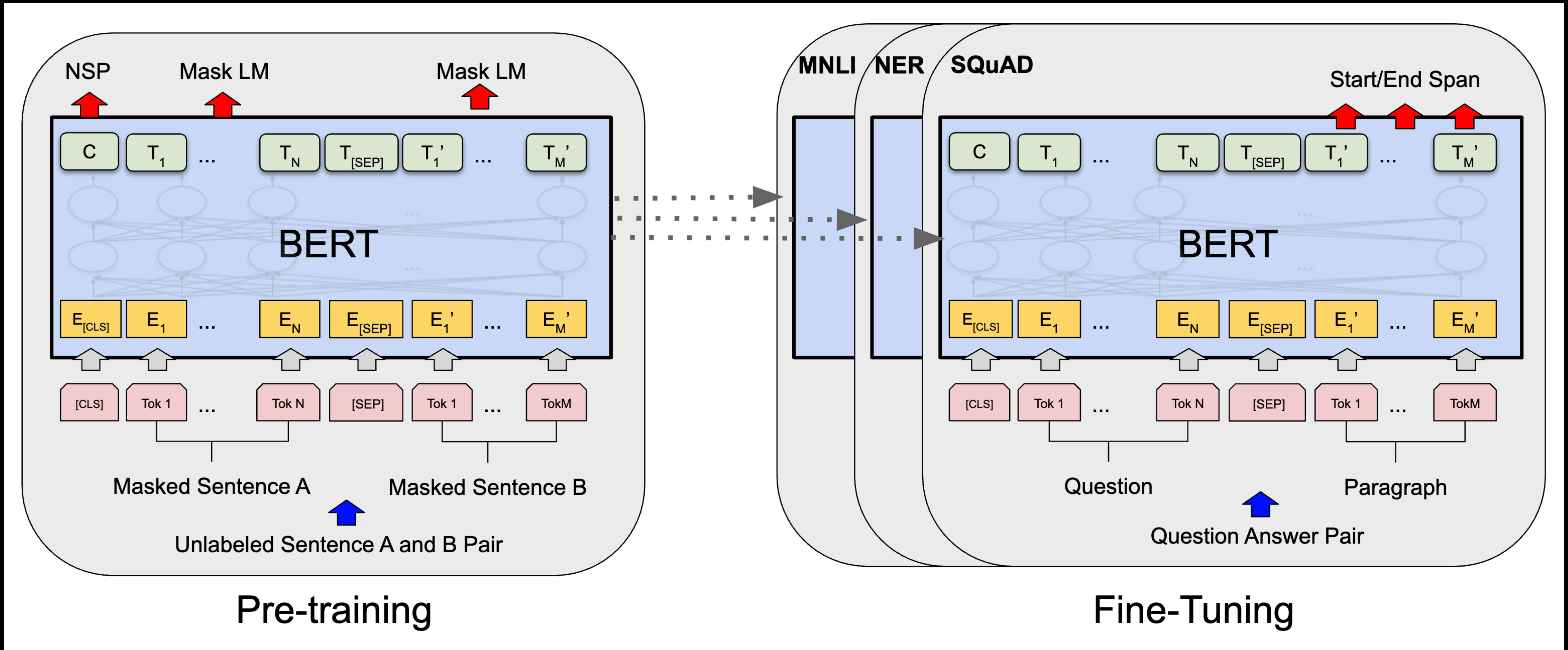
- Objective — Self-supervised learning
- Predictive task: Masked modeling (modality agnostic)
- Inter-sample (instance) task: Contrastive learning
    - Positive pair: image—caption (*human-annotated* like MS-COCO, or *alt-text* like CC, LAION-2B/5B)
    - Negative pair: others

Vision-and-Language Pre-training

BERT-based approach

BERT (2018)

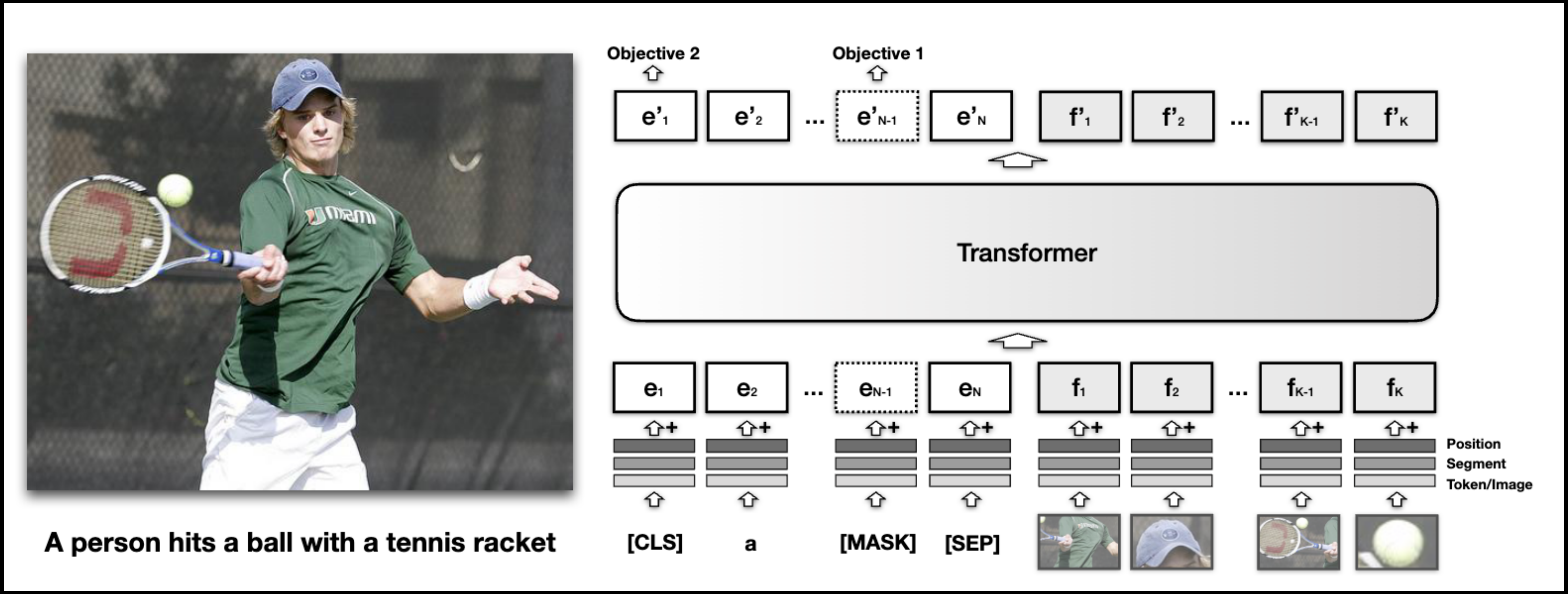
- Self-supervised learning: Masked language modeling, Next sentence prediction



"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" (2018)

VisualBERT (Aug, 2019)

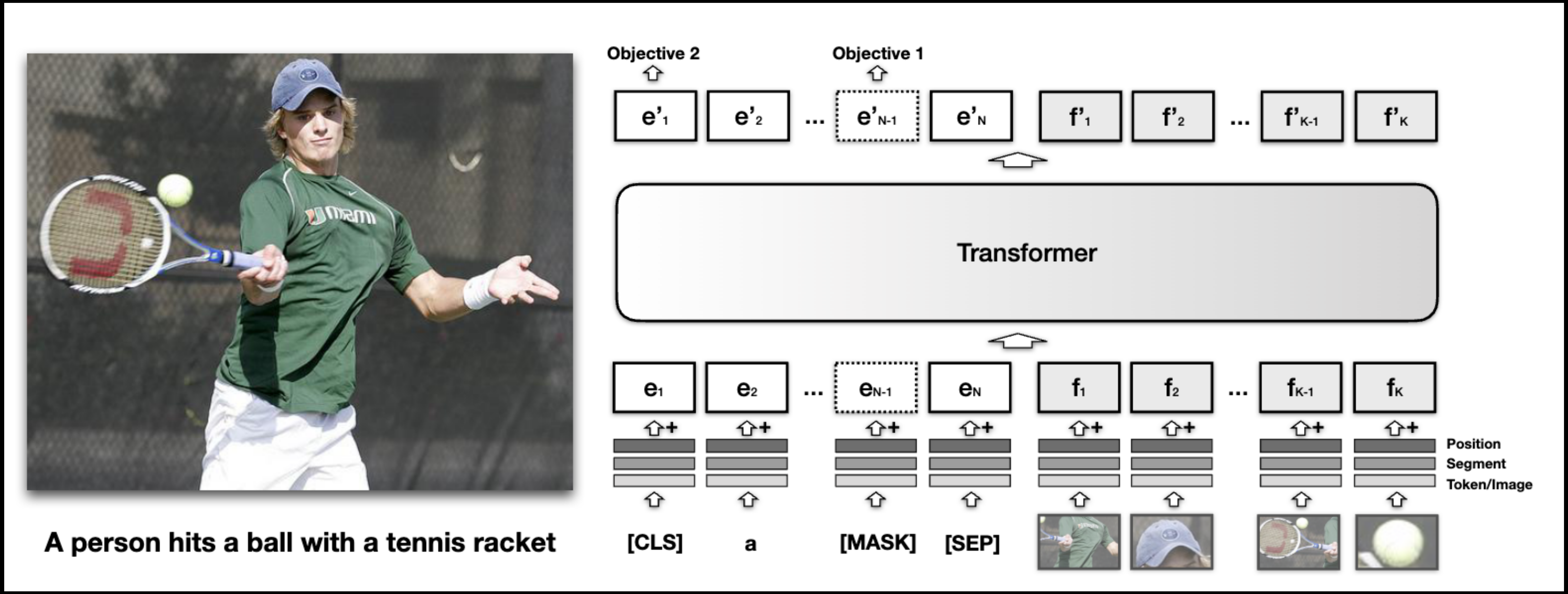
- Simply concatenate visual region features and word embeddings as input
- Use self-attention to implicitly align elements of the text and image regions



"VisualBERT: A Simple and Performant Baseline for Vision and Language" (2019)

VisualBERT (Aug, 2019)

- Objectives: masked language modeling (*MLM*), image-text matching (*ITM*)



"VisualBERT: A Simple and Performant Baseline for Vision and Language" (2019)

ViLBERT (Aug, 2019)

- Objectives: image-text matching (*ITM*), masked language modeling (*MLM*), masked region modeling (*MRM*)

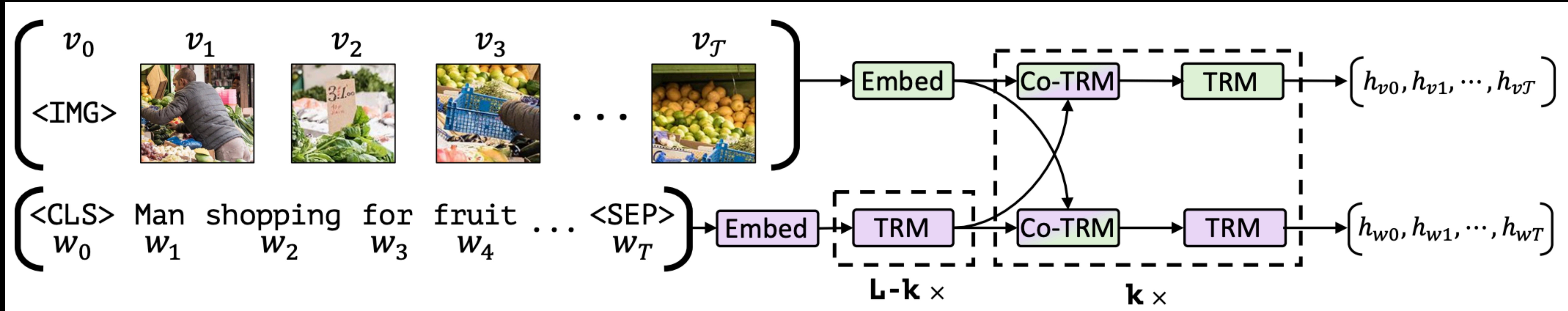
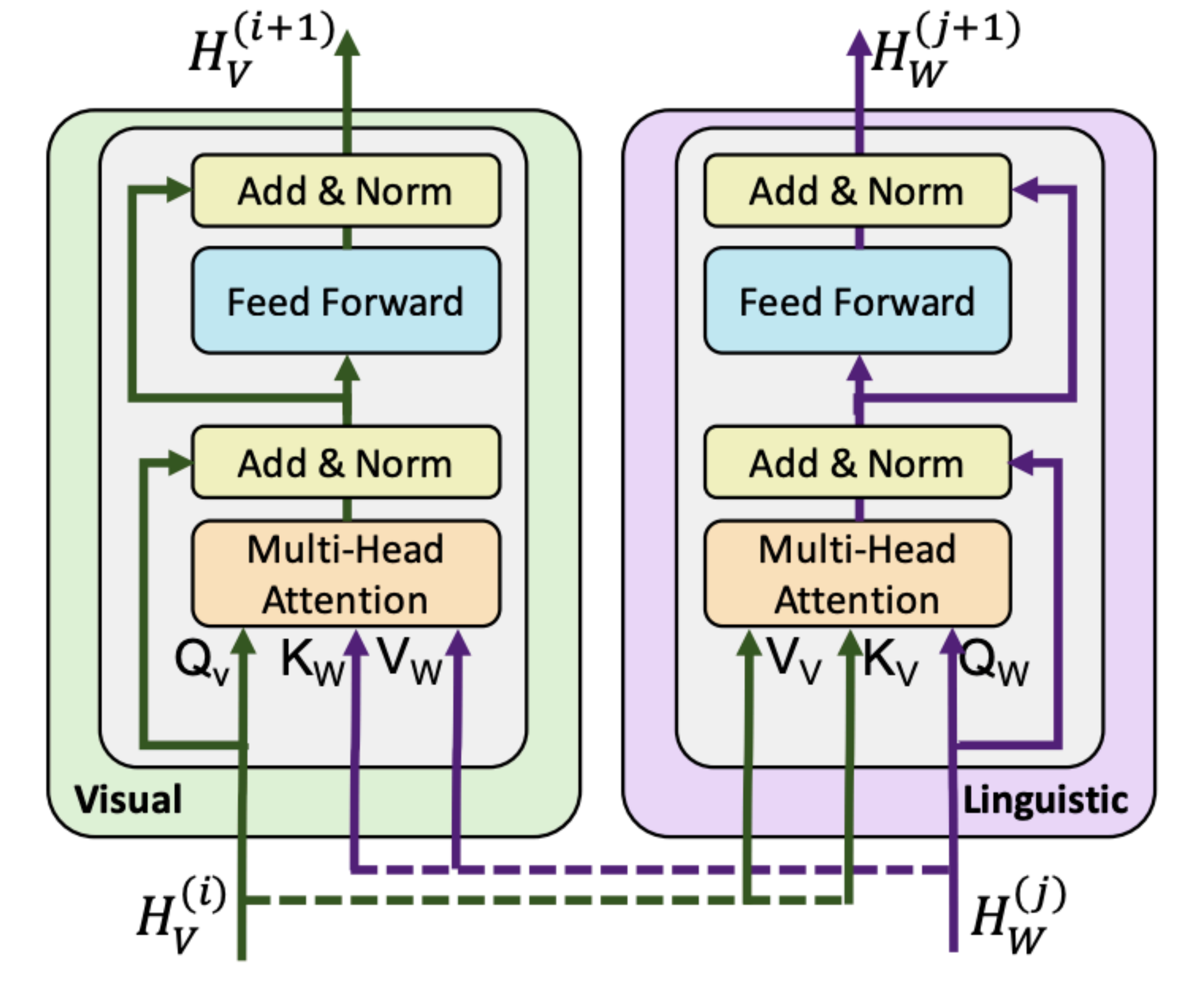


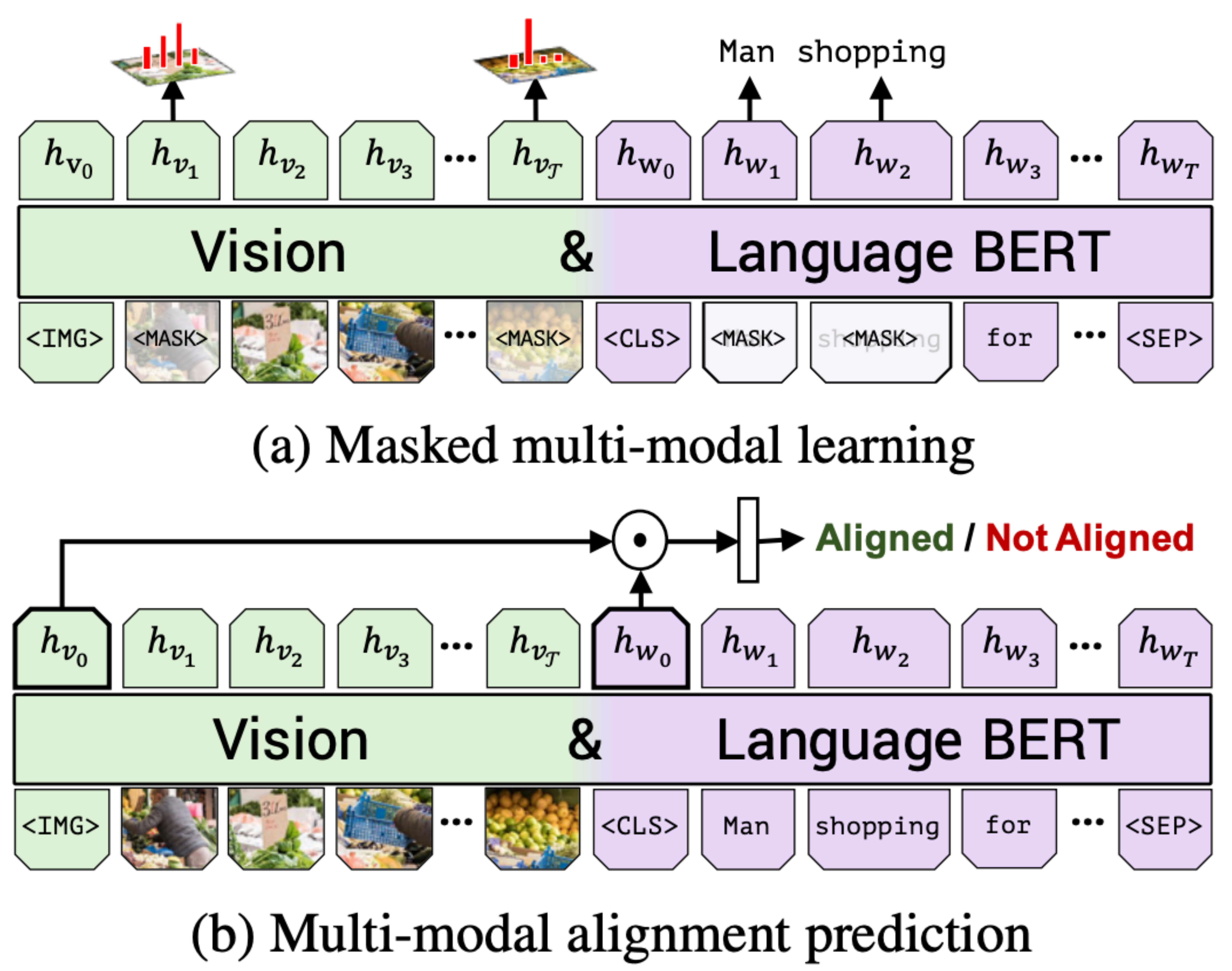
Figure 1: Our ViLBERT model consists of two parallel streams for visual (green) and linguistic (purple) processing that interact through novel co-attentional transformer layers. This structure allows for variable depths for each modality and enables sparse interaction through co-attention. Dashed boxes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



(b) Our co-attention transformer layer

ViLBERT (Aug, 2019)

- *MLM*
  - Follows MLM in standard BERT
- *MRM*
  - Predict a distribution over semantic classes for the corresponding image region
- *ITM (Binary classification task)*
  - Similar to NSP in BERT
  - Positive (matched) or negative (unmatched)

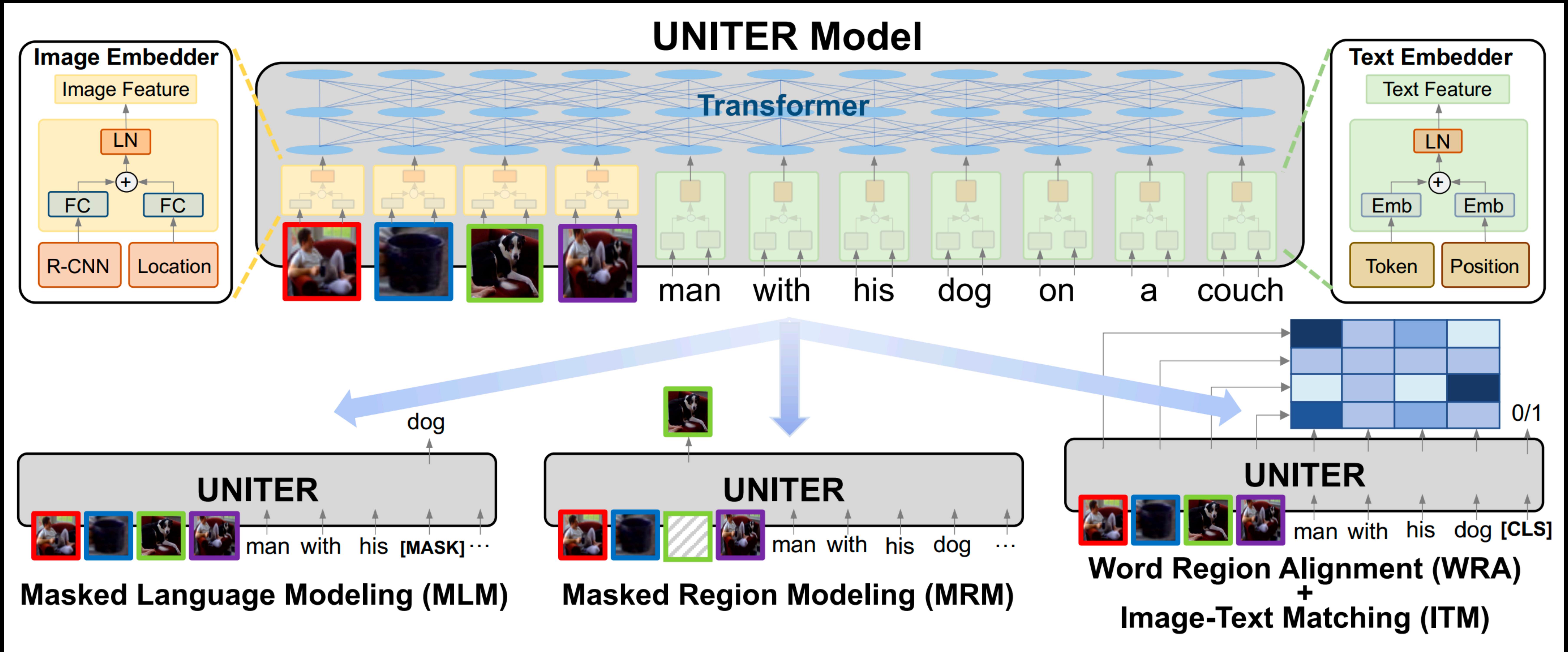


(a) Masked multi-modal learning

(b) Multi-modal alignment prediction

UNITER (2020)

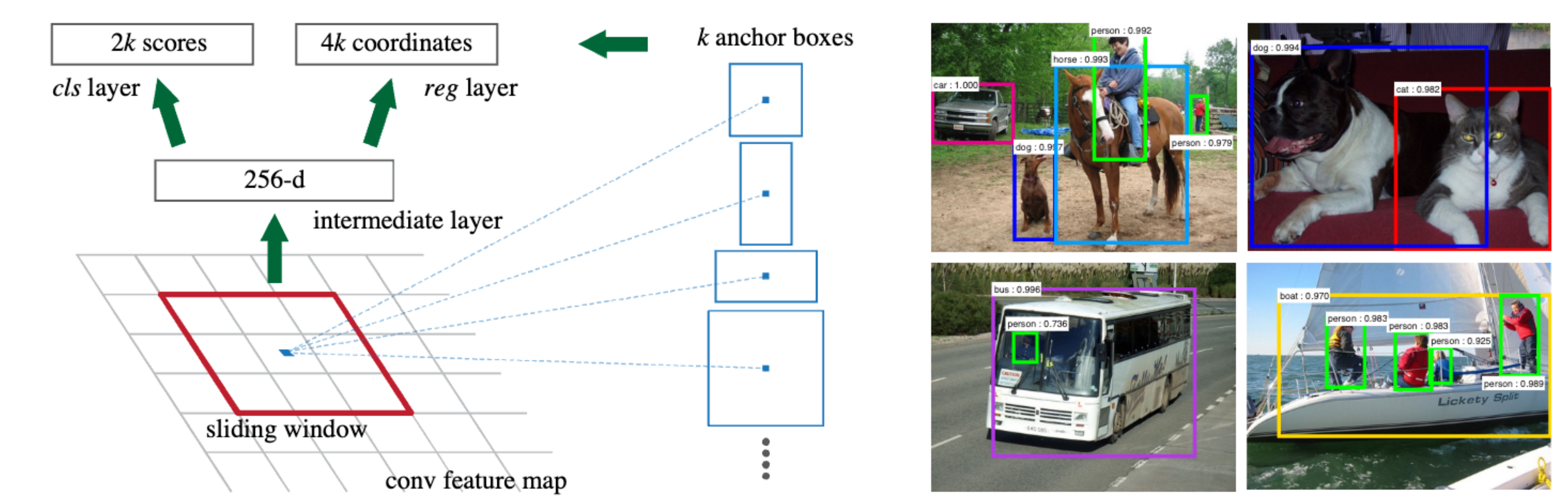
- Objectives: image-text matching (ITM), masked language modeling (MLM), masked region modeling (MRM), *word region alignment (WRA)*



“UNITER: Universal image-text representation learning” ECCV (2020)

Relying on external visual encoder

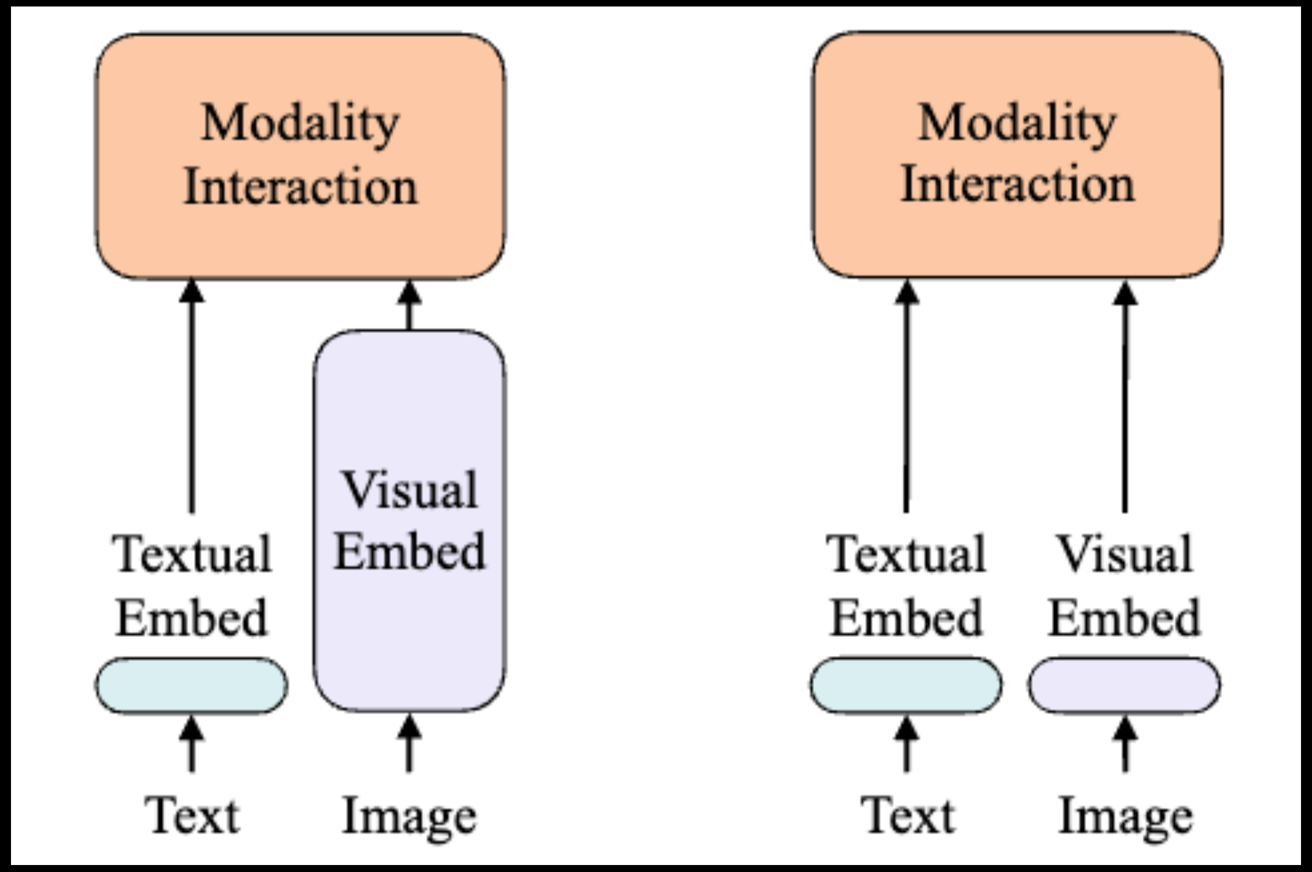
- "We use Faster R-CNN (with ResNet-101 backbone) pretrained on the Visual Genome dataset (...) to extract region features" in ViLBERT (Lu et al, 2019)
- Same in VisualBERT (Li et al., 2019), UNITER (Chen et al, 2019), OSCAR (Li et al., 2020), VinVL (Zhang et al., 2021)

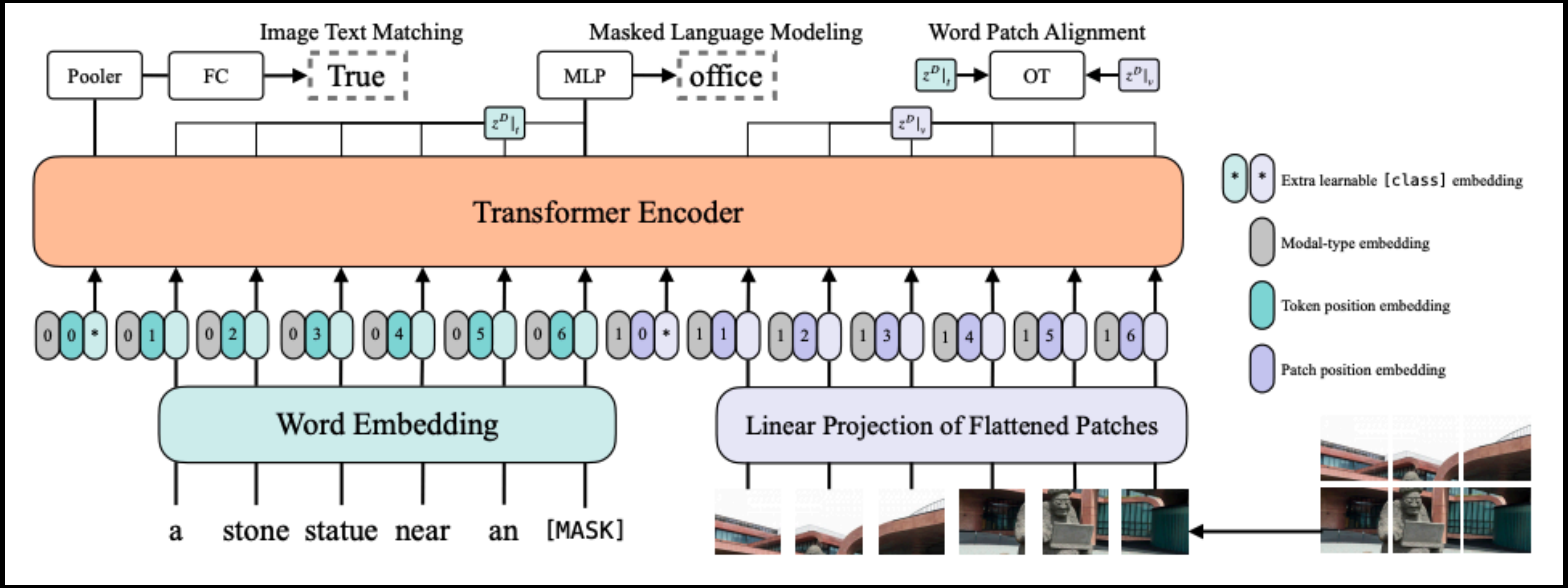


"Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks" (2015)

Beyond regional features — ViLT

- Objectives: image-text matching (ITM), masked language modeling (MLM), word patch alignment (WPA)





“ViLT: Vision-and-language transformer without convolution or region supervision” ICML (2021)

ViLT

- Objectives: image-text matching (ITM), masked language modeling (MLM), word patch alignment (WPA)

Visual Embed	Model	Time (ms)	VQAv2 test-dev	NLVR2 dev	NLVR2 test-P
Region	w/o VLP SOTA	~900	70.63	54.80	53.50
	ViLBERT	~920	70.55	-	-
	VisualBERT	~925	70.80	67.40	67.00
	LXMERT	~900	72.42	74.90	74.50
	UNITER-Base	~900	72.70	75.85	75.80
	OSCAR-Base†	~900	73.16	78.07	78.36
	VinVL-Base†‡	~650	75.95	82.05	83.08
Grid	Pixel-BERT-X152	~160	74.45	76.50	77.20
	Pixel-BERT-R50	~60	71.35	71.70	72.40
Linear	ViLT-B/32	~15	70.33	74.41	74.57
	ViLT-B/32ⓐ	~15	70.85	74.91	75.57
	ViLT-B/32ⓐ⊕	~15	71.26	75.70	76.13

VQA & Visual reasoning

Visual Embed	Model	Time (ms)	Text Retrieval						Image Retrieval					
			Flickr30k (1K)			MSCOCO (5K)			Flickr30k (1K)			MSCOCO (5K)		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Region	w/o VLP SOTA	~900	67.4	90.3	95.8	50.4	82.2	90.0	48.6	77.7	85.2	38.6	69.3	80.4
	ViLBERT-Base	~920	-	-	-	-	-	-	58.2	84.9	91.5	-	-	-
	Unicoder-VL	~925	86.2	96.3	99.0	62.3	87.1	92.8	71.5	91.2	95.2	48.4	76.7	85.9
	UNITER-Base	~900	85.9	97.1	98.8	64.4	87.4	93.1	72.5	92.4	96.1	50.3	78.5	87.2
	OSCAR-Base†	~900	-	-	-	70.0	91.1	95.5	-	-	-	54.0	80.8	88.5
	VinVL-Base†‡	~650	-	-	-	74.6	92.6	96.3	-	-	-	58.1	83.2	90.1
Grid	Pixel-BERT-X152	~160	87.0	98.9	99.5	63.6	87.5	93.6	71.5	92.1	95.8	50.1	77.6	86.2
	Pixel-BERT-R50	~60	75.7	94.7	97.1	59.8	85.5	91.6	53.4	80.4	88.5	41.1	69.7	80.5
Linear	ViLT-B/32	~15	81.4	95.6	97.6	61.8	86.2	92.6	61.9	86.8	92.8	41.3	72.0	82.5
	ViLT-B/32ⓐ	~15	83.7	97.2	98.1	62.9	87.1	92.7	62.2	87.6	93.2	42.6	72.8	83.4
	ViLT-B/32ⓐ⊕	~15	83.5	96.7	98.6	61.5	86.3	92.7	64.4	88.7	93.8	42.7	72.9	83.1

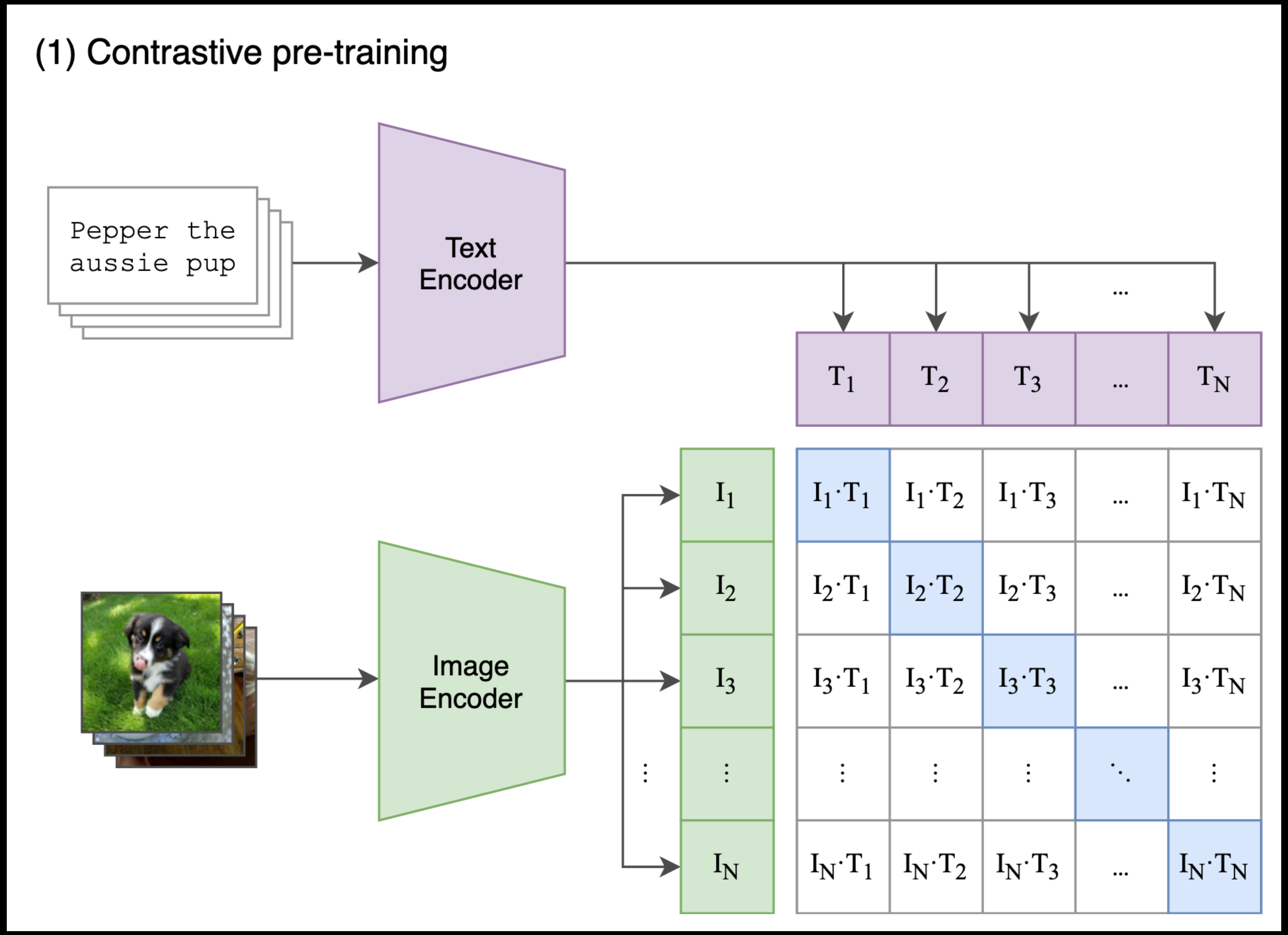
Retrieval

Vision-and-Language Pre-training

Contrastive learning

CLIP

Contrastive Language-Image Pre-training (CLIP)



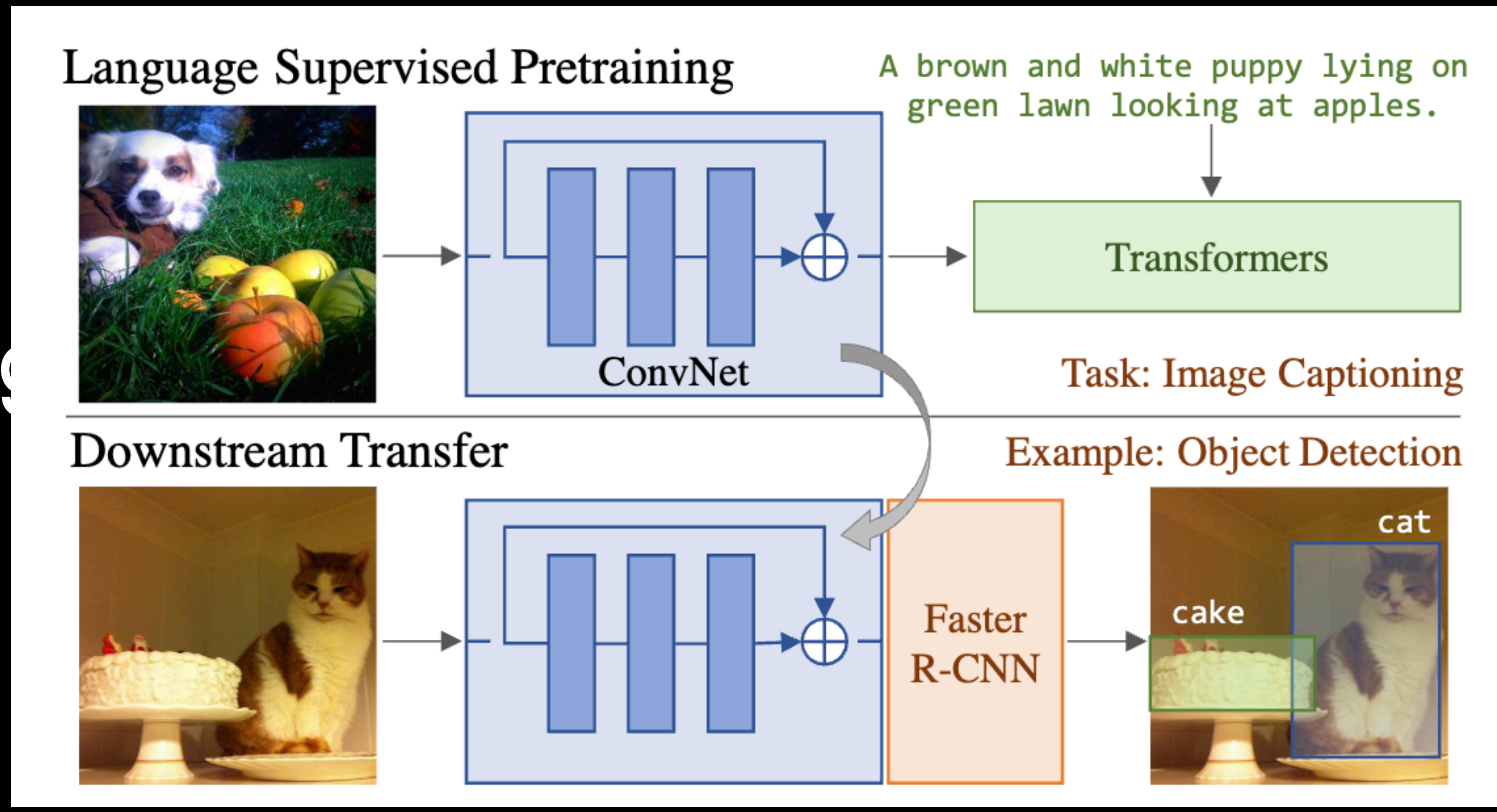
Radford et al., "Learning transferable visual models from natural language supervision," ICML (2021).

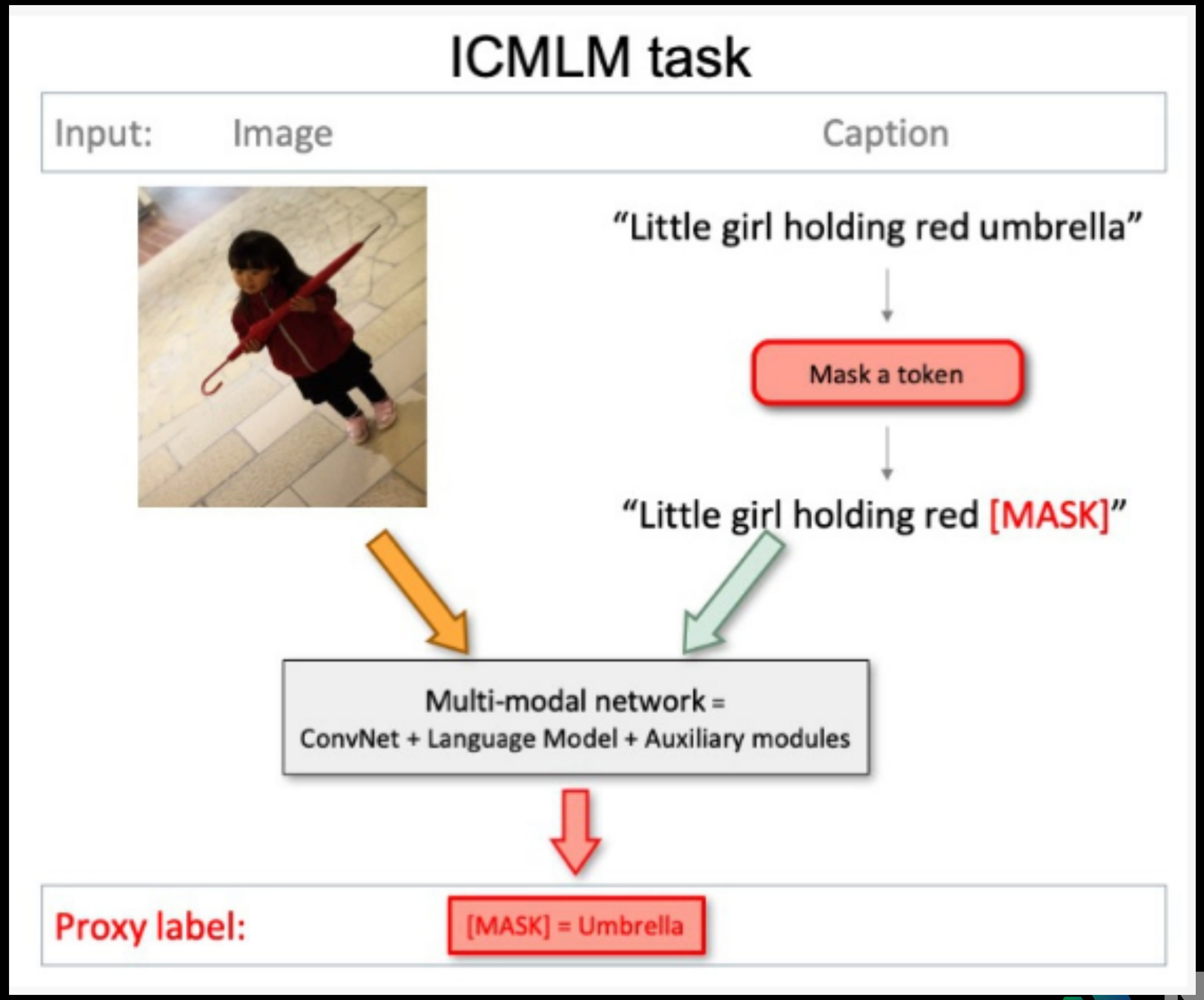
CLIP — motivation

- We saw NLP tasks can be solved in a zero-shot manner (GPT family)
  - But, in vision tasks: zero-shot 11.5% accuracy on ImageNet in 2017

CLIP — motivation

- We saw NLP tasks can be solved in a zero-shot ma
  - But, in vision tasks: zero-shot 11.5% accuracy on Imag
- Similar methods
  - VirTex (Desai & Johnson, 2020)
  - ICMLM (Bulent Sariyildiz et al., 2020)
  - ConVIRT (Zhang et al., 2020)
  - But, small-scale training (< 1 million images)



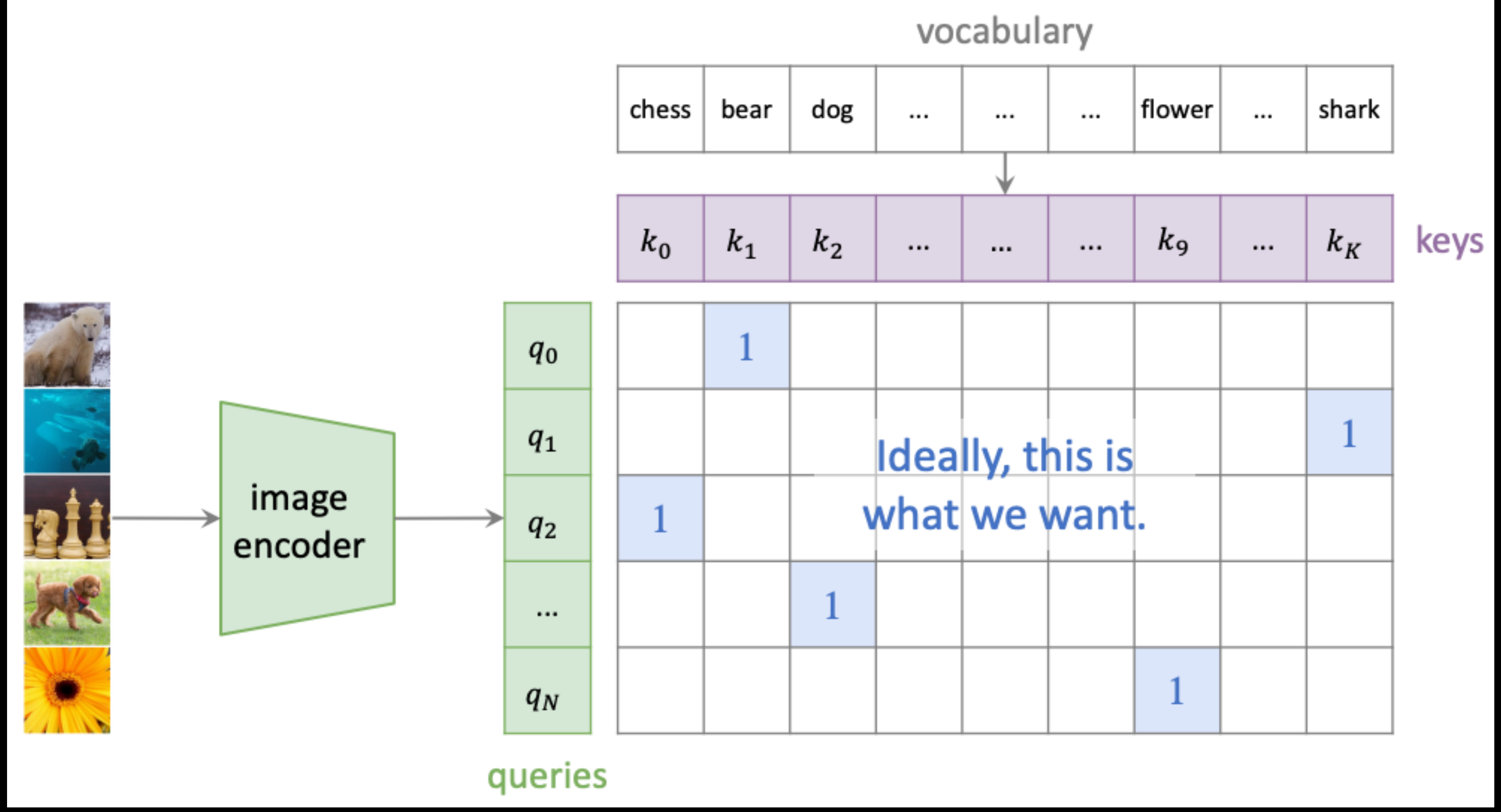


CLIP — motivation

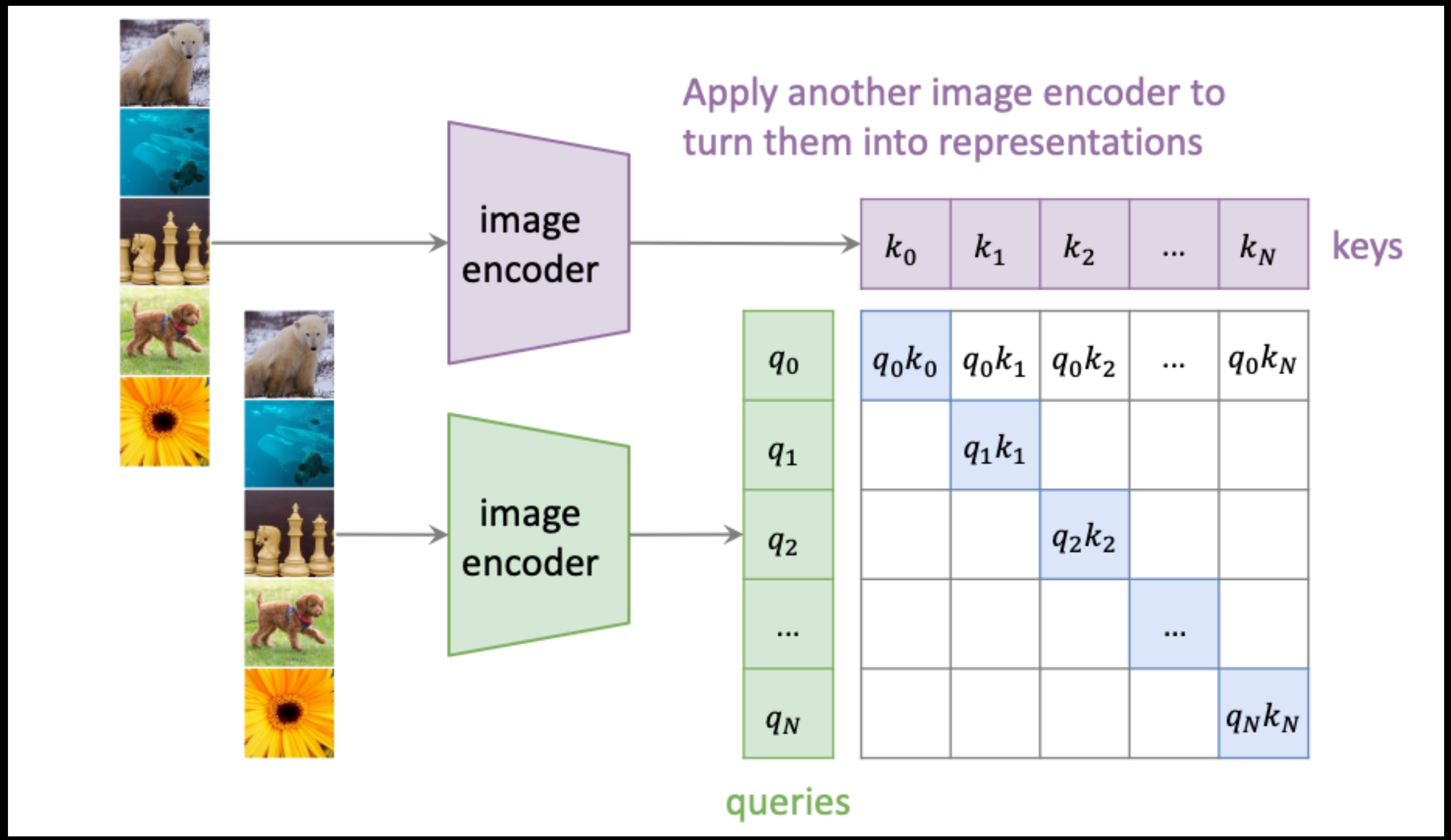
- Solution: *Scaling-up*
  - Larger data size: 400 million image-text pairs
  - Larger model size: ViT-Base/Large (with architectural change from Conv to ViT)

CLIP — contrastive learning

Supervised learning is contrastive learning



Contrastive learning with two views (e.g., SimCLR)

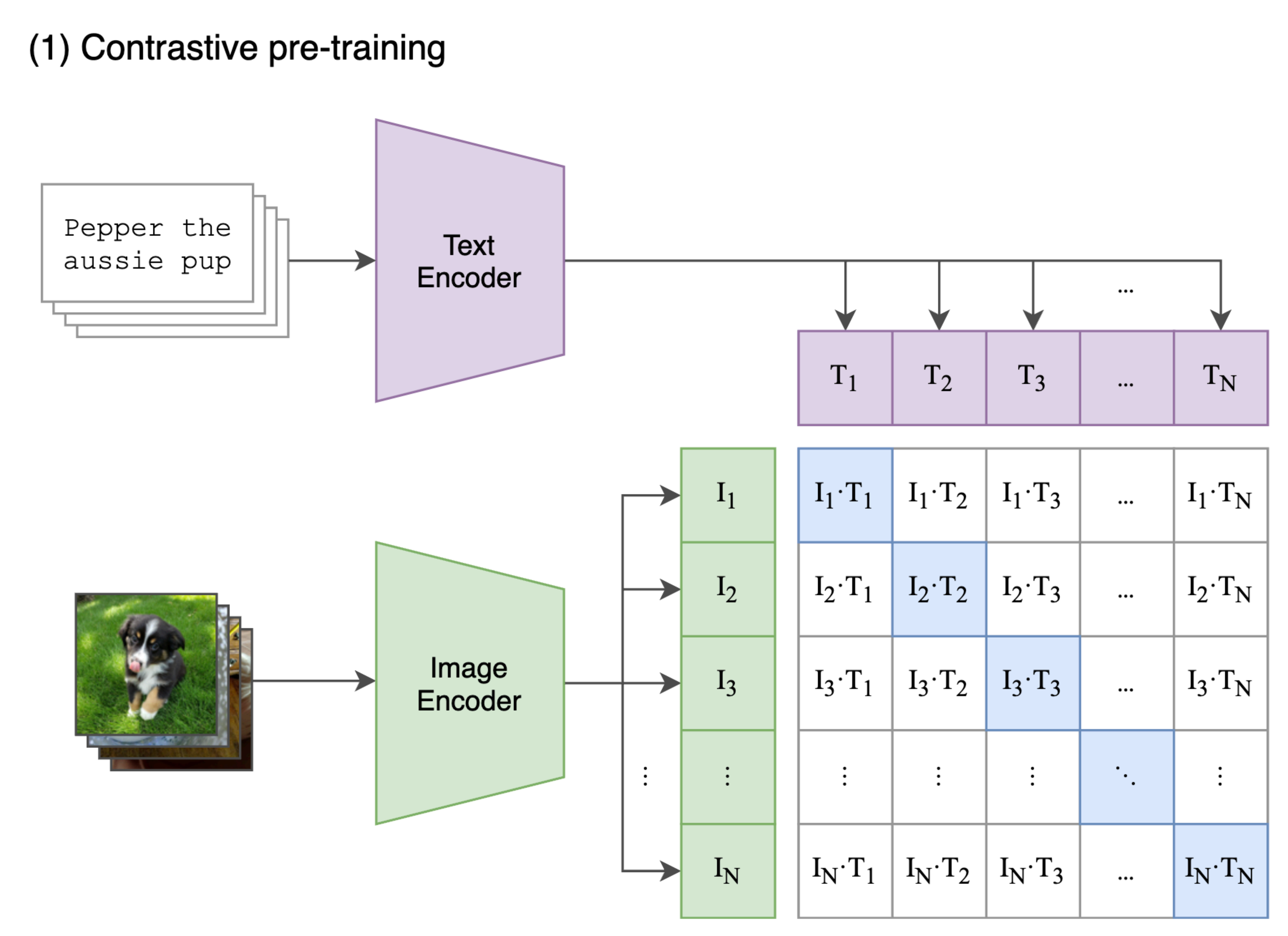


From lecture 12 of Advances in Computer Vision, (MIT CSAIL, Fall 2024)

NAVER AI LAB

CLIP — pre-training

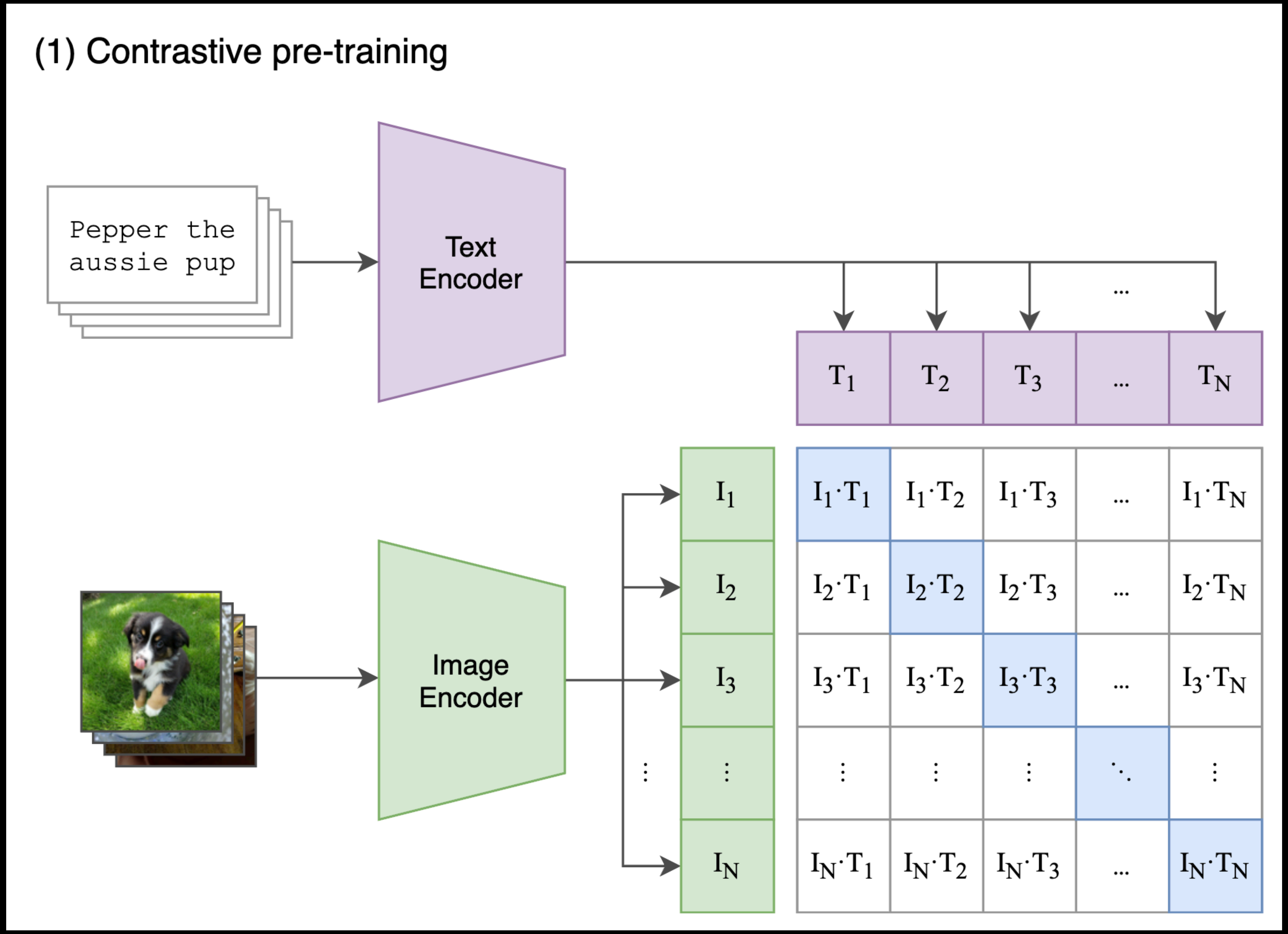
- Learning Transferable *Visual Models* From *Natural Language Supervision*



Radford et al., "Learning transferable visual models from natural language supervision," ICML (2021).

CLIP — pre-training

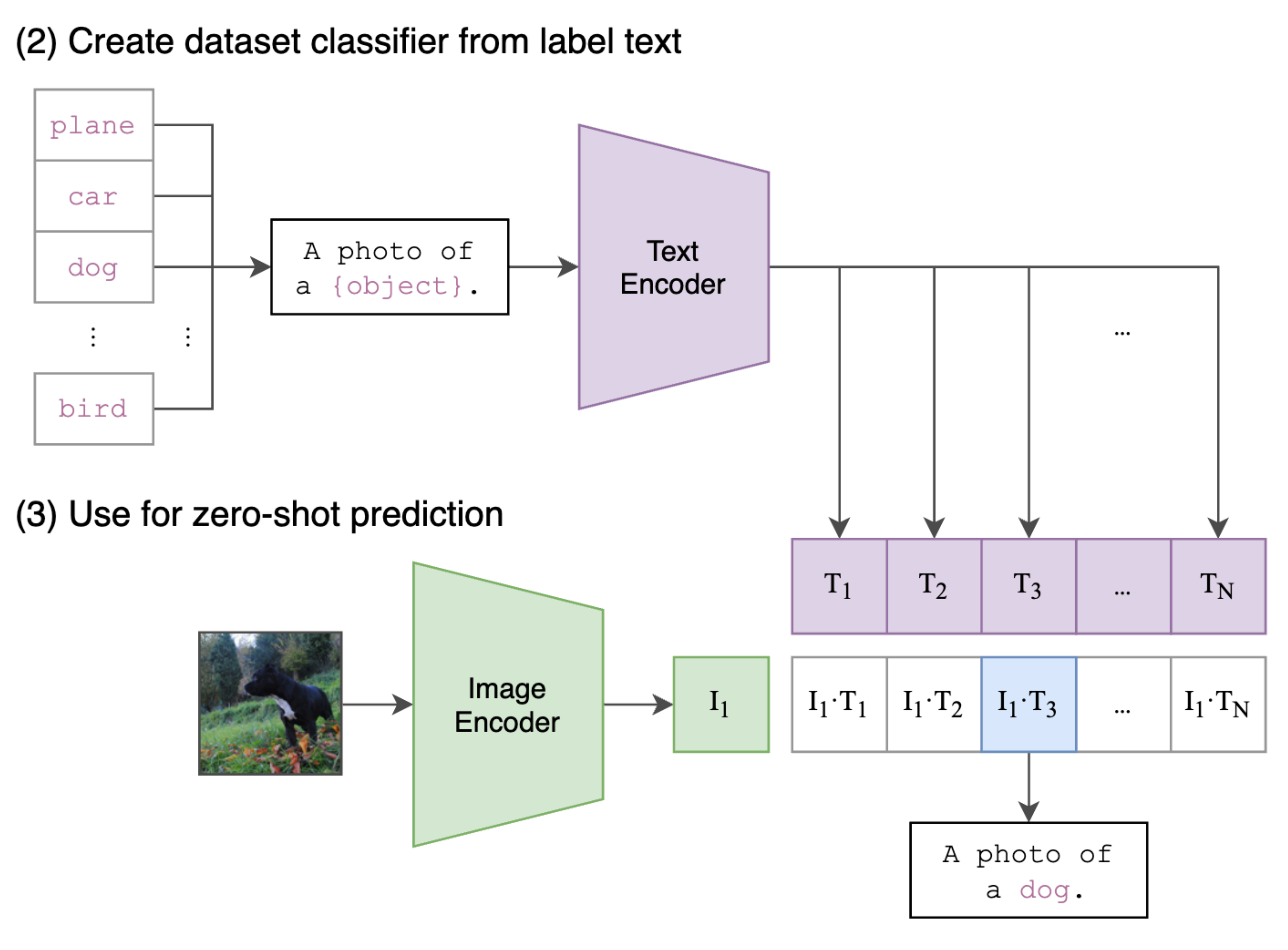
- Data: WIT-400M
- Image encoder: ResNets or ViT-B/L
- Text encoder: Transformer

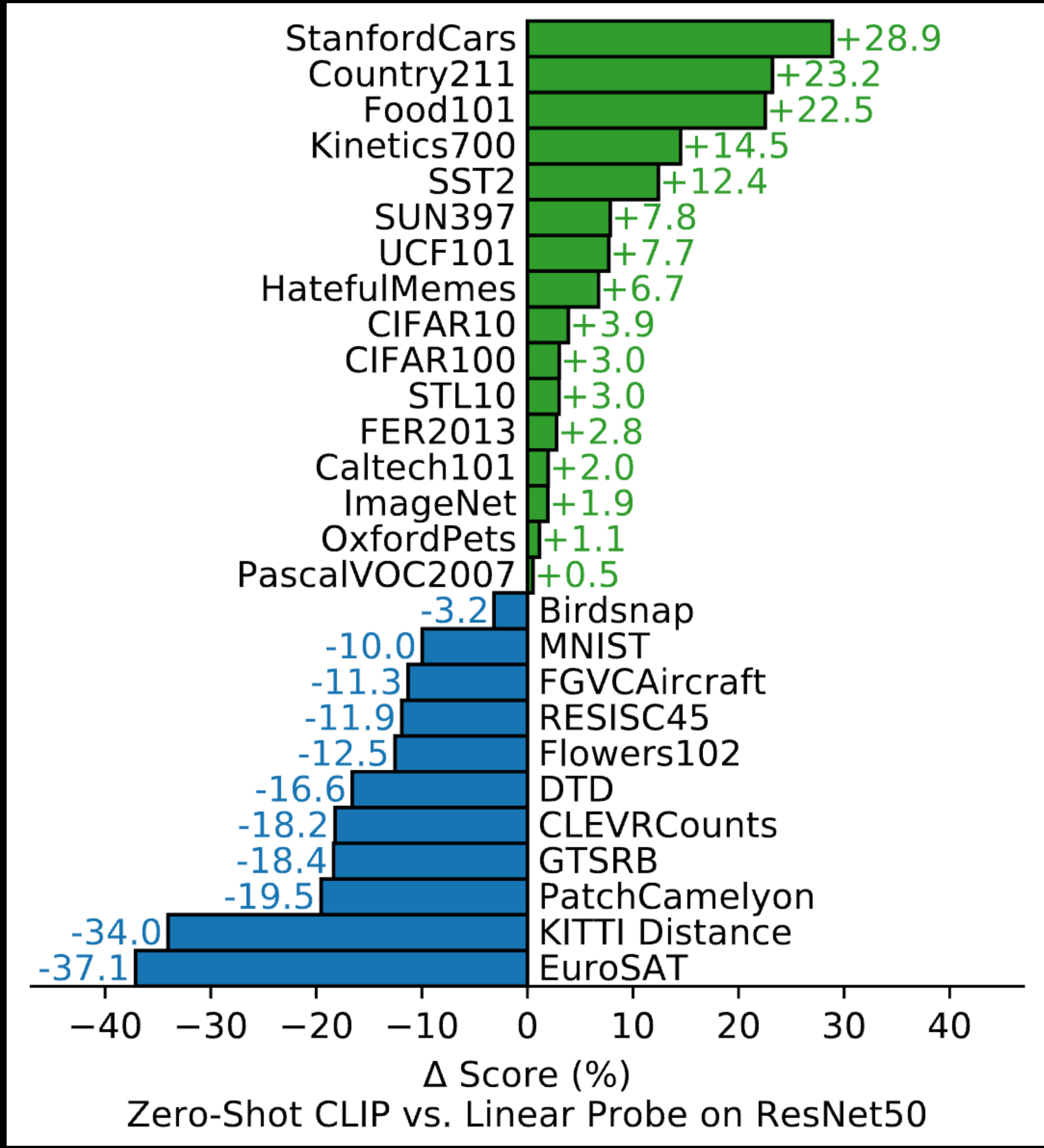



```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Radford et al., "Learning transferable visual models from natural language supervision," ICML (2021).

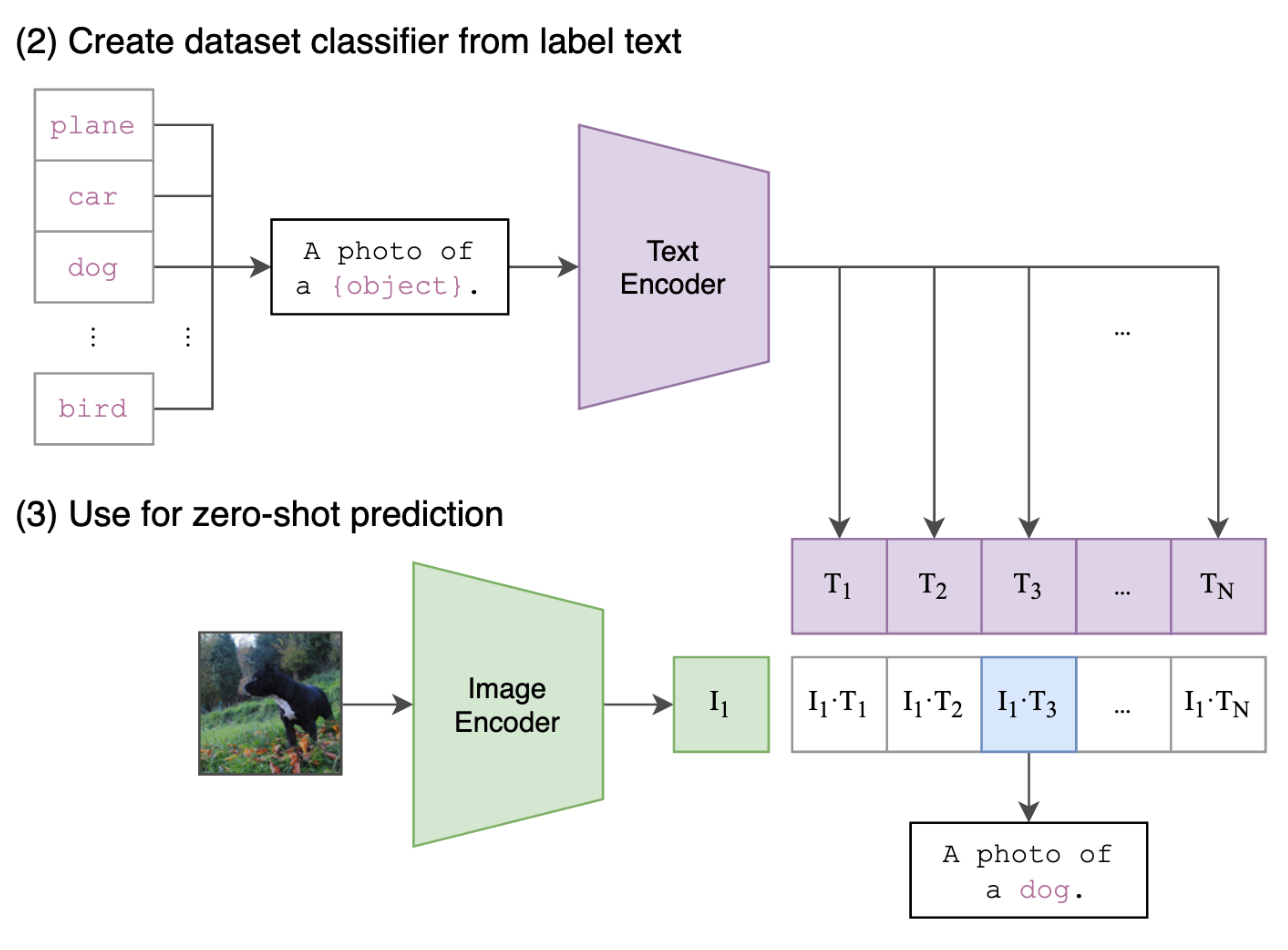
CLIP — inference (zero-shot)

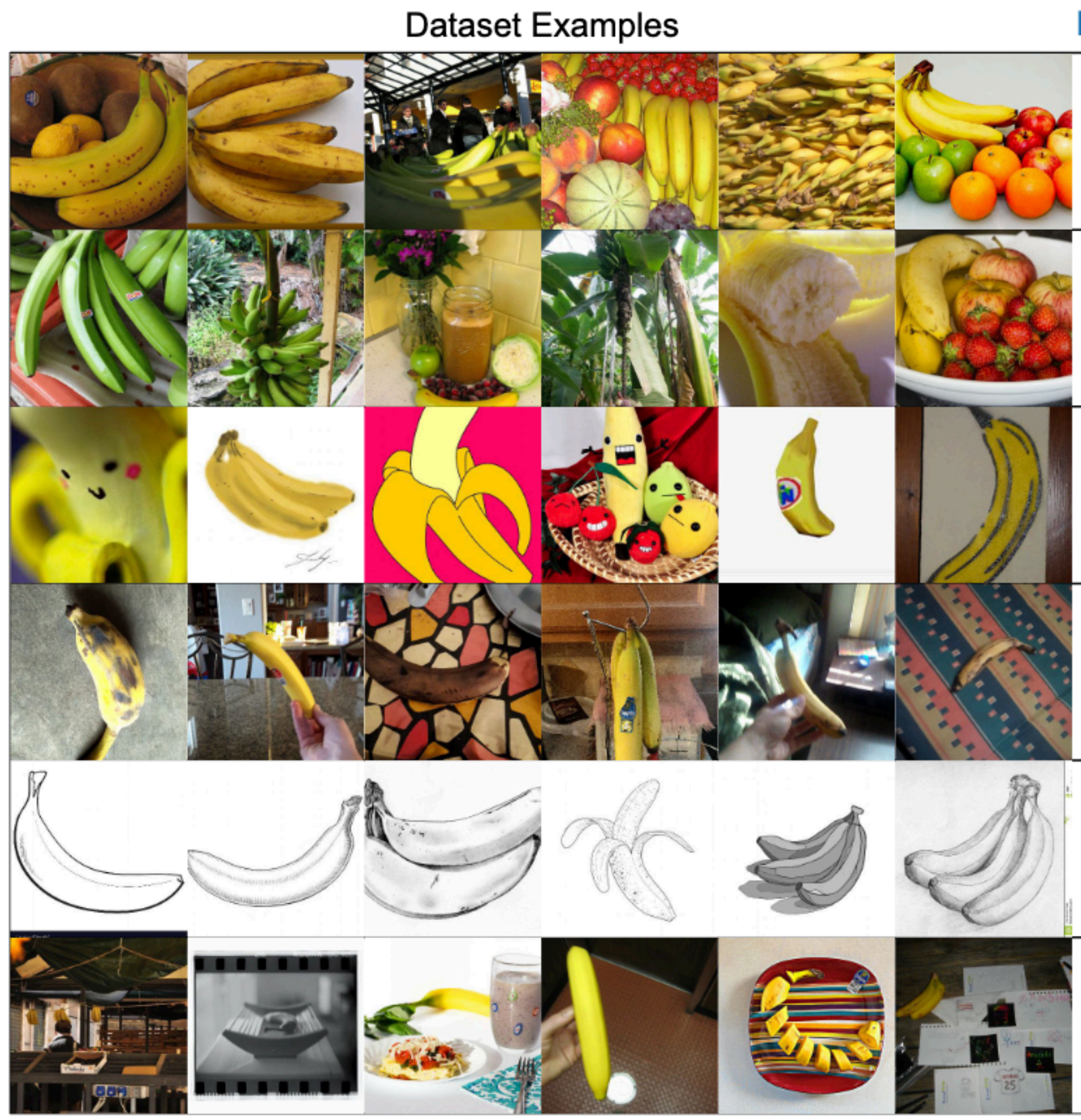




Radford et al., "Learning transferable visual models from natural language supervision," ICML (2021).

CLIP — inference (zero-shot)



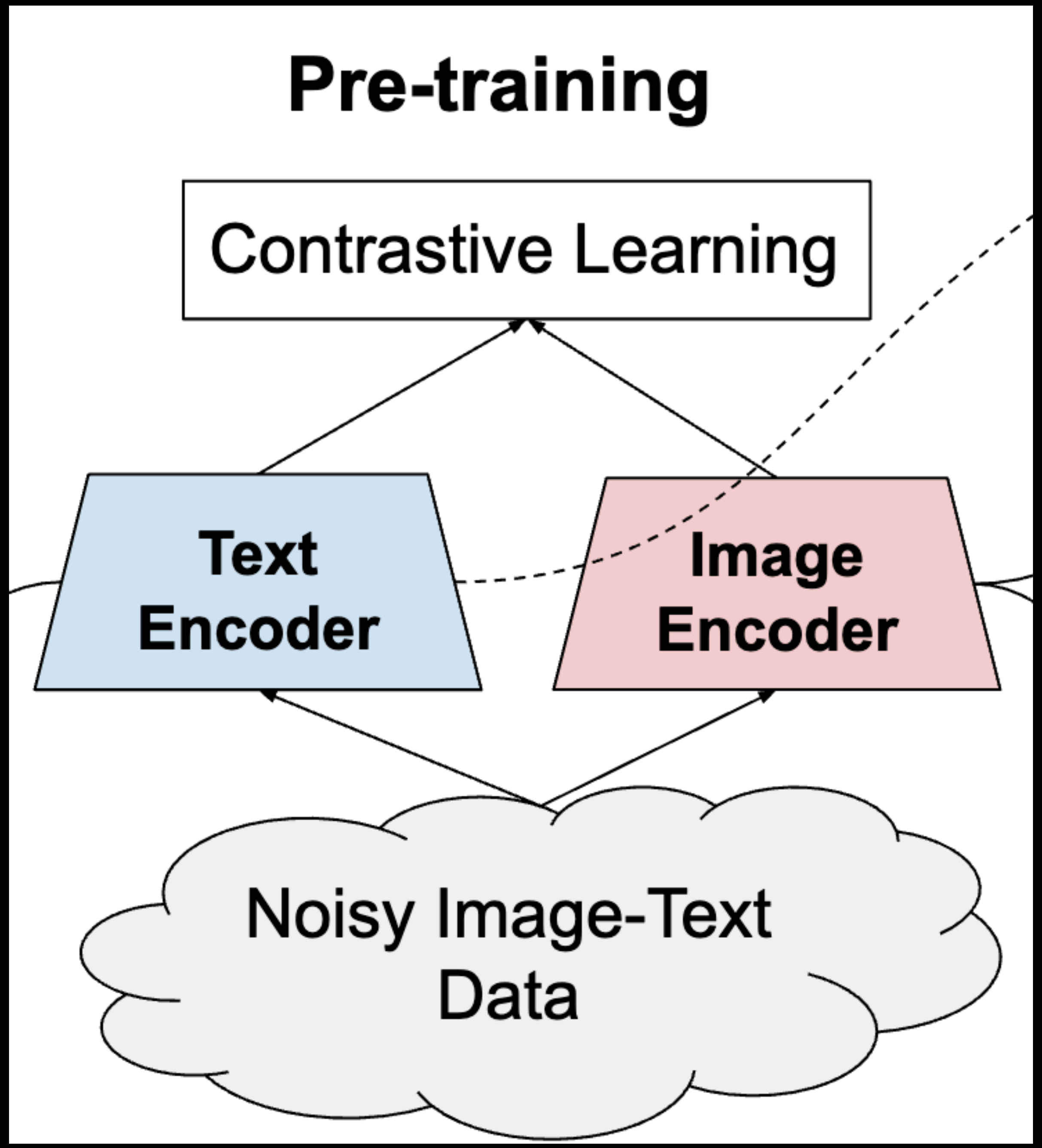


	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet	**76.2**	**76.2**	0%
ImageNetV2	64.3	**70.1**	+5.8%
ImageNet-R	37.7	**88.9**	+51.2%
ObjectNet	32.6	**72.3**	+39.7%
ImageNet Sketch	25.2	**60.2**	+35.0%
ImageNet-A	2.7	**77.1**	+74.4%

Radford et al., "Learning transferable visual models from natural language supervision," ICML (2021).

ALIGN

- Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision (by Google)



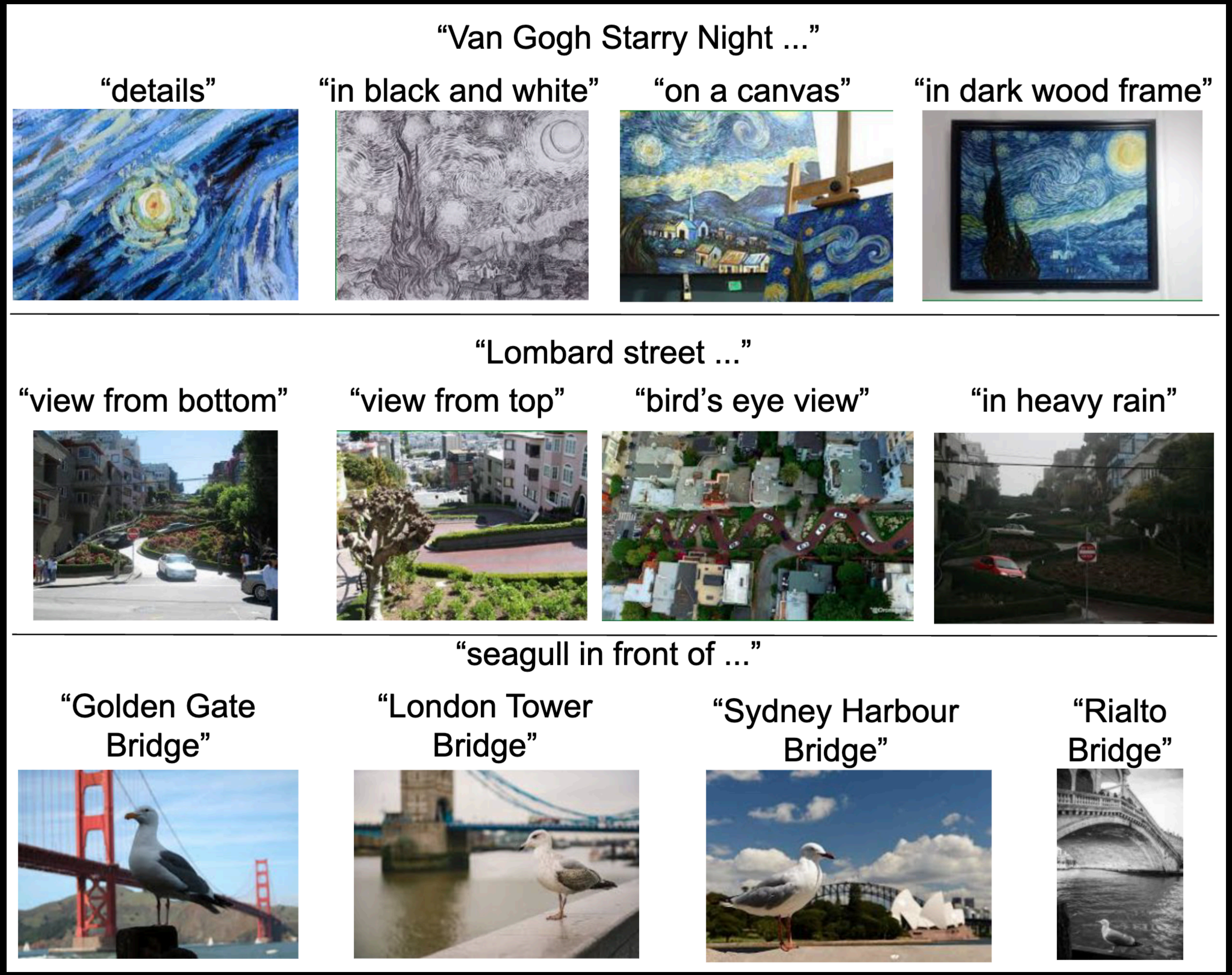
- Data: ALIGN 1.8B
- Image encoder: EfficientNet-L2
- Text encoder: BERT Transformer
- (From scratch)

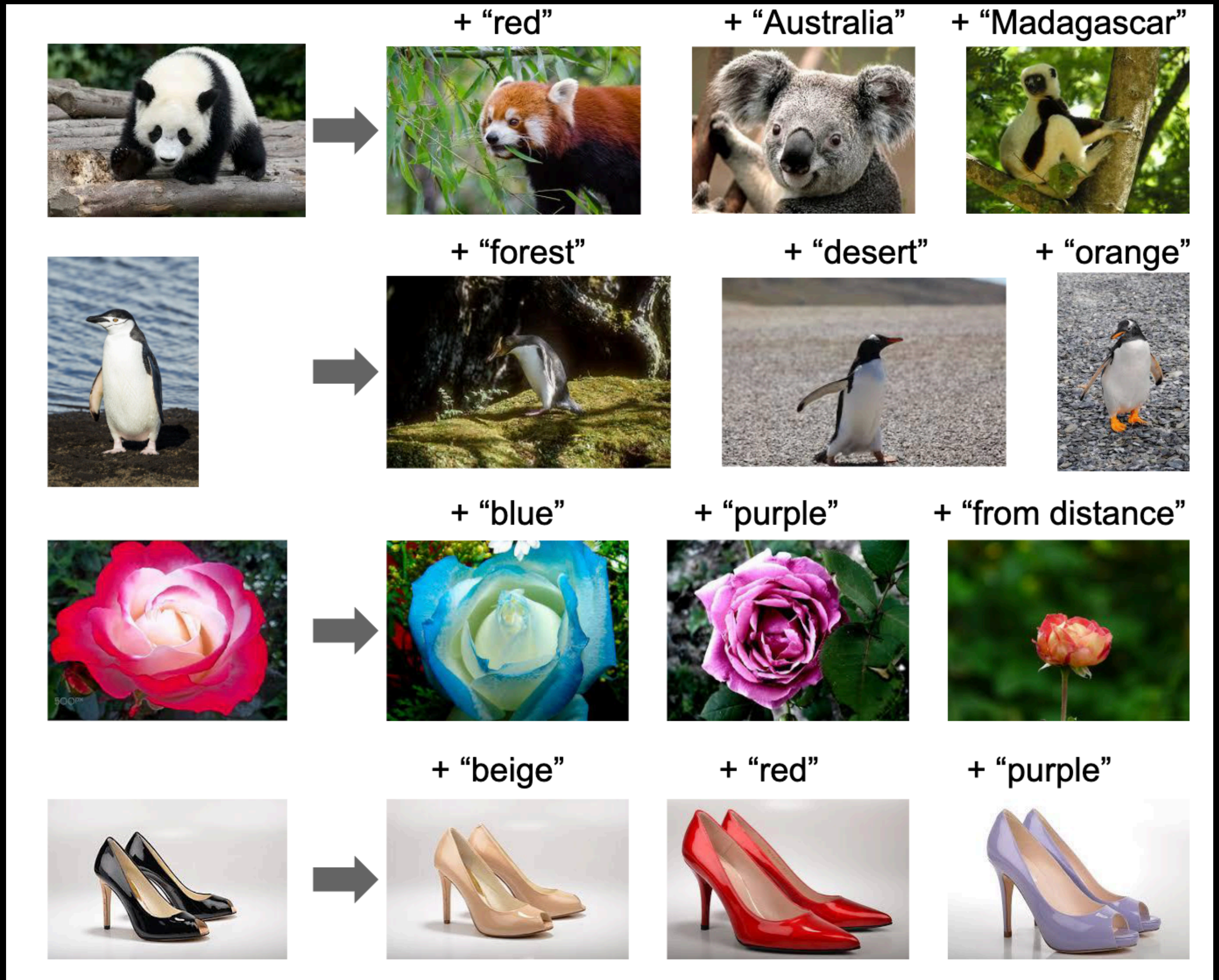
Table 4. Top-1 Accuracy of zero-shot transfer of ALIGN to image classification on ImageNet and its variants.

Model	ImageNet	ImageNet-R	ImageNet-A	ImageNet-V2
CLIP	76.2	88.9	**77.2**	**70.1**
ALIGN	**76.4**	**92.2**	75.8	**70.1**

ALIGN

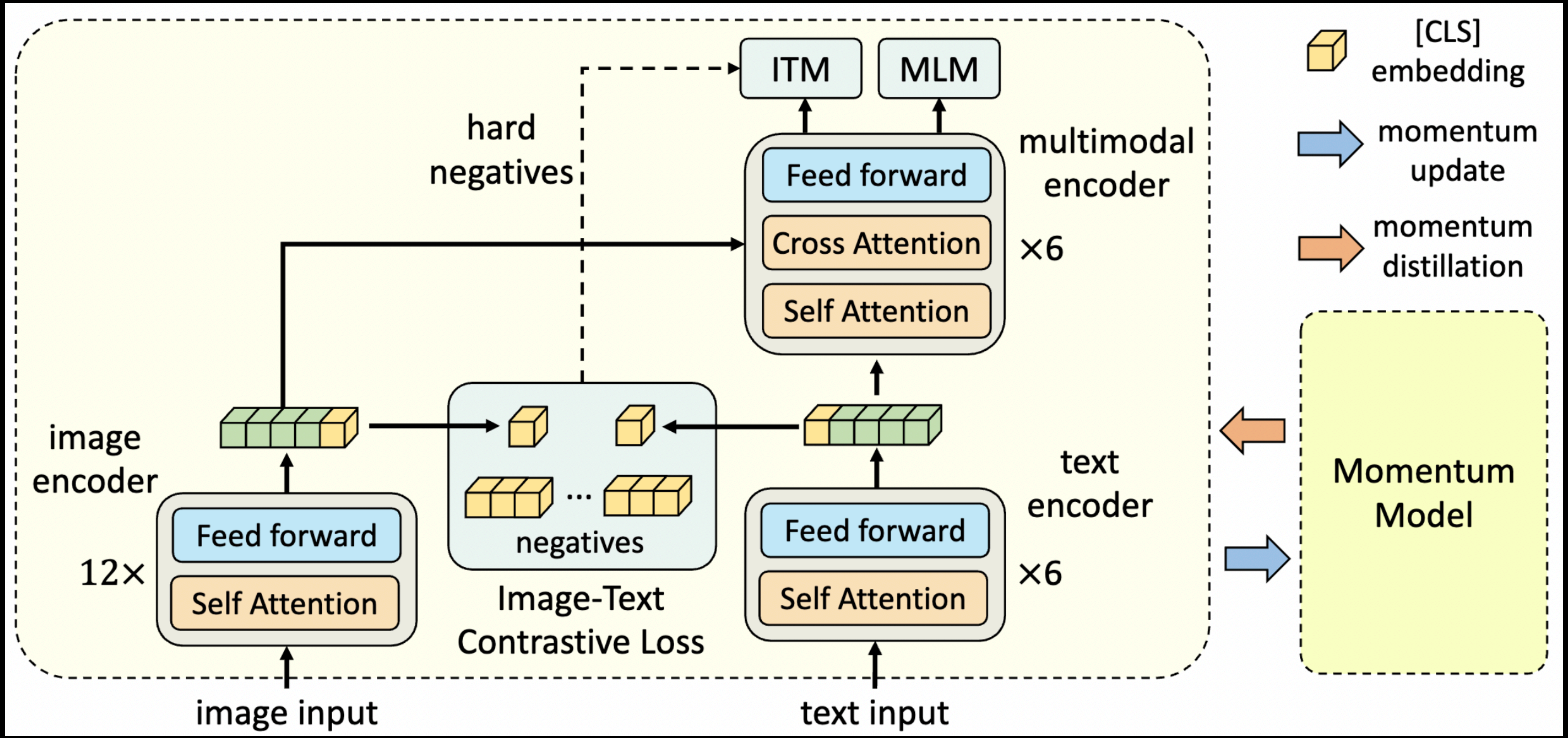
- Analysis of the embedding space





Mix with contrastive learning: ALBEF

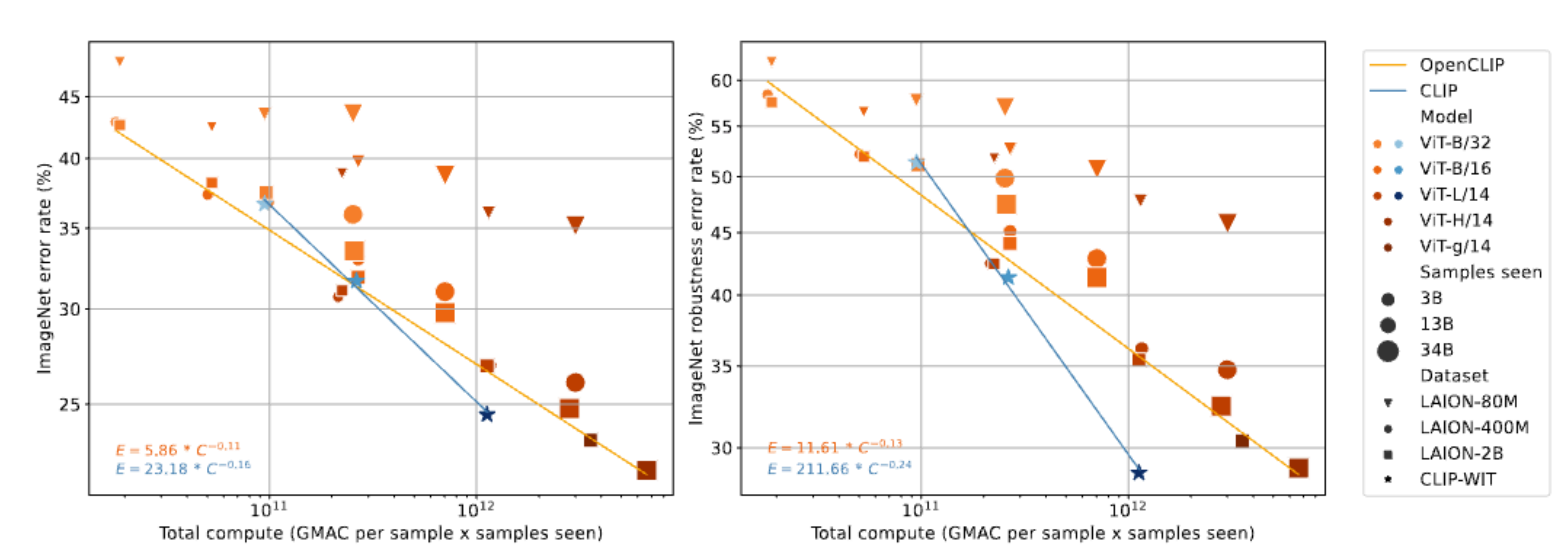
- Objectives: *image-text contrastive* (ITC), image-text matching (ITM), masked language modeling (MLM)



Data quality for CLIP

Reproducing CLIP

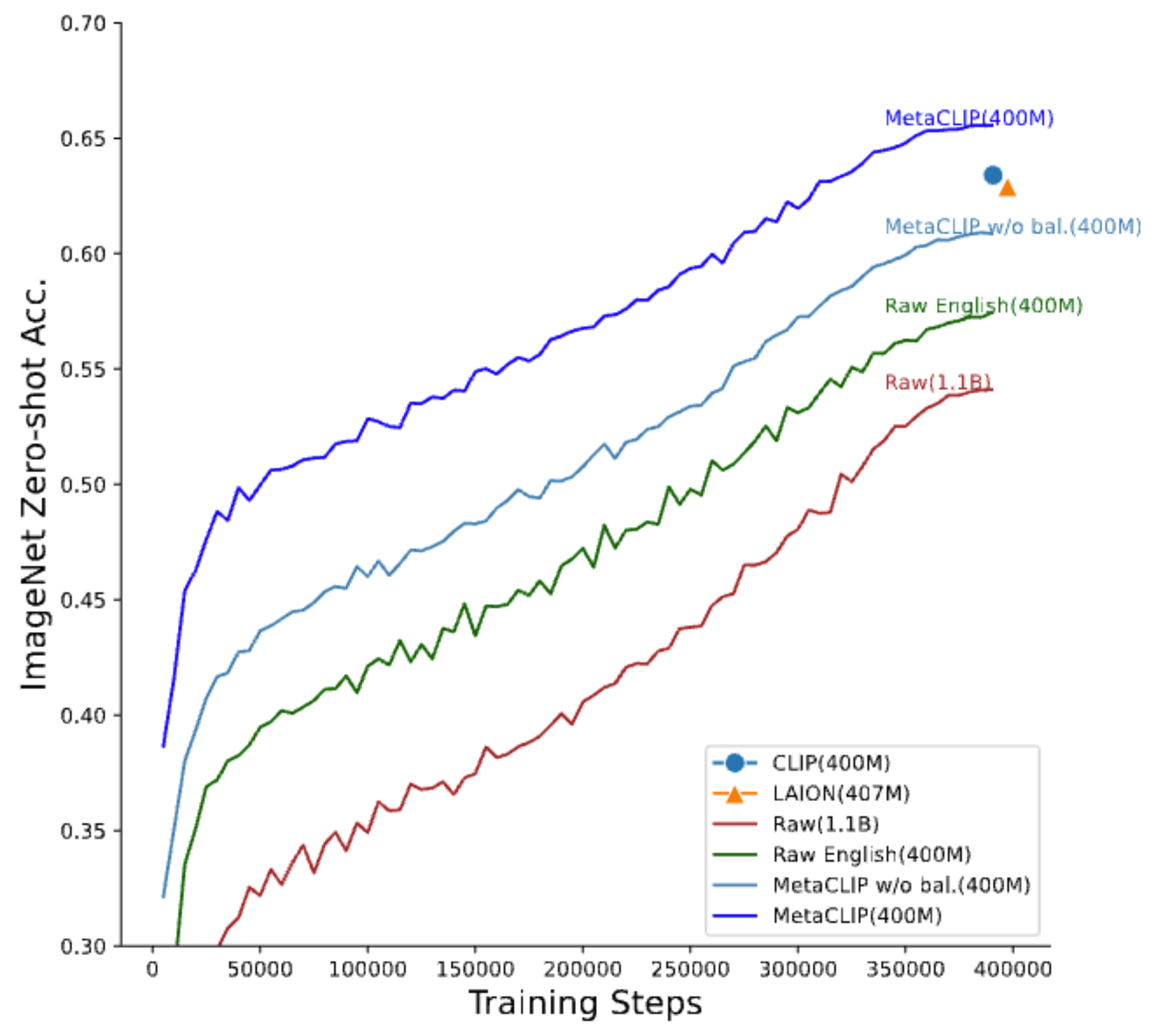
- CLIP (Radford et al., 2021) vs. Open-CLIP (Cherti et al., 2022)
- CLIP trained with (private) WIT-400M
- Open-CLIP trained with (open) LAION-400M (later, LAION-2B)



Cherti et al., "Reproducible scaling laws for contrastive language-image learning" (2022)

Beyond CLIP data

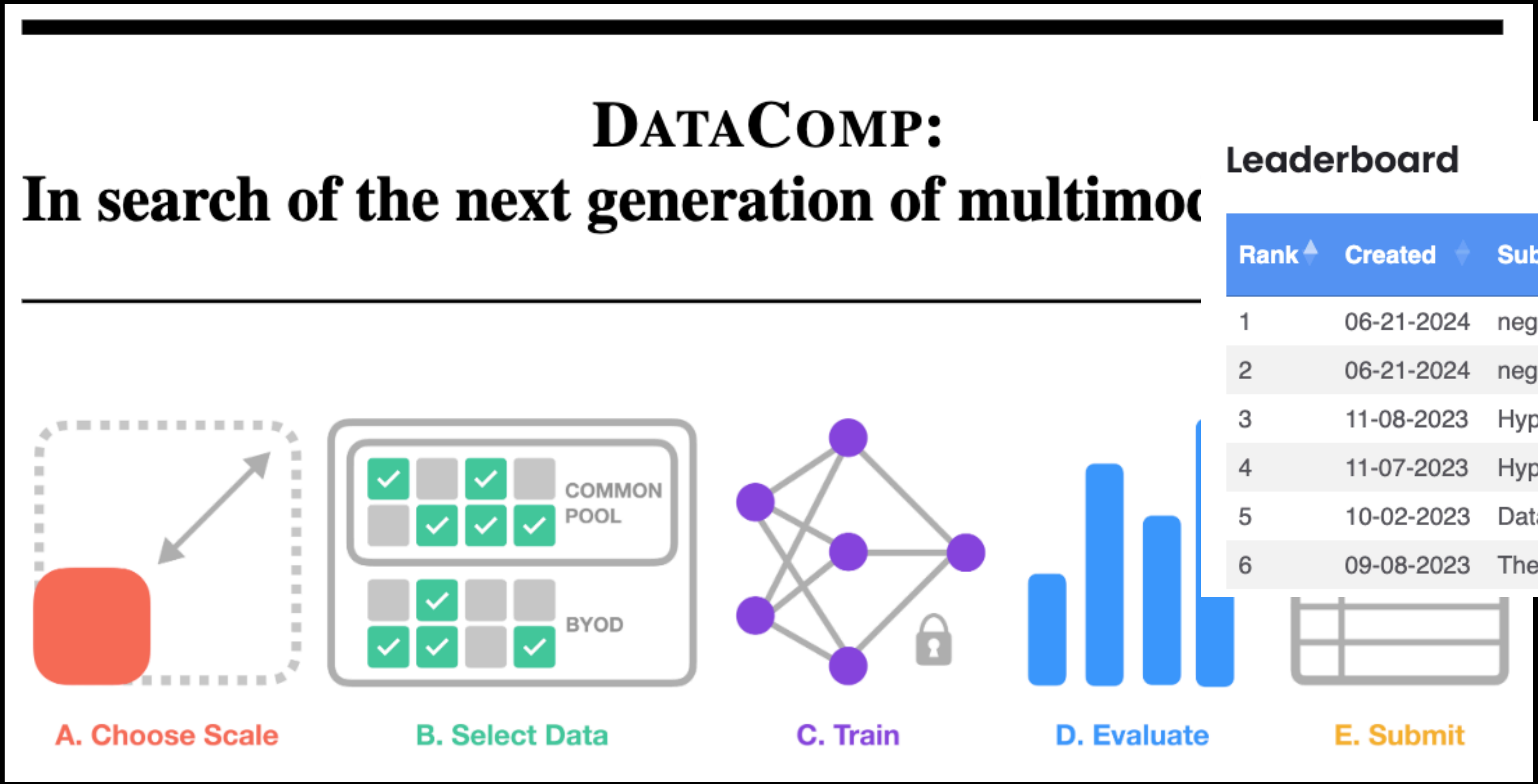
- Demystifying CLIP Data (MetaCLIP)



Beyond web-crawled (CommonCrawl),
They claim the importance of metadata curation and balancing

Beyond CLIP data — DataComp

- Beyond LAION datasets, multimodal data *curation* and *filtering* are the keys



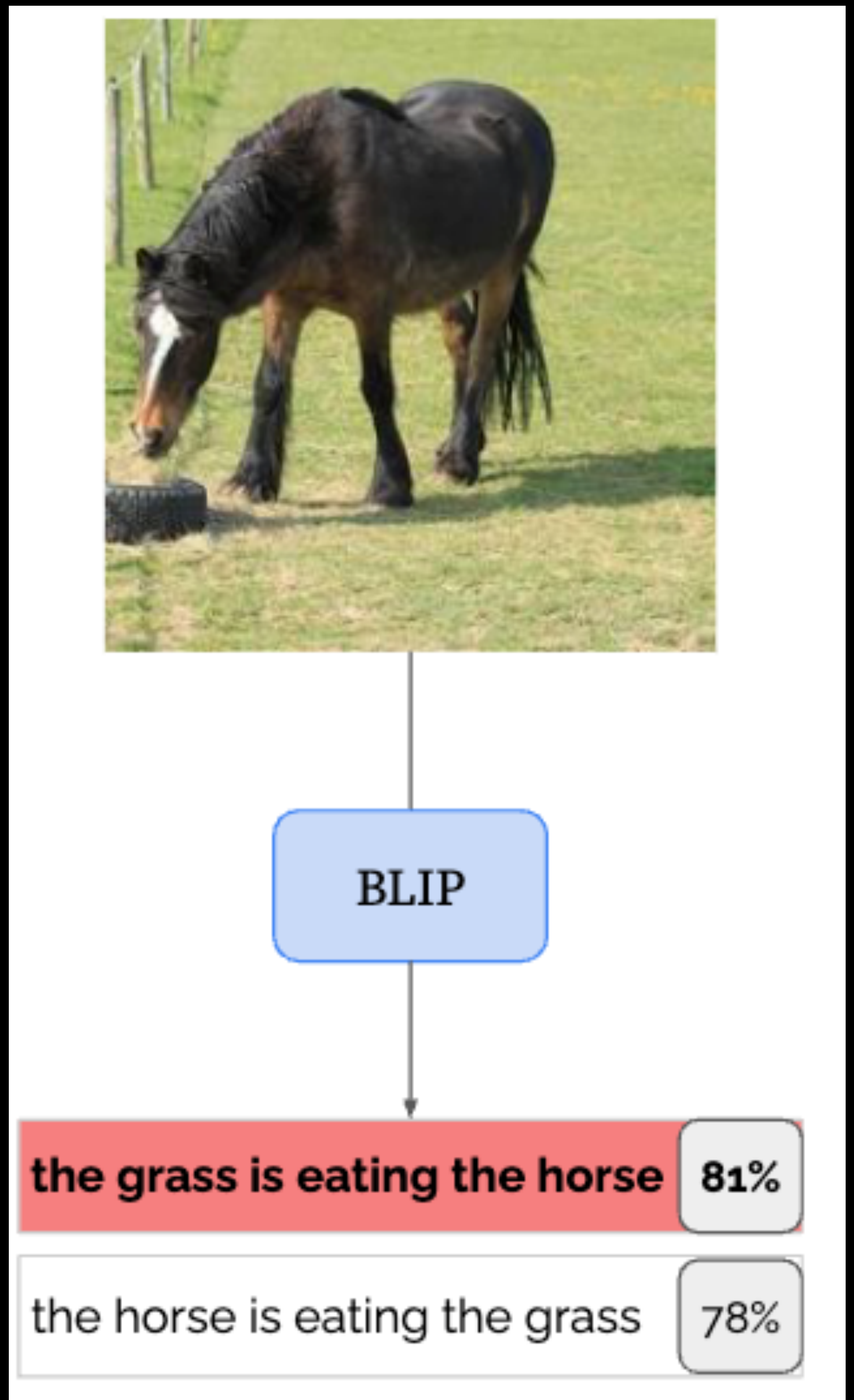
Leaderboard

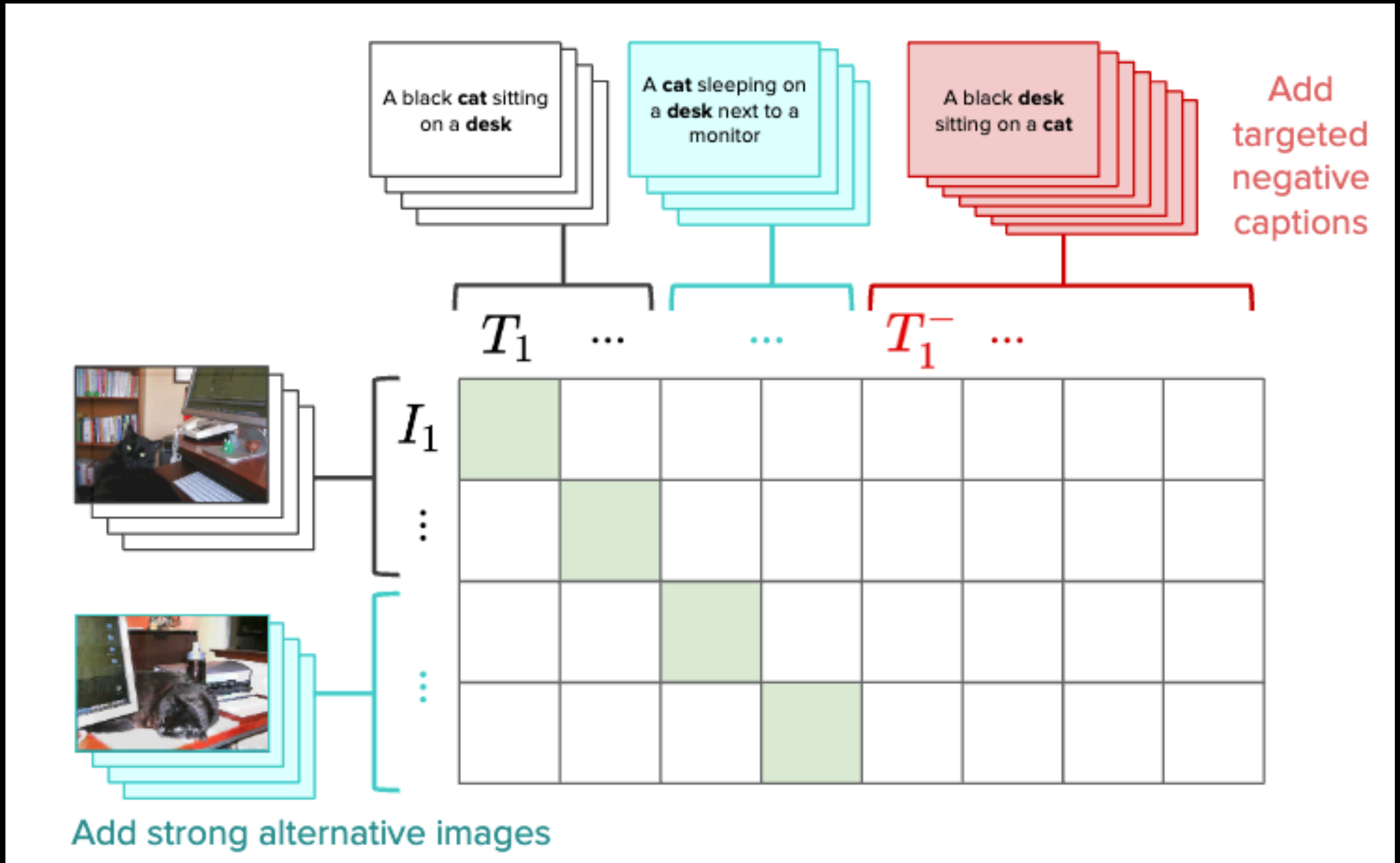
Rank	Created	Submission	ImageNet acc.	Average perf.
1	06-21-2024	negCLIPLoss & NormSim + DFN + HYPE	0.382	0.388
2	06-21-2024	negCLIPLoss & NormSim + DFN	0.375	0.386
3	11-08-2023	Hype sampler + DFN	0.382	0.379
4	11-07-2023	Hype sampler	0.346	0.373
5	10-02-2023	Data Filtering Networks	0.371	0.373
6	09-08-2023	The Devil Is in the Details	0.320	0.371

DataComp: In search of the next generation of multimodal datasets (2023)

Behave like bag-of-words

- Do they understand the structure and composition of the query?

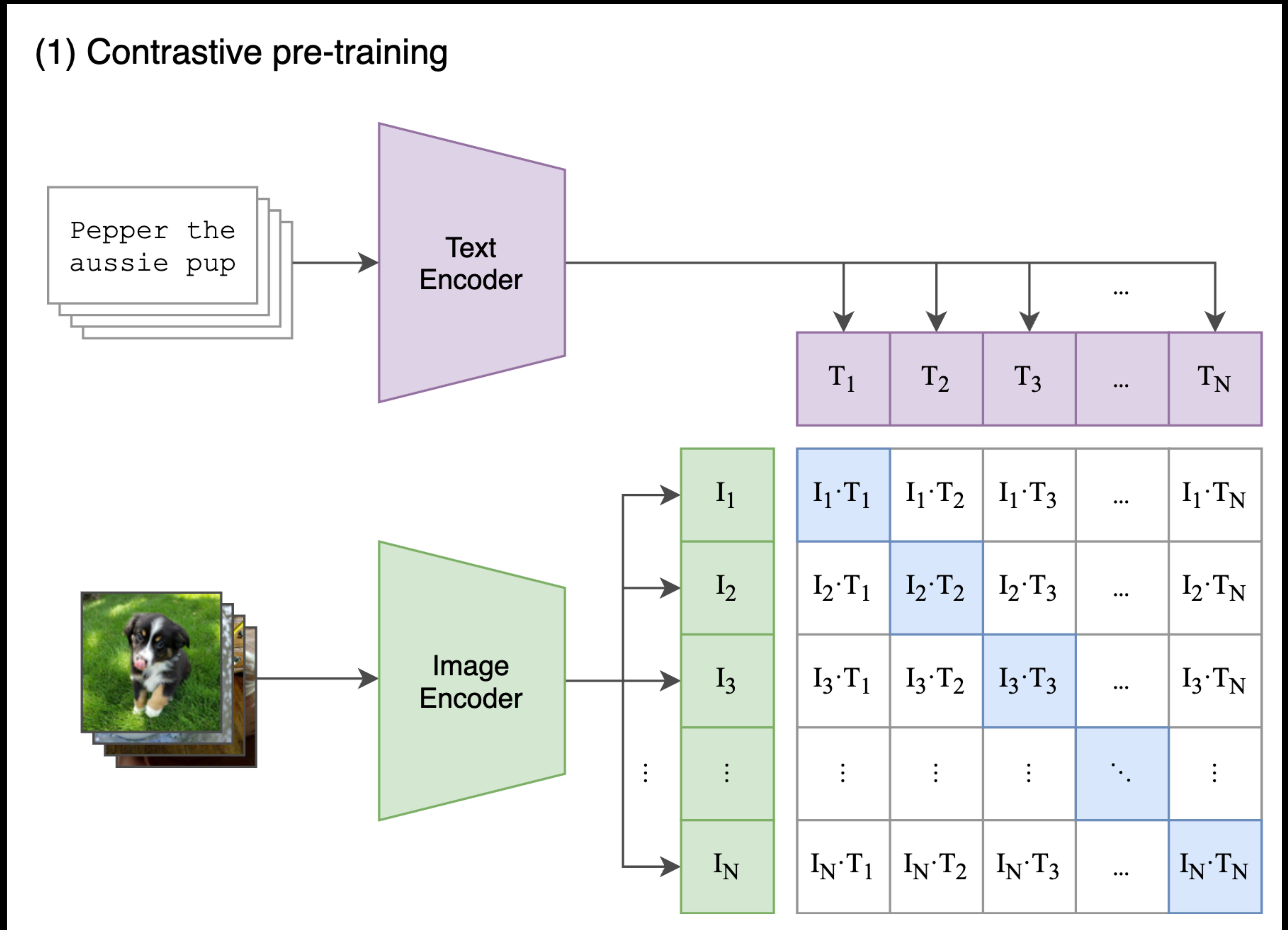




SigLIP: Scaling up training with Sigmoid loss

Revisit CLIP

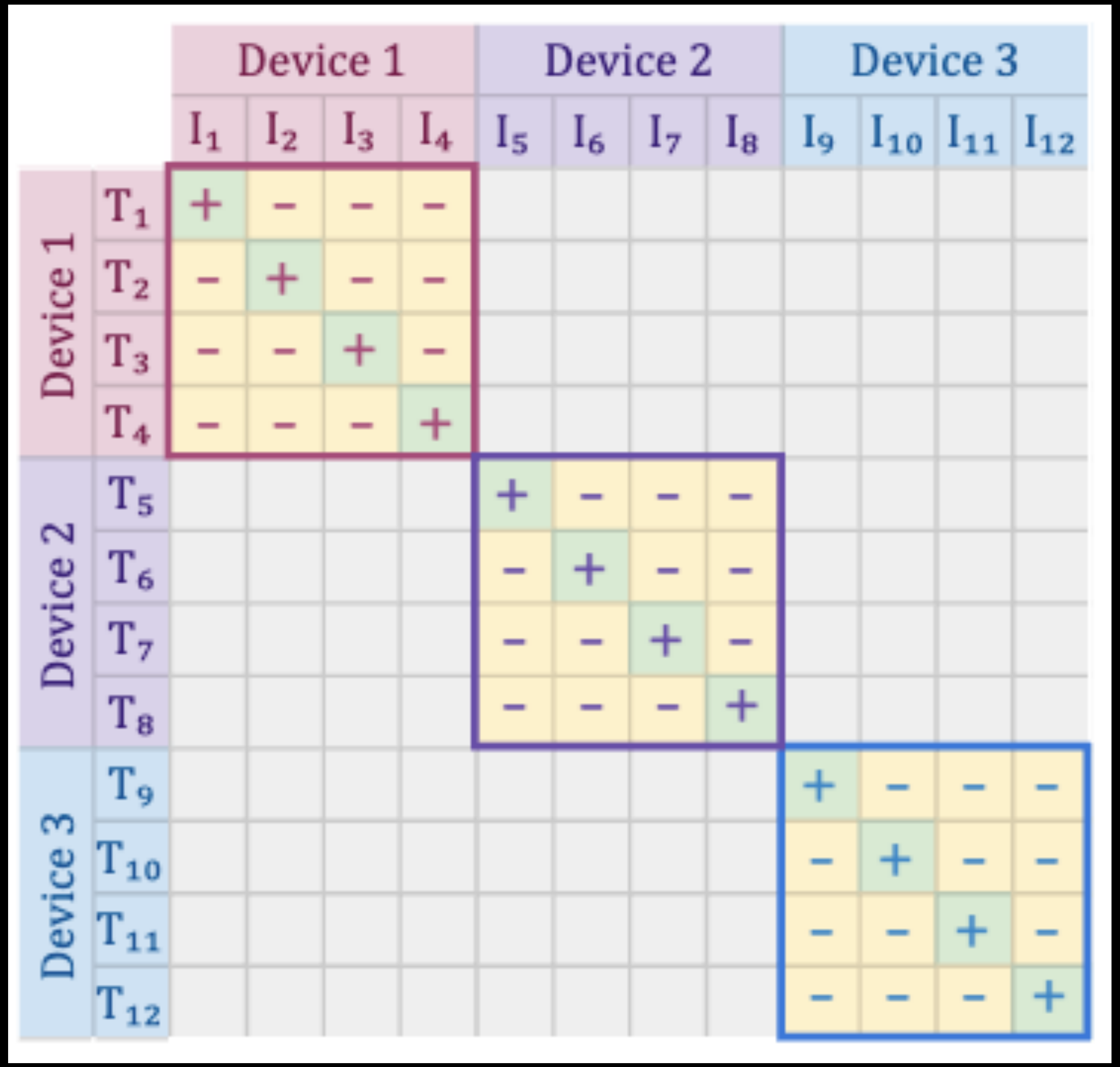
- Softmax-based contrastive objective (CLIP's objective)



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

Revisit CLIP

- Softmax-based contrastive objective (CLIP's objective)
- Contrastive learning requires large batch-size (e.g., 16K, 32K)
- To compute the denominator term, gather all image features (and text features)



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

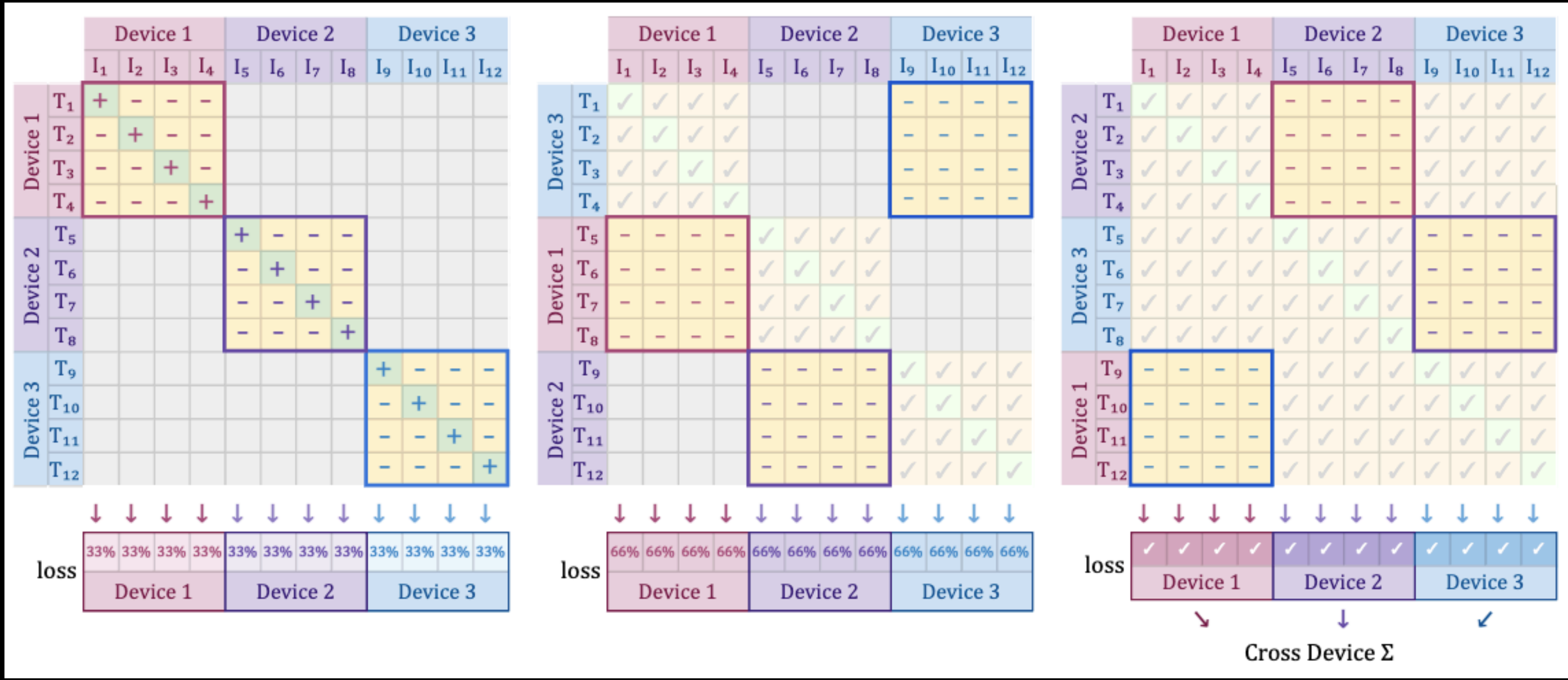
SigLIP: Sigmoid-based loss

$$-\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_i\cdot\mathbf{y}_j}}}^{\text{image}\rightarrow\text{text softmax}}+\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_j\cdot\mathbf{y}_i}}}^{\text{text}\rightarrow\text{image softmax}}\right) \Rightarrow -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\underbrace{\log\frac{1}{1+e^{z_{ij}(-t\mathbf{x}_i\cdot\mathbf{y}_j+b)}}}_{\mathcal{L}_{ij}}$$

- z_{ij} is the label for the image (x_i) and text (y_j) — 1 if paired, otherwise 0
- Compared to Softmax, sigmoid loss simplifies the problem to *binary classification*

SigLIP: Sigmoid-based loss

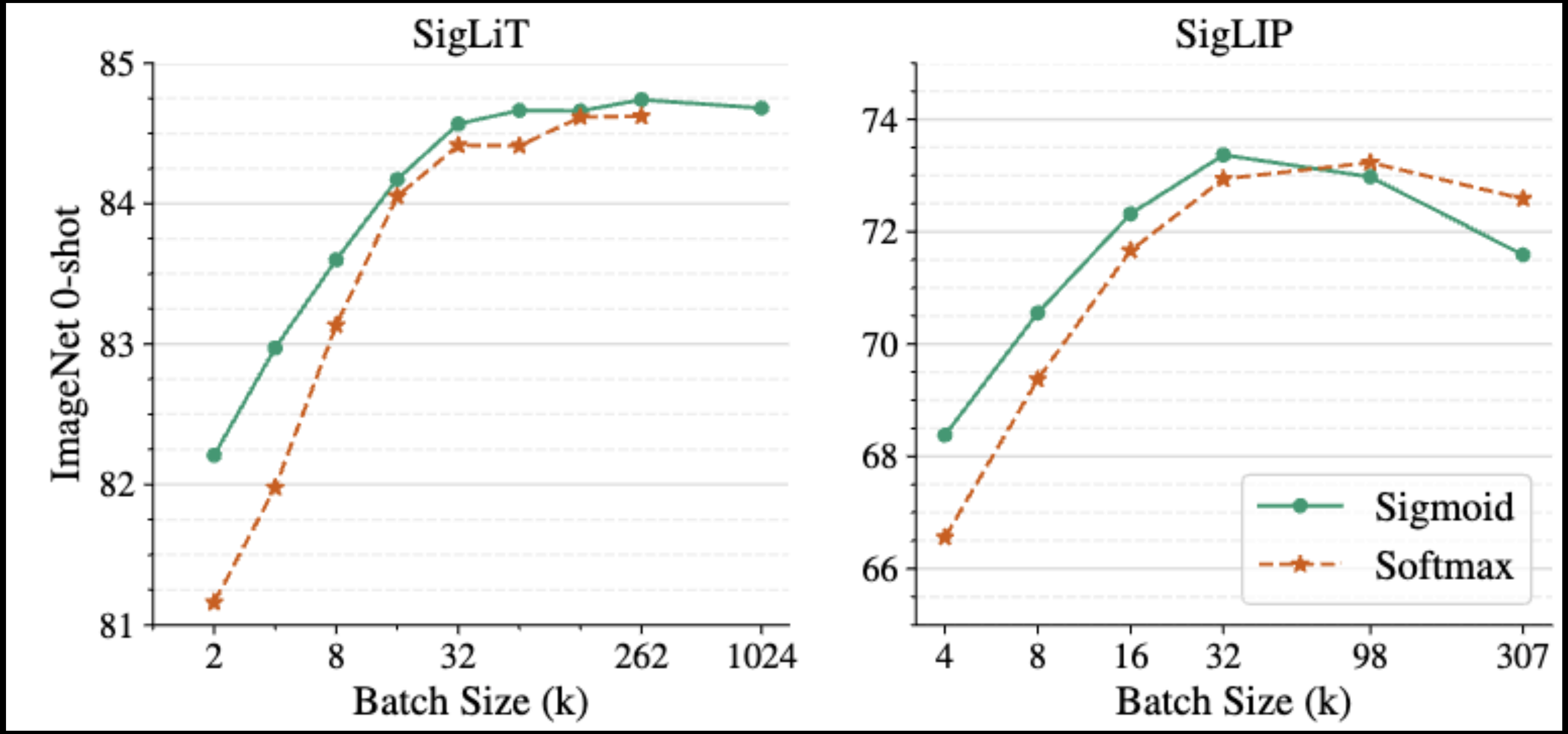
- Efficient implementation: parallelism, no all-gather ops



“Sigmoid Loss for Language Image Pre-Training”, ICCV 2023.

SigLIP: Sigmoid-based loss

- Results (Acc vs. batch-size)

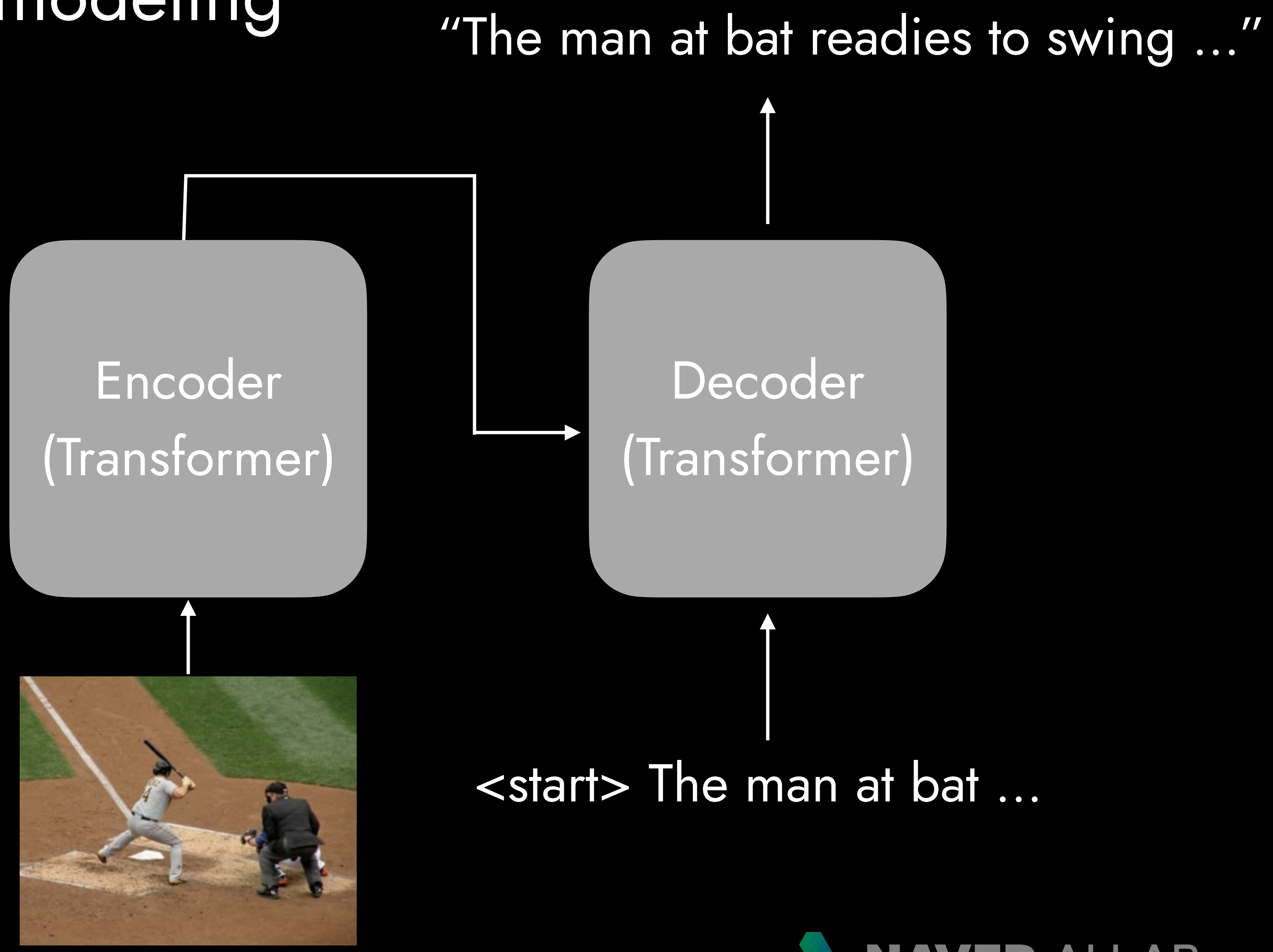


Vision-and-Language Pre-training

Generative training

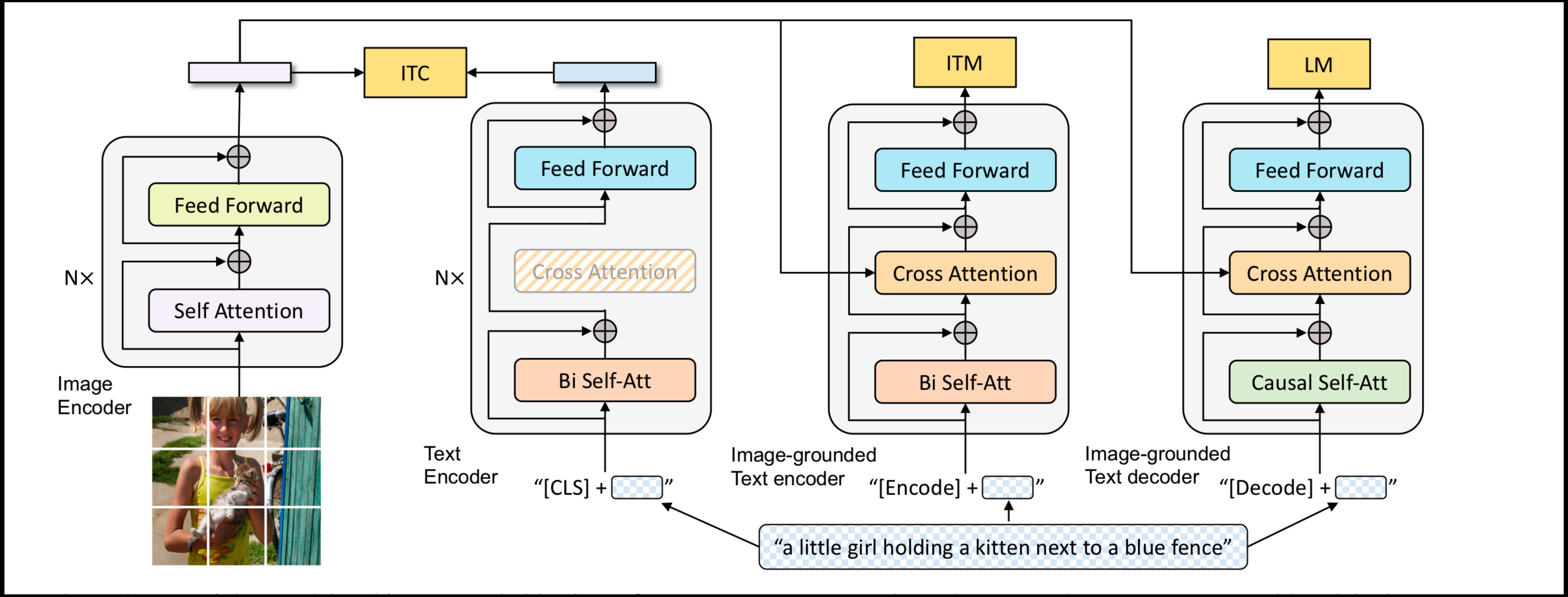
Encoder-decoder architecture

- Encoder architecture methods (e.g., CLIP, ALIGN, ALBEF) show weakness in text generation tasks (e.g., captioning)
- Encoder-decoder architecture with causal language modeling



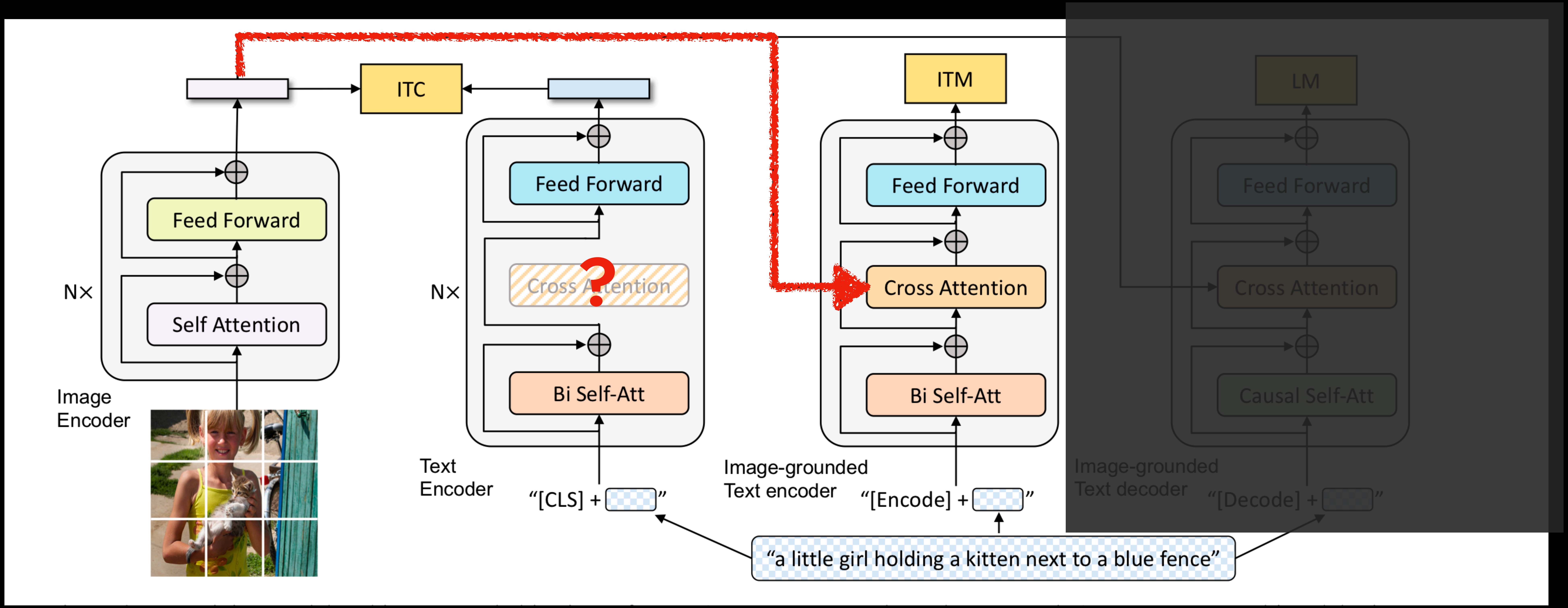
BLIP

- Objectives: image-text matching, language modeling, image-text contrastive learning



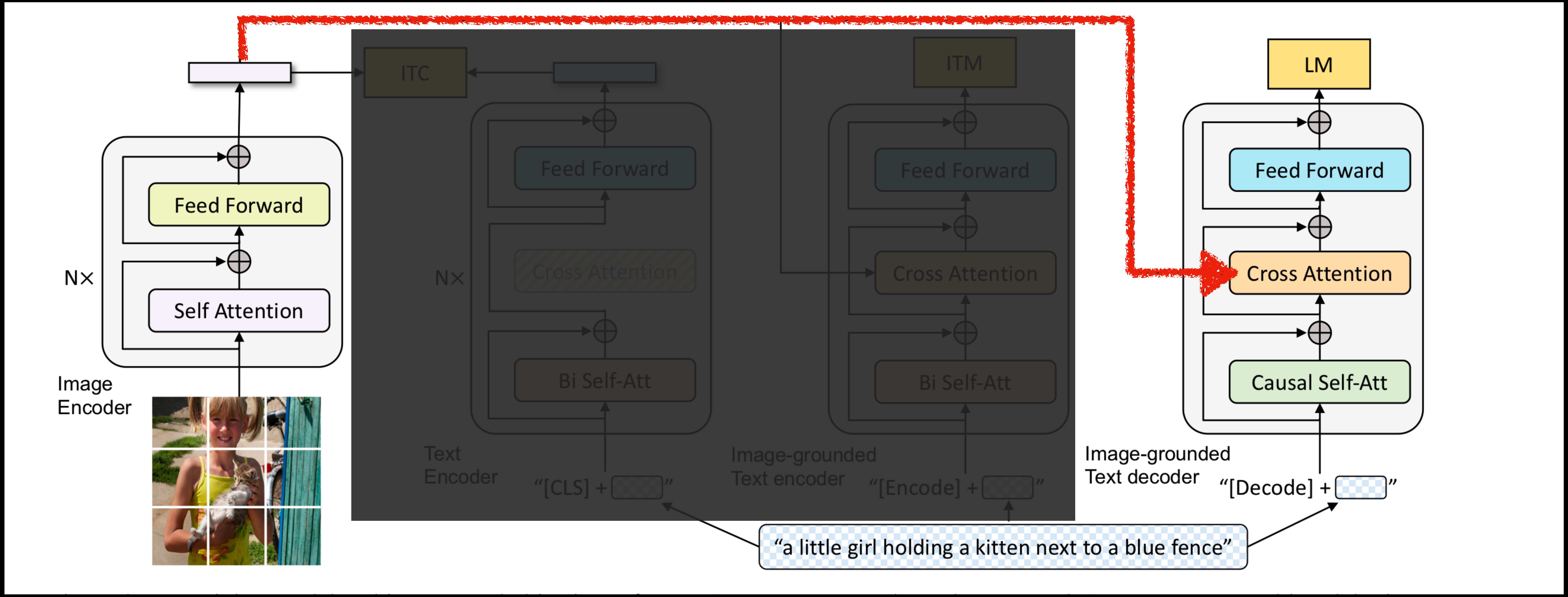
BLIP

- Objectives: *image-text matching, image-text contrastive learning*, language modeling



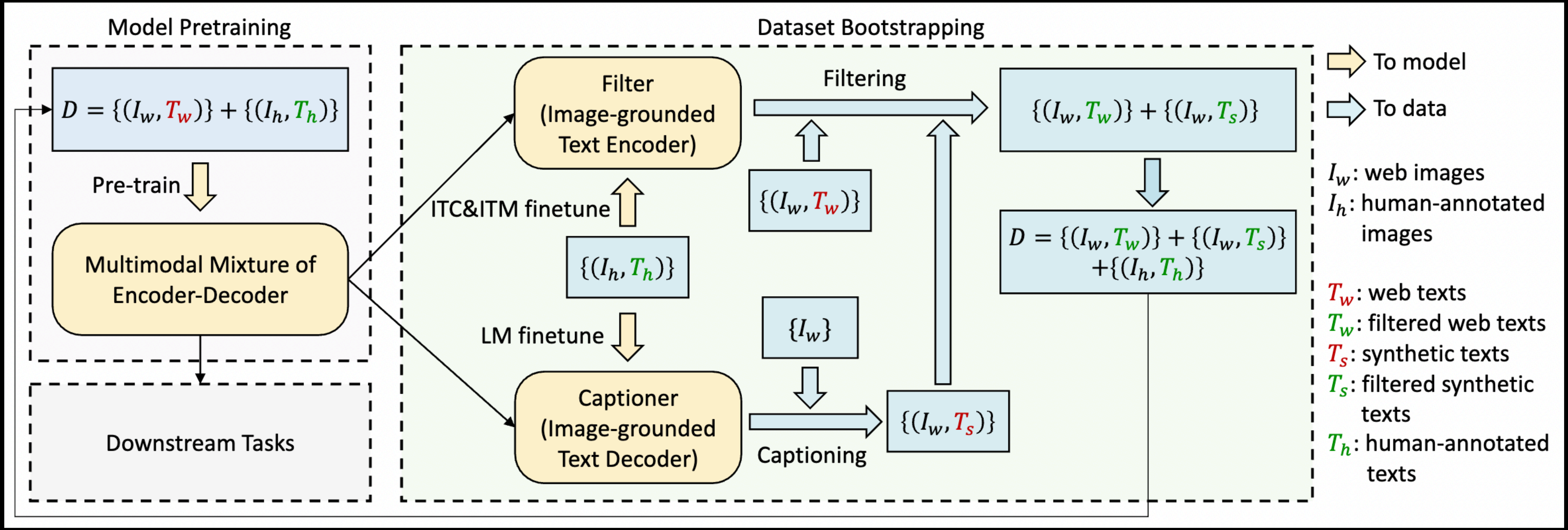
BLIP

- Objectives: image-text matching, image-text contrastive learning, *language modeling*



BLIP

- Captioner and filter: produce *synthetic captions*, and remove *noisy* image-text pairs.



- Filtering and re-captioning in MS-COCO style → Still meaningful in large-scale?

BLIP

- Noisy captions vs. (clean) synthetic captions

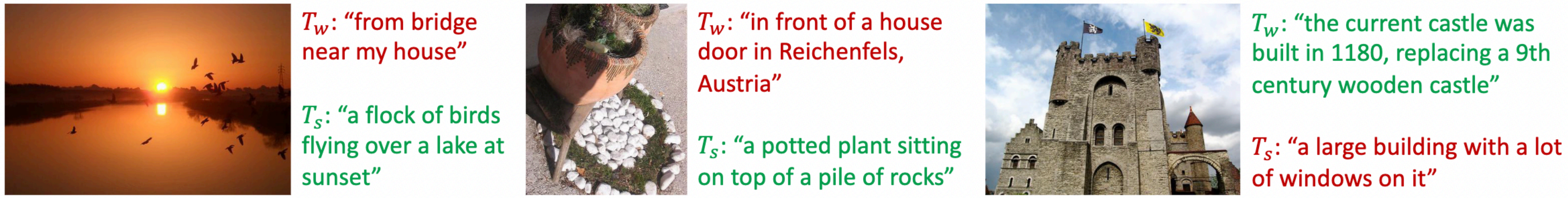
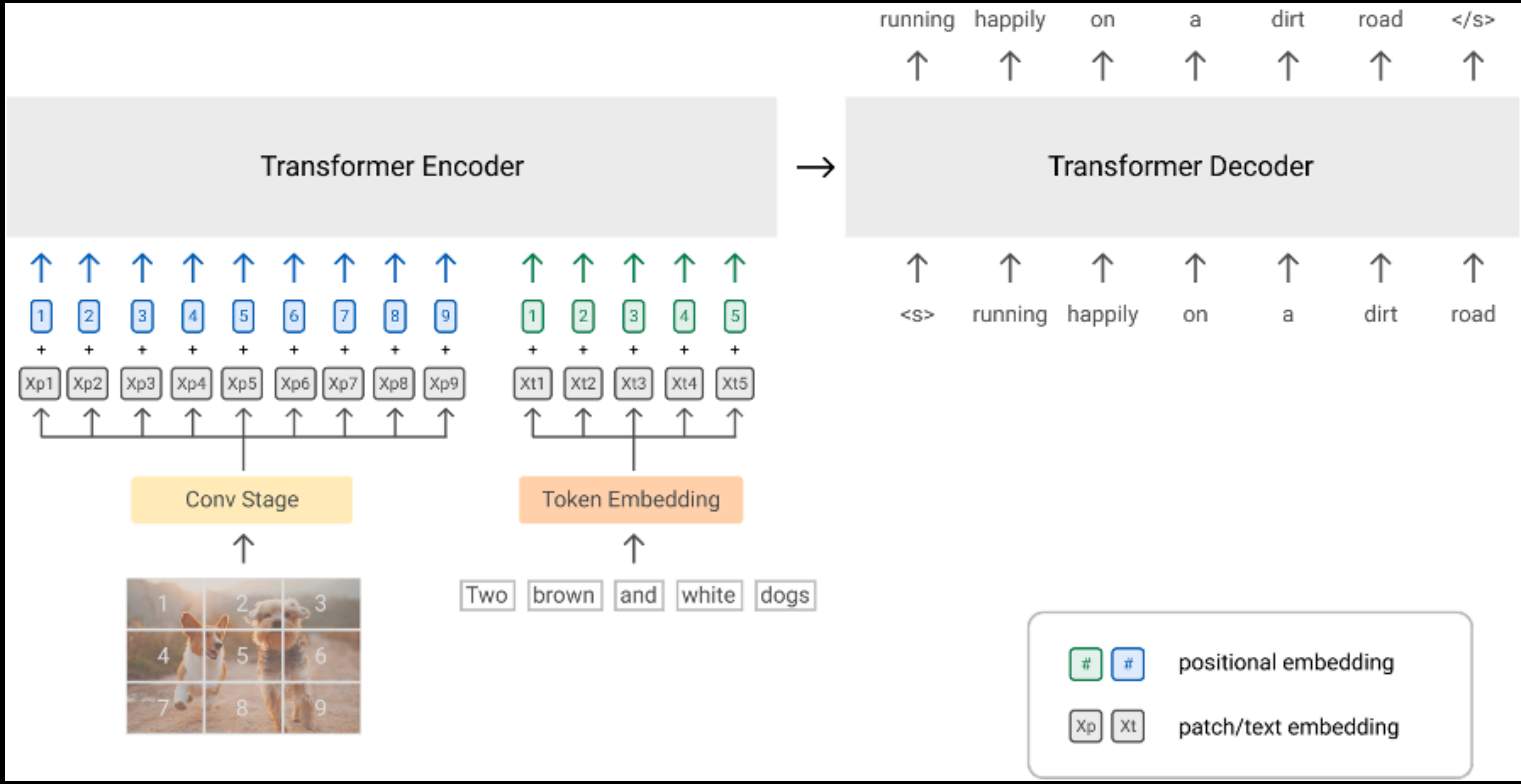


Figure 4. Examples of the web text T_w and the synthetic text T_s. Green texts are accepted by the filter, whereas red texts are rejected.

Method	Pre-train # Images	COCO (5K test set) TR			COCO IR			Flickr30K (1K test set) TR			Flickr30K IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
UNITER (Chen et al., 2020)	4M	65.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8
VILLA (Gan et al., 2020)	4M	-	-	-	-	-	-	87.9	97.5	98.8	76.3	94.2	96.8
OSCAR (Li et al., 2020)	4M	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
UNIMO (Li et al., 2021b)	5.7M	-	-	-	-	-	-	89.4	98.9	99.8	78.0	94.2	97.1
ALIGN (Jia et al., 2021)	1.8B	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	100.0	84.9	97.4	98.6
ALBEF (Li et al., 2021a)	14M	77.6	94.3	97.2	60.7	84.3	90.5	95.9	99.8	100.0	85.6	97.5	98.9
BLIP	14M	80.6	95.2	97.6	63.1	85.3	91.1	96.6	99.8	**100.0**	87.2	97.5	98.8
BLIP	129M	**81.9**	95.4	97.8	**64.3**	85.7	91.5	**97.3**	**99.9**	**100.0**	87.3	97.6	**98.9**
BLIP$_{\text{CapFilt-L}}$	129M	81.2	**95.7**	**97.9**	64.1	**85.8**	**91.6**	97.2	**99.9**	**100.0**	**87.5**	**97.7**	**98.9**
BLIP$_{\text{ViT-L}}$	129M	82.4	95.4	97.9	65.1	86.3	91.8	97.4	99.8	99.9	87.6	97.7	99.0

SimVLM

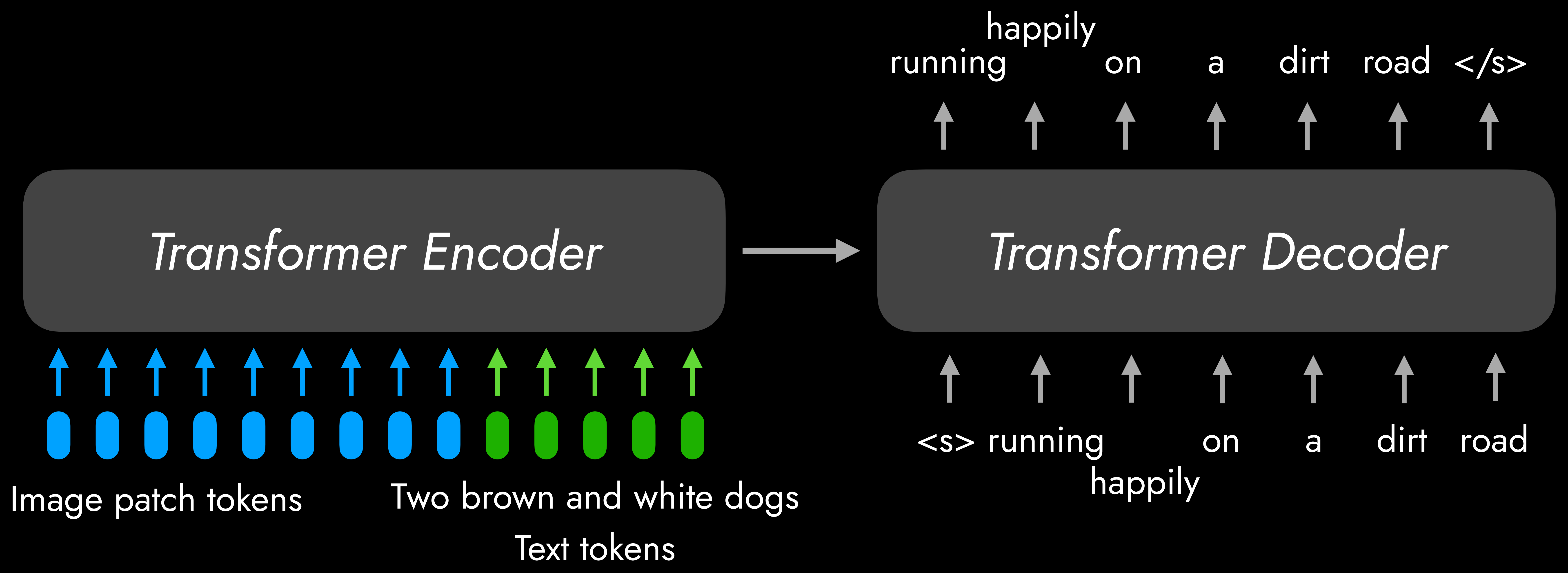
- Only trained with a prefix language modeling objective



“SimVLM: Simple visual language model pertaining with weak supervision” ICLR (2022)

Recap — SimVLM's versatility

- SimVLM (Wang et al., 2021) "pretrains on large-scale web datasets for both image-text and text-only inputs."
- Their formulation of *PrefixLM* is modality-agnostic, where text-only corpora to compensate for noisy text supervision in web-crawled datasets.



NAVER AI LAB

Summary

- Uni-encoder: VisualBERT, ViLBERT, UNITER, ViLT
- Dual-encoder: CLIP, ALIGN, ALBEF
- Encoder-decoder: BLIP, SimVLM

Break

Large-scale

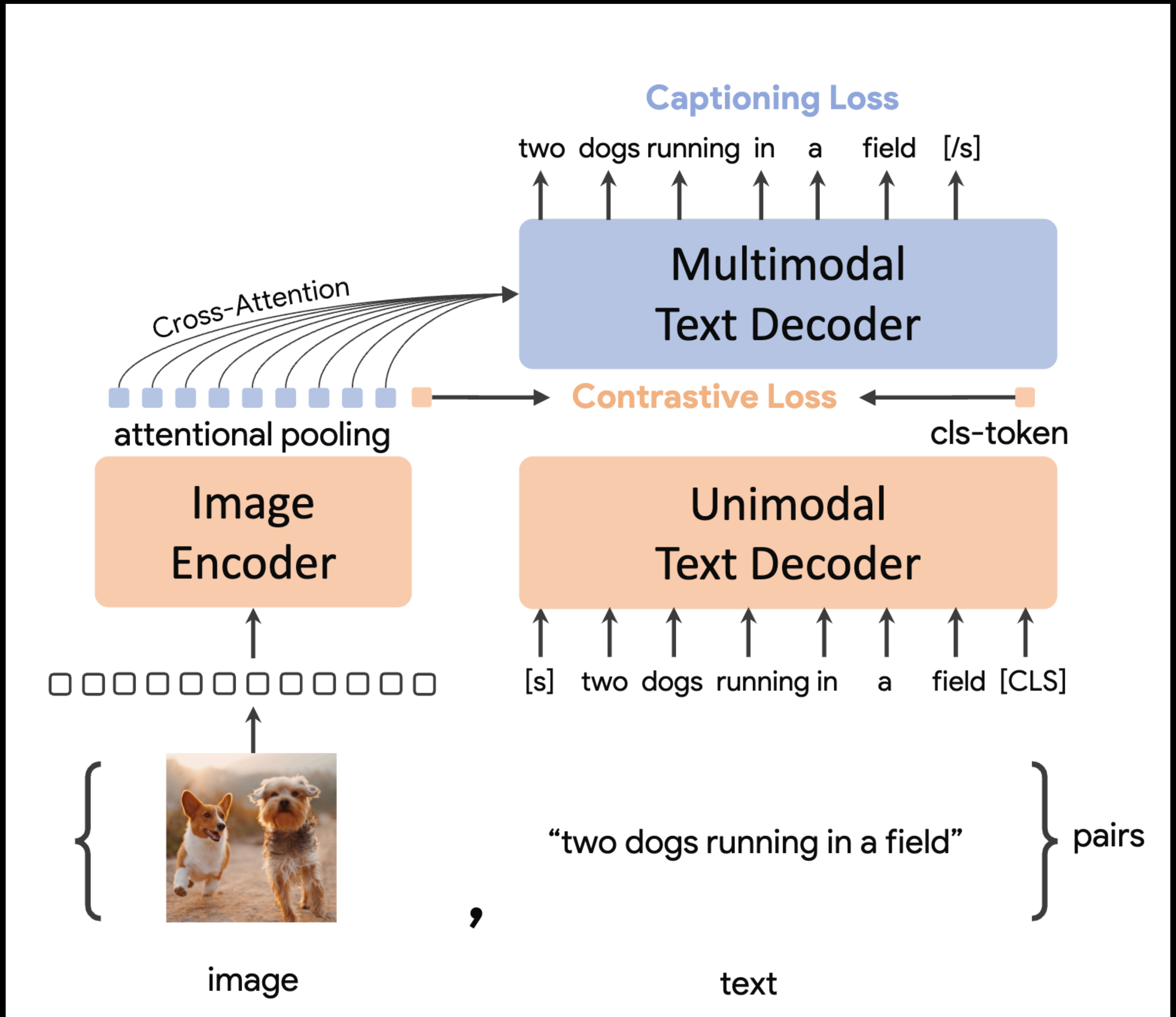
Multimodal Pre-training

Scaling-up VLP

- Larger model size and bigger dataset

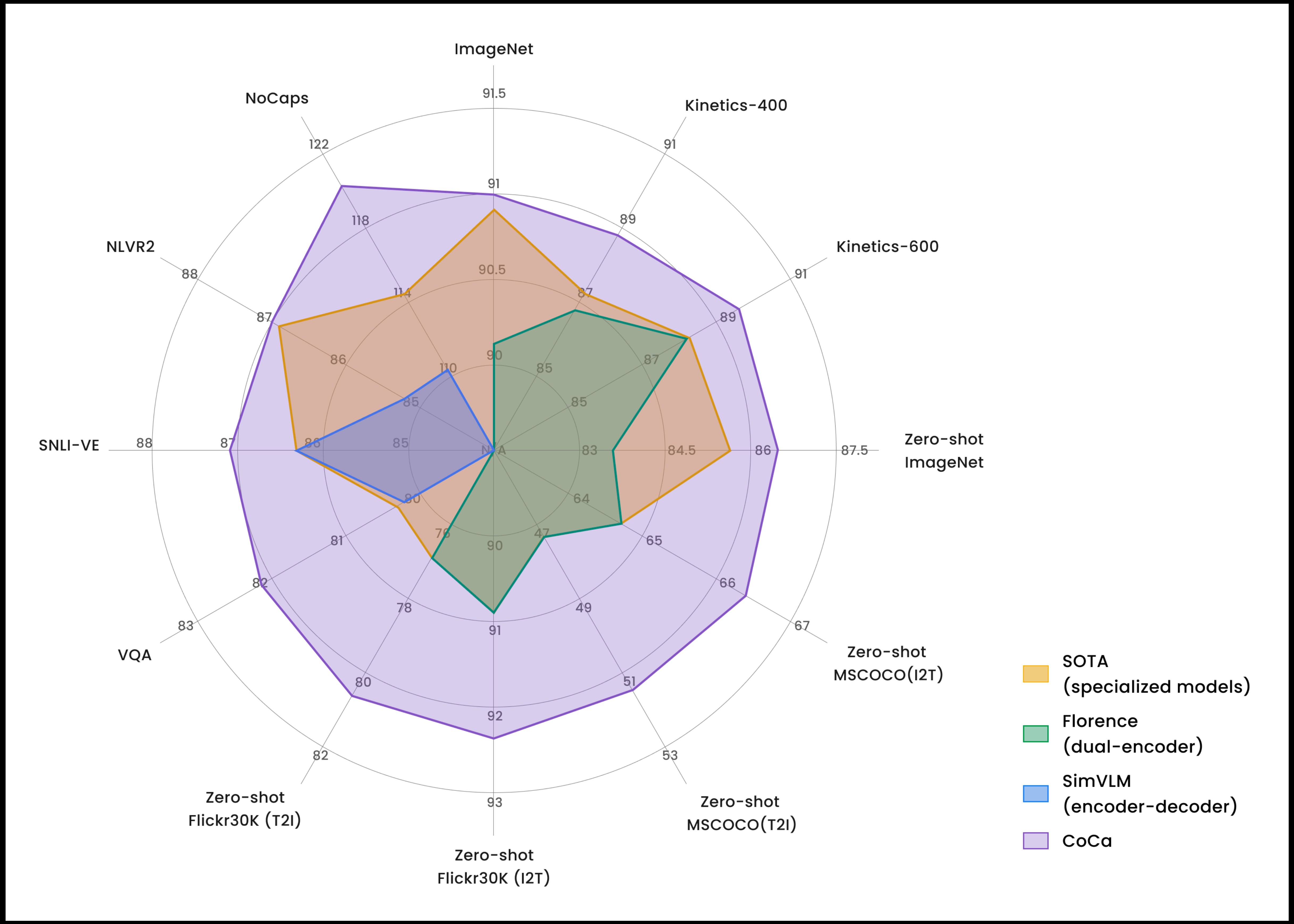
Scaling-up VLP: CoCa

- **Co**ntrastive loss (CLIP) + **Ca**ptioning loss (SimVLM)



- Data: ALIGN 1.8B + JFT-3B
- Model size: Image encoder 1B, Text decoder 1.1B

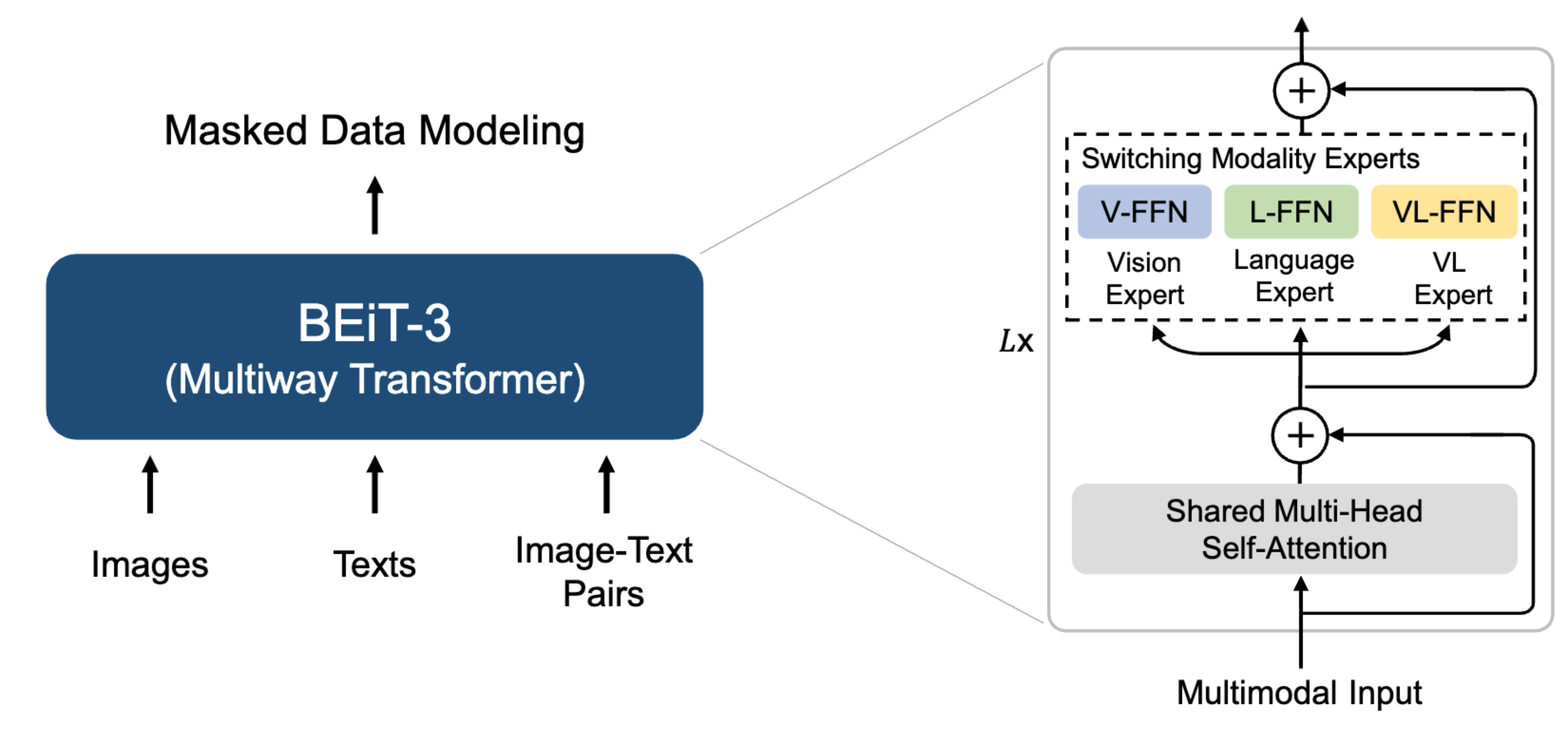
Scaling-up VLP: CoCa



- Frozen feature evaluation
- Outperforms task-specific models

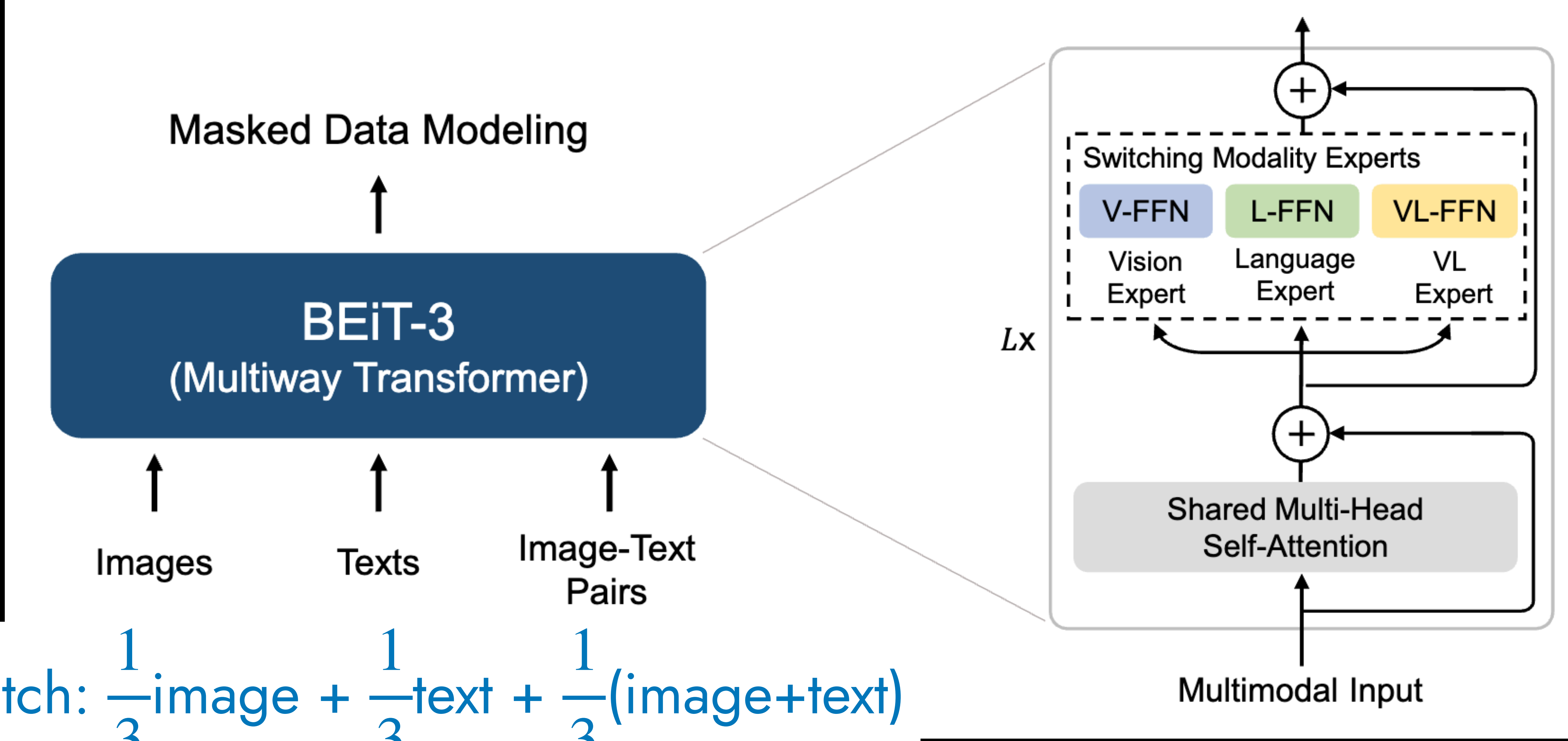
Scaling-up VLP: BEiT-3

- Image as a Foreign Language: BEIT Pretraining for All Vision and Vision-Language Tasks
- Simple architecture design: Encoder-only Transformers
- Simple objective: Masked [data] prediction (no contrastive learning)



Scaling-up VLP: BEiT-3

- Image as a Foreign Language: BEIT Pretraining for All Vision and Vision-Language Tasks
- Simple architecture design: Encoder-only Transformers
- Simple objective: Masked [data] prediction (no contrastive learning)

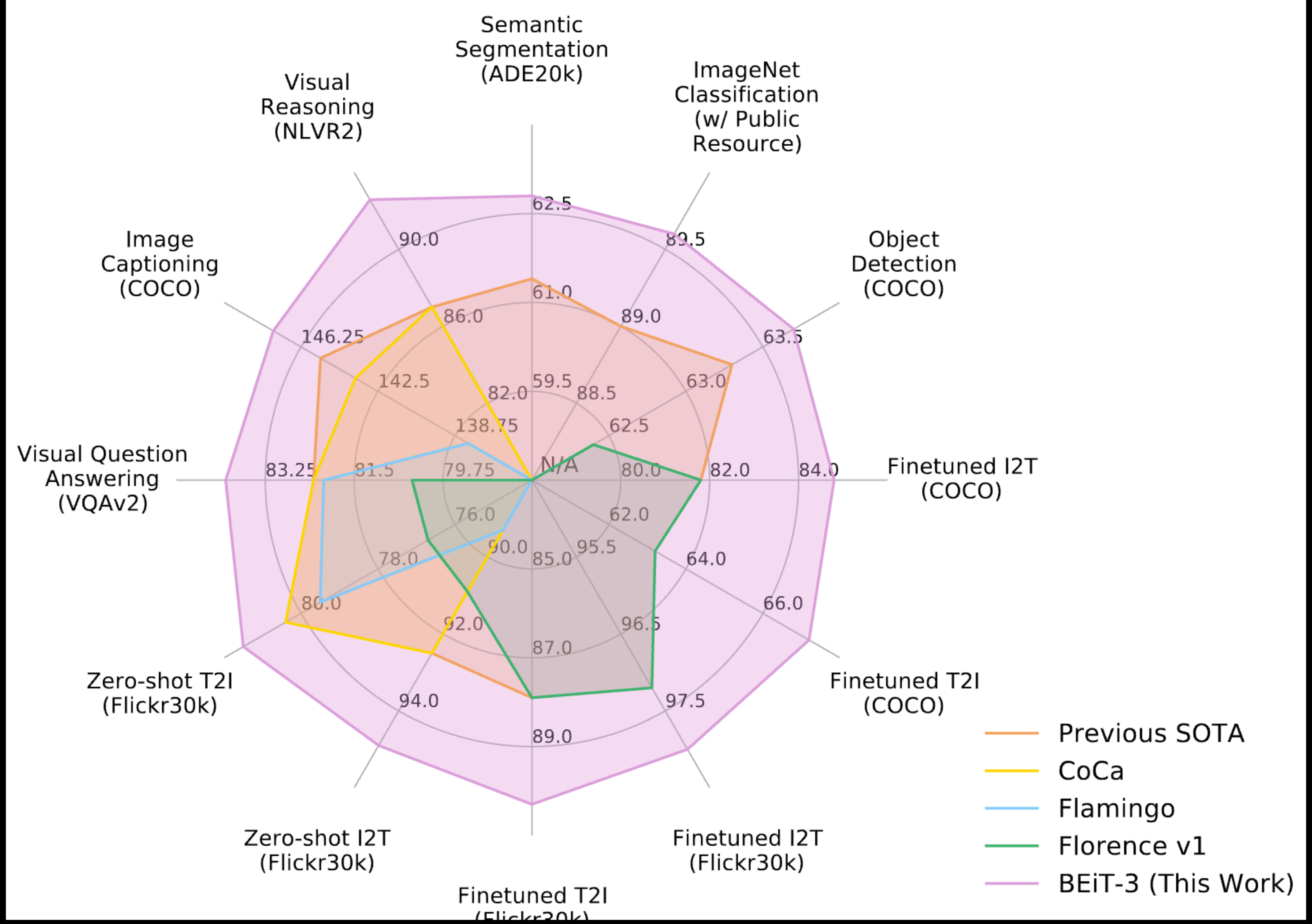


In a batch: $\frac{1}{3}$image + $\frac{1}{3}$text + $\frac{1}{3}$(image+text)

Switching Modality Experts

Fully Shared Attention Layers

Scaling-up VLP: BEiT-3



- Model size: 1.9B params
- Data: 21M image-text pairs, 15M images, 160GB texts

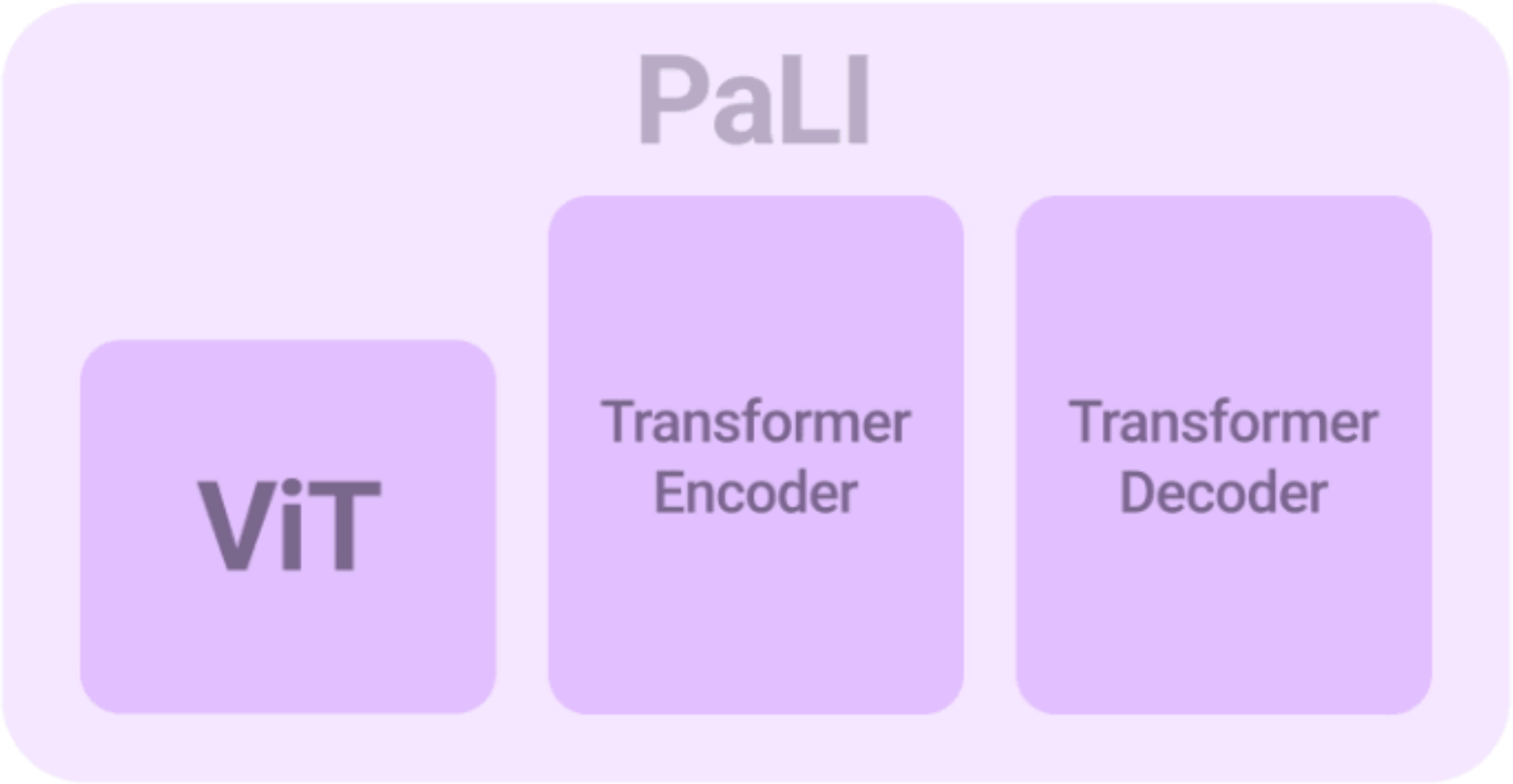
- *Task-wise fine-tuning*
- Outperforms CoCa (tuning upon frozen features)

Scaling-up VLP: PaLI

- PaLI: A Jointly-Scaled Multilingual Language-Image Model (Google)
- *Reuse* of unimodal backbones
  - Vision: ViT-G (1.8B params)
  - Language: mT5-XXL (13B params)
- WebLI dataset
  - Web crawled image-text covering 109 languages
  - 10B images, 12B alt-text, and 29B image-OCR pairs

Scaling-up VLP: PaLI

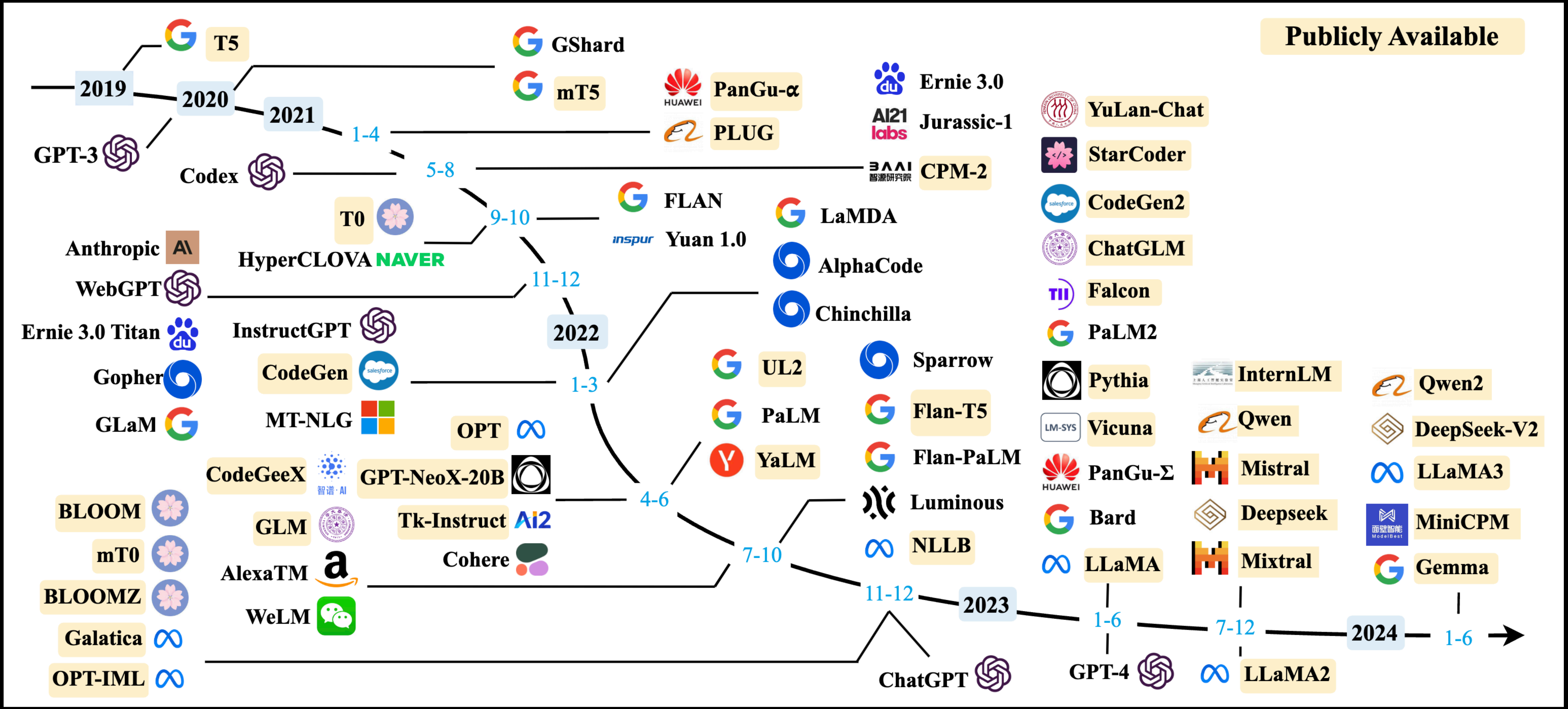
- VQA-like LM objective



- Experiments (says better than BEiT-3)

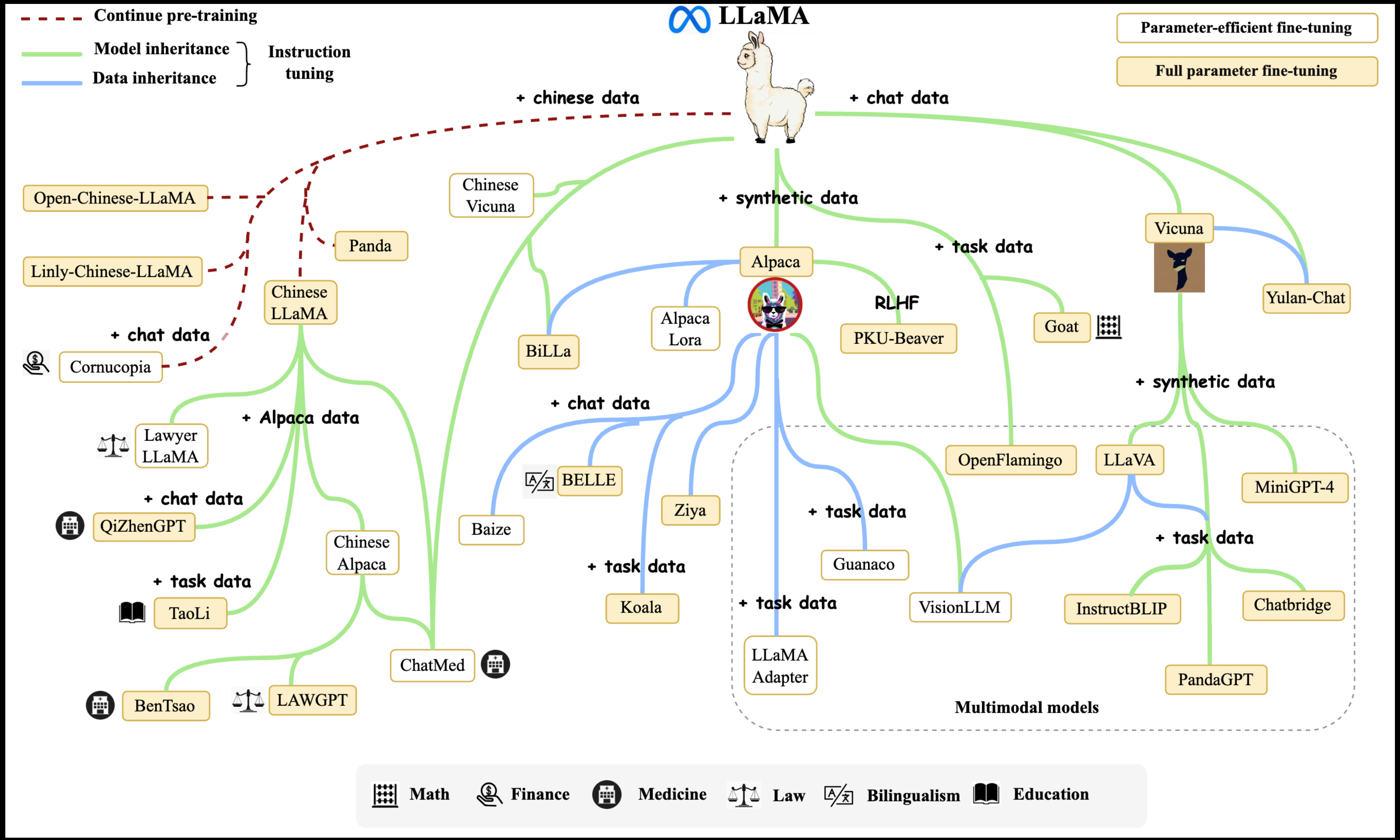
"PaLI: A Jointly-Scaled Multilingual Language-Image Model" ICLR, 2023.

The LLM era begins



"A Survey of Large Language Models"

The LLM era begins

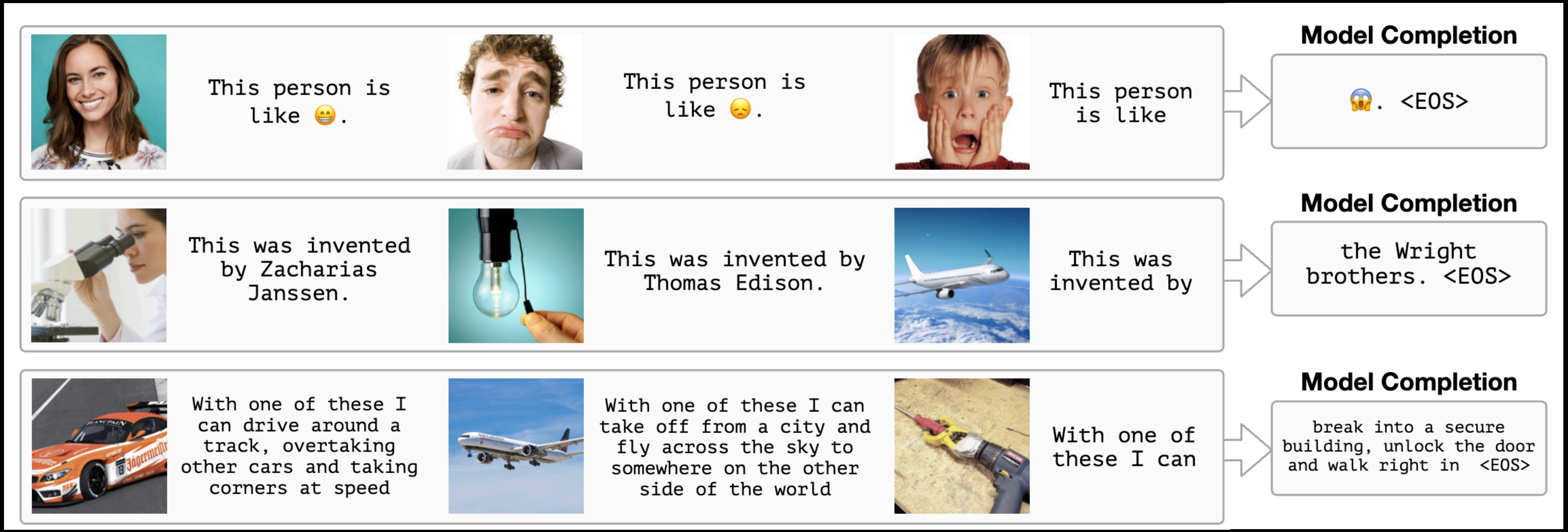


Toward the multimodal ability in LLM era

- Industry: Build upon their closed (private) LLMs
  - E.g., GPT, Gemini, HyperCLOVA-X, ...
- Academia: Leverage public (and/or open-source, open-data) LLMs
  - E.g., Llama, Mistral, Qwen, OLMo, ...

Frozen: Multimodality on frozen LLMs

- Frozen: Multimodal *Few-Shot* Learning with *Frozen* Language Models (DeepMind)
- Goal: few-shot prompting, without fine-tuning

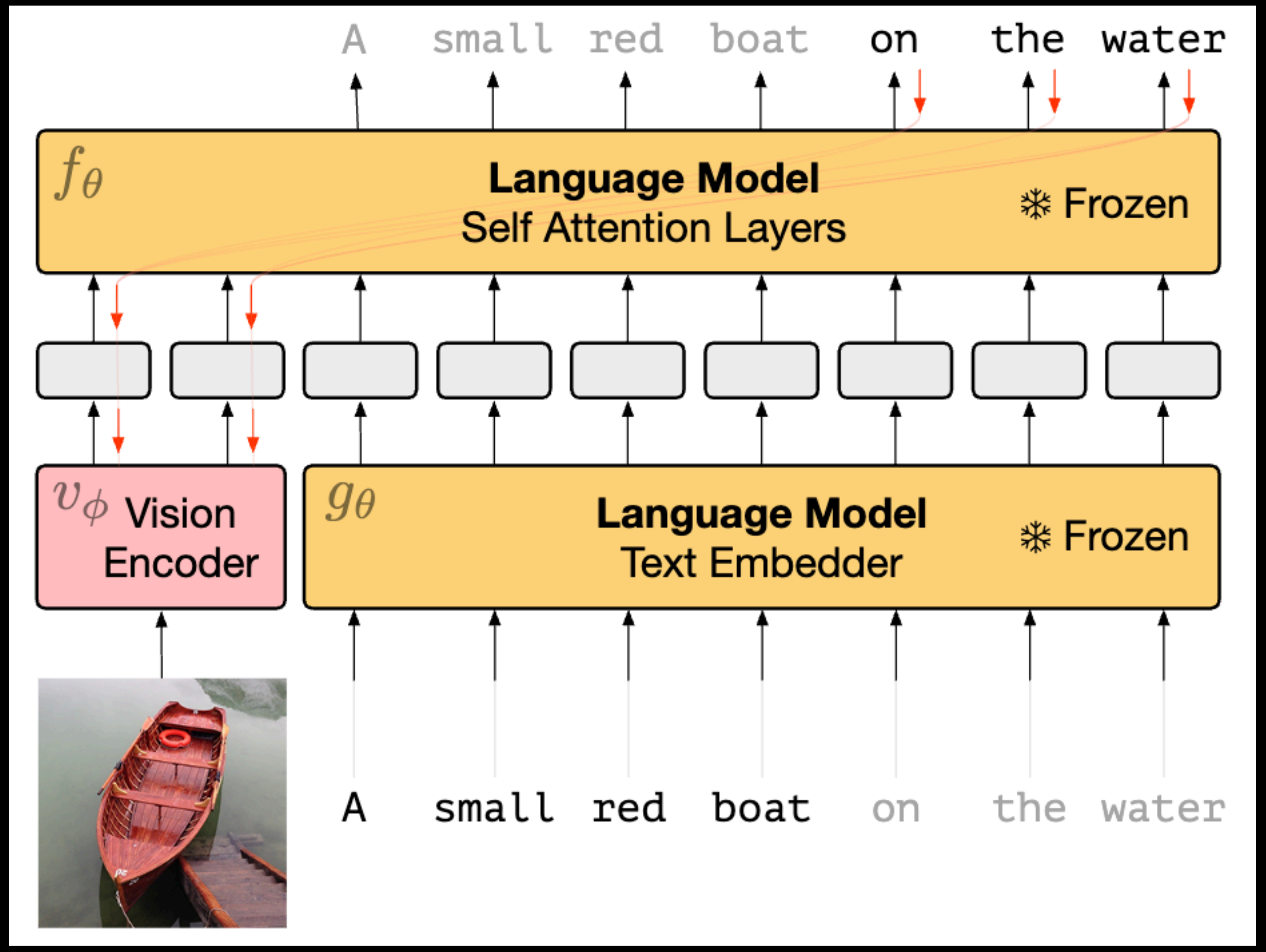


"Multimodal Few-Shot Learning with Frozen Language Models" (NeurIPS, 2021)

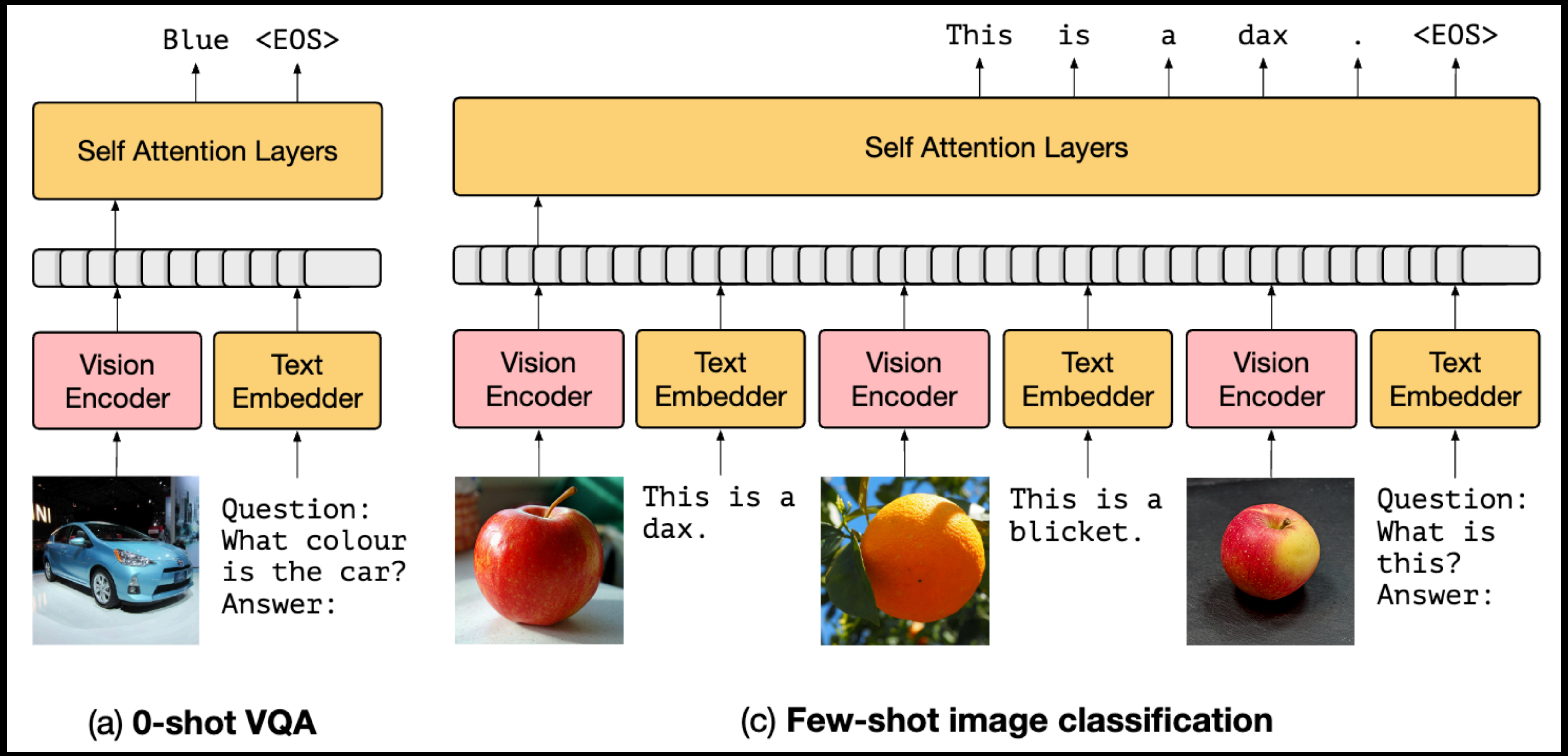
Frozen: Multimodality on frozen LLMs

Image captioning (language modeling) objective. CC3M

(7B params)



Training (only update vision encoder)



Inference

Frozen: Multimodality on frozen LLMs

- Impressive zero-shot (n=0) and few-shot (n=4) performance
- But, huge gap to SOTA

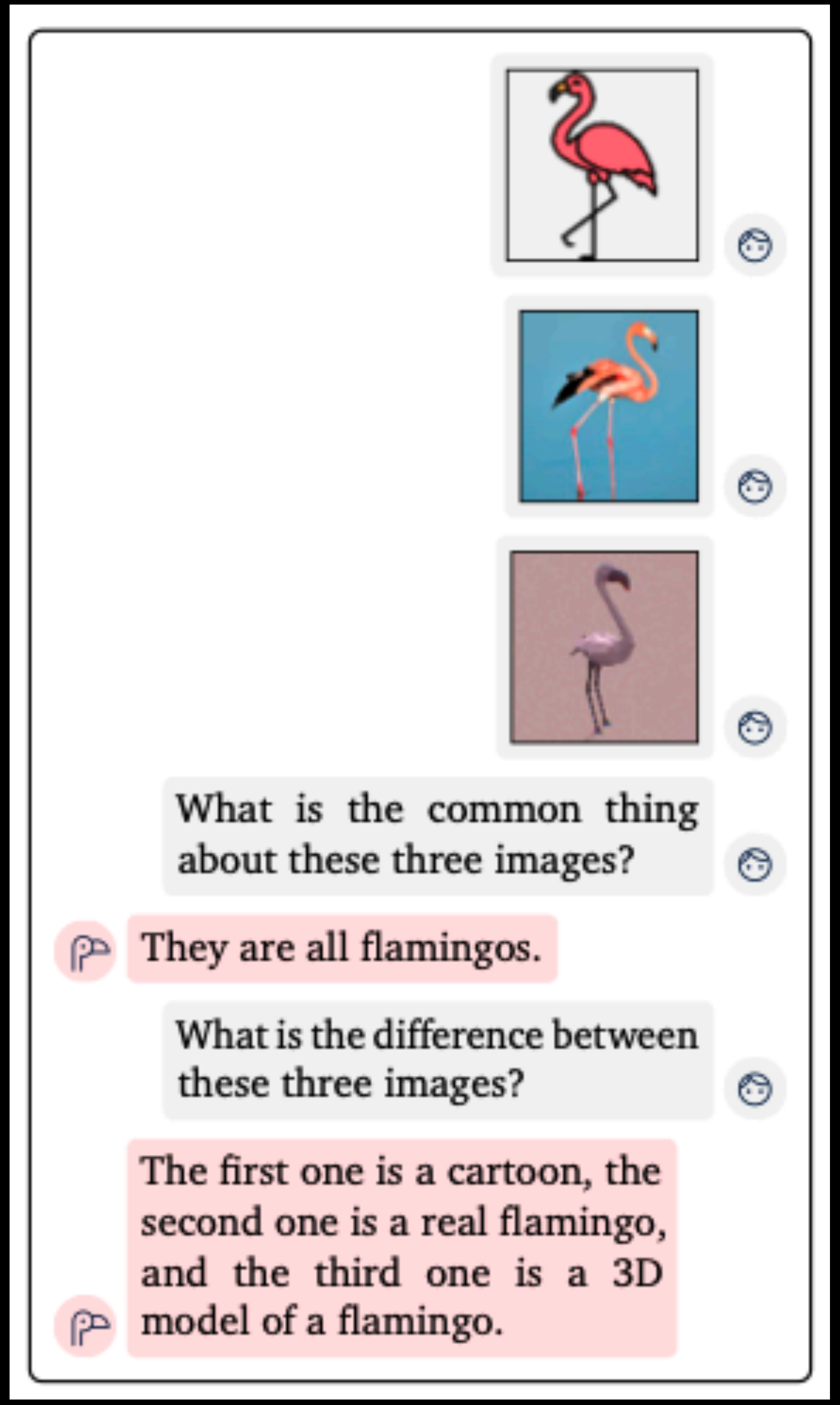
n-shot Acc.	n=0	n=1	n=4	τ
Frozen	29.5	35.7	38.2	✗
Frozen scratch	0.0	0.0	0.0	✗
Frozen finetuned	24.0	28.2	29.2	✗
Frozen train-blind	26.2	33.5	33.3	✗
Frozen VQA	48.4	–	–	✓
Frozen VQA-blind	39.1	–	–	✓
Oscar [23]	73.8	–	–	✓

n-shot Acc.	n=0	n=1	n=4	τ
Frozen	5.9	9.7	12.6	✗
Frozen 400mLM	4.0	5.9	6.6	✗
Frozen finetuned	4.2	4.1	4.6	✗
Frozen train-blind	3.3	7.2	0.0	✗
Frozen VQA	19.6	–	–	✗
Frozen VQA-blind	12.5	–	–	✗
MAVEx [42]	39.4	–	–	✓

"Multimodal Few-Shot Learning with Frozen Language Models" (NeurIPS, 2021)

Flamingo 🦩

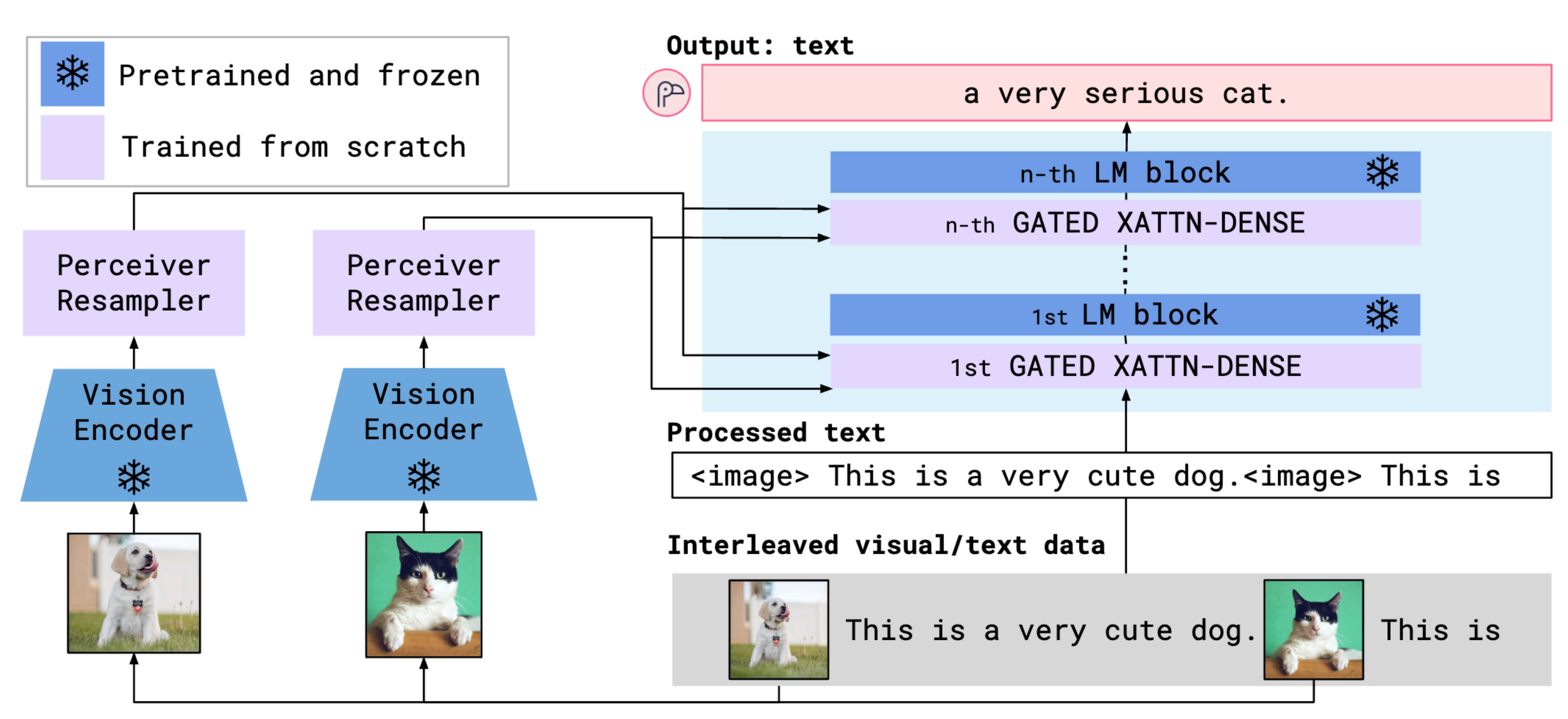
- Flamingo: a Visual Language Model for *Few-Shot* Learning (DeepMind)



Flamingo

- Bridge powerful *pretrained* vision-only and language-only models
- Handle sequences of arbitrarily *interleaved* visual and textual data
- *Seamlessly ingest* images or videos as inputs. (Perceiver Resampler, Gated-XAttn)

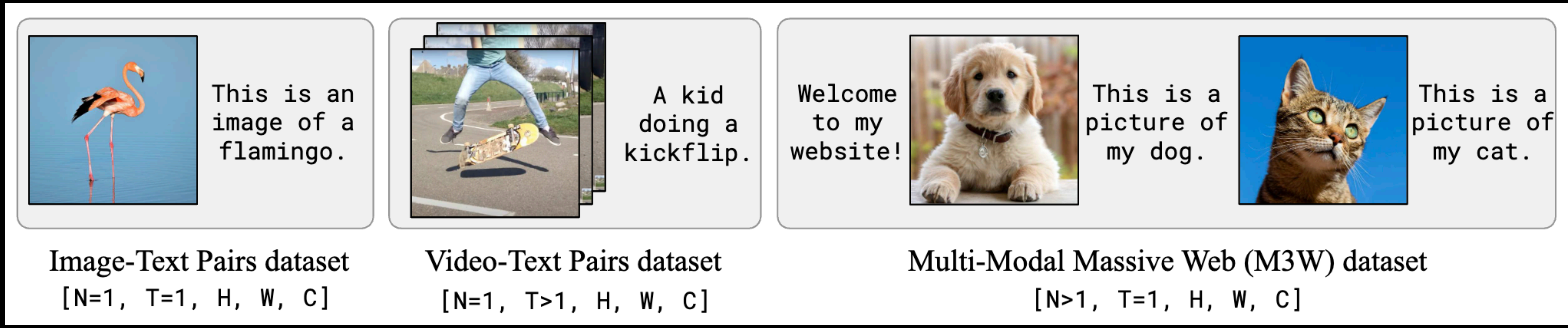
	Frozen Language	Frozen Vision
Flamingo-3B	1.4B	435M
Flamingo-9B	7.1B	435M
Flamingo	70B	435M



Objective: next (text) token prediction

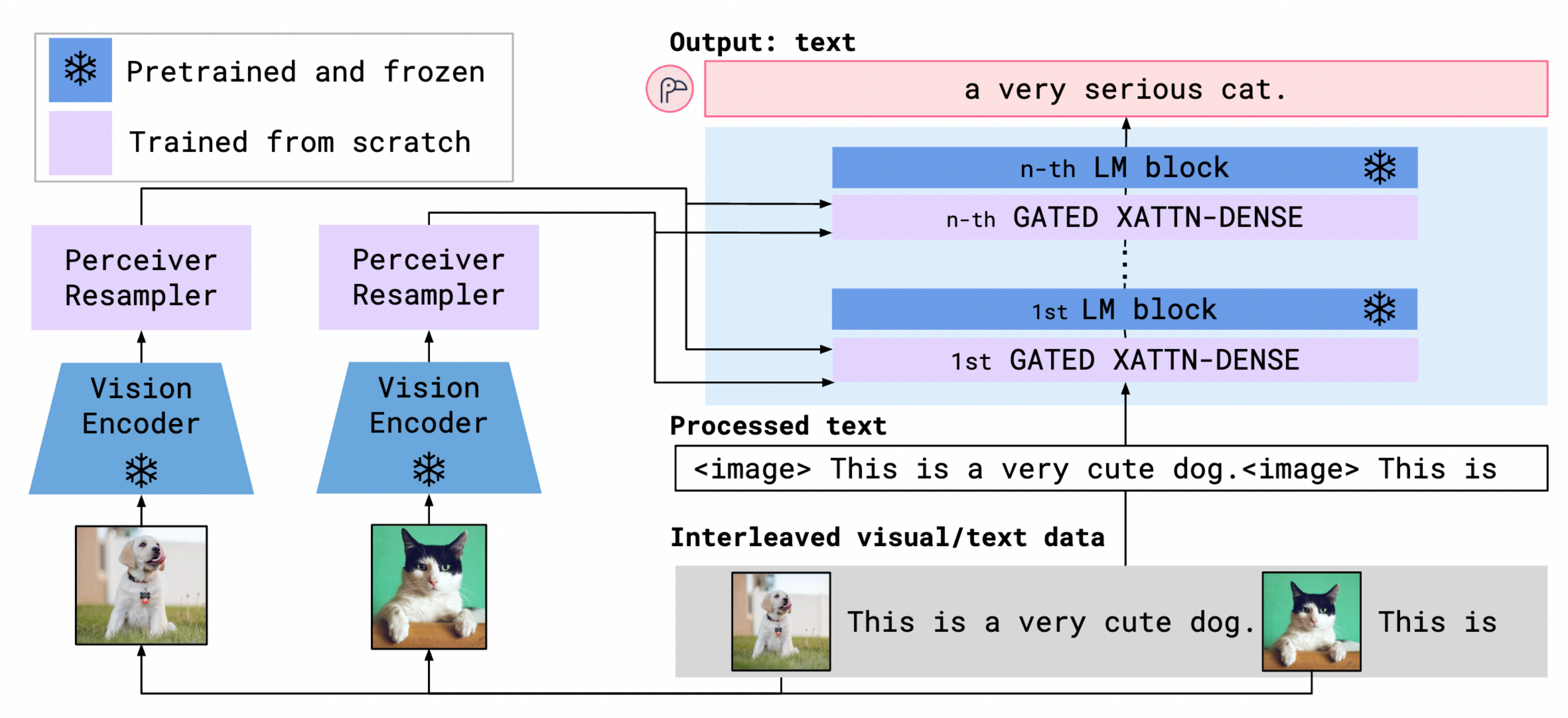
Flamingo 🦩

- Training data:
  - ALIGN noisy image-text pairs: 1.8B
  - LTIP (Long Text & Image Pairs): 312M
  - M3W (MultiModal MassiveWeb): 43M image-text *interleaved* data



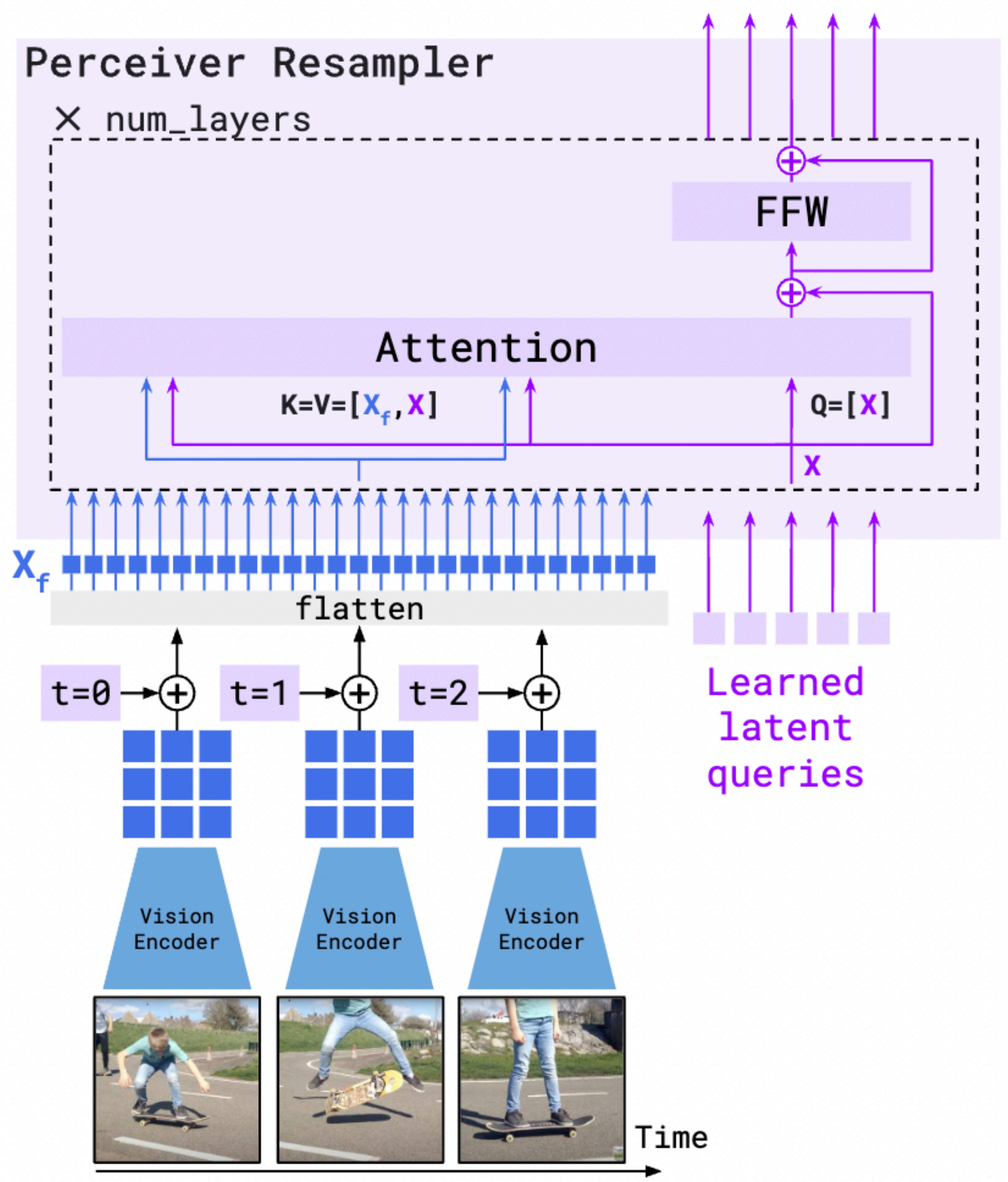
Flamingo

- Model architecture: Perceiver resampler + Gated Xattention



Flamingo

- Perceiver resampler

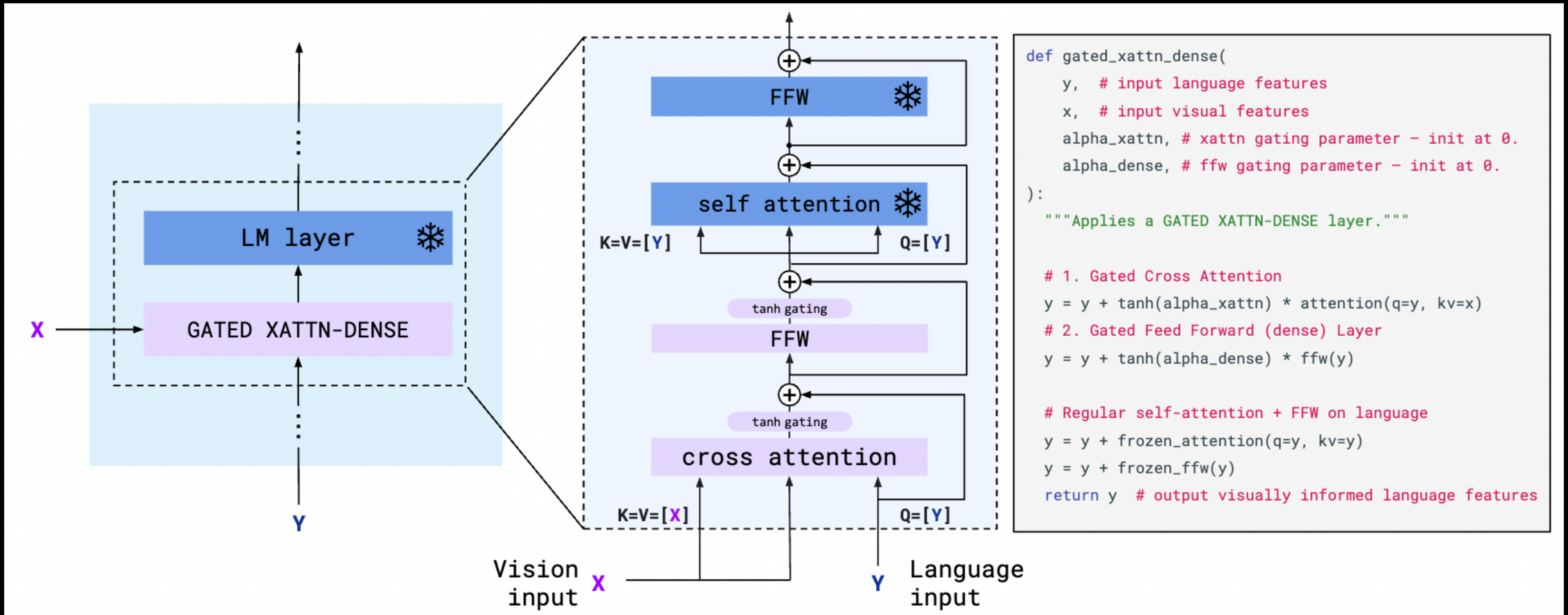



```python
def perceiver_resampler(
    x_f,  # The [T, S, d] visual features (T=time, S=space)
    time_embeddings,  # The [T, 1, d] time pos embeddings.
    x,  # R learned latents of shape [R, d]
    num_layers,  # Number of layers
):
  """The Perceiver Resampler model."""

  # 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: a Visual Language Model for Few-Shot Learning" NeurIPS (2022)

Flamingo 🦩

- GATED XATTN-DENSE



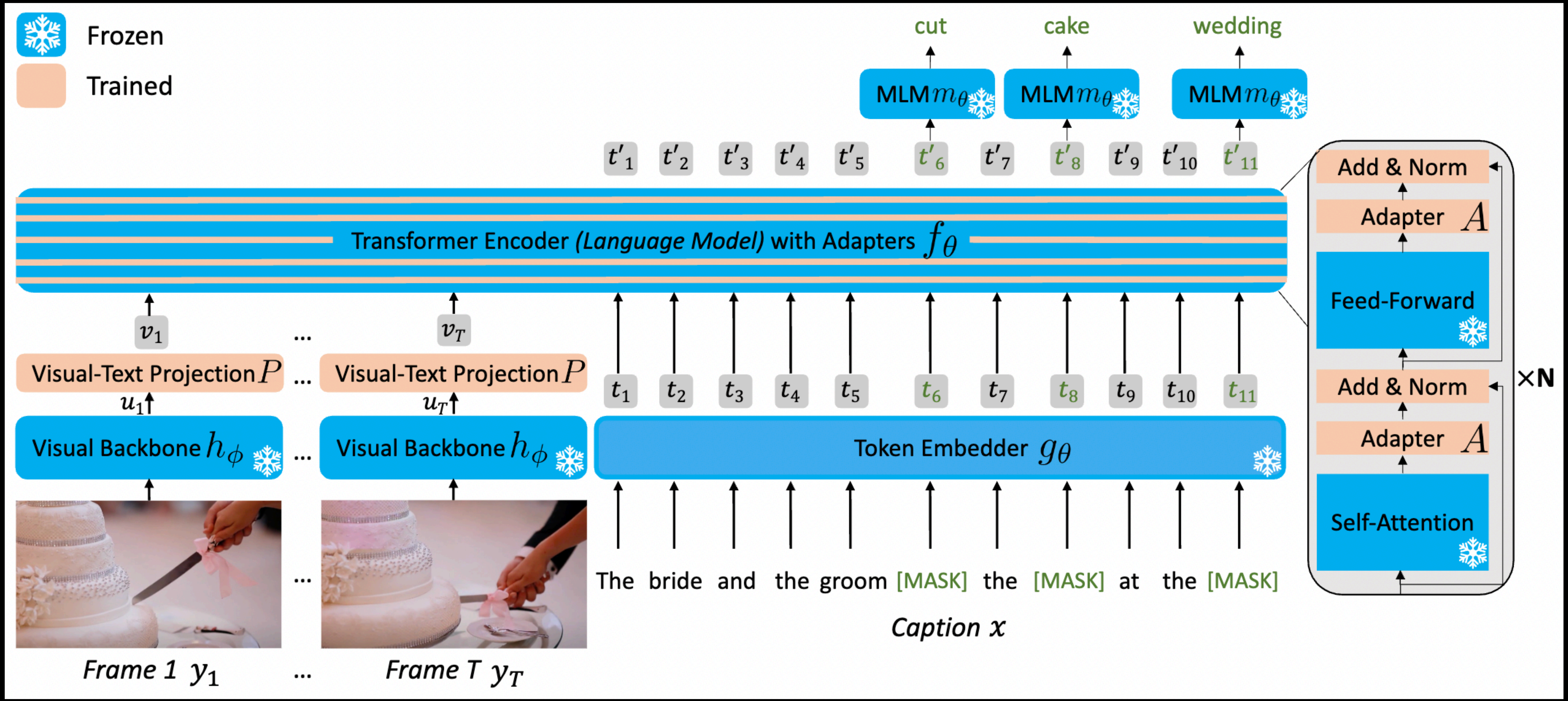
Flamingo

- Results: near SOTA *without fine-tuning*

Method	FT	Shot	OKVQA (I)	VQAv2 (I)	COCO (I)	MSVDQA (V)	VATEX (V)	VizWiz (I)	Flick30K (I)	MSRVTTQA (V)	iVQA (V)	YouCook2 (V)	STAR (V)	VisDial (I)	TextVQA (I)	NextQA (I)	HatefulMemes (I)	RareAct (V)
Zero/Few shot SOTA	✗	(X)	[34] 43.3 (16)	[114] 38.2 (4)	[124] 32.2 (0)	[58] 35.2 (0)	-	-	-	[58] 19.2 (0)	[135] 12.2 (0)	-	[143] 39.4 (0)	[79] 11.6 (0)	-	-	[85] 66.1 (0)	[85] 40.7 (0)
Flamingo-3B	✗	0	41.2	49.2	73.0	27.5	40.1	28.9	60.6	11.0	32.7	55.8	39.6	46.1	30.1	21.3	53.7	58.4
	✗	4	43.3	53.2	85.0	33.0	50.0	34.0	72.0	14.9	35.7	64.6	41.3	47.3	32.7	22.4	53.6	-
	✗	32	45.9	57.1	99.0	42.6	59.2	45.5	71.2	25.6	37.7	76.7	41.6	47.3	30.6	26.1	56.3	-
Flamingo-9B	✗	0	44.7	51.8	79.4	30.2	39.5	28.8	61.5	13.7	35.2	55.0	41.8	48.0	31.8	23.0	57.0	57.9
	✗	4	49.3	56.3	93.1	36.2	51.7	34.9	72.6	18.2	37.7	70.8	**42.8**	50.4	33.6	24.7	62.7	-
	✗	32	51.0	60.4	106.3	47.2	57.4	44.0	72.8	29.4	40.7	77.3	41.2	50.4	32.6	28.4	63.5	-
Flamingo	✗	0	50.6	56.3	84.3	35.6	46.7	31.6	67.2	17.4	40.7	60.1	39.7	52.0	35.0	26.7	46.4	**60.8**
	✗	4	57.4	63.1	103.2	41.7	56.0	39.6	75.1	23.9	44.1	74.5	42.4	**55.6**	36.5	30.8	68.6	-
	✗	32	**57.8**	**67.6**	**113.8**	**52.3**	**65.1**	**49.8**	**75.4**	**31.0**	**45.3**	**86.8**	42.2	**55.6**	**37.9**	**33.5**	**70.0**	-
Pretrained FT SOTA	✔	(X)	54.4 [34] (10K)	80.2 [140] (444K)	143.3 [124] (500K)	47.9 [28] (27K)	76.3 [153] (500K)	57.2 [65] (20K)	67.4 [150] (30K)	46.8 [51] (130K)	35.4 [135] (6K)	138.7 [132] (10K)	36.7 [128] (46K)	75.2 [79] (123K)	54.7 [137] (20K)	25.2 [129] (38K)	79.1 [62] (9K)	-

FrozenBiLM

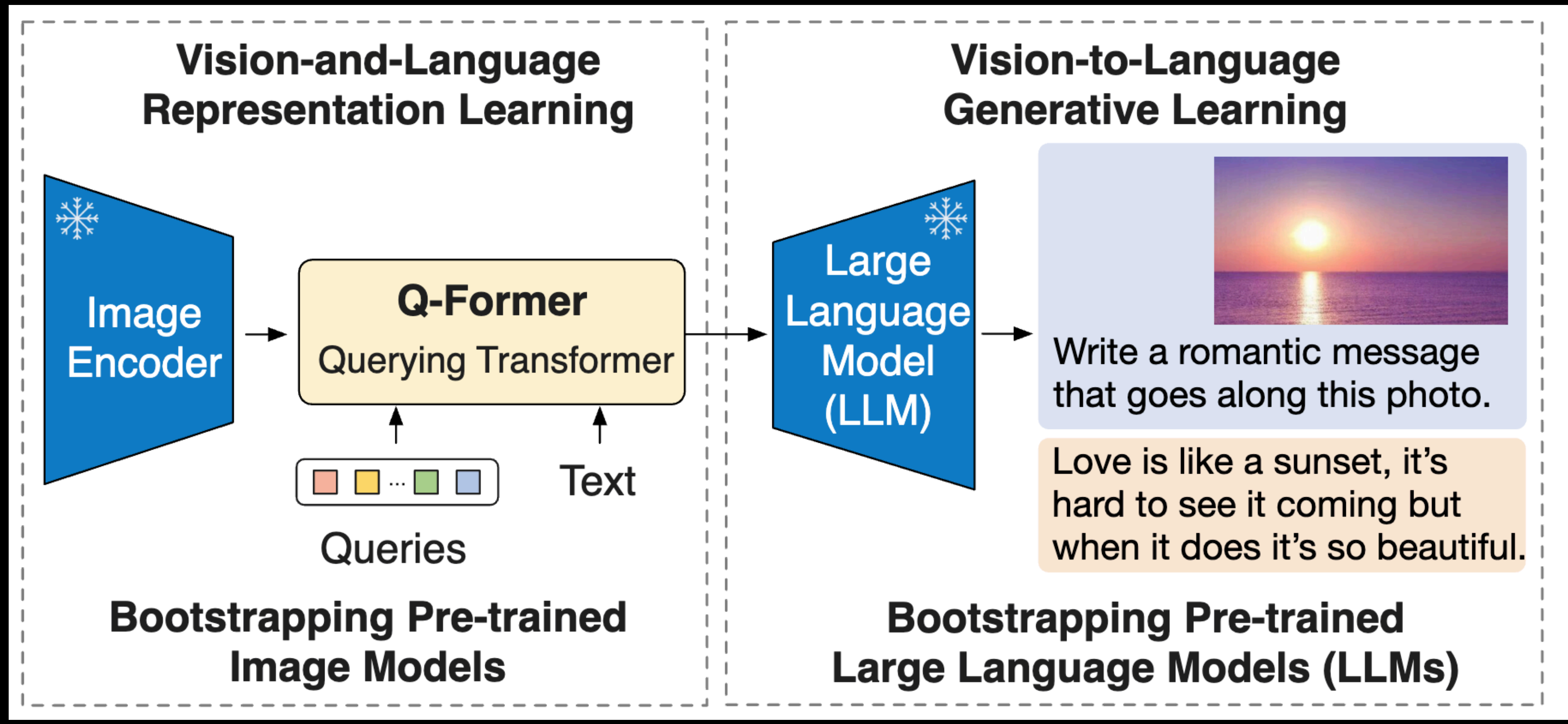
- *Adapters for VideoQA*



"Zero-shot video question answering via frozen bidirectional language models" NeurIPS (2022)

BLIP-2

- BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

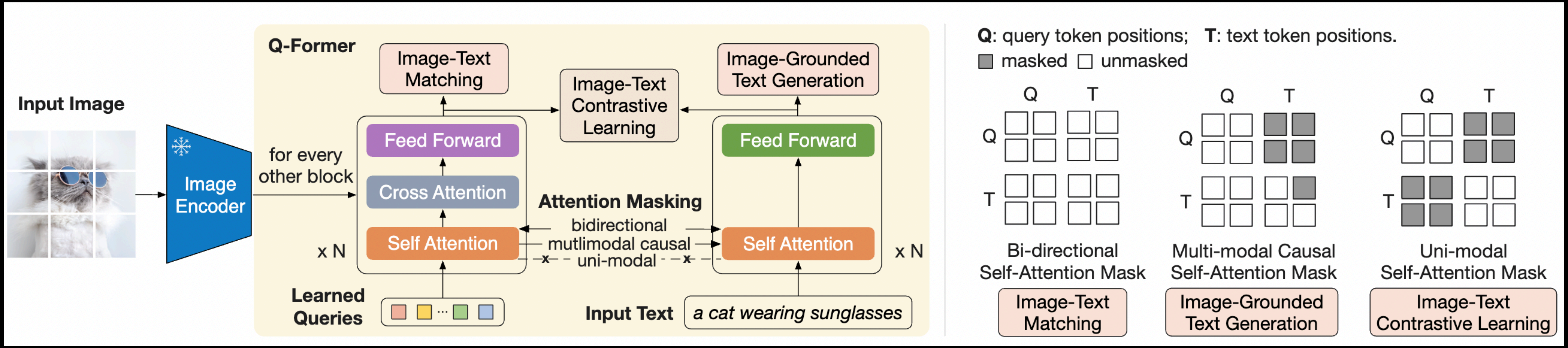


- Training (vision encoder) + (language model) is *heavy* (e.g., Frozen, Flamingo)
- *Lightweight* way to bridge two modalities

"BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models", ICML 2023

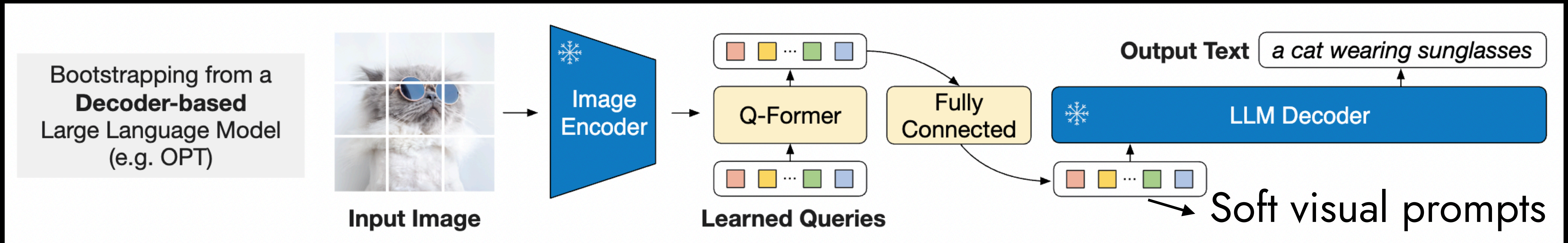
BLIP-2

- BLIP-2 has two-stage training: (1) training Q-Former (2) Integrating with LLMs
- Q-Former: similar to Perceiver Resampler of Flamingo 🦩
- Objectives: image-text matching, text generation, image-text contrastive learning

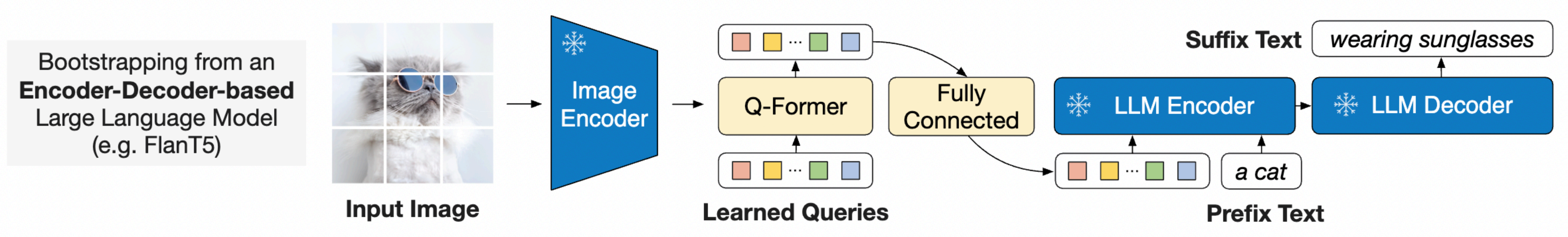


BLIP-2

- Connect Q-Former (with the frozen image encoder attached) to a frozen LLM



Decoder-based LLM (e.g. OPT) - *language modeling loss*



Encoder-Decoder-based LLM (e.g. FlanT5) - *prefix language modeling loss*

"BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models", ICML 2023

BLIP-2

- Performance

Models	#Trainable Params	NoCaps Zero-shot (validation set) in-domain C	in-domain S	near-domain C	near-domain S	out-domain C	out-domain S	overall C	overall S	COCO Fine-tuned Karpathy test B@4	C
OSCAR (Li et al., 2020)	345M	-	-	-	-	-	-	80.9	11.3	37.4	127.8
VinVL (Zhang et al., 2021)	345M	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5	38.2	129.3
BLIP (Li et al., 2022)	446M	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8	40.4	136.7
OFA (Wang et al., 2022a)	930M	-	-	-	-	-	-	-	-	**43.9**	145.3
Flamingo (Alayrac et al., 2022)	10.6B	-	-	-	-	-	-	-	-	-	138.1
SimVLM (Wang et al., 2021b)	∼1.4B	113.7	-	110.9	-	115.2	-	112.2	-	40.6	143.3
BLIP-2 ViT-g $\text{OPT}_{2.7B}$	1.1B	123.0	15.8	117.8	15.4	123.4	**15.1**	119.7	15.4	43.7	**145.8**
BLIP-2 ViT-g $\text{OPT}_{6.7B}$	1.1B	**123.7**	15.8	119.2	15.3	124.4	14.8	121.0	15.3	43.5	145.2
BLIP-2 ViT-g FlanT5_{XL}	1.1B	**123.7**	**16.3**	**120.2**	**15.9**	**124.8**	**15.1**	**121.6**	**15.8**	42.4	144.5

Models	#Trainable Params	VQAv2 test-dev	test-std
Open-ended generation models			
ALBEF (Li et al., 2021)	314M	75.84	76.04
BLIP (Li et al., 2022)	385M	78.25	78.32
OFA (Wang et al., 2022a)	930M	82.00	82.00
Flamingo80B (Alayrac et al., 2022)	10.6B	82.00	82.10
BLIP-2 ViT-g FlanT5_{XL}	1.2B	81.55	81.66
BLIP-2 ViT-g $\text{OPT}_{2.7B}$	1.2B	81.59	81.74
BLIP-2 ViT-g $\text{OPT}_{6.7B}$	1.2B	**82.19**	**82.30**
Closed-ended classification models			
VinVL	345M	76.52	76.60
SimVLM (Wang et al., 2021b)	∼1.4B	80.03	80.34
CoCa (Yu et al., 2022)	2.1B	82.30	82.30
BEIT-3 (Wang et al., 2022b)	1.9B	**84.19**	**84.03**

- Relatively small model size, open-sourced model

Production-level Large VLMs

Example of GPT-4 visual input:

User What is funny about this image? Describe it panel by panel.

Source: `https://www.reddit.com/r/hmmm/comments/ubab5v/hmmm/`

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

Benchmark	GPT-4 Evaluated few-shot	Few-shot SOTA	SOTA Best external model (includes benchmark-specific training)
VQAv2 VQA score (test-dev)	77.2% 0-shot	67.6% Flamingo 32-shot	84.3% PaLI-17B
TextVQA VQA score (val)	78.0% 0-shot	37.9% Flamingo 32-shot	71.8% PaLI-17B
ChartQA Relaxed accuracy (test)	78.5%[A]	-	58.6% Pix2Struct Large
AI2 Diagram (AI2D) Accuracy (test)	78.2% 0-shot	-	42.1% Pix2Struct Large
DocVQA ANLS score (test)	88.4% 0-shot (pixel-only)	-	88.4% ERNIE-Layout 2.0
Infographic VQA ANLS score (test)	75.1% 0-shot (pixel-only)	-	61.2% Applica.ai TILT
TVQA Accuracy (val)	87.3% 0-shot	-	86.5% MERLOT Reserve Large
LSMDC Fill-in-the-blank accuracy (test)	45.7% 0-shot	31.0% MERLOT Reserve 0-shot	52.9% MERLOT

GPT-4 Technical Report (2023)

LLaVA

- LLaVA: Large Language and Vision Assistant
- What is missing in previous works?
  - Lack of *Instruction-following ability*
  - Lack of instruction-related data
- Then, how to collect the vision-language instruction dataset?
  - Human annotation is too costly (e.g., read text and see image, then write the output)
  - *Leverage LLMs to generate instruction data*

GPT-4 visual input example, Extreme Ironing:

User What is unusual about this image?

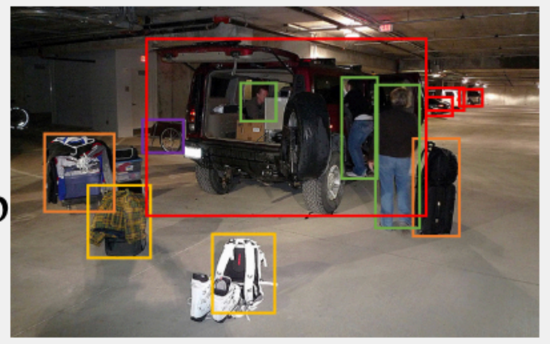
Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

GPT-4 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.

LLaVA

- Use language-only GPT-4 as strong teacher
- Creating dataset: 158K image-text instruction dataset

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>

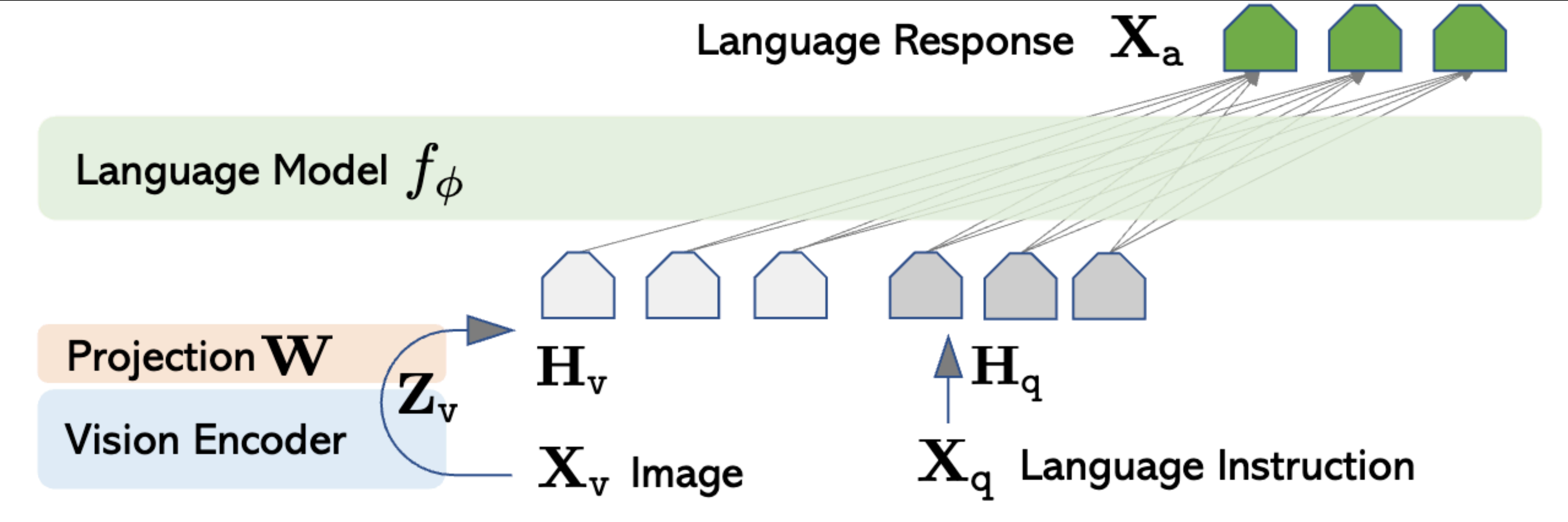
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They 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>

"Large Language and Vision Assistant", NeurIPS (2023)

LLaVA

- Model architecture
  - Vision encoder: CLIP ViT-L/14
  - LLM: Vicuna-7B



Visual input example, Extreme Ironing:

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

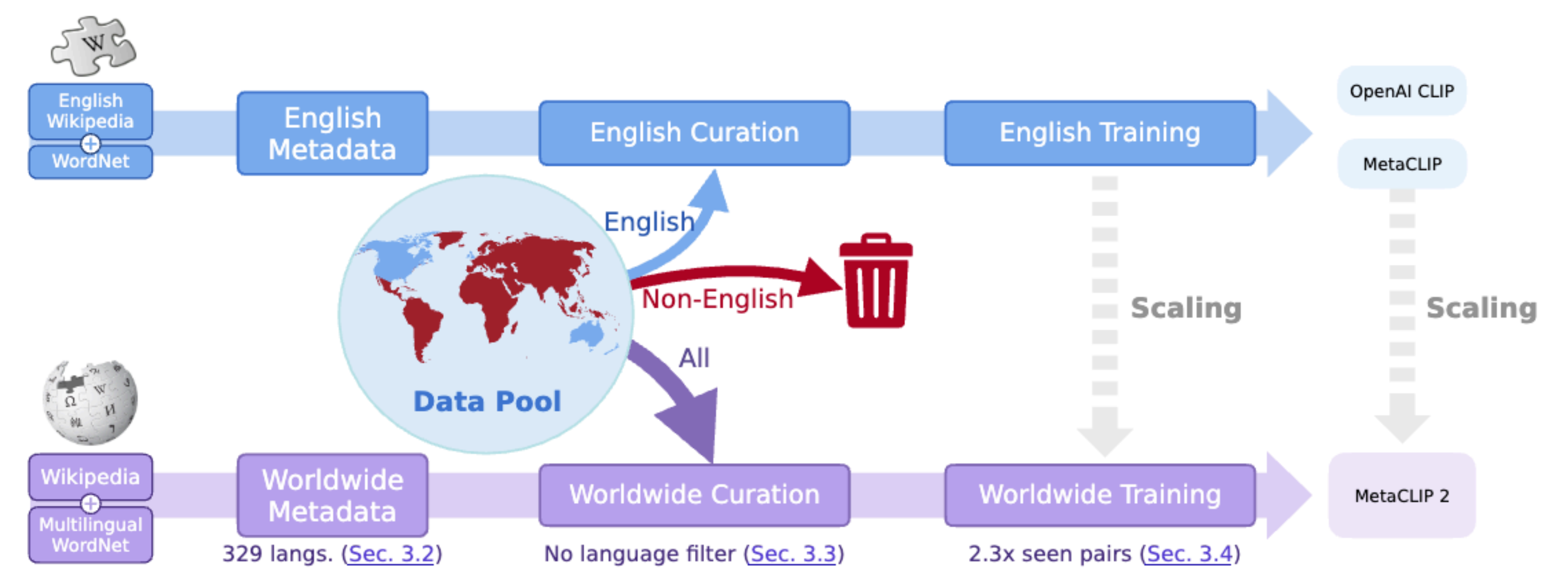
User	What is unusual about this image?
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing 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 [36]	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	What is unusual about this image?
BLIP-2	a man is sitting on the back of a yellow cab
User	What is unusual about this image?
OpenFlamingo	The man is drying his clothes on the hood of his car.

"Large Language and Vision Assistant", NeurIPS (2023)

Vision-language Models 2024~

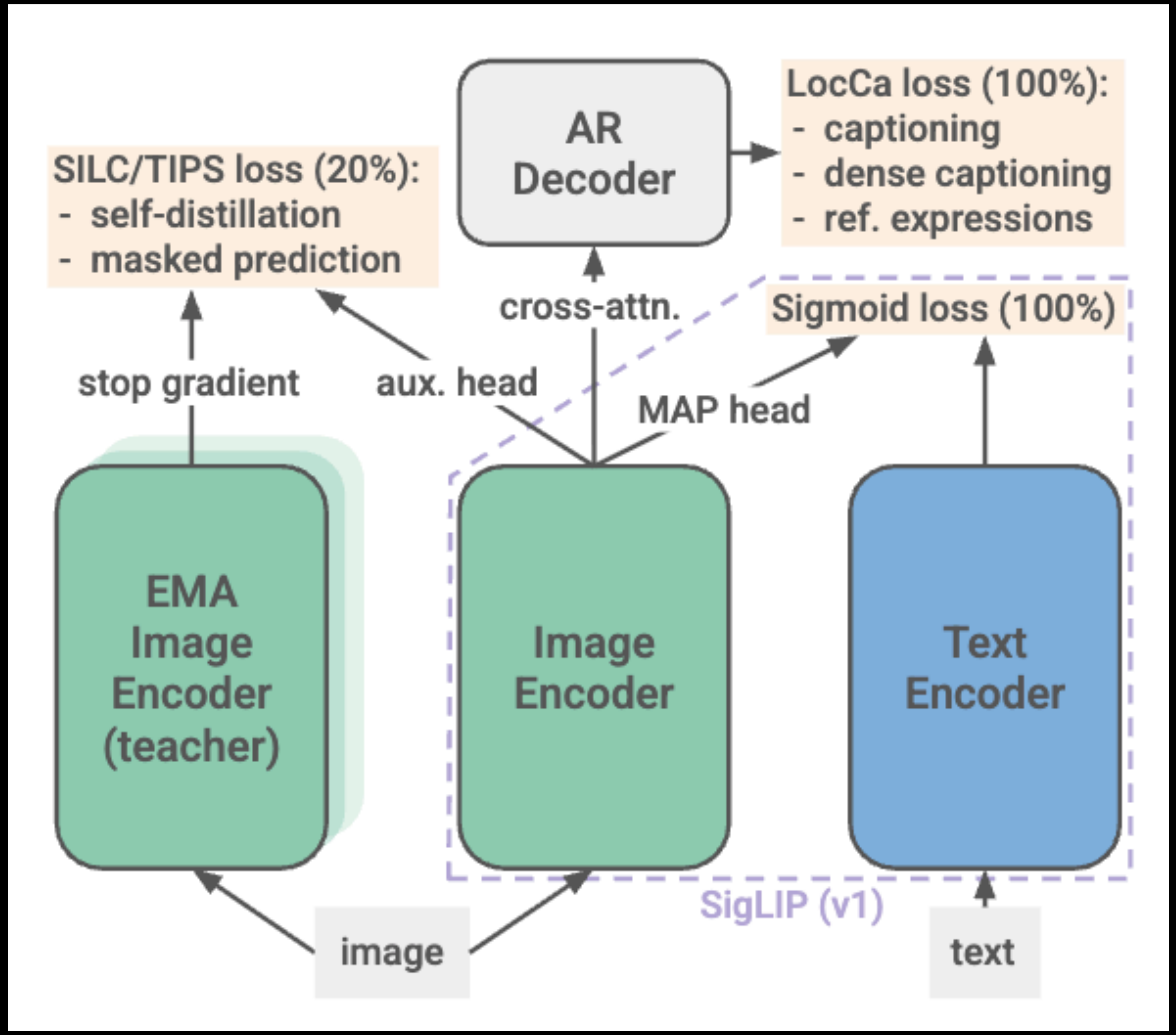
Meta CLIP 2 (2025)

- Previous filtering: English-based, data with non-English will be removed
- This causes multi-linguality problem in CLIP



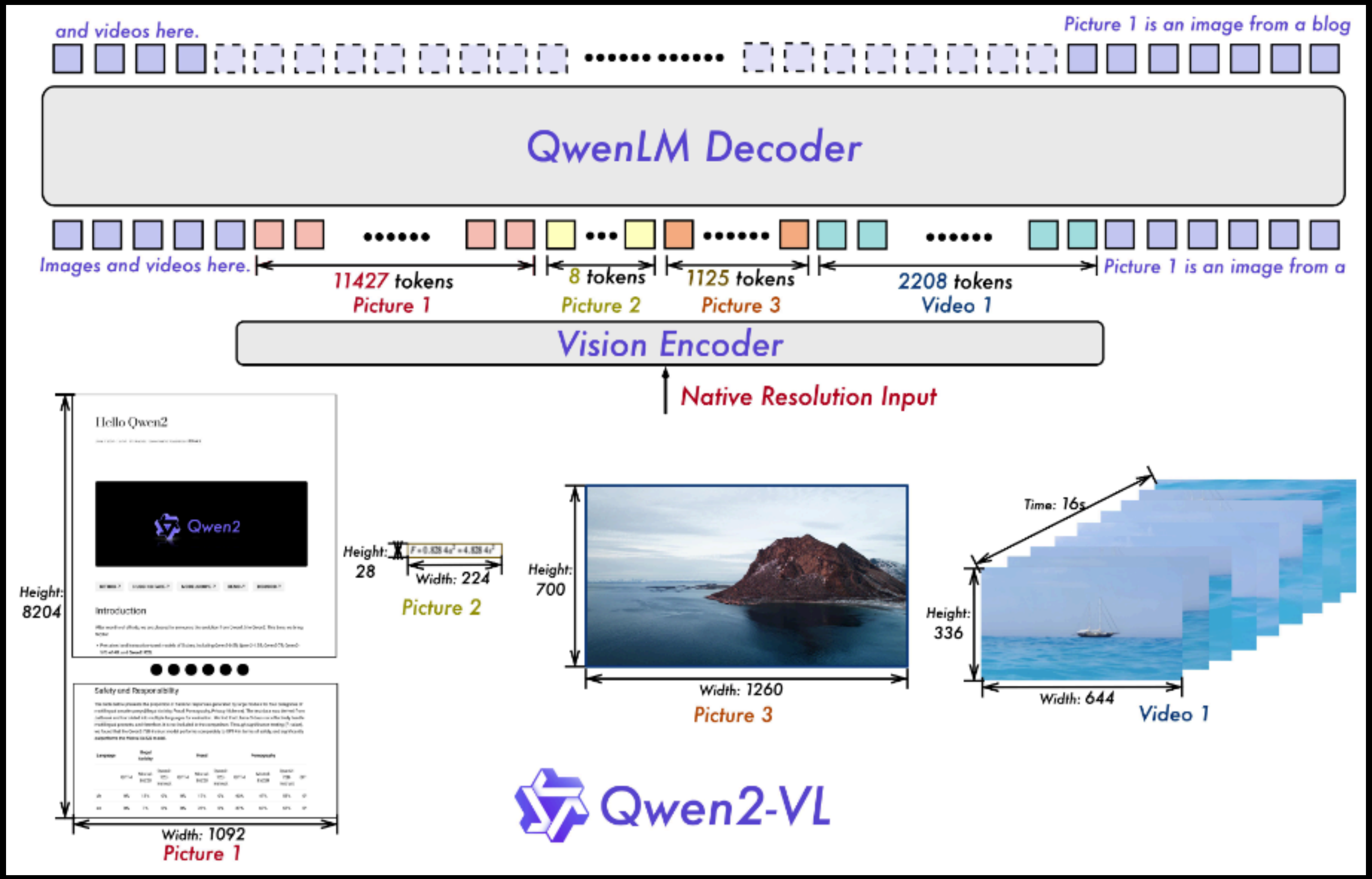
SigLIP 2 (2025)

- Better localization and dense prediction tasks (semantic segmentation, etc.)
- Vision-inspired techniques on vision encoders
- Add captioning loss

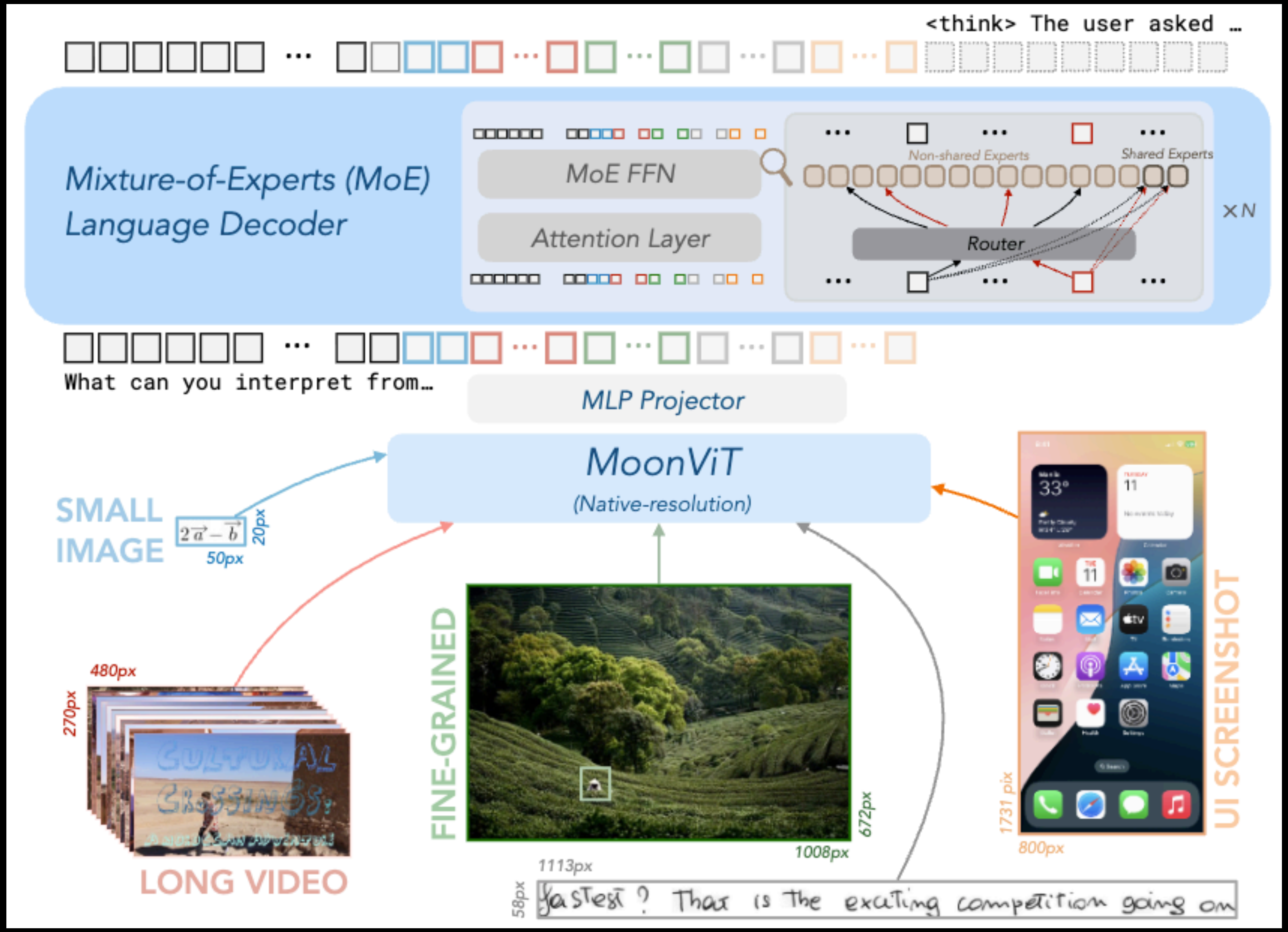


"SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features, 2025

Qwen-VL / Kimi-VL



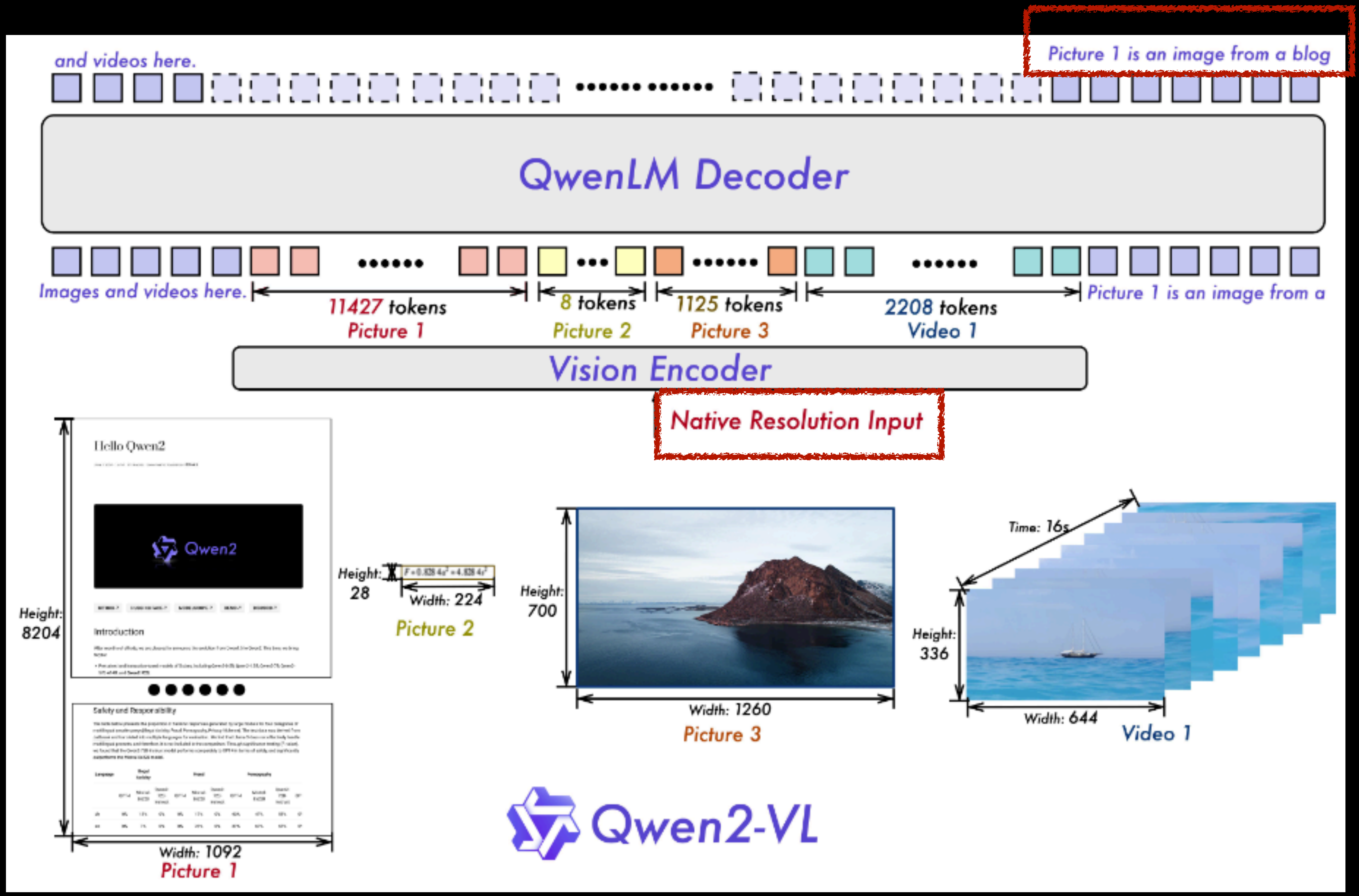
Qwen2-VL (https://arxiv.org/abs/2409.12191)



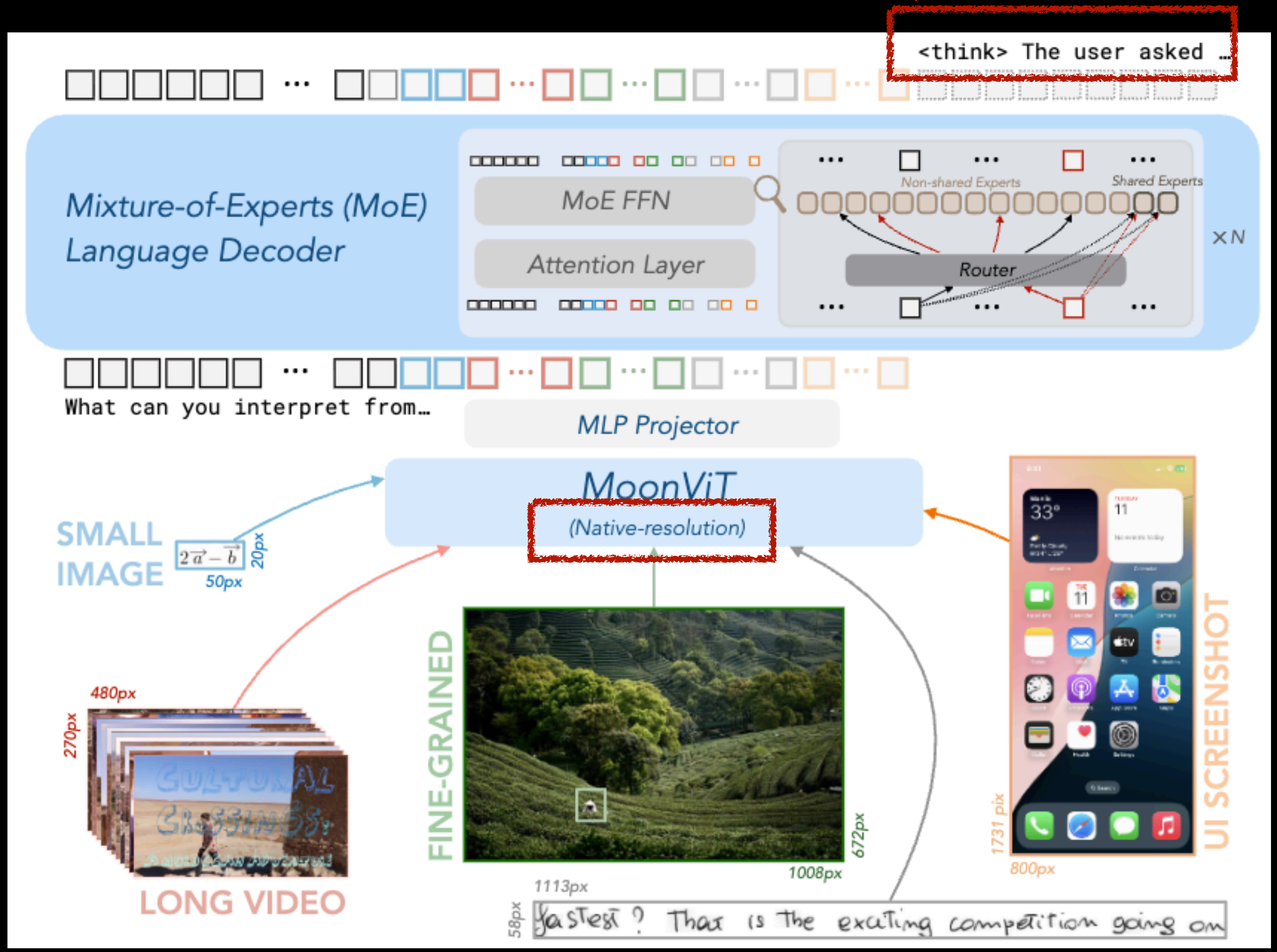
Kimi-VL (https://arxiv.org/abs/2504.07491)

Qwen-VL / Kimi-VL

Next-token (text) prediction loss



Qwen2-VL (https://arxiv.org/abs/2409.12191)



Kimi-VL (https://arxiv.org/abs/2504.07491)

Conclusion

- Scaling-up multimodal model size
- Scaling-up multimodal dataset size
- Simplified architecture and objective

Thank You!

Reference materials

- https://cs.uwaterloo.ca/~wenhuche/teaching/cs886/
- https://advances-in-vision.github.io/schedule.html