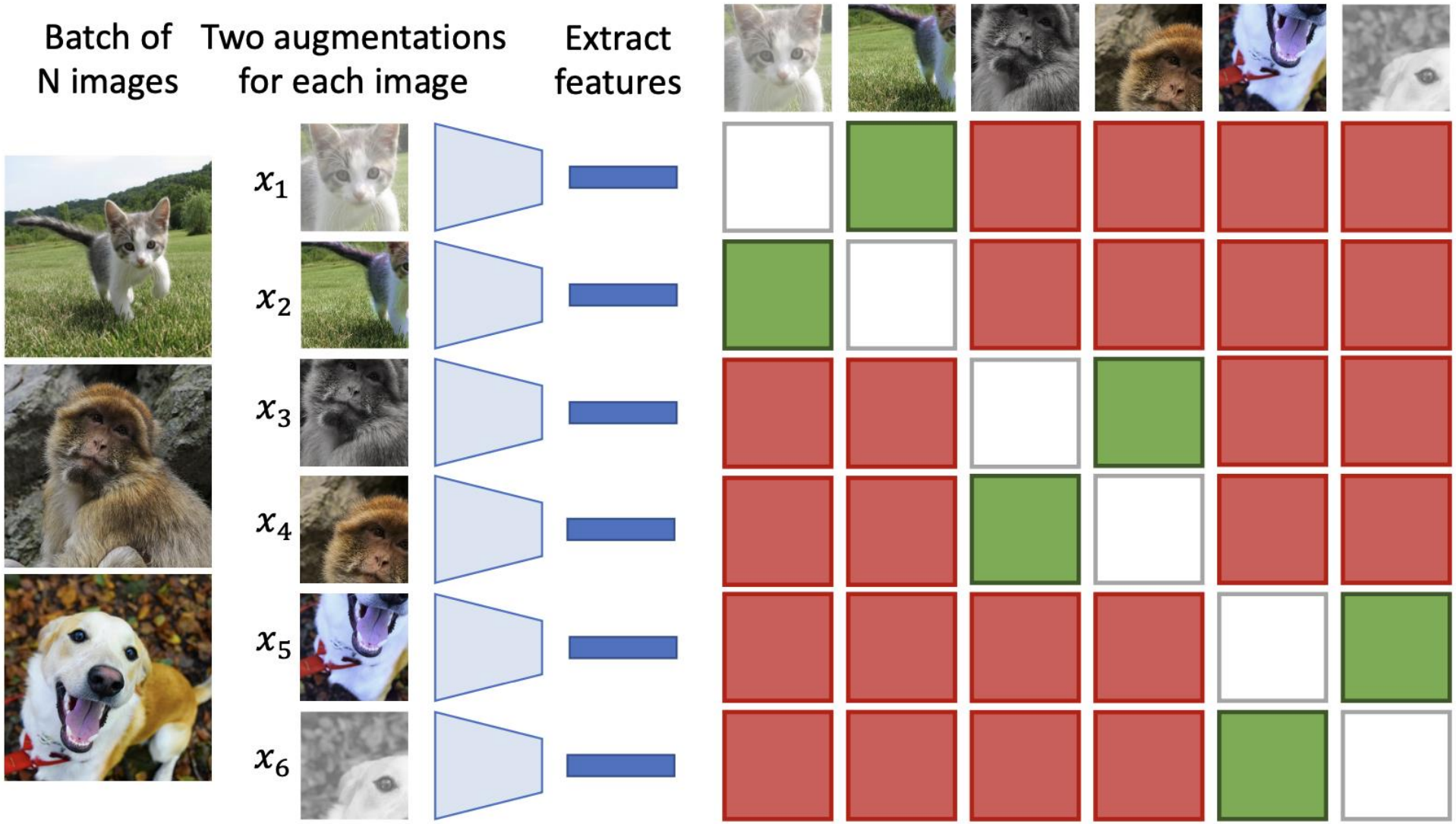
Contrastive Learning

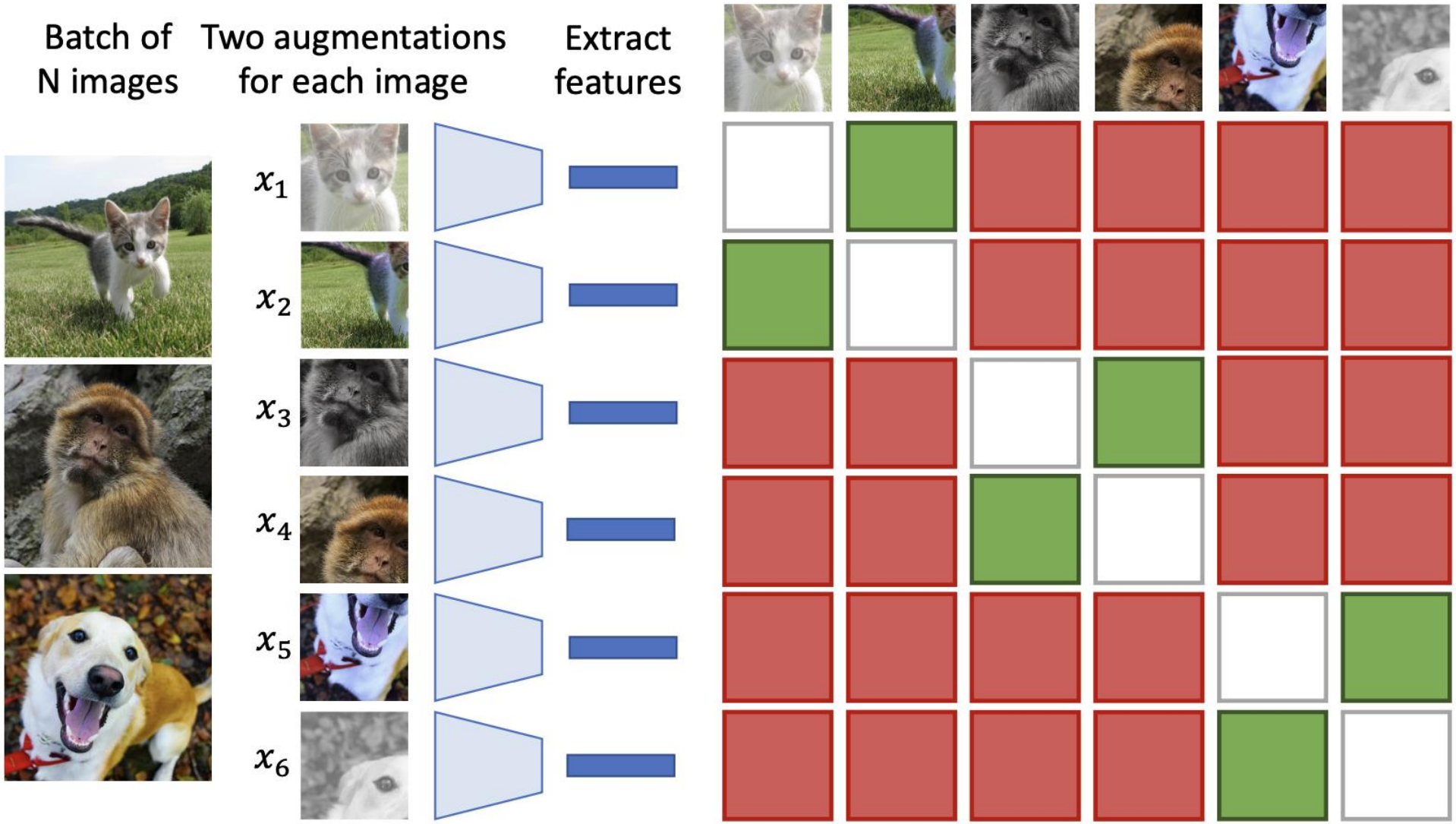


Hadsell et al, "Dimensionality Reduction by Learning and Invariant Mapping", CVPR 2006
Wu et al, "Unsupervised Feature Learning by Non-Parametric Instance-Level Discrimination", CVPR 2018
Van den Oord et al, "Representation Learning with Contrastive Predictive Coding", NeurIPS 2018

Hjelm et al, "Learning deep representations by mutual information estimation and maximization", ICLR 2019
Bachman et al, "Learning Representations by Maximizing Mutual Information Across Views", NeurIPS 2019
Henaff et al, "Data-Efficient Image Recognition with Contrastive Predictive Coding", ICML 2020

Tian et al, "Contrastive Multiview Coding", ECCV 2020
He et al, "Momentum Contrast for Unsupervised Visual Representation Learning", CVPR 2020
Chen et al, "A Simple Framework for Contrastive Learning of Visual Representations", ICML 2020

Contrastive Learning



Similarity between x_i and x_j:

$$s_{i,j} = \frac{\phi(x_i)^T \phi(x_j)}{\|\phi(x_i)\| \cdot \|\phi(x_i)\|}$$

If (x_i, x_j) is a positive pair, then loss for x_i is:

$$L_i = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{\substack{k=1 \\ k \neq i}}^{2N} \exp(s_{i,k}/\tau)}$$

(τ is a *temperature*)

Hadsell et al, "Dimensionality Reduction by Learning and Invariant Mapping", CVPR 2006
Wu et al, "Unsupervised Feature Learning by Non-Parametric Instance-Level Discrimination", CVPR 2018
Van den Oord et al, "Representation Learning with Contrastive Predictive Coding", NeurIPS 2018

Hjelm et al, "Learning deep representations by mutual information estimation and maximization", ICLR 2019
Bachman et al, "Learning Representations by Maximizing Mutual Information Across Views", NeurIPS 2019
Henaff et al, "Data-Efficient Image Recognition with Contrastive Predictive Coding", ICML 2020

Tian et al, "Contrastive Multiview Coding", ECCV 2020
He et al, "Momentum Contrast for Unsupervised Visual Representation Learning", CVPR 2020
Chen et al, "A Simple Framework for Contrastive Learning of Visual Representations", ICML 2020

Multimodal SSL

Don't learn from isolated images -- take images together with some **context**

Video: Image together with adjacent video frames

Agrawal et al, "Learning to See by Moving", ICCV 2015
Wang et al, "Unsupervised Learning of Visual Representations using Videos", ICCV 2015
Pathak et al, "Learning Features by Watching Objects Move", CVPR 2017

Sound: Image with audio track from video

Owens et al, "Ambient Sound Provides Supervision for Visual Learning", ECCV 2016
Arandjelovic and Zisserman, "Look, Listen and Learn", ICCV 2017

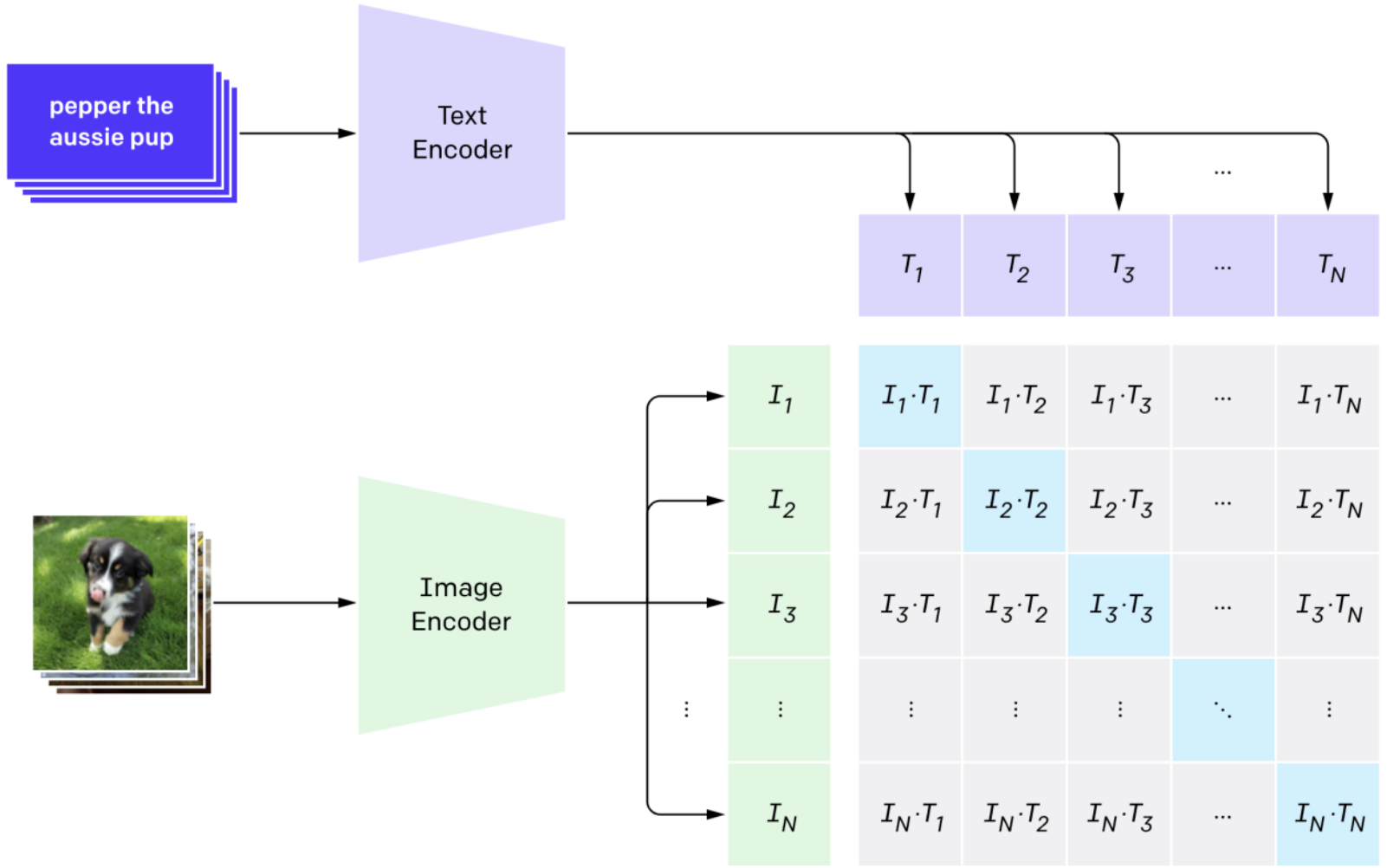
3D: Image with depth map or point cloud

Xie et al, "PointContrast: Unsupervised Pre-training for 3D Point Cloud Understanding", ECCV 2020
Zhang et al, "Self-supervised pretraining of 3D features on any point-cloud", CVPR 2021

Language: Image with natural-language text

Sariyildiz et al, "Learning Visual Representations with Caption Annotations", ECCV 2020
Desai and Johnson, "VirTex: Learning Visual Representations from Textual Annotations", CVPR 2021
Radford et al, "Learning Transferable Visual Models form Natural Language Supervision", ICML 2021
Jia et al, "Scaling up Visual and Vision-Language Representation Learning with Noisy Text Supervision", ICLR 2021
Desai et al, "RedCaps: Web-curated Image-Text data created by the people, for the people", NeurIPS 2021

CLIP



Contrastive loss: Each image predicts which caption matches

CLIP

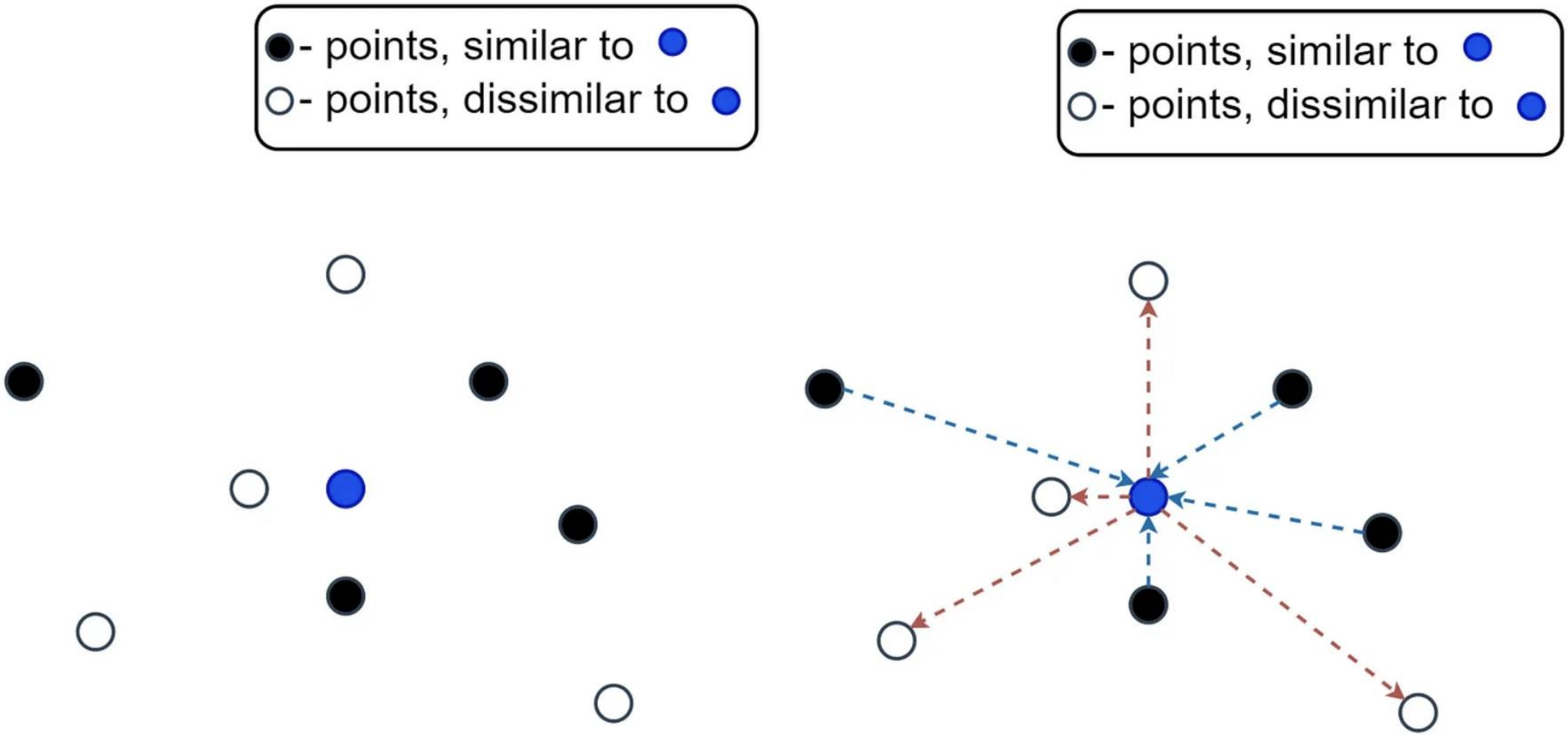
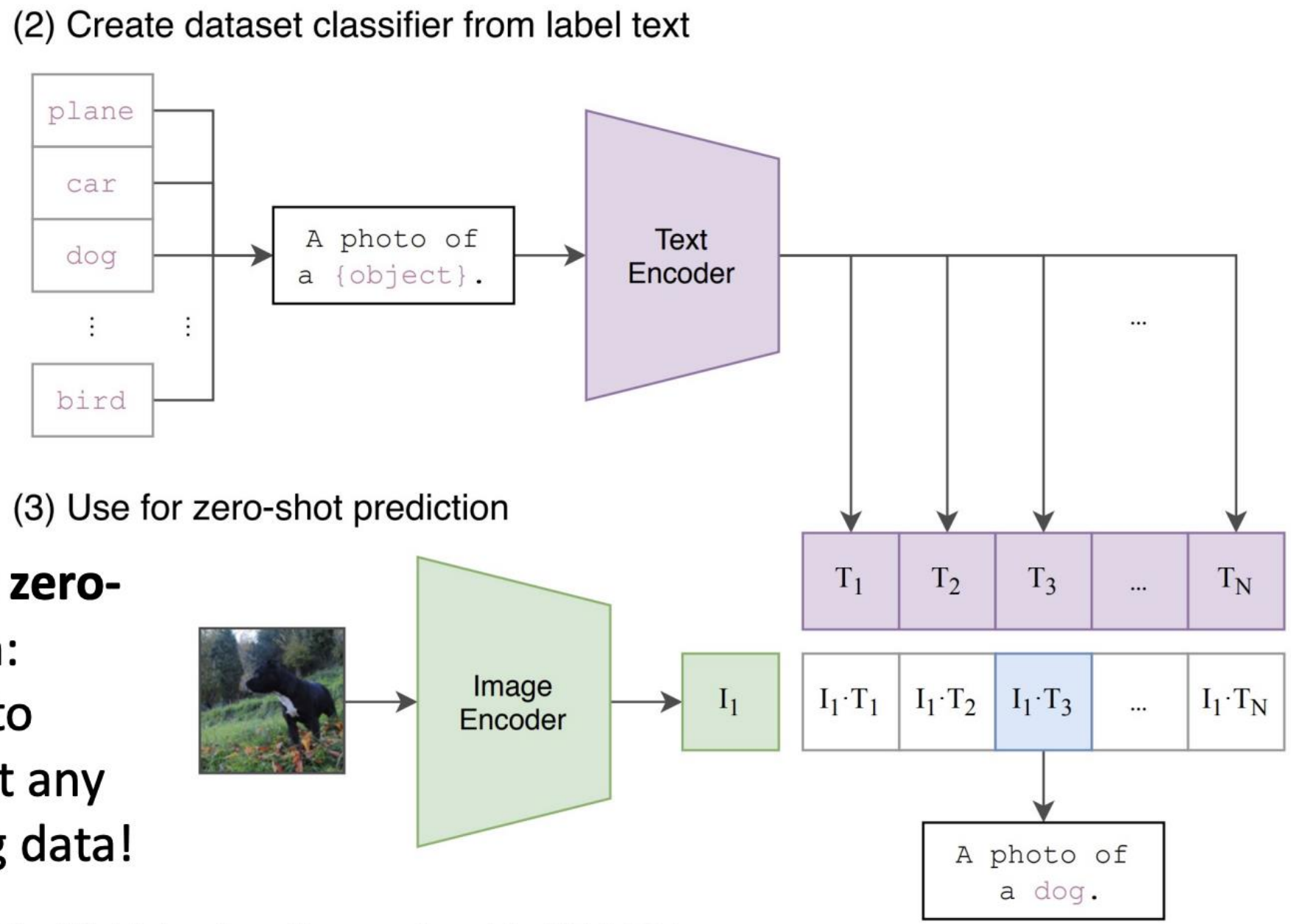


Figure 4 — We would like to bring black dots closer to the blue one, and push white dots away.

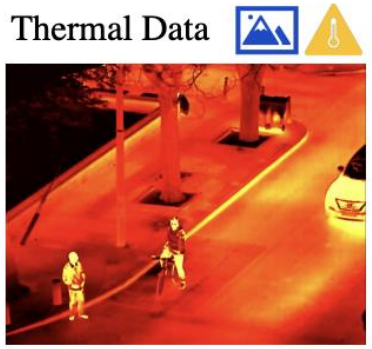
CLIP: Zero-shot classification

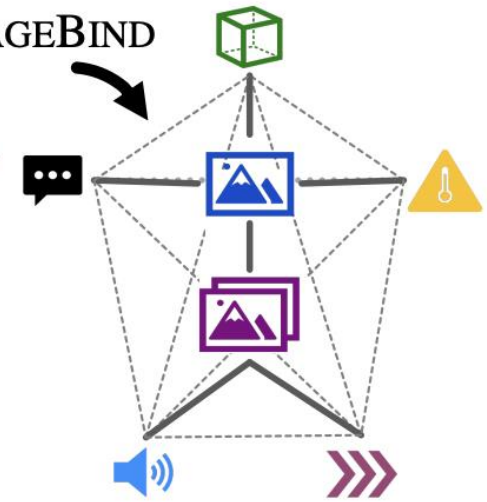


Language enables **zero-shot classification**: Classify images into categories without any additional training data!

Beyond CLIP: ImageBind

$$L_{\mathcal{I},\mathcal{M}} = -\log \frac{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau)}{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau) + \sum_{j \neq i} \exp(\mathbf{q}_i^\intercal \mathbf{k}_j/\tau)}$$

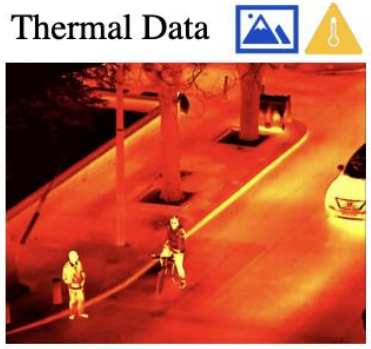


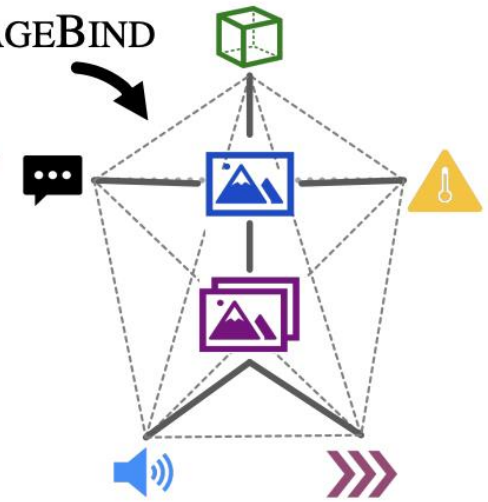


[Girdhar CVPR 2023]

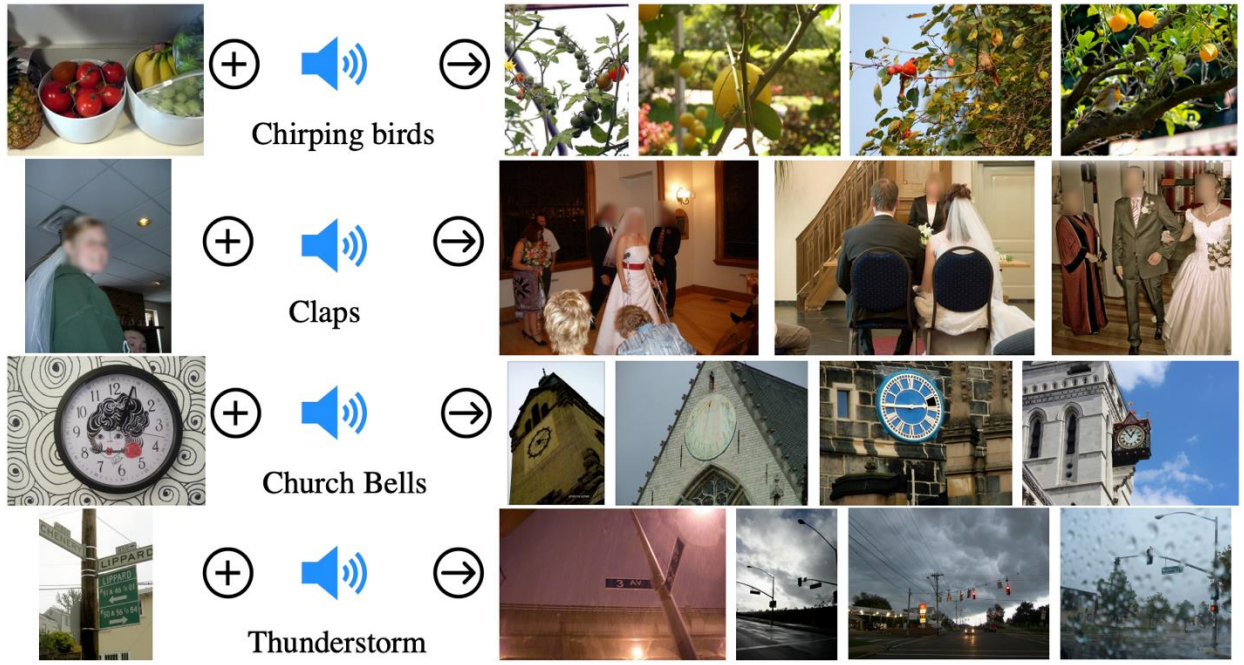
Beyond CLIP: ImageBind

$$L_{\mathcal{I},\mathcal{M}} = -\log \frac{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau)}{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau) + \sum_{j \neq i} \exp(\mathbf{q}_i^\intercal \mathbf{k}_j/\tau)}$$





Embedding arithmetics



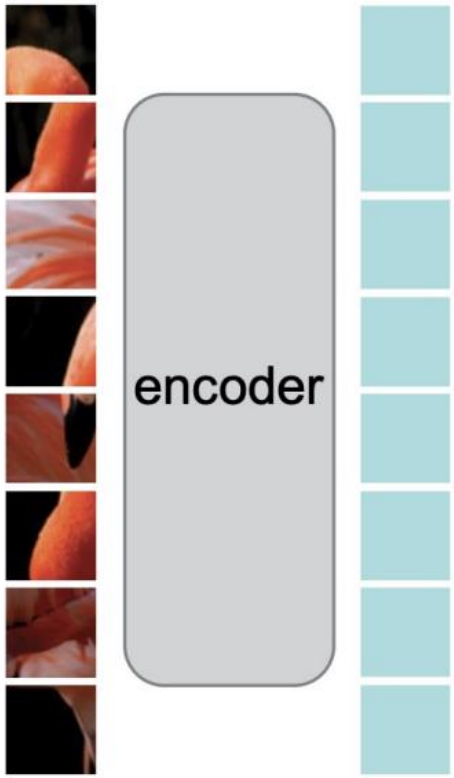
[Girdhar CVPR 2023]

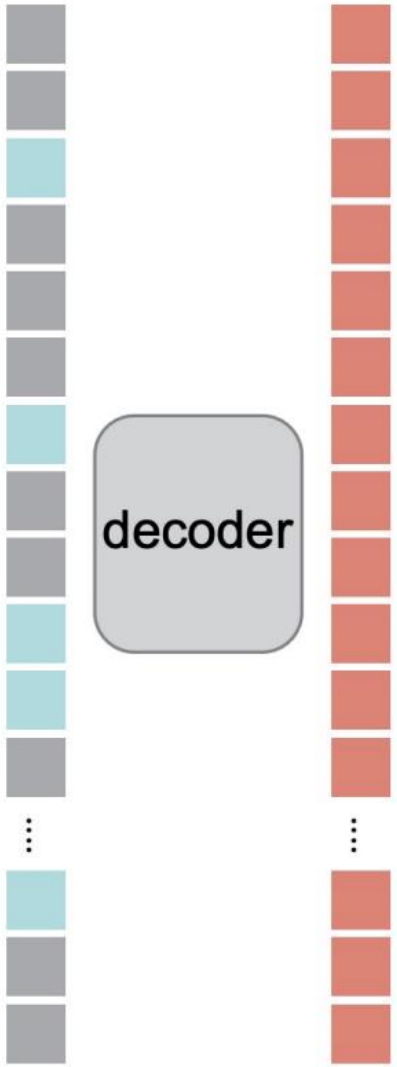
MAE: Masked Auto Encoders

Divide image into nonoverlapping patches, discard most of them

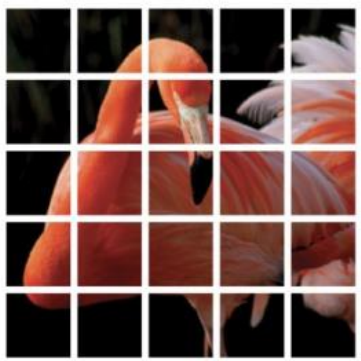
input

Encode remaining patches with a ViT





Decoder is a small ViT that predicts pixel values of the masked patches



target

Hadsell et al, "Dimensionality Reduction by Learning and Invariant Mapping", CVPR 2006
Wu et al, "Unsupervised Feature Learning by Non-Parametric Instance-Level Discrimination", CVPR 2018
Van den Oord et al, "Representation Learning with Contrastive Predictive Coding", NeurIPS 2018

Hjelm et al, "Learning deep representations by mutual information estimation and maximization", ICLR 2019
Bachman et al, "Learning Representations by Maximizing Mutual Information Across Views", NeurIPS 2019
Henaff et al, "Data-Efficient Image Recognition with Contrastive Predictive Coding", ICML 2020

Tian et al, "Contrastive Multiview Coding", ECCV 2020
He et al, "Momentum Contrast for Unsupervised Visual Representation Learning", CVPR 2020
Chen et al, "A Simple Framework for Contrastive Learning of Visual Representations", ICML 2020

Multimodal MAE

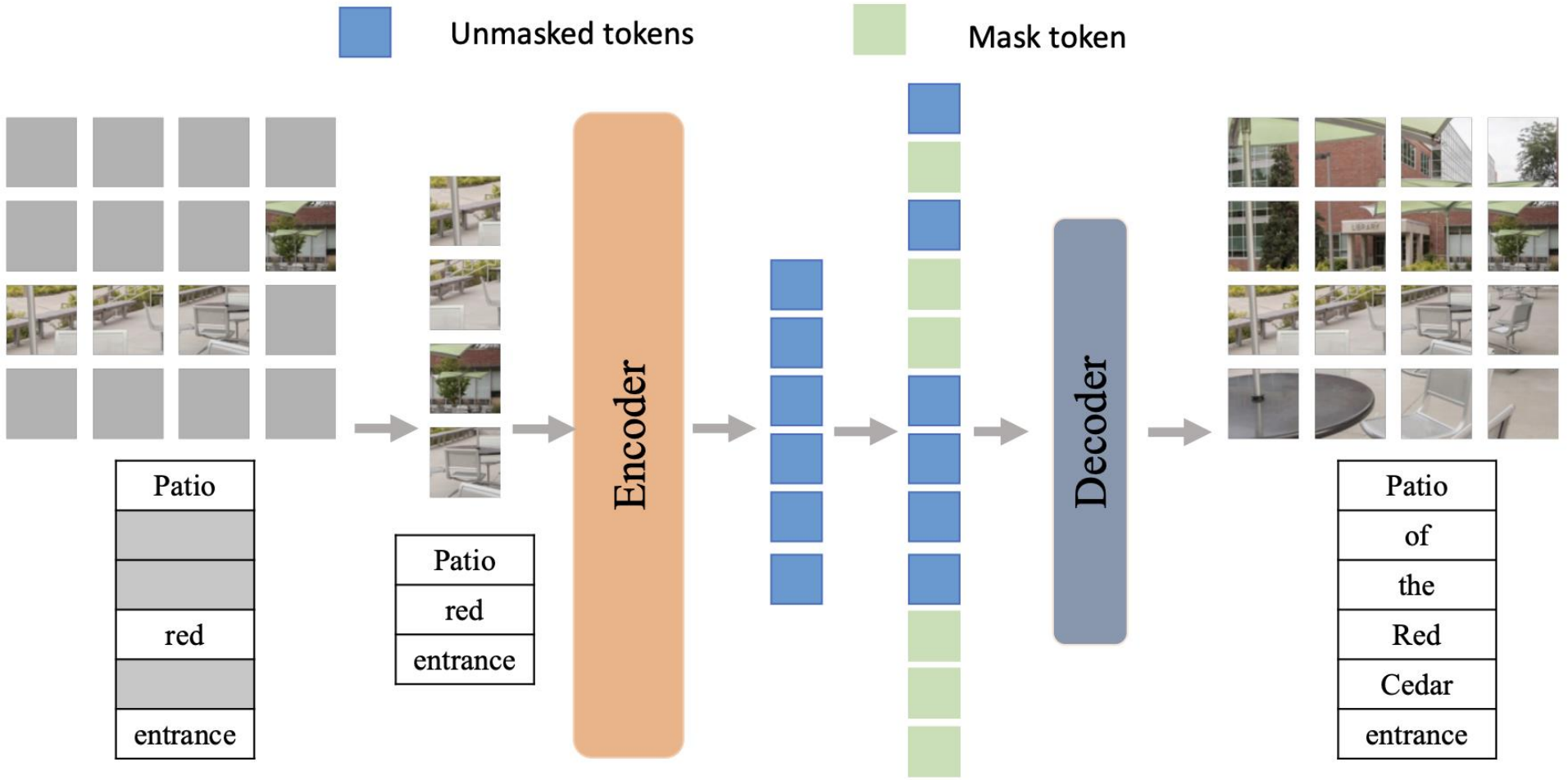
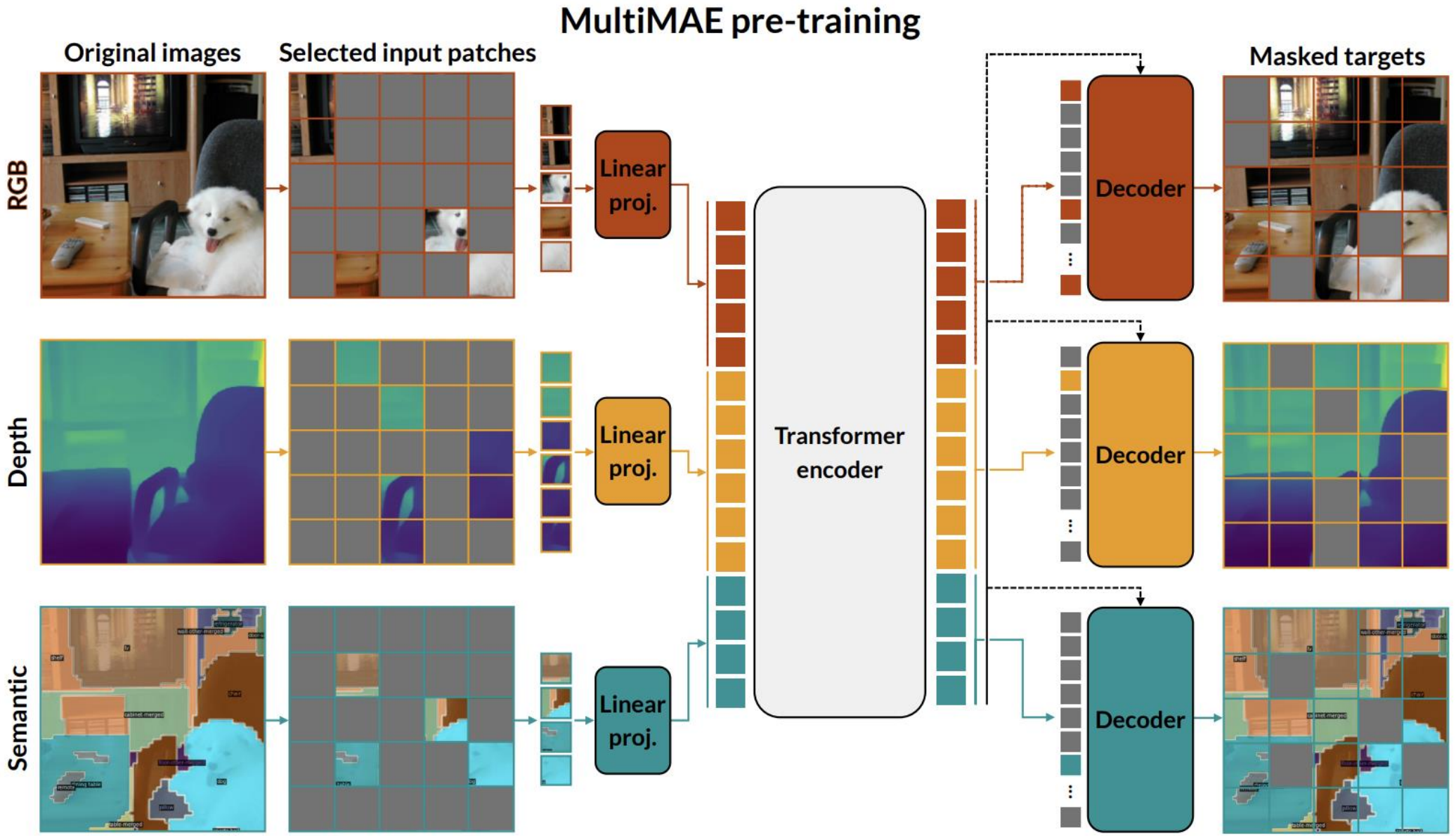


Figure 1: Multimodal masked autoencoder (M3AE) consists of an encoder that maps language tokens and image patches to a shared representation space, and a decoder that reconstructs the original image and language from the representation.

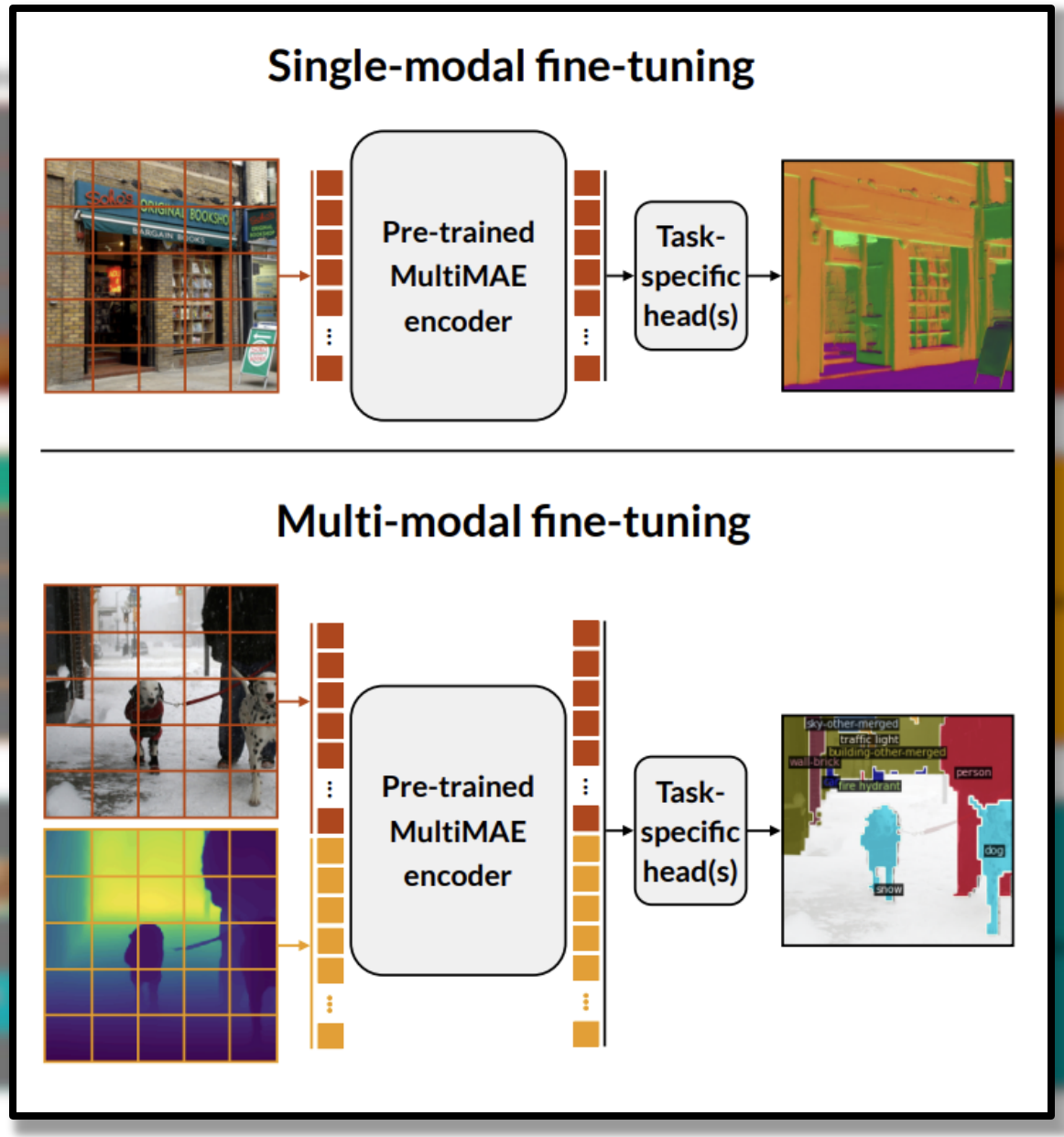
[Geng et al 2022, Multimodal Masked Autoencoders Learn Transferable Representations]

MultiMAE



[Bachmann et al ECCV 2022]

MultiMAE

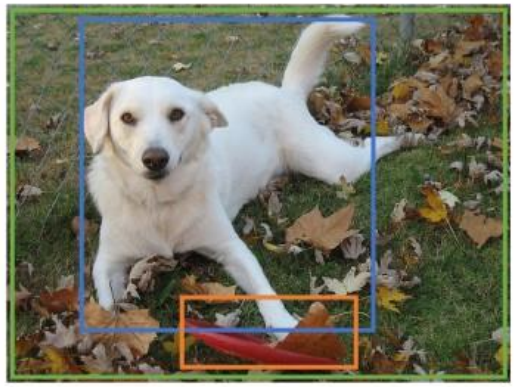


[Bachmann et al ECCV 2022]

Multimodal tasks (=Vision and Language Tasks)

- Multimodal Classification
- Image-Text Retrieval
- **Visual Grounding**
- Visual Question Answering and Visual Reasoning
- Image Captioning
- Text-to-image Generation

Visual grounding



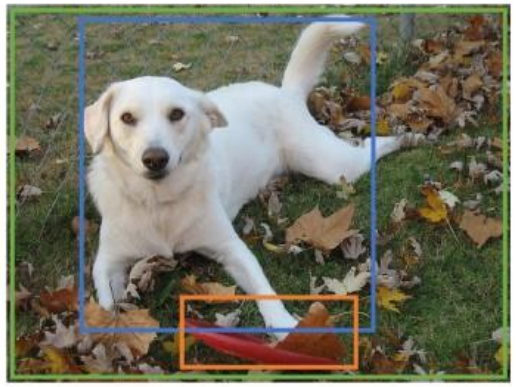
A dog is lying on the grass next to a frisbee.

(a) Phrase grounding.

The red frisbee next to the dog.

(b) Referring expression comprehension.

Visual grounding



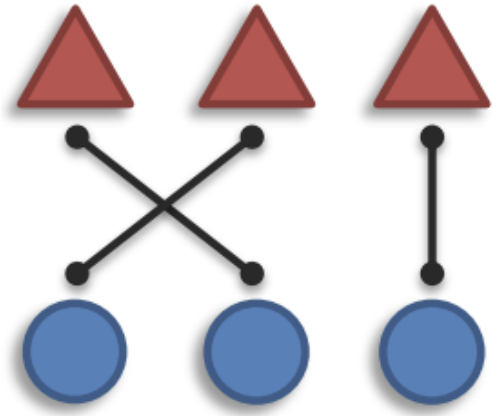
A dog is lying on the grass next to a frisbee.

(a) Phrase grounding.

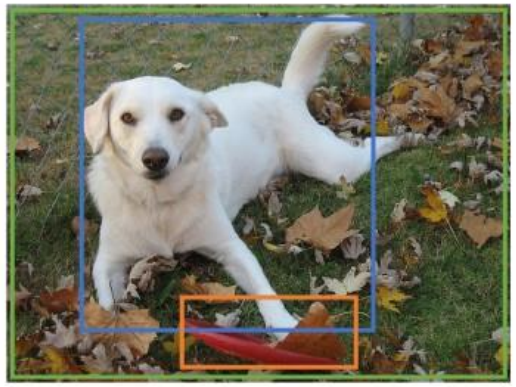
The red frisbee next to the dog.

(b) Referring expression comprehension.

Alignment



Visual grounding



A dog is lying on the grass next to a frisbee.

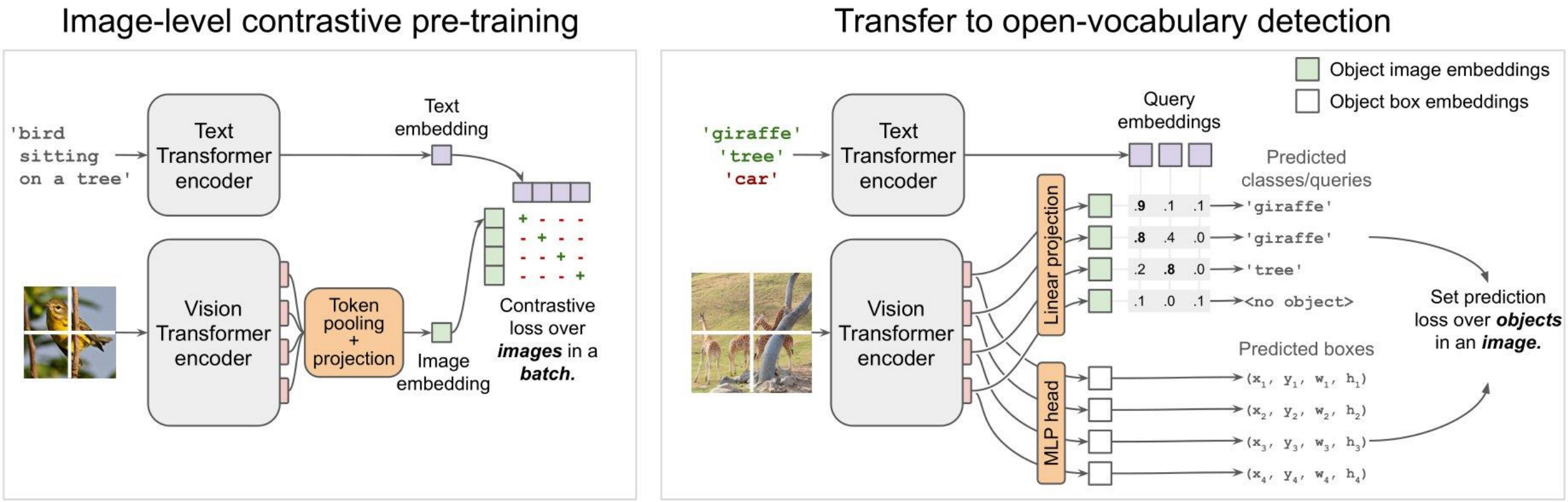
(a) Phrase grounding.

The red frisbee next to the dog.

(b) Referring expression comprehension.

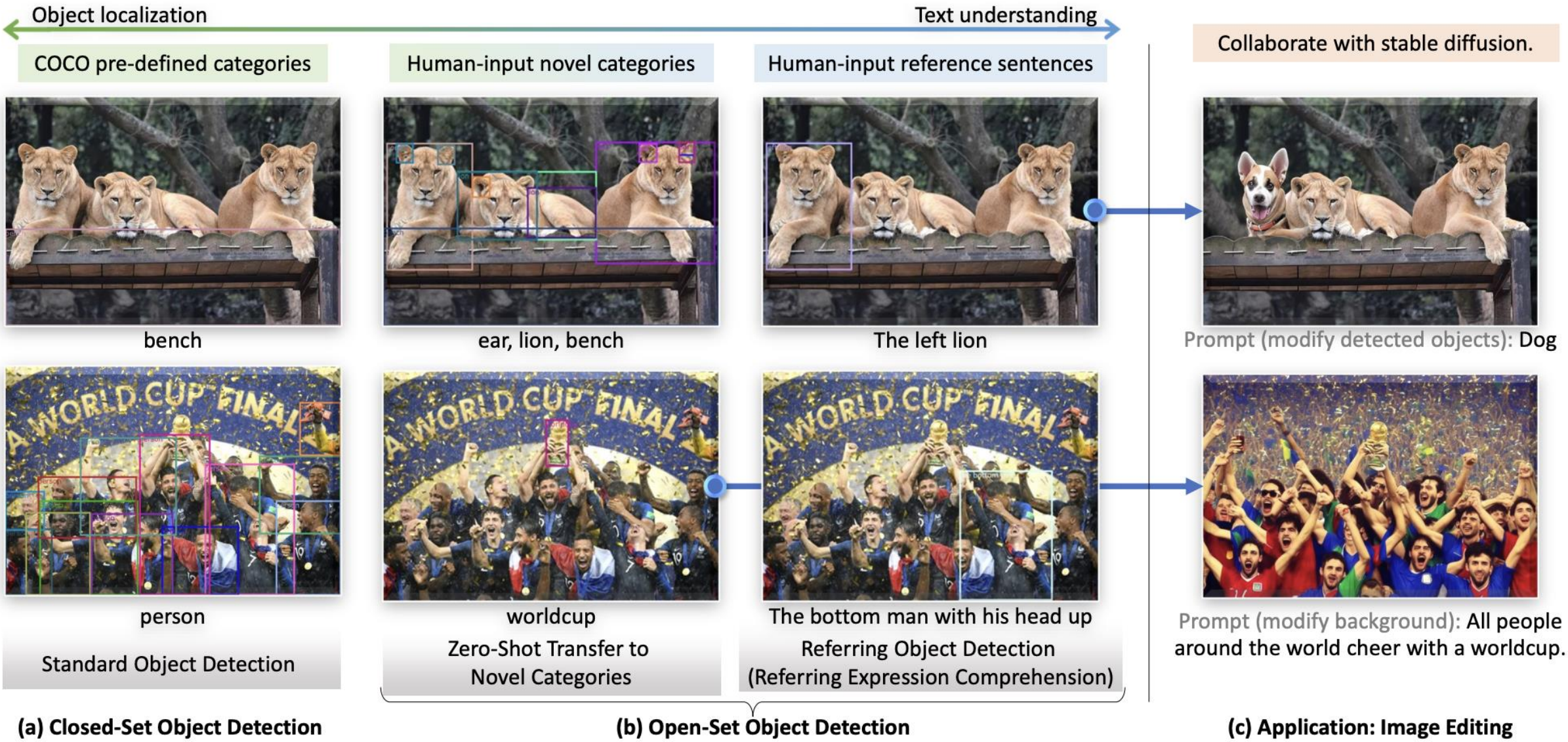
- **Inputs:**
  - Image: A visual representation of a scene or object.
  - Natural language query: A text description or question that refers to a specific part of the image.
- **Output:** Bounding box or segmentation mask: A spatial region within the image that corresponds to the object or area described in the query. This is typically represented as coordinates or a highlighted region.
- **Task:** Locating the relevant object or region: The model must correctly identify the part of the image that matches the query. This involves understanding both the visual content of the image and the linguistic meaning of the query.

OWL-VIT (Vision Transformer for Open-World Localization)



[Minderer et al, Simple Open-Vocabulary Object Detection with Vision Transformers, ECCV 2022]

Grounding DINO



[Liu et al, ECCV 2024]

DINO: Self-supervised Vision Transformers

[Caron et al, Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021]
[Oquab et al, DINOv2: Learning Robust Visual Features without Supervision, TMLR 2024]

DINO

Depth Estimation

State-of-the-art results and strong generalization on estimating depth from a single image.

Semantic Segmentation

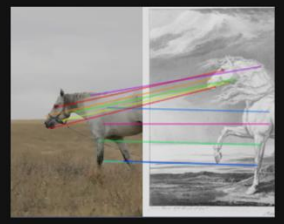
Competitive results without any fine-tuning on clustering an images into object classes.

Instance Retrieval

Directly use frozen features to find art pieces similar to an image from a large art collection.

Dense Matching

Consistently map all parts of an image without supervision.



Sparse Matching

Compare DINOv2 patch features across two images to match their most similar parts.

[Caron et al, Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021]
[Oquab et al, DINOv2: Learning Robust Visual Features without Supervision, TMLR 2024]

DINO

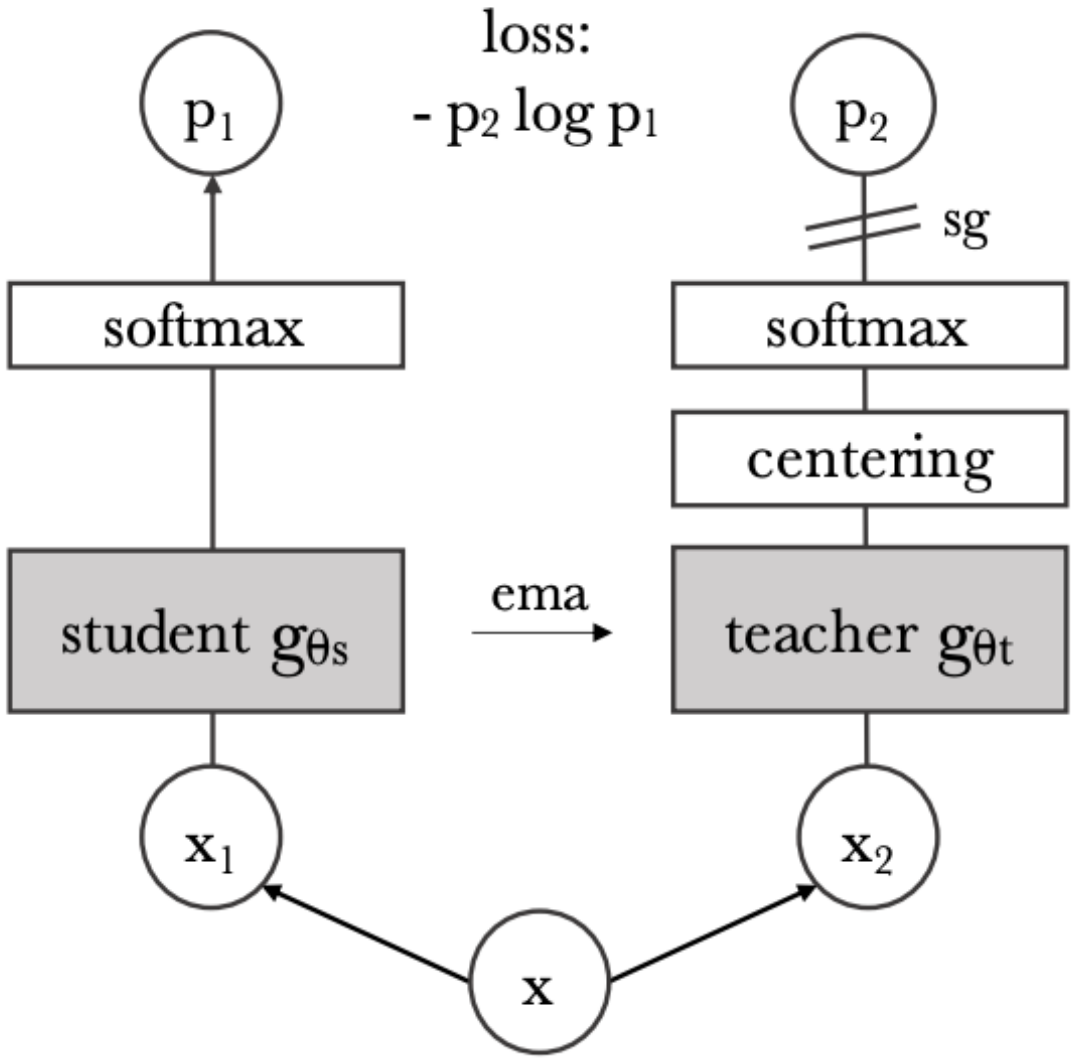


Figure 2: **Self-distillation with no labels.** We illustrate DINO in the case of one single pair of views (x_1, x_2) for simplicity. The model passes two different random transformations of an input image to the student and teacher networks. Both networks have the same architecture but different parameters. The output of the teacher network is centered with a mean computed over the batch. Each networks outputs a K dimensional feature that is normalized with a temperature softmax over the feature dimension. Their similarity is then measured with a cross-entropy loss. We apply a stop-gradient (sg) operator on the teacher to propagate gradients only through the student. The teacher parameters are updated with an exponential moving average (ema) of the student parameters.

[Caron et al, Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021]
[Oquab et al, DINOv2: Learning Robust Visual Features without Supervision, TMLR 2024]

DINO

[Caron et al, Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021]
[Oquab et al, DINOv2: Learning Robust Visual Features without Supervision, TMLR 2024]

DINO: Results

Method	Arch.	ADE20k (62.9) lin.	ADE20k (62.9) +ms	CityScapes (86.9) lin.	CityScapes (86.9) +ms	Pascal VOC (89.0) lin.	Pascal VOC (89.0) +ms
OpenCLIP	ViT-G/14	39.3	46.0	60.3	70.3	71.4	79.2
MAE	ViT-H/14	33.3	30.7	58.4	61.0	67.6	63.3
DINO	ViT-B/8	31.8	35.2	56.9	66.2	66.4	75.6
iBOT	ViT-L/16	44.6	47.5	64.8	74.5	82.3	84.3
DINOv2	ViT-S/14	44.3	47.2	66.6	77.1	81.1	82.6
	ViT-B/14	47.3	51.3	69.4	80.0	82.5	84.9
	ViT-L/14	47.7	**53.1**	70.3	80.9	82.1	86.0
	ViT-g/14	**49.0**	53.0	**71.3**	**81.0**	**83.0**	**86.2**

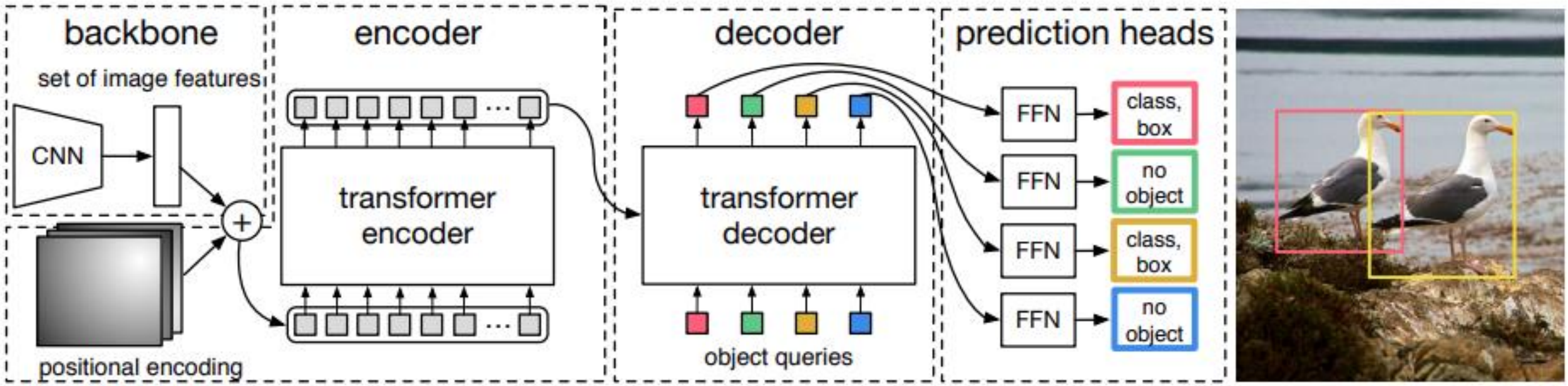
Segmentation results (mIoU) obtained on ADE20K, Cityscapes and Pascal VOC with frozen features extracted DINOv2 and alternatives with a linear classifier. Results with theMask2Former (M2F) pipeline are also shown.

Method	Arch.	NYUd (0.330) lin. 1	NYUd (0.330) lin. 4	NYUd (0.330) DPT	KITTI (2.10) lin. 1	KITTI (2.10) lin. 4	KITTI (2.10) DPT	NYUd → SUN-RGBd (0.421) lin. 1	NYUd → SUN-RGBd (0.421) lin. 4	NYUd → SUN-RGBd (0.421) DPT
OpenCLIP	ViT-G/14	0.541	0.510	0.414	3.57	3.21	2.56	0.537	0.476	0.408
MAE	ViT-H/14	0.517	0.483	0.415	3.66	3.26	2.59	0.545	0.523	0.506
DINO	ViT-B/8	0.555	0.539	0.492	3.81	3.56	2.74	0.553	0.541	0.520
iBOT	ViT-L/16	0.417	0.387	0.358	3.31	3.07	2.55	0.447	0.435	0.426
DINOv2	ViT-S/14	0.449	0.417	0.356	3.10	2.86	2.34	0.477	0.431	0.409
	ViT-B/14	0.399	0.362	0.317	2.90	2.59	2.23	0.448	0.400	0.377
	ViT-L/14	0.384	0.333	0.293	2.78	2.50	2.14	0.429	0.396	0.360
	ViT-g/14	**0.344**	**0.298**	**0.279**	**2.62**	**2.35**	**2.11**	**0.402**	**0.362**	**0.338**

Depth estimation results (RMSE, lower is better) on NYU Depth, KITTI and SUN RGB-D with a linear classifier on top of one or four transformer layers and with a DPT decoder. Underlined results outperform the state of the art.

[Caron et al, Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021]
[Oquab et al, DINOv2: Learning Robust Visual Features without Supervision, TMLR 2024]

DETR


backbone
set of image features
CNN
positional encoding
encoder
transformer encoder
decoder
transformer decoder
object queries
prediction heads
FFN
FFN
FFN
FFN
class, box
no object
class, box
no object

Grounding DINO

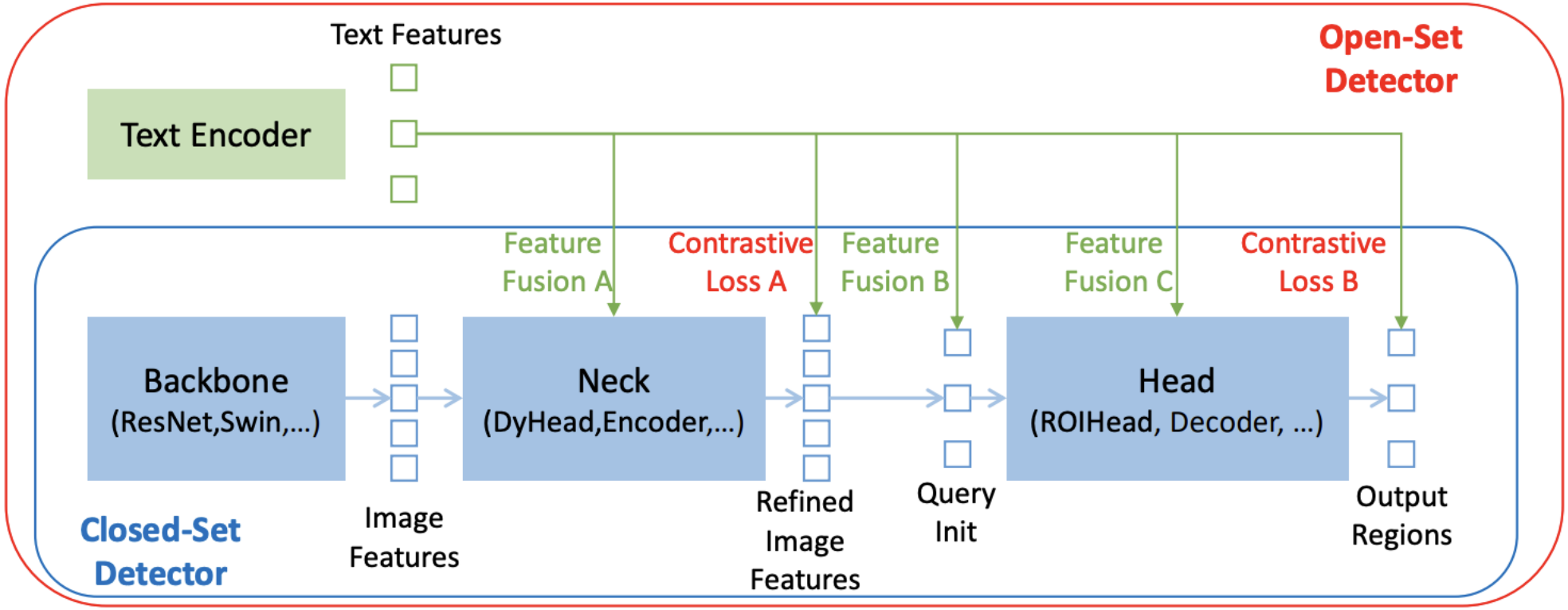
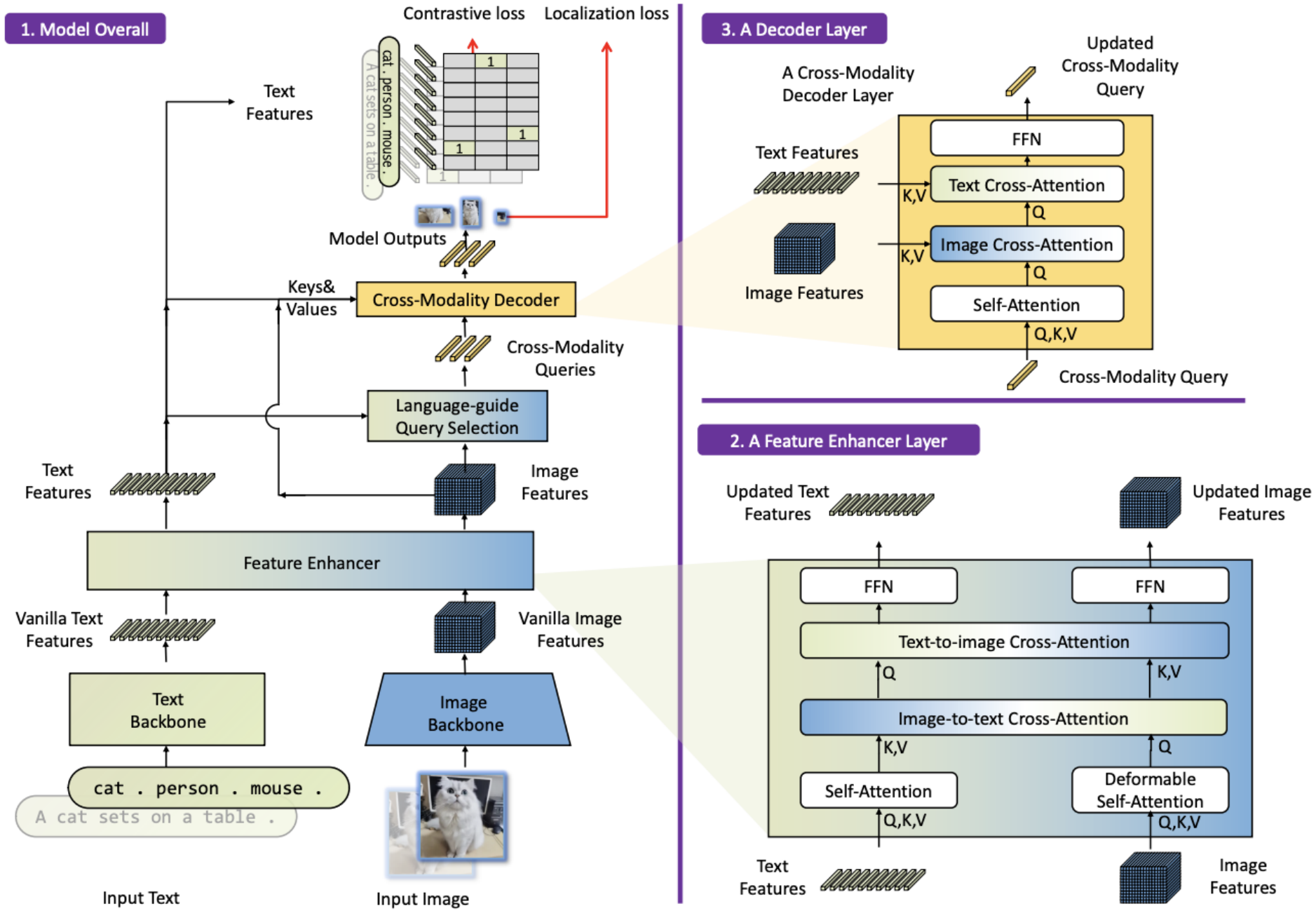


Fig. 2: Extending closed-set detectors to open-set scenarios.

[Liu et al, ECCV 2024]

Grounding DINO



[Liu et al, ECCV 2024]

Multimodal tasks (=Vision and Language Tasks)

- Multimodal Classification
- Image-Text Retrieval
- Visual Grounding
- **Image Captioning**
- Visual Question Answering and Visual Reasoning
- Text-to-image Generation

Image captioning

Image Captioning (Paragraph)
Caption: There is a white dog lying on a grass field. There are a lot of leaves on the grass field. There is a chain-link fence next to the dog. There is a red frisbee under the dog's left-front paw.

Image Captioning (Single Sentence)
Caption: A dog tries to catch a yellow, flying frisbee.

Generation

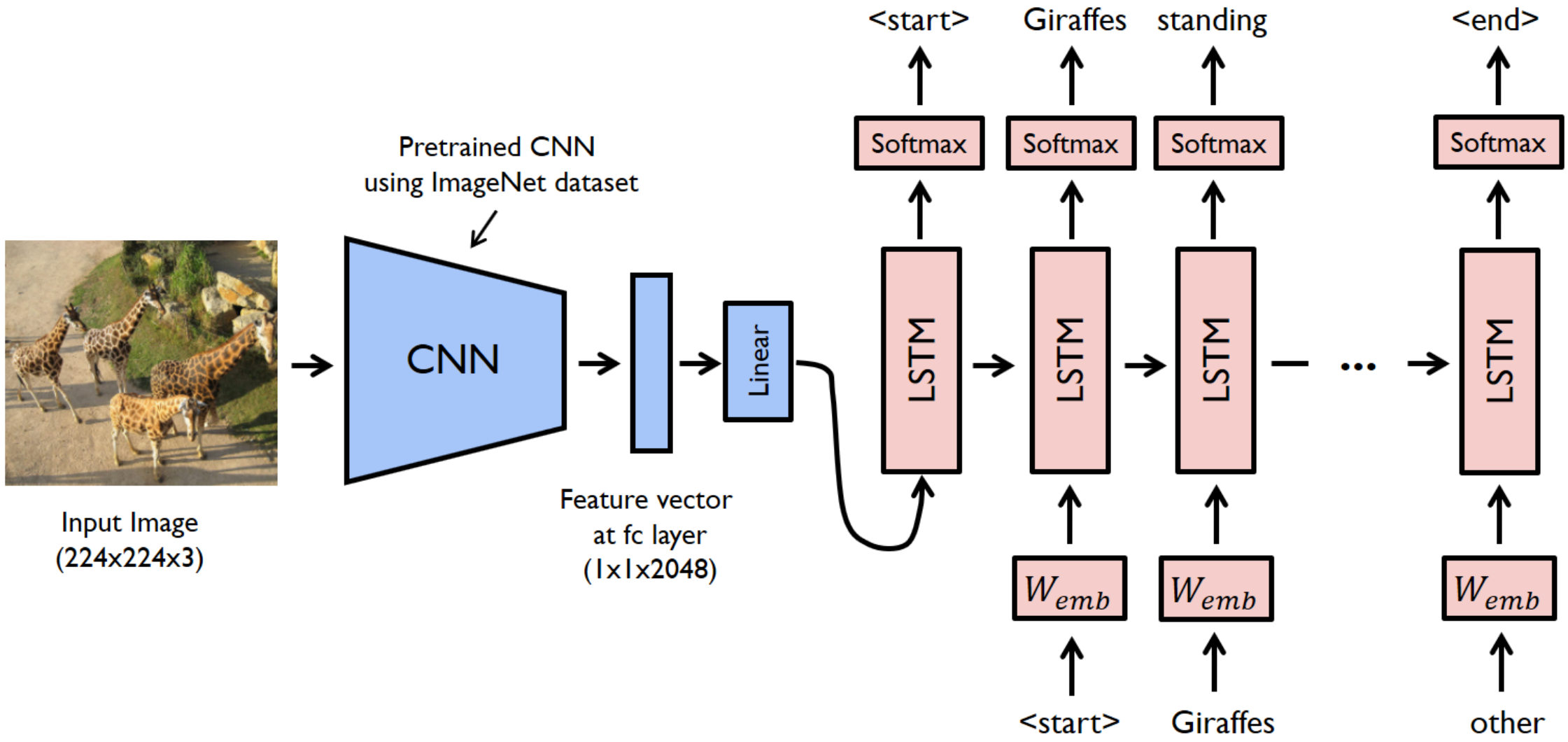
Image captioning

Image Captioning (Paragraph)
Caption: There is a white dog lying on a grass field. There are a lot of leaves on the grass field. There is a chain-link fence next to the dog. There is a red frisbee under the dog's left-front paw.

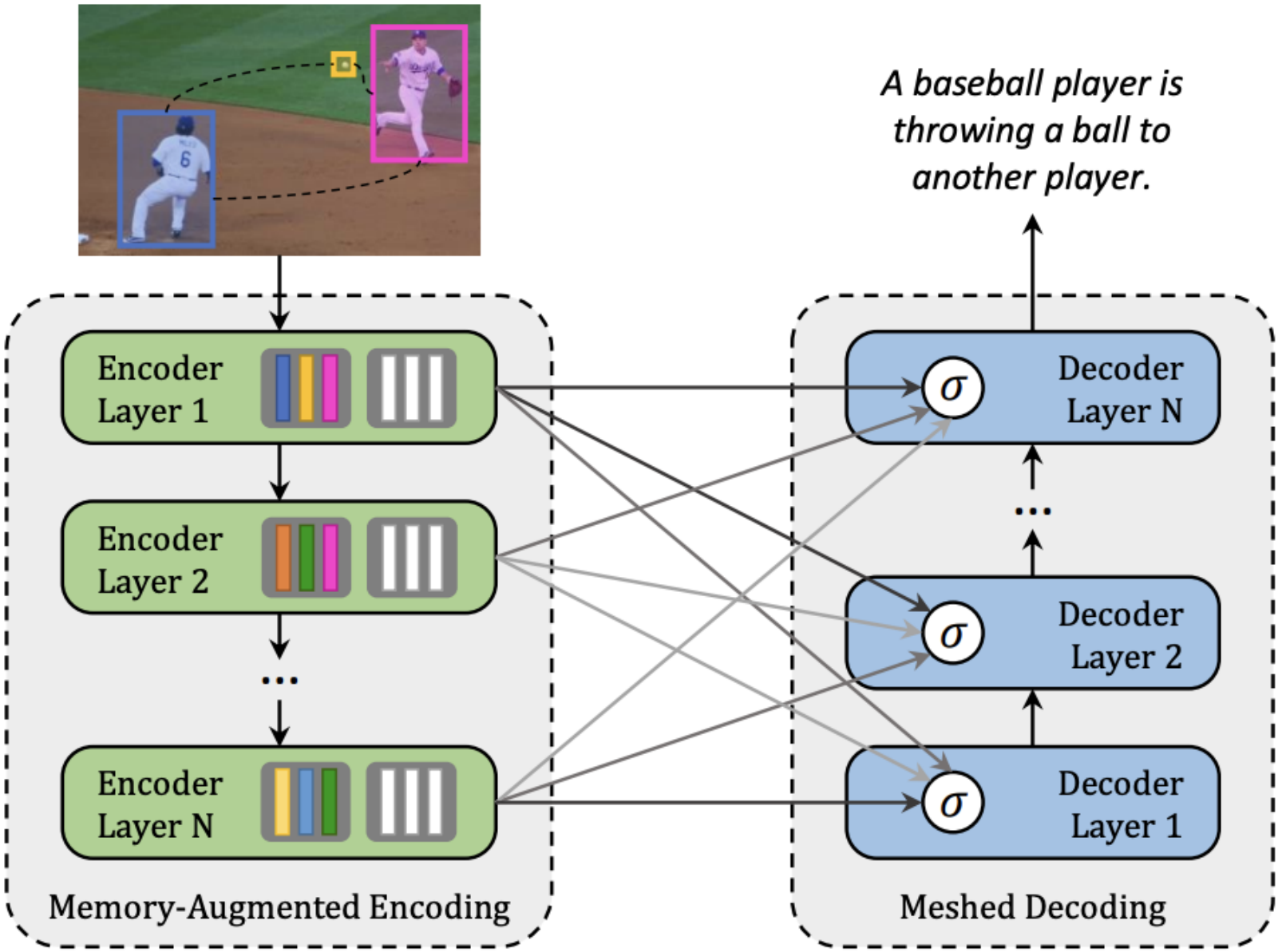
Image Captioning (Single Sentence)
Caption: A dog tries to catch a yellow, flying frisbee.

- **Inputs:**
  - Image
  - Pre-trained image feature extractor (optional): A pre-trained neural network that can extract meaningful features from images, such as CNN.
- **Outputs:** Textual captions: Single Sentence or Paragraph that accurately describe the content of the input images, capturing objects, actions, relationships, and overall context.
- **Task:** To automatically generate natural language descriptions of images. This involves: (1) Understanding the visual content of the image (objects, actions, relationships). (2) Encoding this information into a meaningful representation. (3) Decoding this representation into a coherent, grammatically correct, and informative sentence or phrase.

In the past…

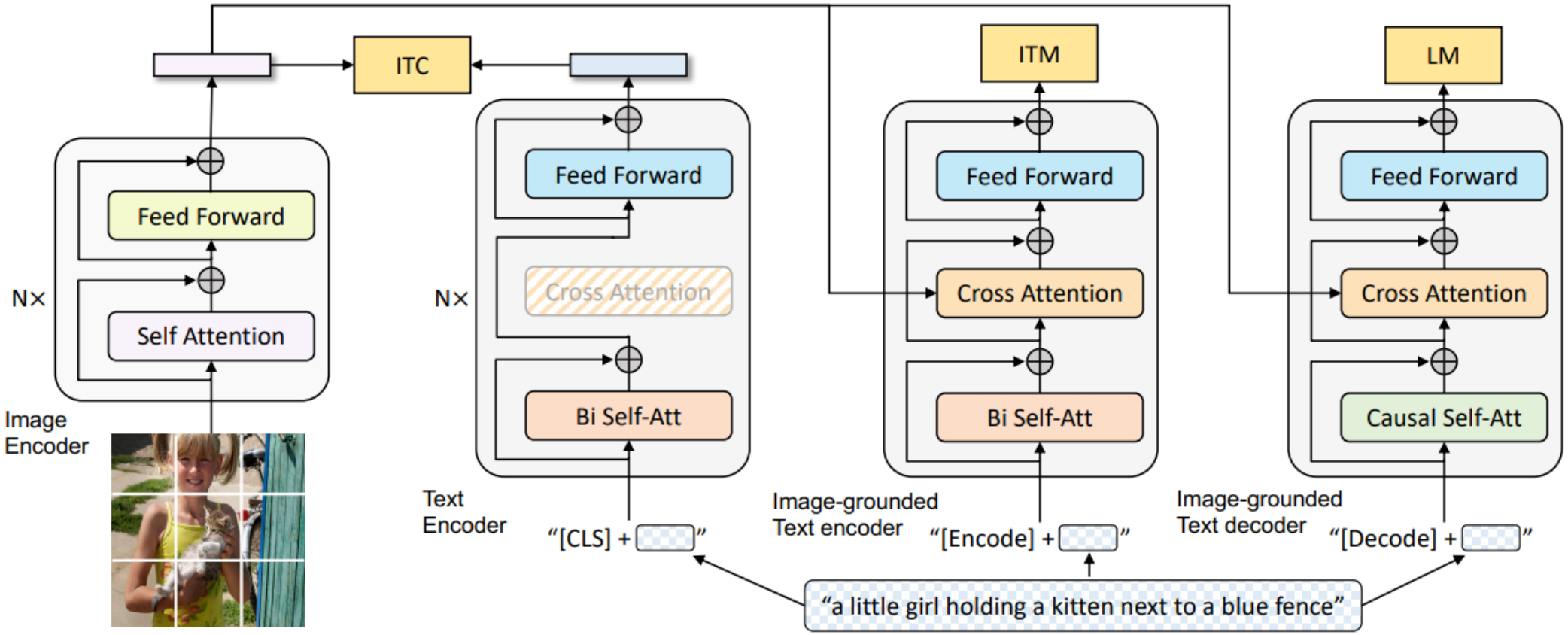

Pretrained CNN
using ImageNet dataset
CNN
Input Image
(224x224x3)
Feature vector
at fc layer
(1x1x2048)
Linear
LSTM
LSTM
LSTM
LSTM
Softmax
Softmax
Softmax
Softmax
<start>
Giraffes
standing
<end>
W_{emb}
W_{emb}
W_{emb}
<start>
Giraffes
other

Meshes Memory Transformer



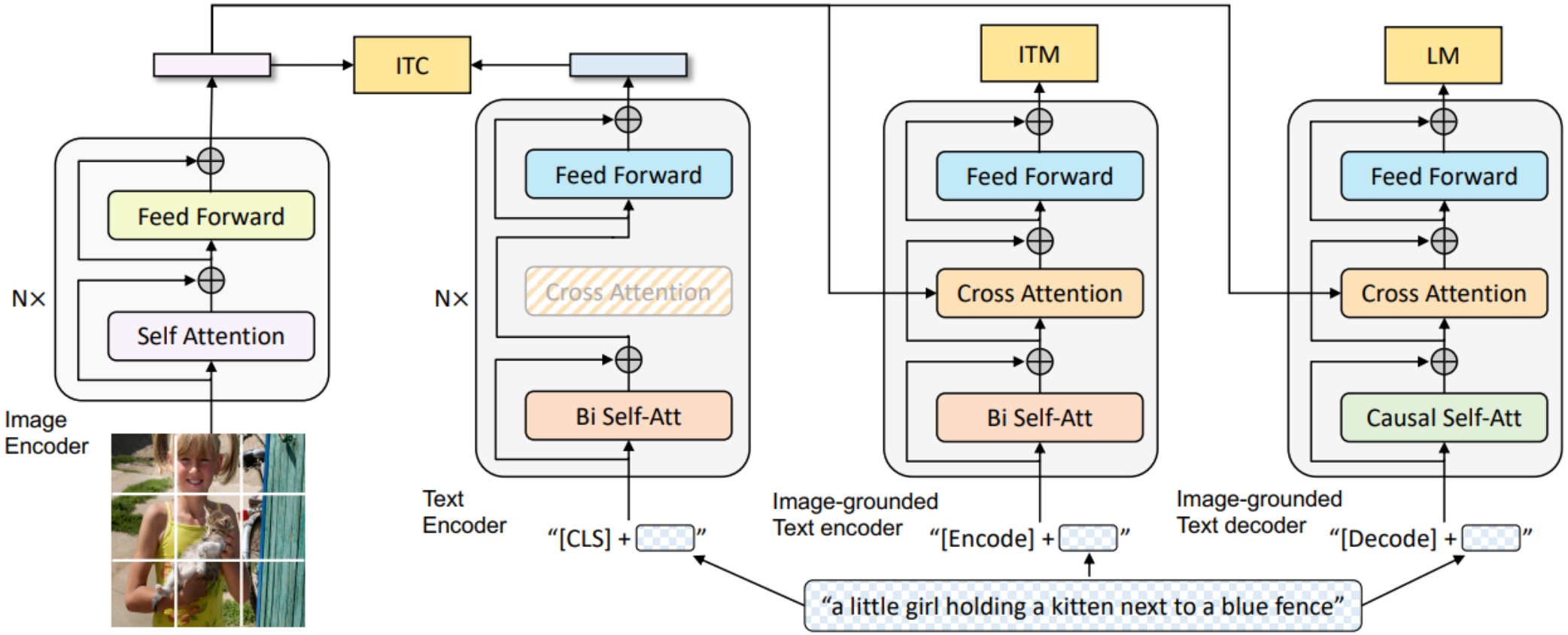
[Cornia CVPR 2020]

BLIP



[Li et al, BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, ICML 2022]

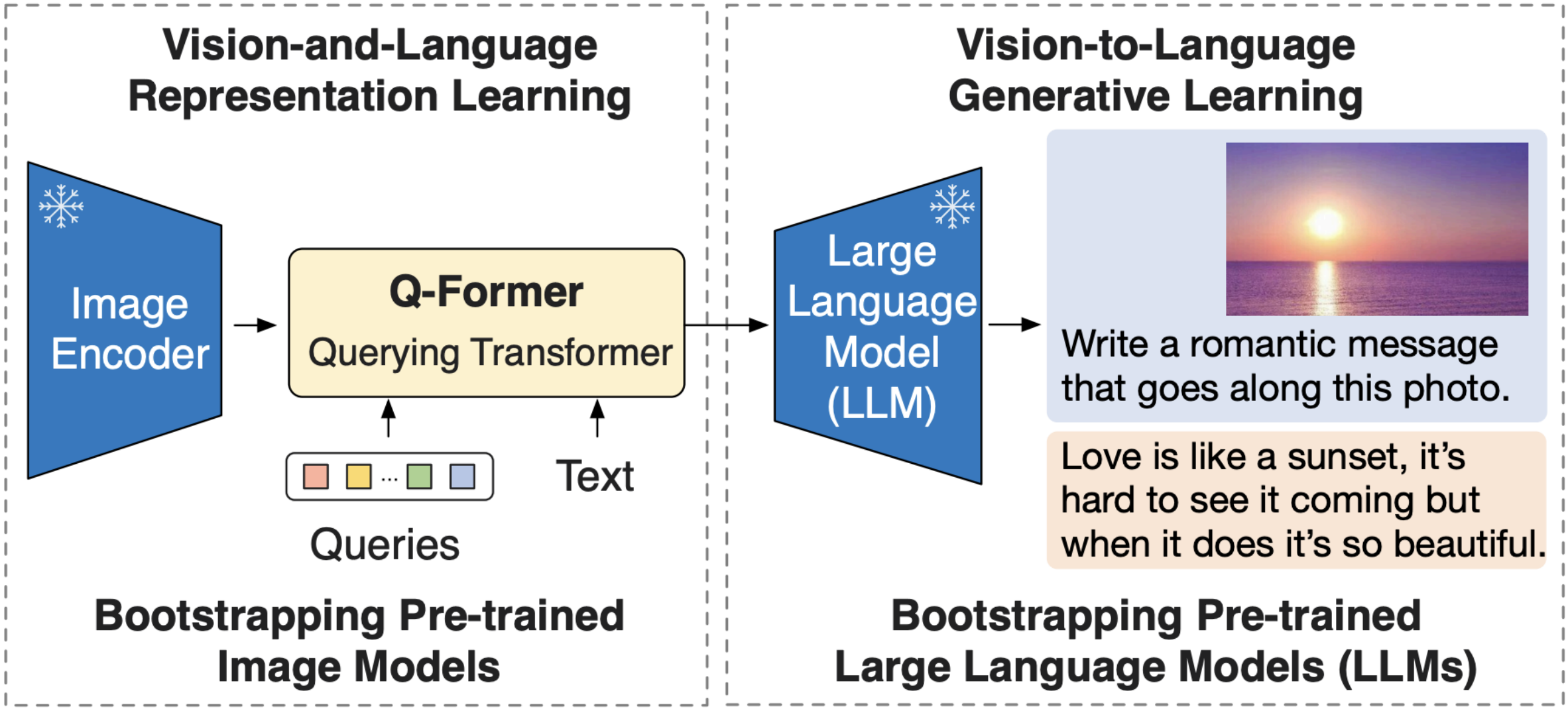
BLIP



- Unimodal encoder is trained with an **image-text contrastive (ITC) loss** to align the vision and language representations
- Image-grounded text encoder uses additional cross-attention layers to model vision-language interactions and is trained with an **image-text matching (ITM) loss** to distinguish between positive and negative image-text pairs.
- Image-grounded text decoder replaces the bi-directional self-attention layers with causal self-attention layers and shares the same cross-attention layers and feed forward networks as the encoder. The decoder is trained with a **language modeling (LM) loss** to generate captions given images.

[Li et al, BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, ICML 2022]

BLIP-2



[Li et al, BLIP-2, ICML 2023]

Multimodal tasks (=Vision and Language Tasks)

- Multimodal Classification
- Image-Text Retrieval
- Visual Grounding
- Image Captioning
- **Visual Question Answering and Visual Reasoning**
- Text-to-image Generation

VQA and Visual Reasoning

Visual Question Answering
Q: What is the dog holding with its paws?
A: Frisbee.

Visual Reasoning
Q: Is the dog in the air **AND** is the frisbee in the air?
A: Yes

Generation

VQA and Visual Reasoning

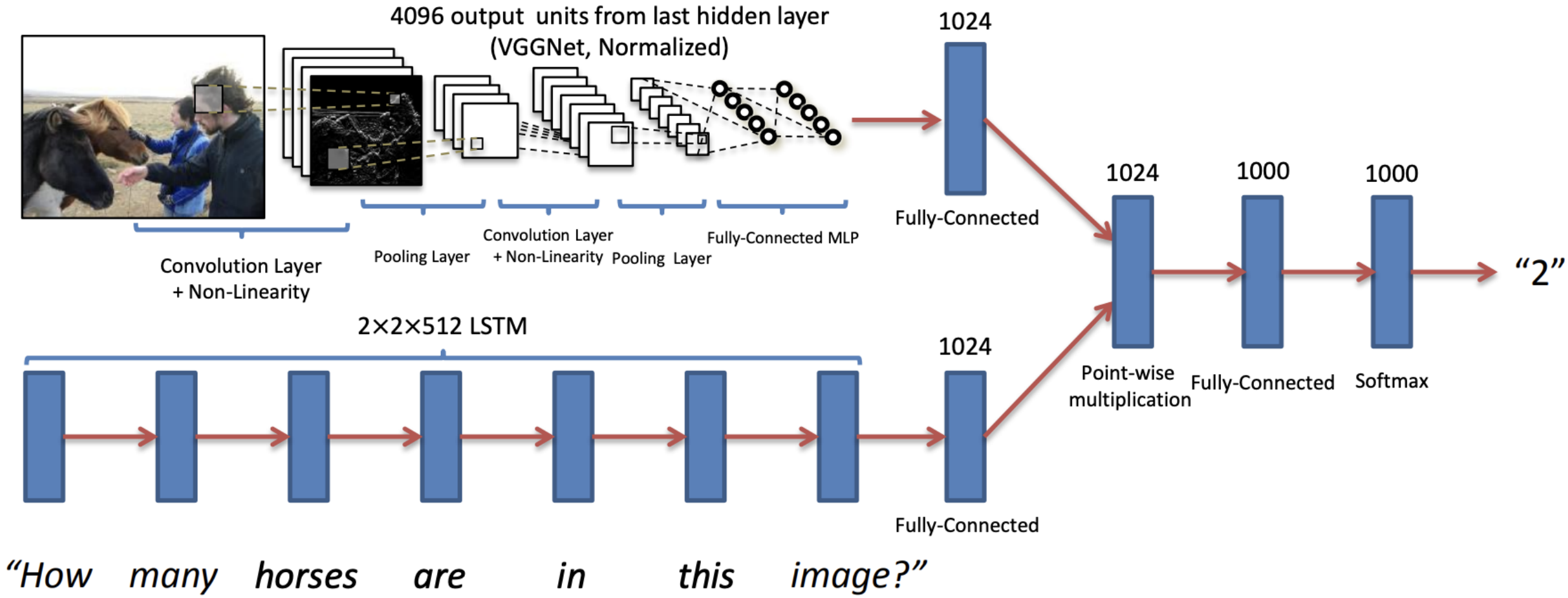
Visual Question Answering
Q: What is the dog holding with its paws?
A: Frisbee.

- **Input:** An image-question pair
- **Output:** In multiple-choice setting: A label corresponding to the correct answer among pre-defined choices. In open-ended setting: A free-form natural language answer based on the image and question.
- **Task:** Answer questions about images. (Most VQA models treat as a classification problem with pre-defined answers)

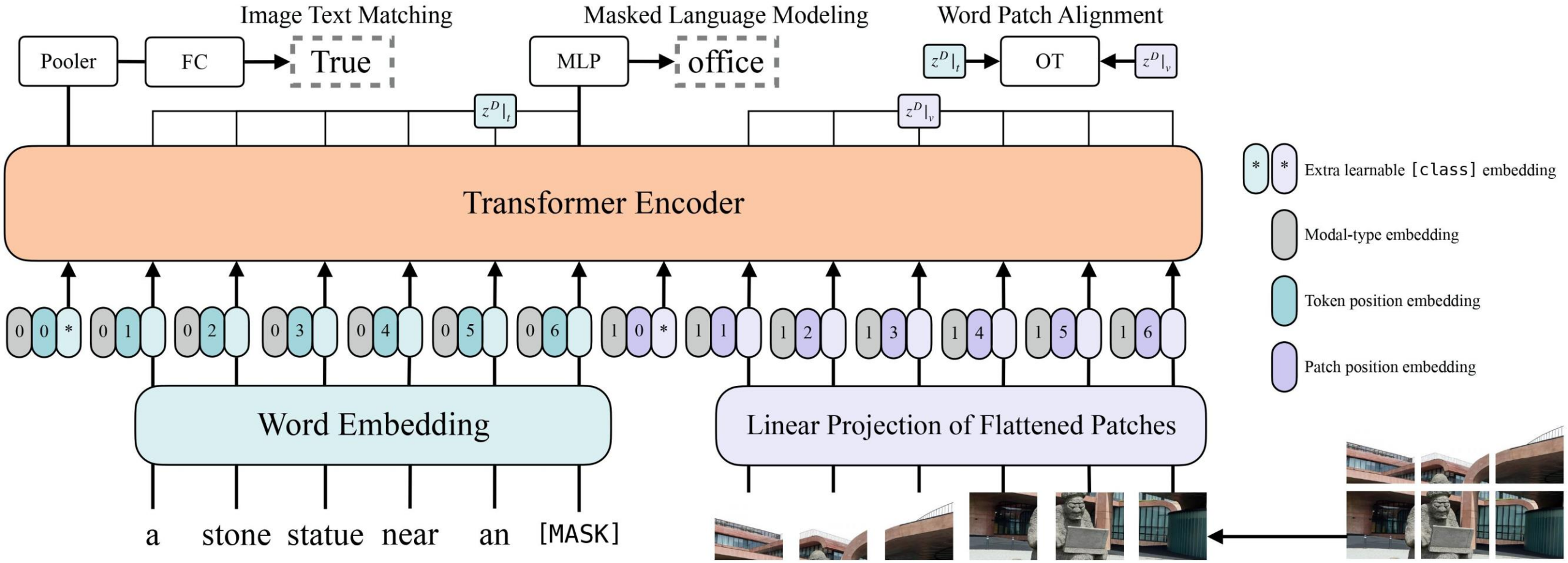
Visual Reasoning
Q: Is the dog in the air **AND** is the frisbee in the air?
A: Yes

- **Output:** Varies depending on the task:
  - VQA: Answers to questions about the image.
  - Matching: True/False for whether the text is true about the image(s).
  - Entailment: Prediction of whether the image semantically entails the text.
  - Sub-question: Answers to the sub-questions related to perception.
- **Task:** Performs various reasoning tasks on images.

In the past…

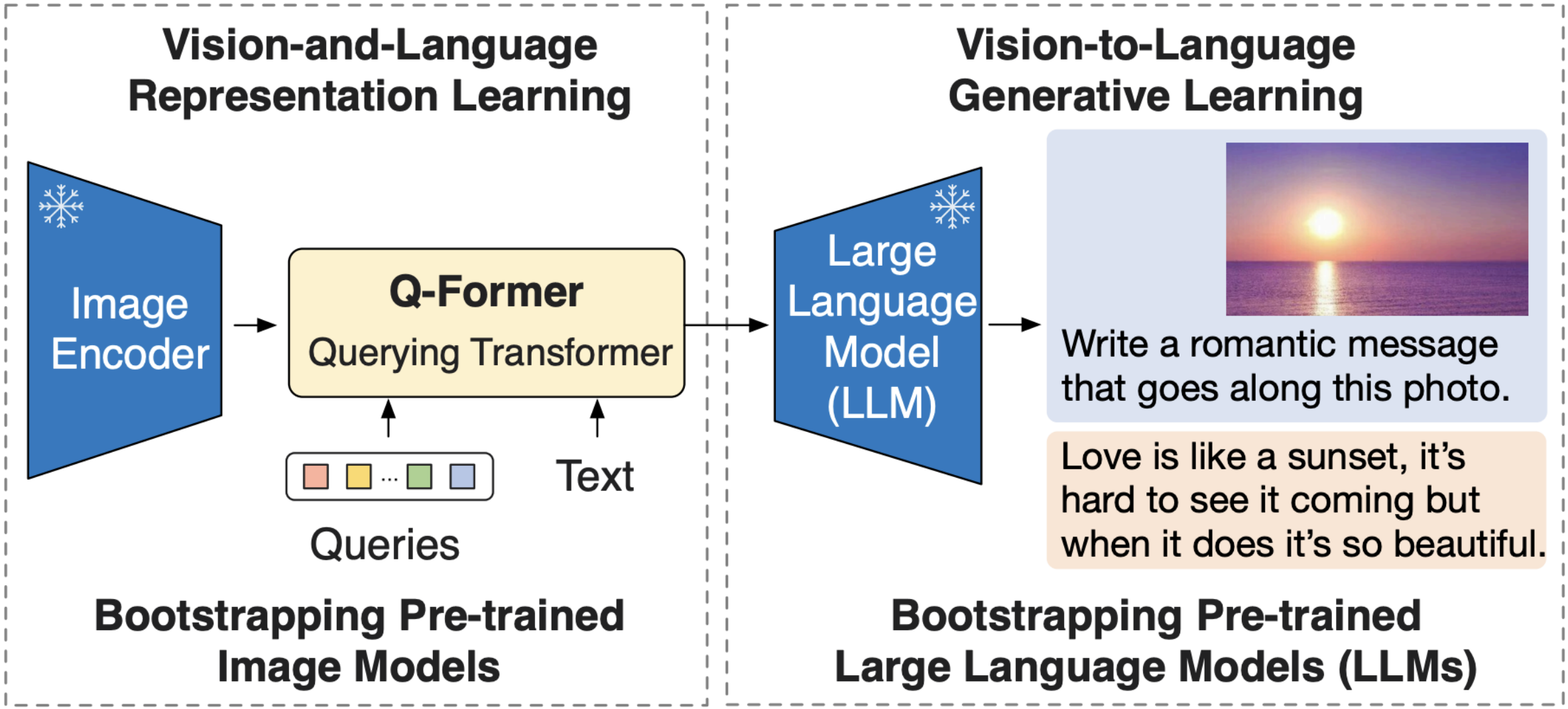

4096 output units from last hidden layer
(VGGNet, Normalized)
1024
Fully-Connected
Convolution Layer
+ Non-Linearity
Pooling Layer
Convolution Layer
+ Non-Linearity
Pooling Layer
Fully-Connected MLP
2×2×512 LSTM
1024
Fully-Connected
1024
1000
1000
Point-wise
multiplication
Fully-Connected
Softmax
"2"
"How many horses are in this image?"

VILT



[Kim et al, ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision, ICML 2021]

BLIP-2



[Li et al, BLIP-2, ICML 2023]

Multimodal tasks (=Vision and Language Tasks)

- **Multimodal Classification**
- **Image-Text Retrieval**
- **Visual Grounding**
- **Image Captioning**
- **Visual Question Answering and Visual Reasoning**
- Text-to-image Generation

Pretraining and Downstream

OD, CNN, ViT, Patch Embedding

Image encoder

A dog lying on the grass next to a frisbee

BERT, RoBERTa, Word Embedding

Text encoder

Merged Attention/ Co-attention/ Dot-product

Multimodal Fusion

Decoder (optional)

Pre-training Objectives

- Masked Language Modeling
- Masked Image Modeling
- Image-Text Matching
- Image-Text Contrastive Learning

Downstream Tasks

- Visual Question Answering
- Image Captioning
- Image-Text Retrieval
- Phrase Grounding

Leveraging LLMs

ChatGPT

What do you see in the image?

The image shows a large group of people gathered outdoors on a grassy area. They are standing and sitting close together, smiling, and raising their hands in the air, giving a celebratory or group photo vibe. The setting appears casual and friendly, possibly for an event like a workshop, summer school, team-building day, or community gathering. In the background,

Generate a bounding box around the tree of the image

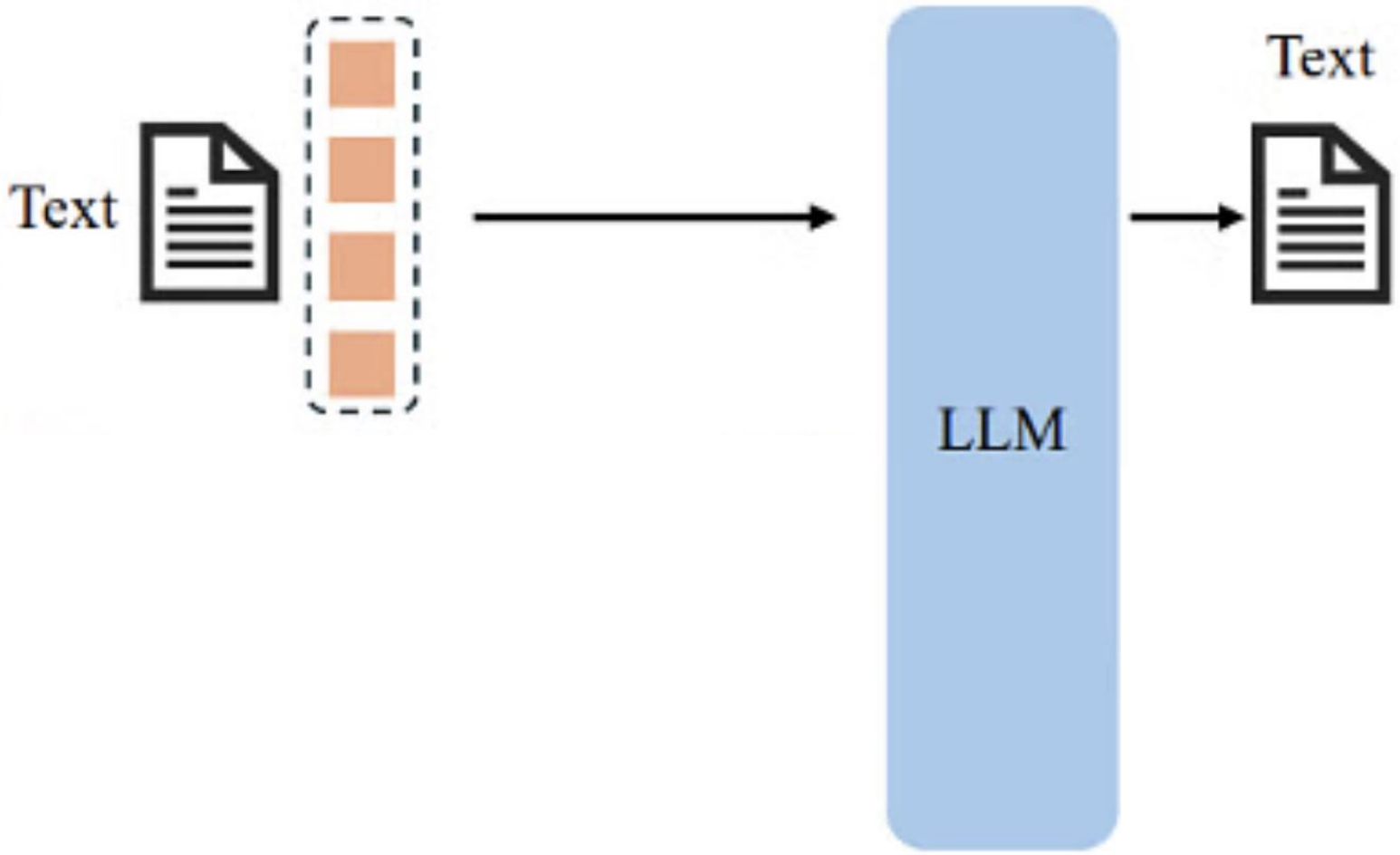
Estimate the number of people in the image

2/2

By counting roughly row by row, it looks like there are about 80–90 people in the image. My closer estimate is around **85 people**.

Generate a similar image with double the amount of people

LLMs



LLMs

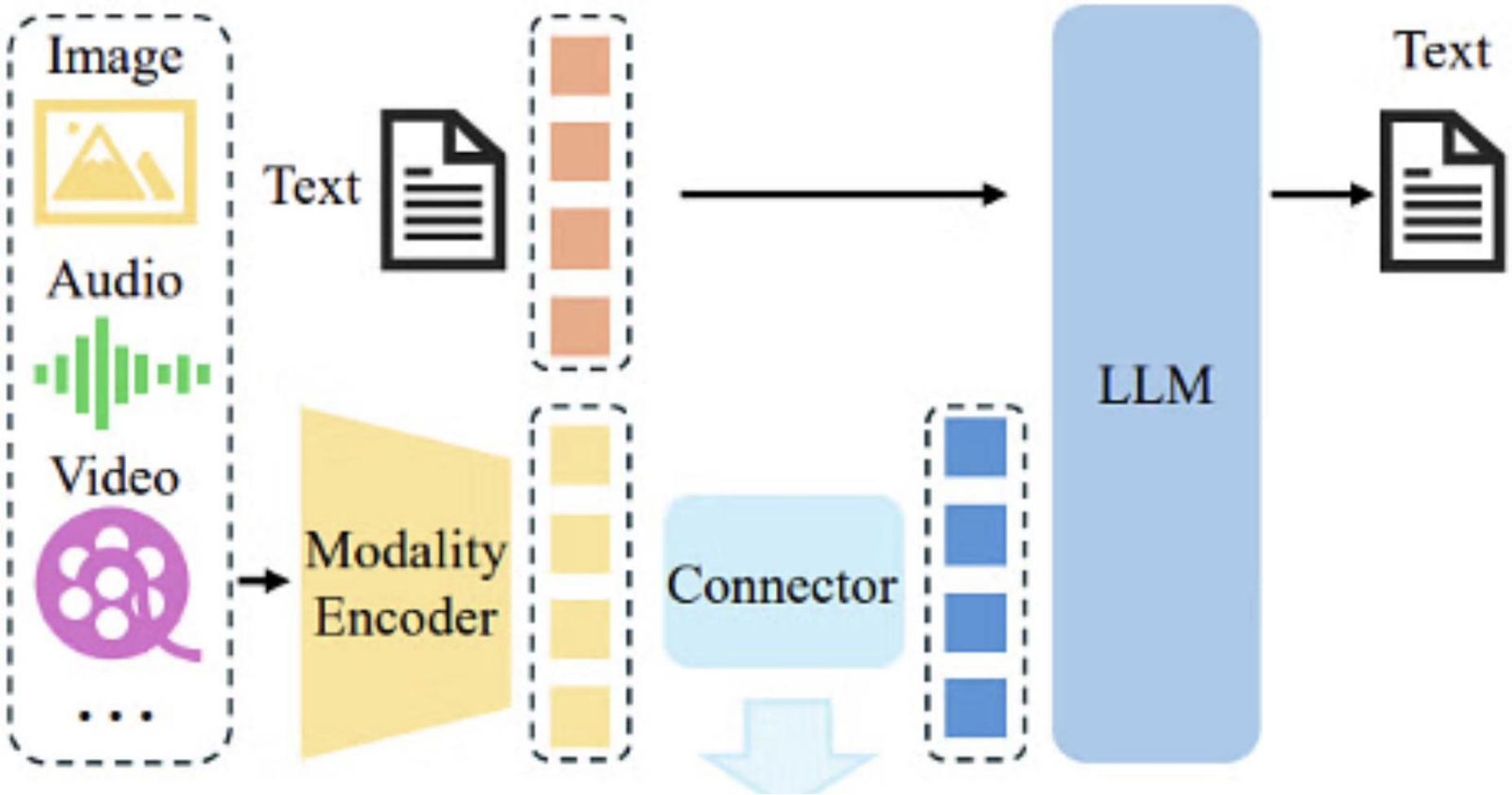
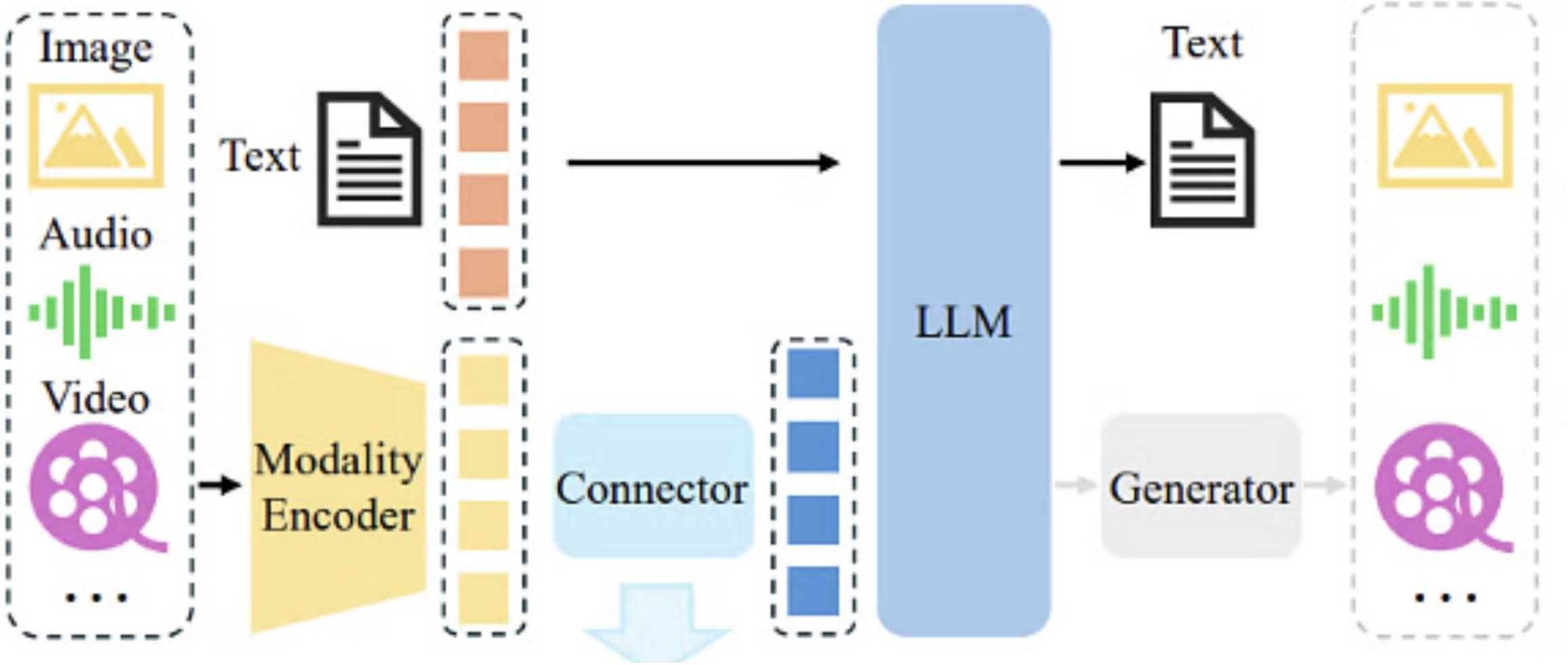
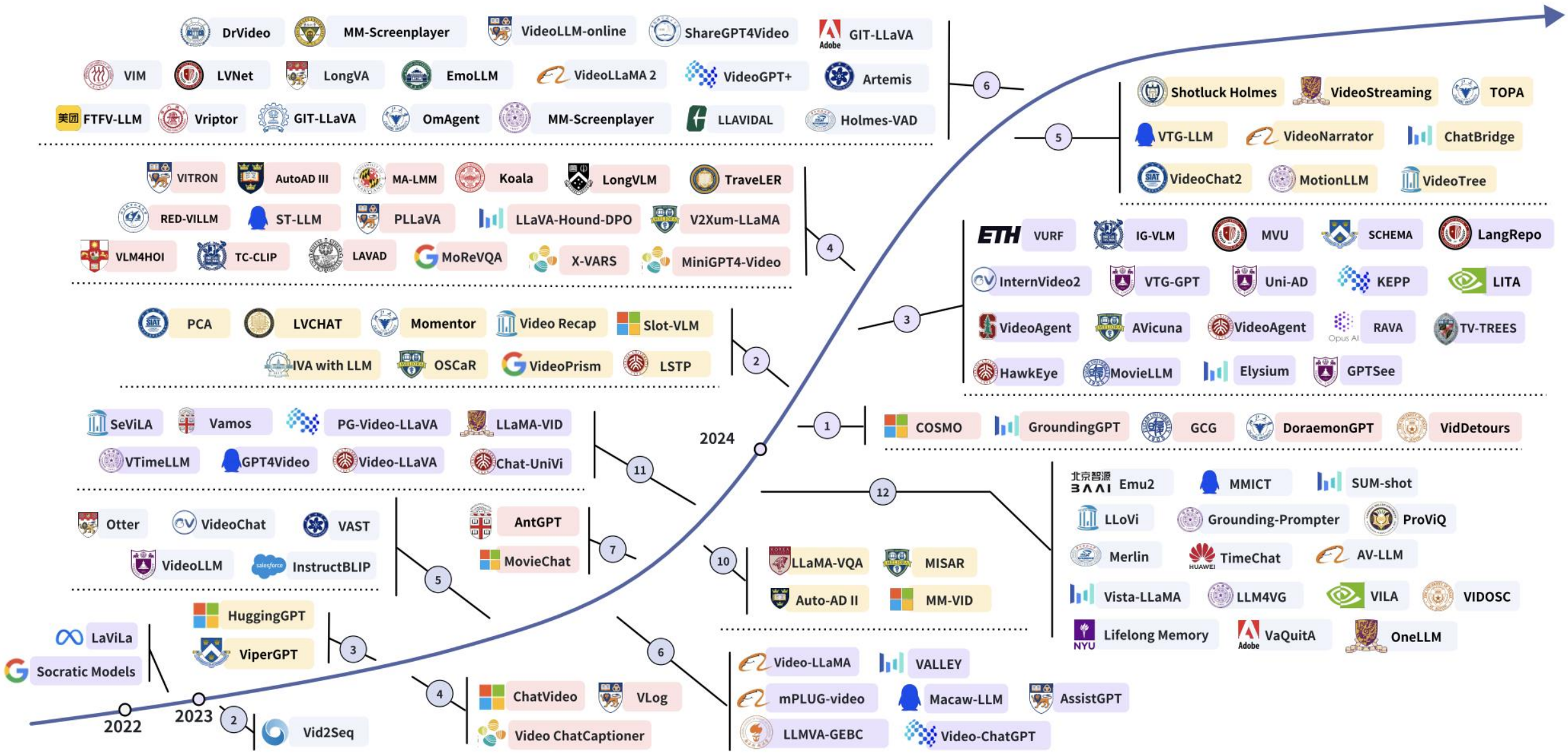


Image
Audio
Video
...
Text
Modality Encoder
Connector
LLM
Text

Leveraging LLMs

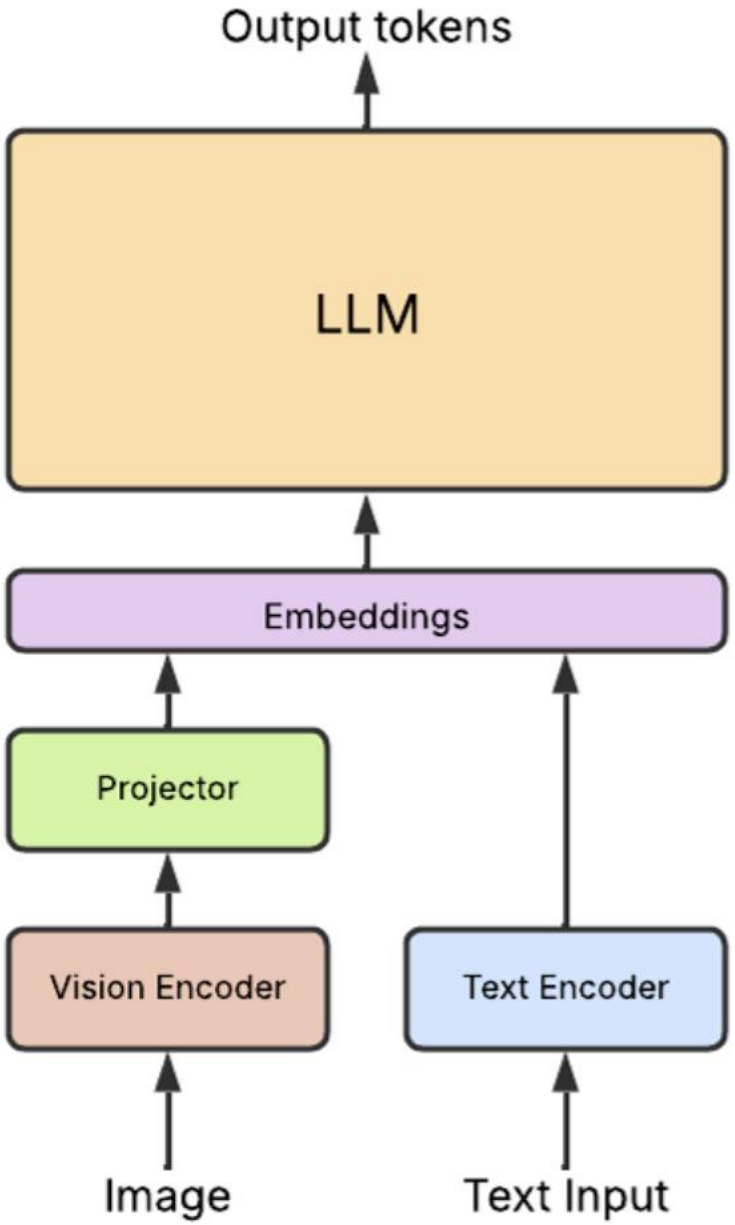


MLLMs (VLMs)



Survey paper: https://arxiv.org/pdf/2312.17432

MLLMs (VLMs)



A typical MLLMs Architecture:

- Text inputs are tokenized
- image inputs are processed through a vision encoder and projector
- Both modalities are mapped into a shared embedding space before being processed by the LLM
- LLM generates output tokens as the final result.

LLava

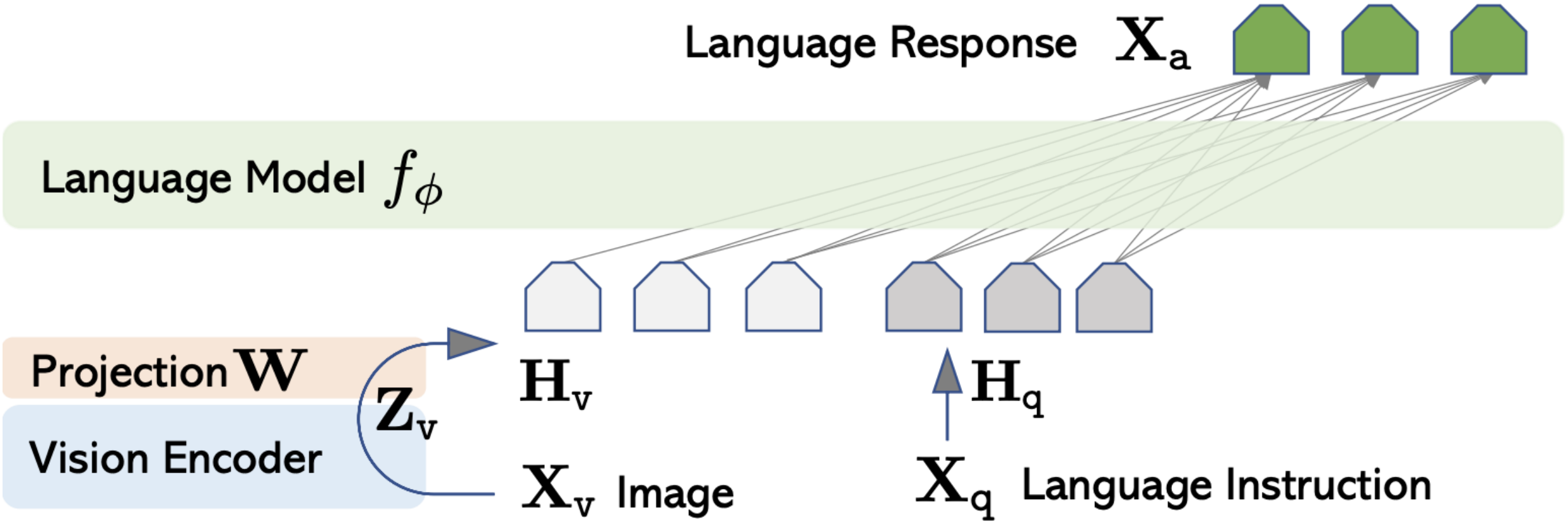


Figure 1: LLaVA network architecture.