Advanced School for Computing and Imaging

University of Amsterdam

Computer Vision by Learning

Cees Snoek, University of Amsterdam

Efstratios Gavves, University of Amsterdam

With an invited tutorial by: Yuki Asano, University of Technology Nuremberg

http://computervisionbylearning.info

amsterdam ai

Abstract

Computer vision has been first revolutionized since the year 2000. Learning from examples became leading. Another revolution happened in 2012, with deep learning from examples. The latest revolution happened in 2022, with the introduction of foundation models.

Progress in computer vision by learning is fast. In the course we will discuss recent methods presented by researchers who are all very active in the field.

The course is supplemented with practical work and is completed with an assignment.

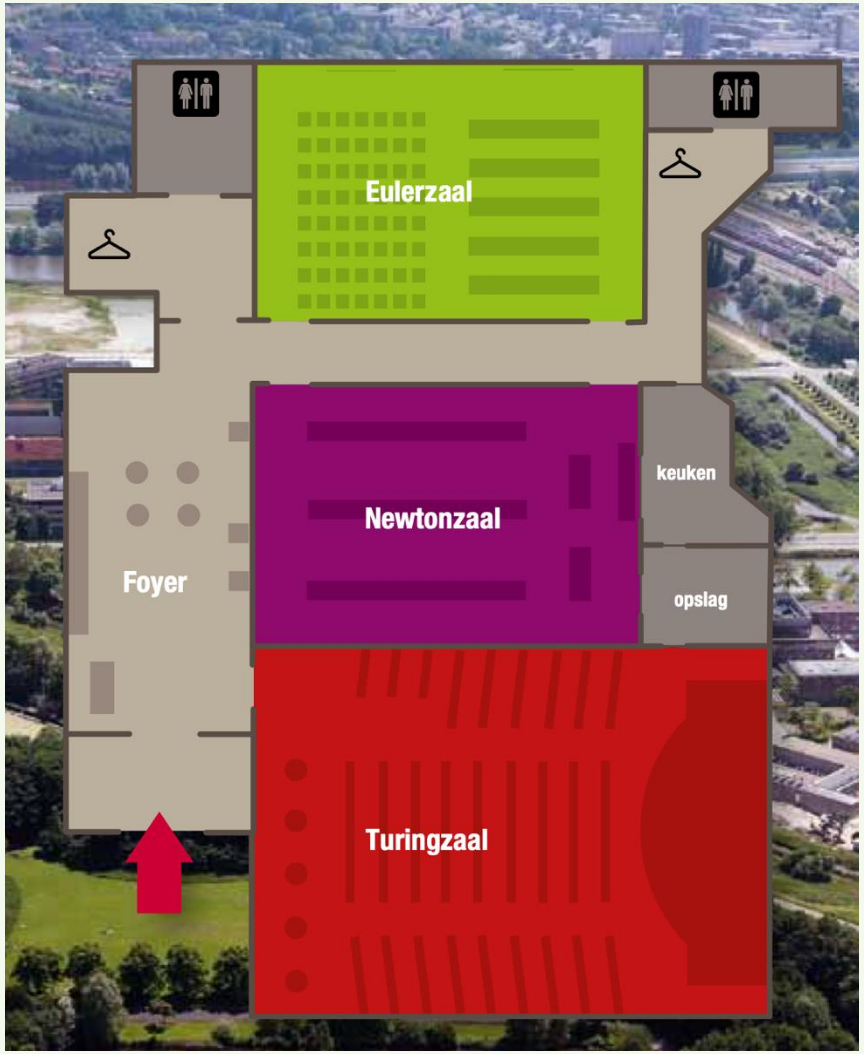
Where and When

Monday 13th to Thursday 16th of January

Lectures	09:30-12:00	Turing
Lunch (included)	12:00-13:30	Newton
Lab	13:30-16:00	Euler

Friday 17th of January

Invited tutorial	09:30-12:00	Turing
Lunch (included)	12:00-13:30	Newton



Program

Monday	Foundations
Tuesday	Machine learning for computer vision
Wednesday	3D vision by learning
Thursday	Computer video by learning
Friday	Invited tutorial by Yuki Asano

Yuki Asano

Guest speakers

Pascal Mettes

Martin Oswald

Dimitris Tzionas

Hazel Doughty

Andrii Zadaianchuk

Lab

Practical 1	Vision by multi-layer perceptron
Practical 2	Vision by convnet
Practical 3	Vision by transformer
Practical 4	Vision by geometric learning **or** Vision by self-supervised learning

TA team every afternoon available for support.

Each **group of 2 students** submits a report about their findings during the practicals. Your report should have roughly 1 page per practical, with a maximum of 8 pages. See lab assignments for all details on format, questions, PyTorch code etc.

Deadline: **January 31th, 2025**

http://computervisionbylearning.info

University of Amsterdam

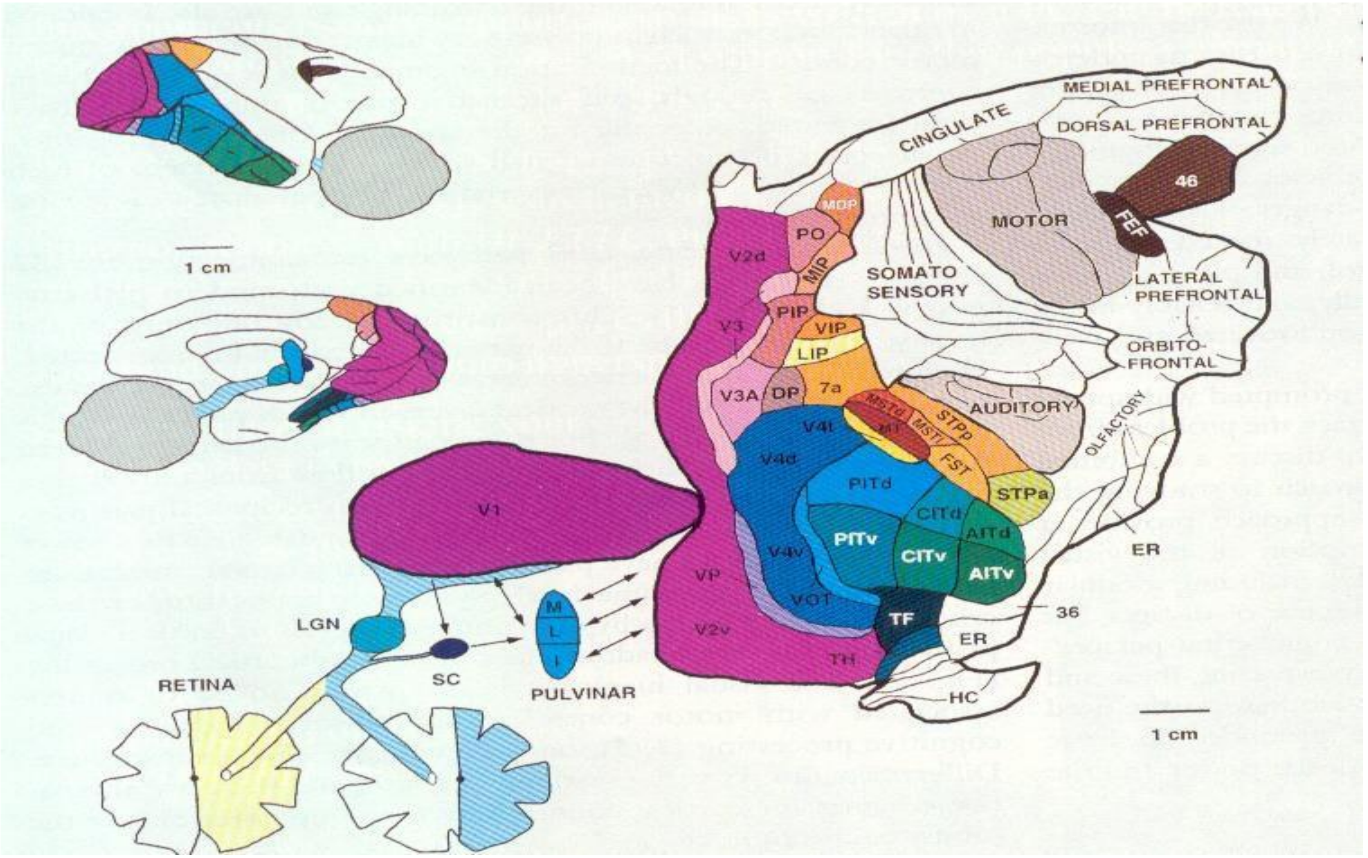
VIS LAB
VIDEO & IMAGE SENSE LAB

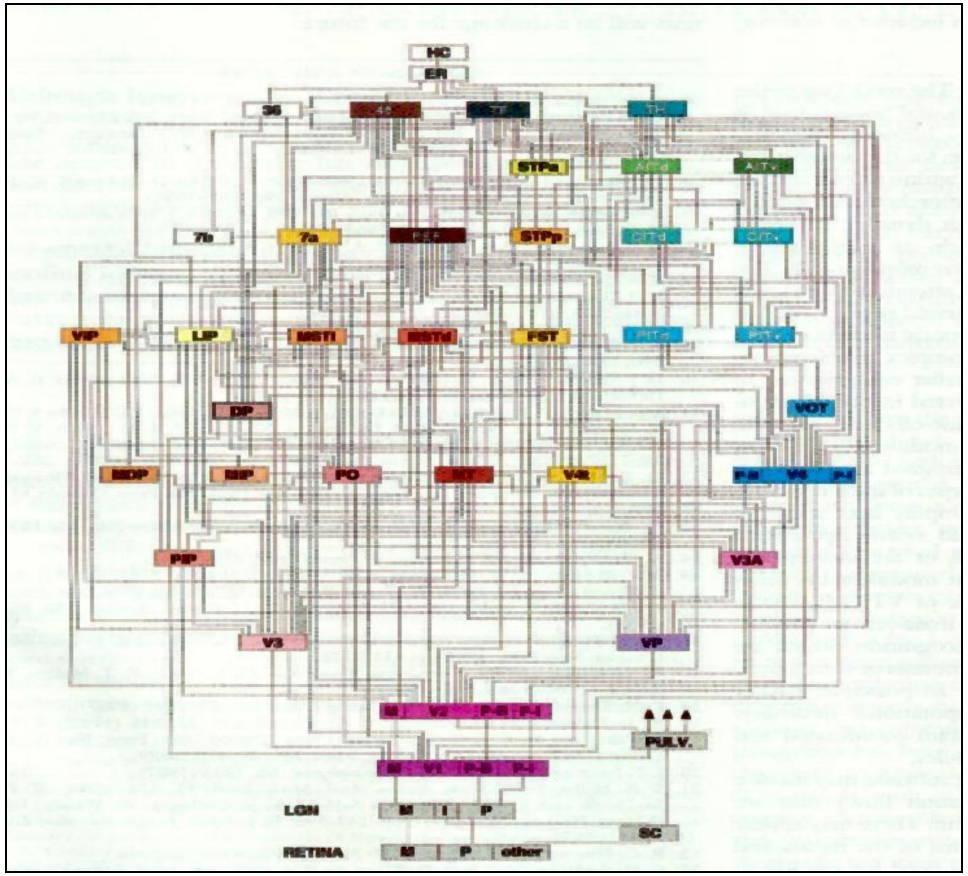
What foundation models *cannot* perceive

Prof. dr. Cees Snoek
University of Amsterdam

Head of Video & Image Sense lab
Scientific Director Amsterdam AI

Human vision consumes 50% brain power

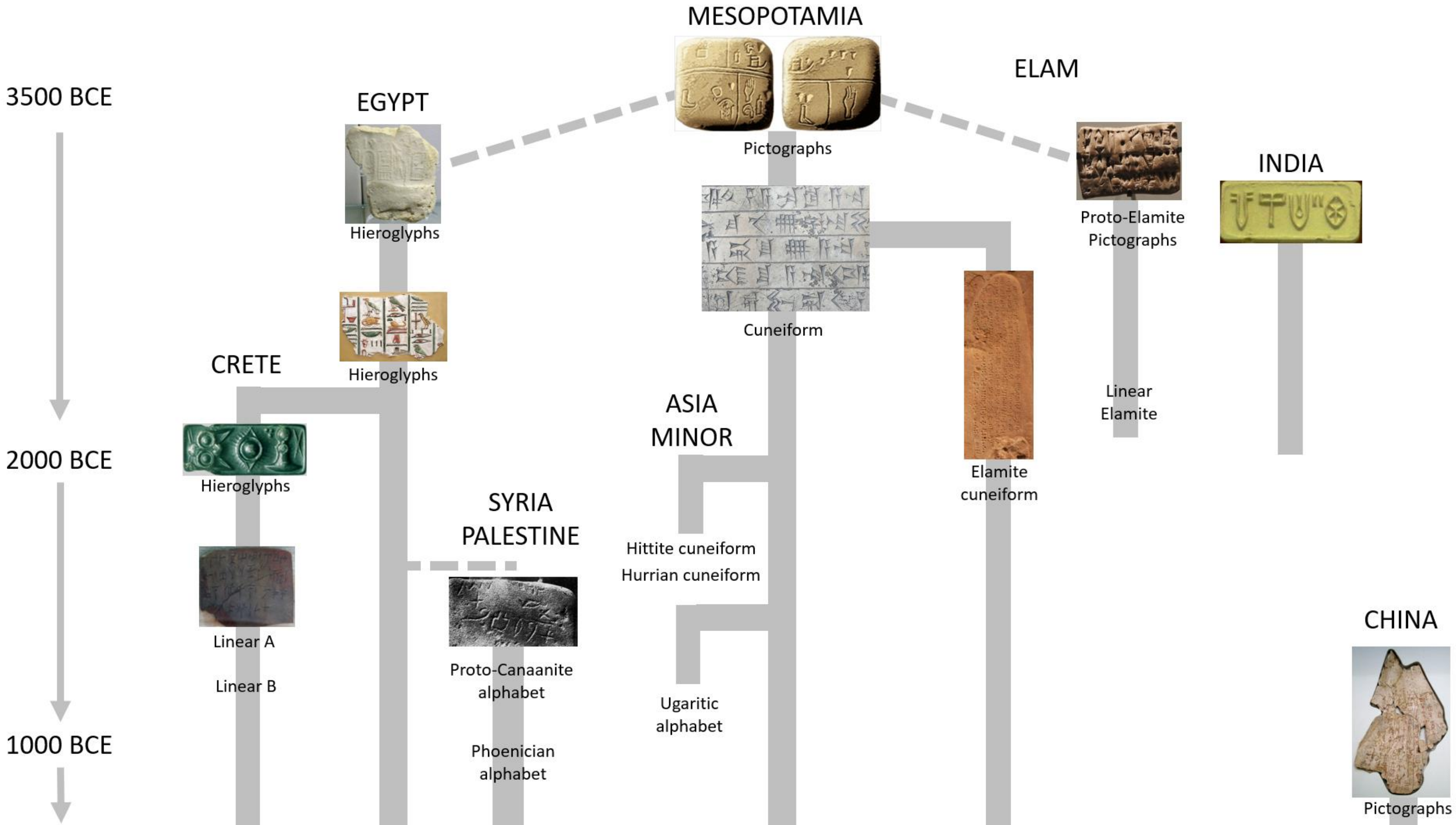




Van Essen, Science 1992

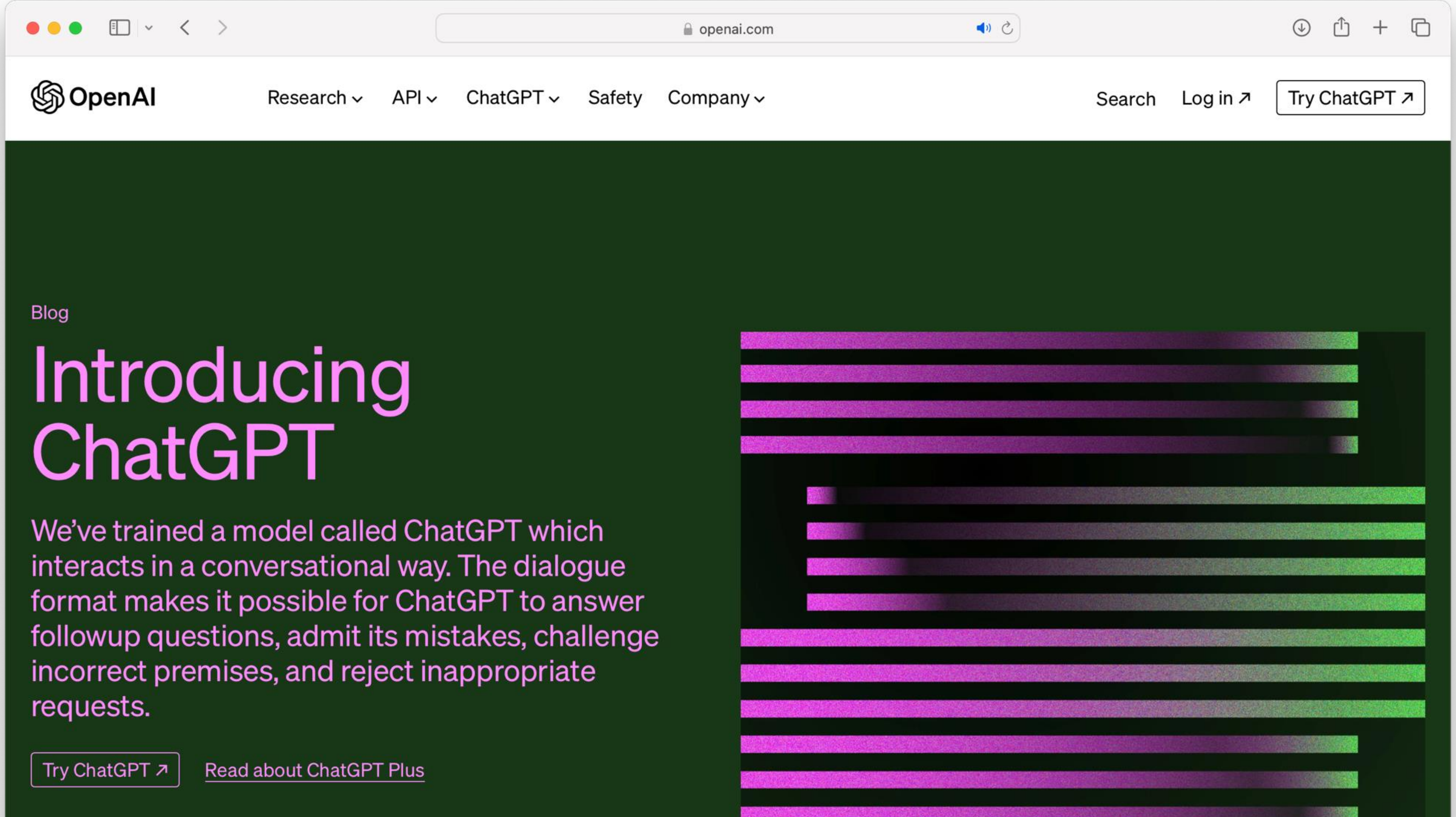
Source: wikipedia

Human invention of written language



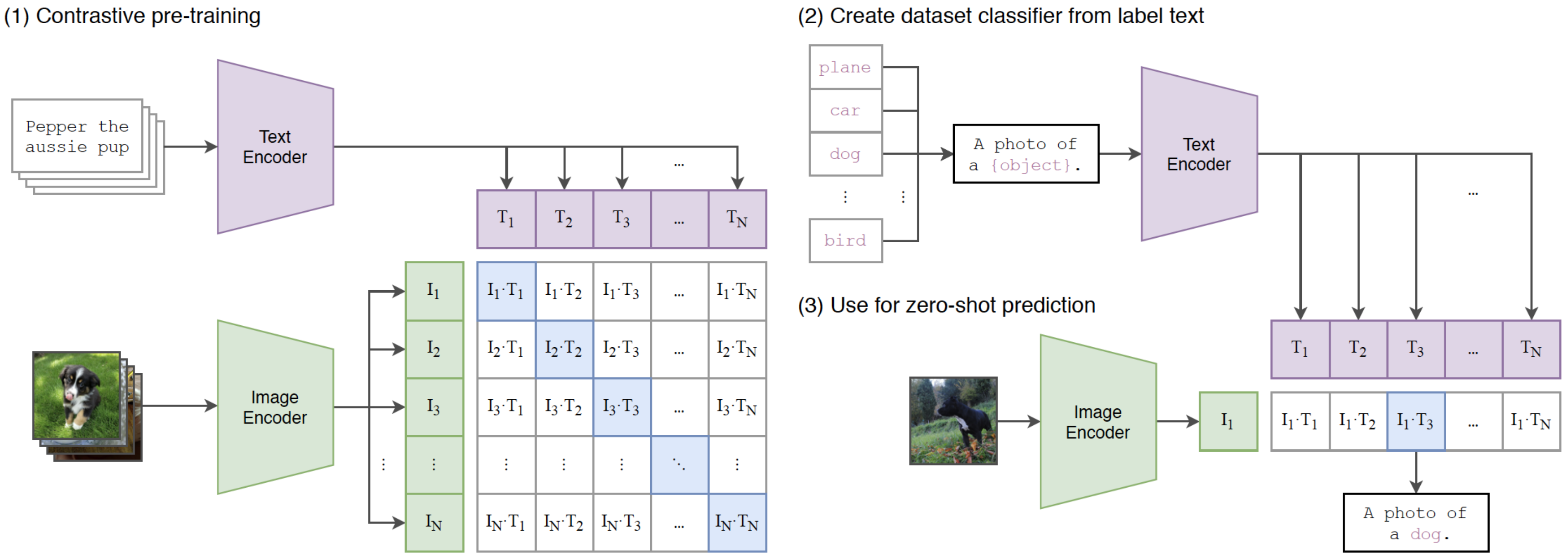
OpenAI, 11/2022

Human invention of ChatGPT



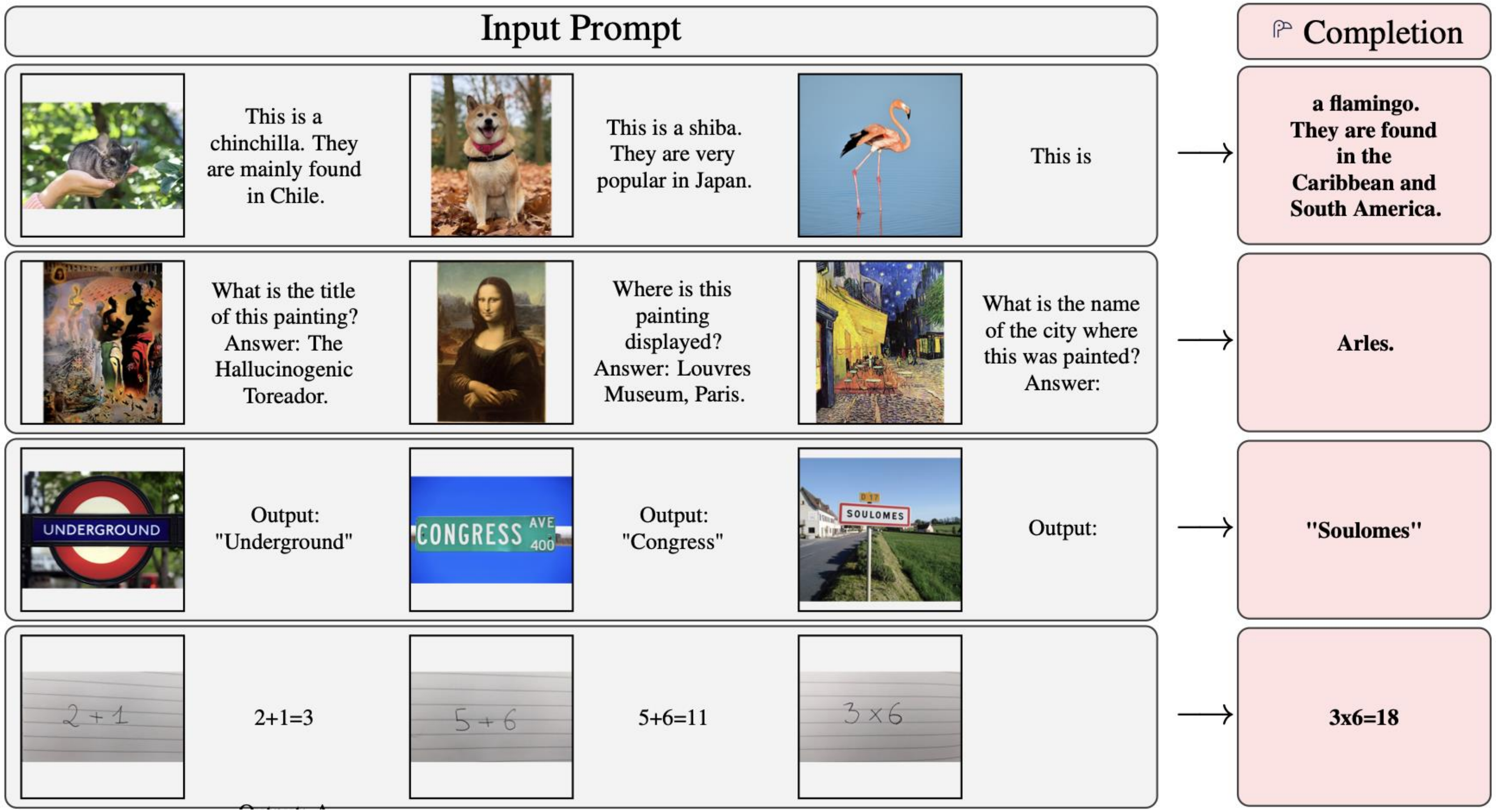
Vision and language even more powerful

1. Collect millions of images and their description from the Internet
2. Learn associations between encoded image and text
3. Amazing zero-shot abilities

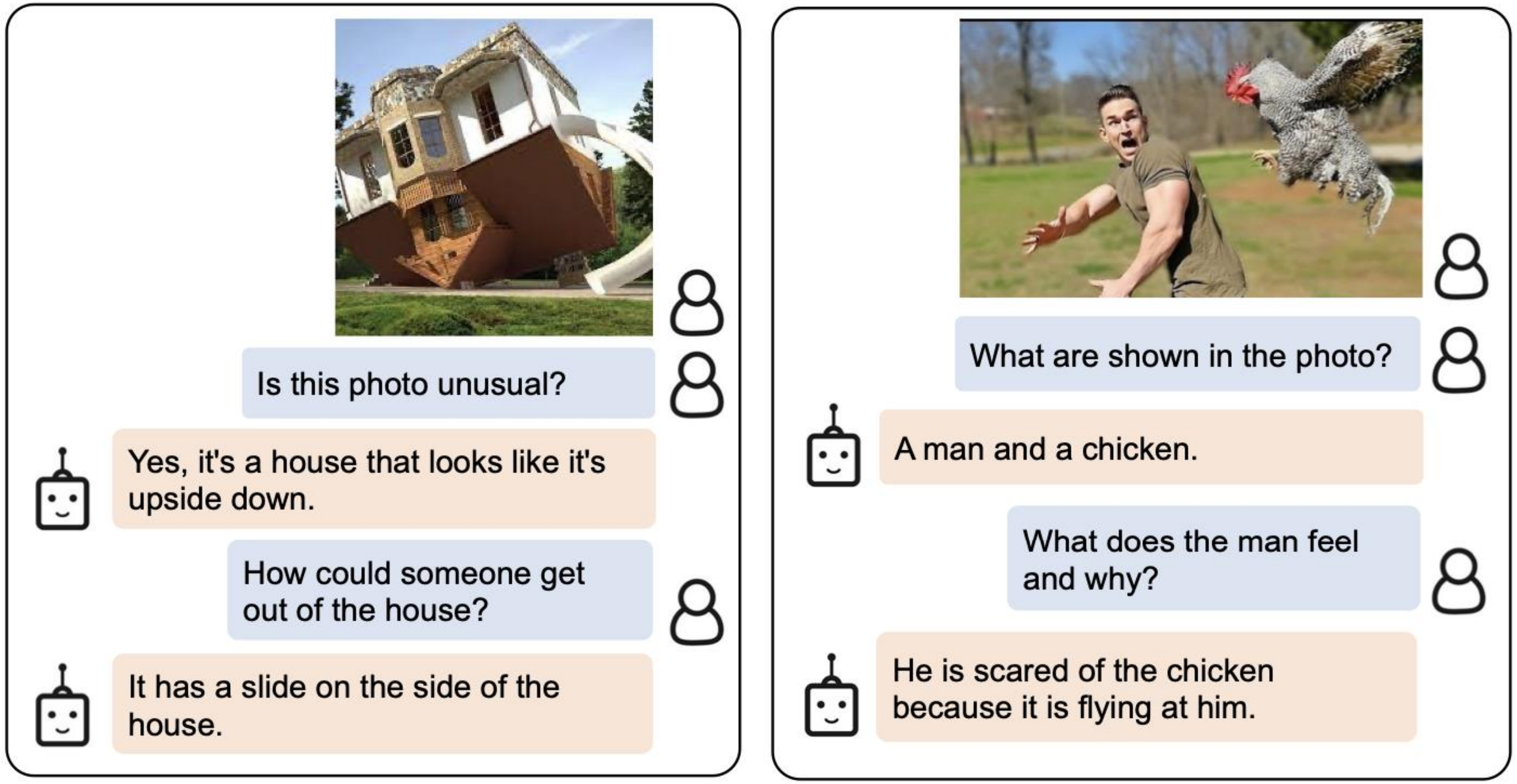


Flamingo, 11/2022

What works well in vision and language?



What works well in vision and language?



This talk

Looks into what multimodal foundation models cannot perceive:

1. Scarcity
2. Space
3. Time
4. Human values

Yunhua Zhang

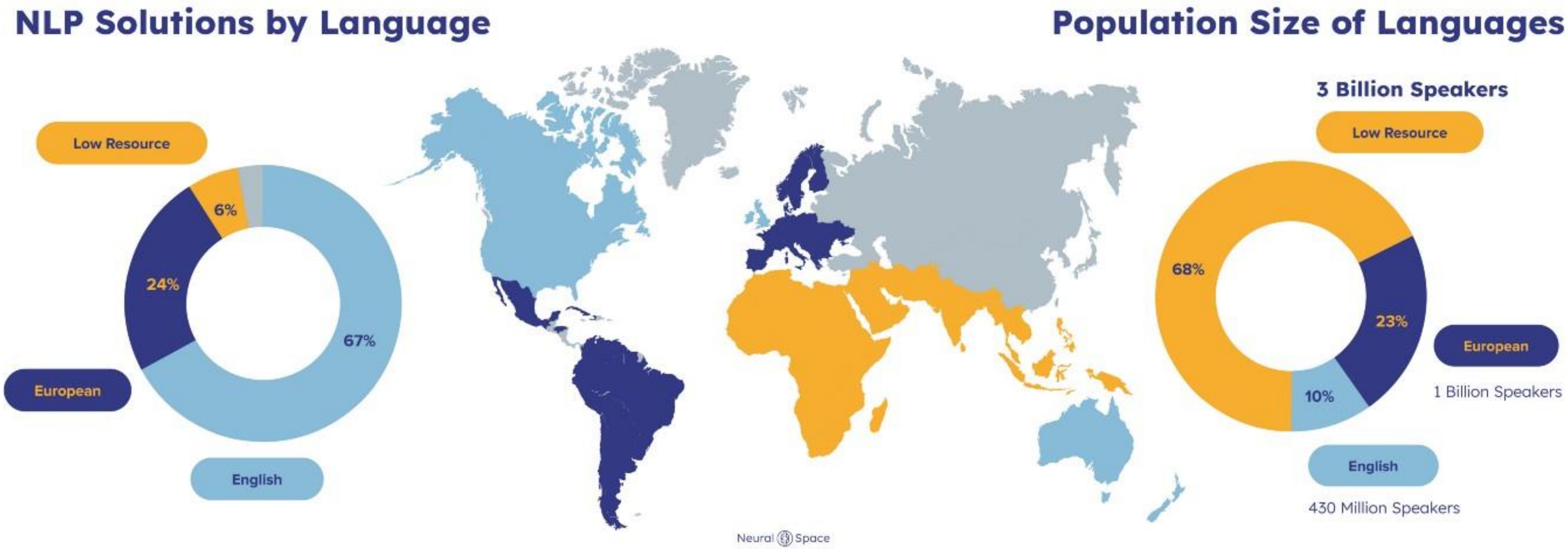
Hazel Doughty

1. Scarcity

Yunhua Zhang, Hazel Doughty, Cees G M Snoek: **Low-Resource Vision Challenges for Foundation Models**. In: CVPR, 2024.

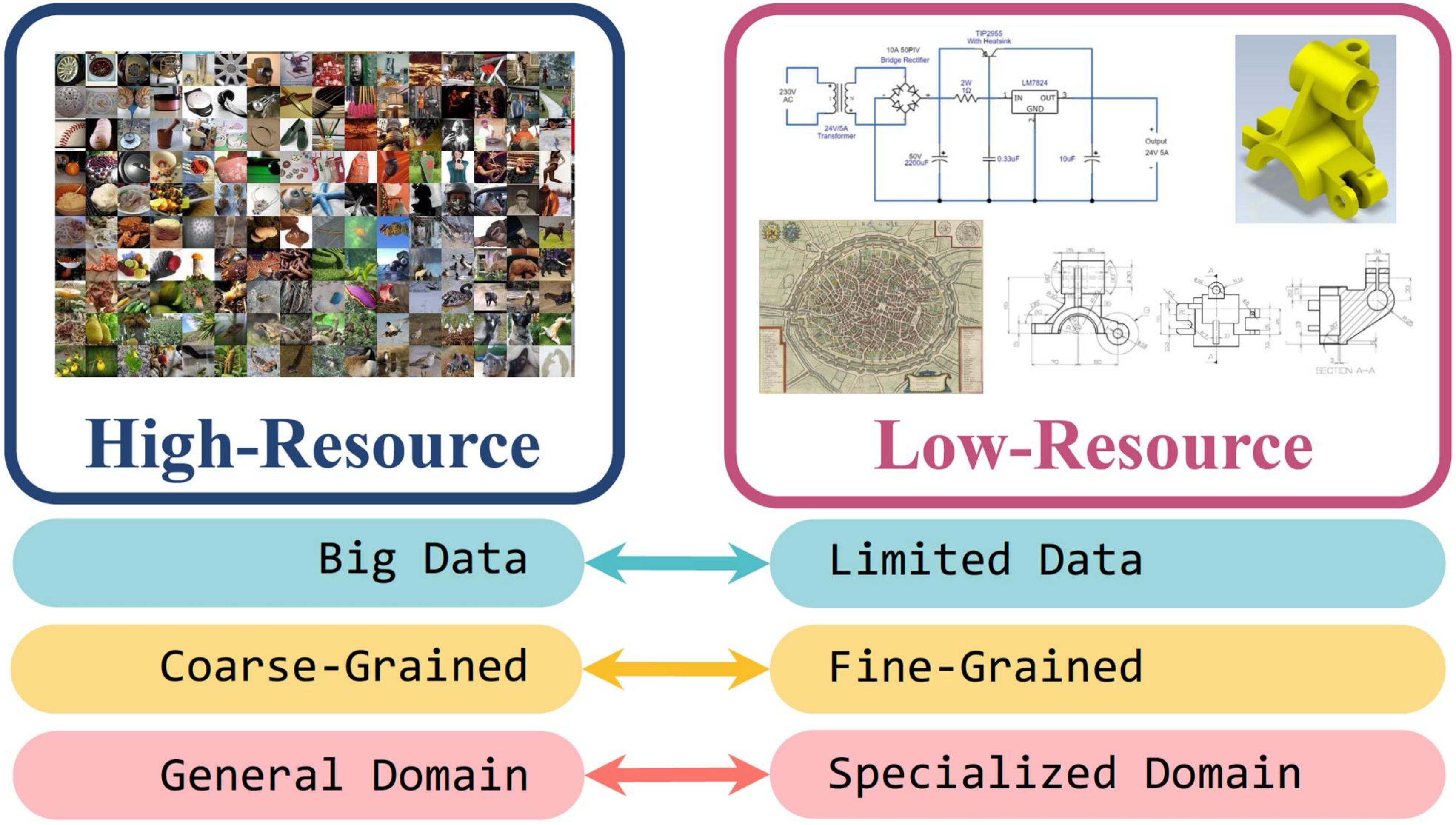
Hedderich et al. ACL 2020

Low-Resource Natural Language Processing

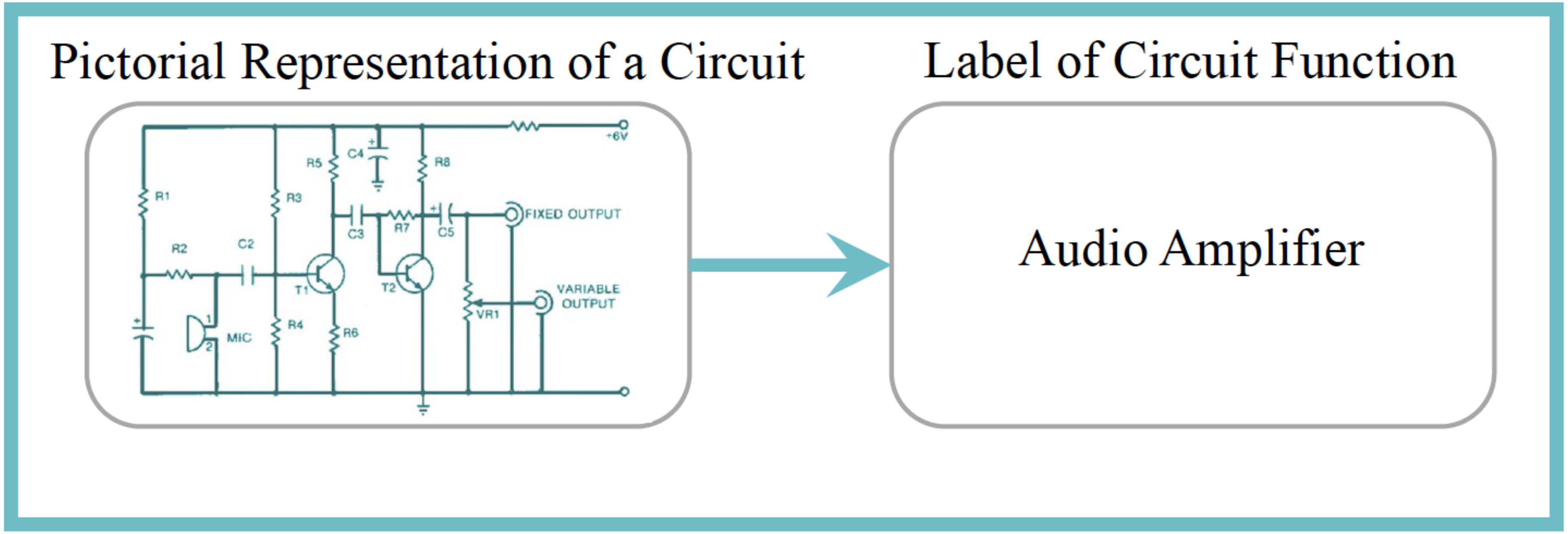


No previous works on low-resource vision tasks.

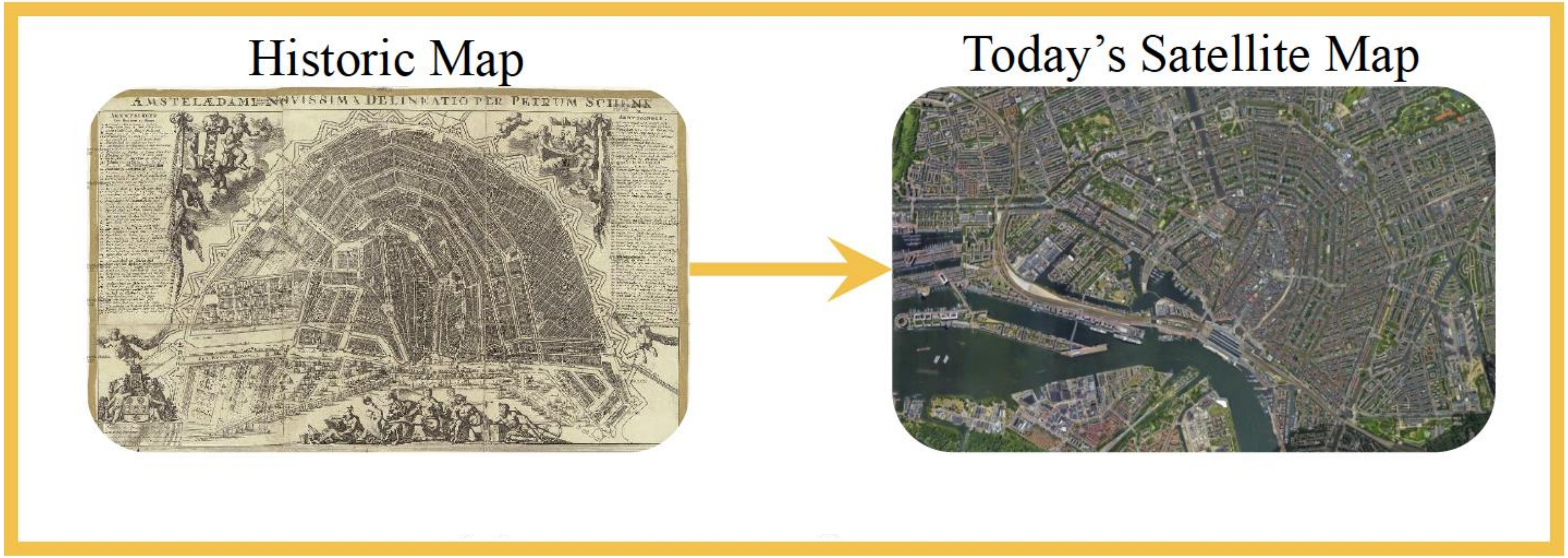
High-resource vs. Low-resource



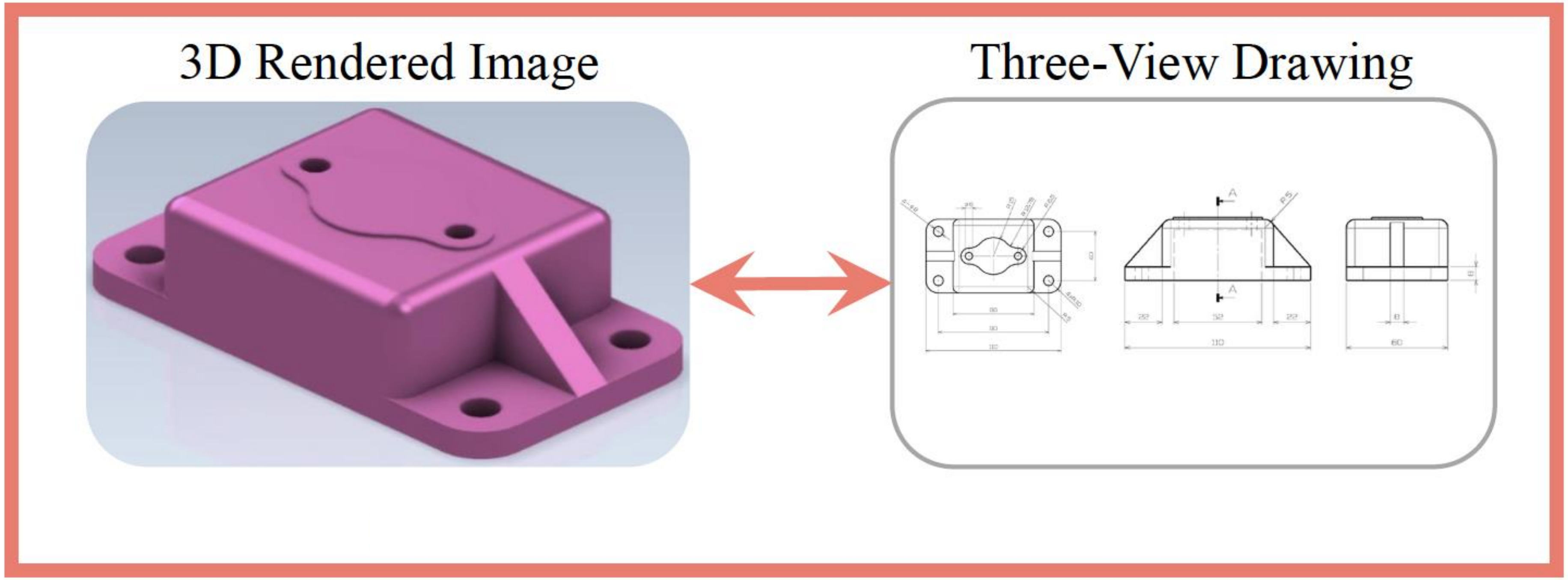
Circuit diagram classification



Historic map retrieval



Mechanical drawing retrieval



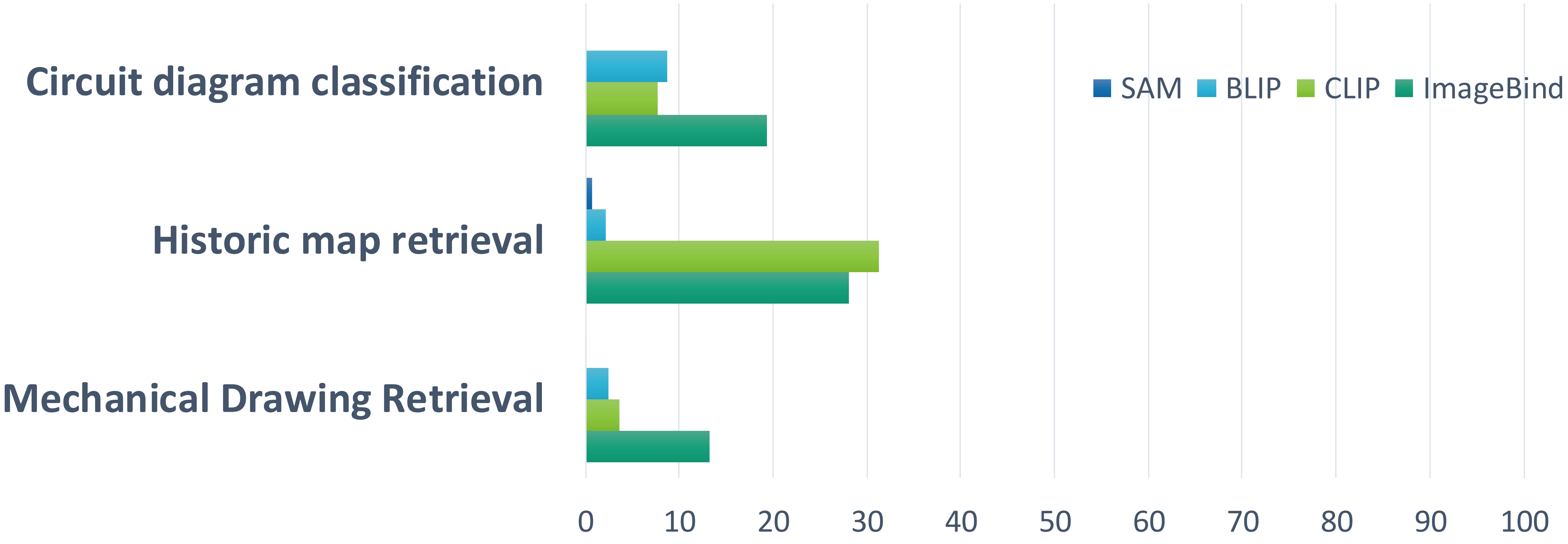
Low-Resource Image Transfer Evaluation

Task	**Formulation**	**Train**	**Val**	**Test**
Circuit Diagram Classification	Image Classification	154	100	1,078
Historic Map Retrieval	Image-to-Image Retrieval	102	140	409
Mechanical Drawing Retrieval	Image-to-Image Retrieval	300	100	754

Number of images (or image pairs) per split

We have collected as much data as we can find **freely available online** for each task, yet, the amount of data is **still incredibly small** showing how low-resource these tasks are.

Poor performance for low-resource vision challenges

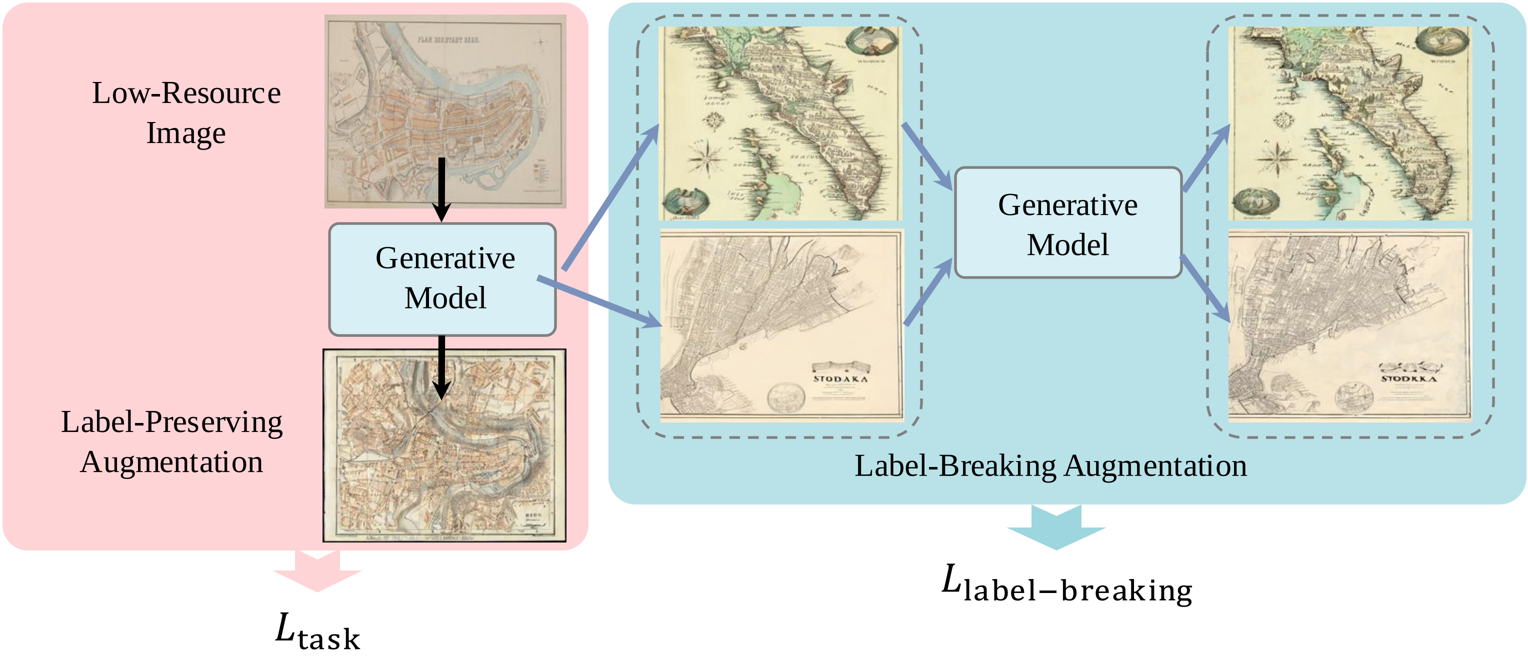


Low-Resource Vision Challenges

Challenge I: Data Scarcity	→	**Baseline I: Generated Data for Data Scarcity**
Challenge II: Fine-Grained	→	**Baseline II: Tokenization for Fine-Grained**
Challenge III: Specialized Domain	→	**Baseline III: Attention for Specialized Domains**

Our goal: adapt foundation models, pre-trained on large-scale datasets, to low-resource tasks.

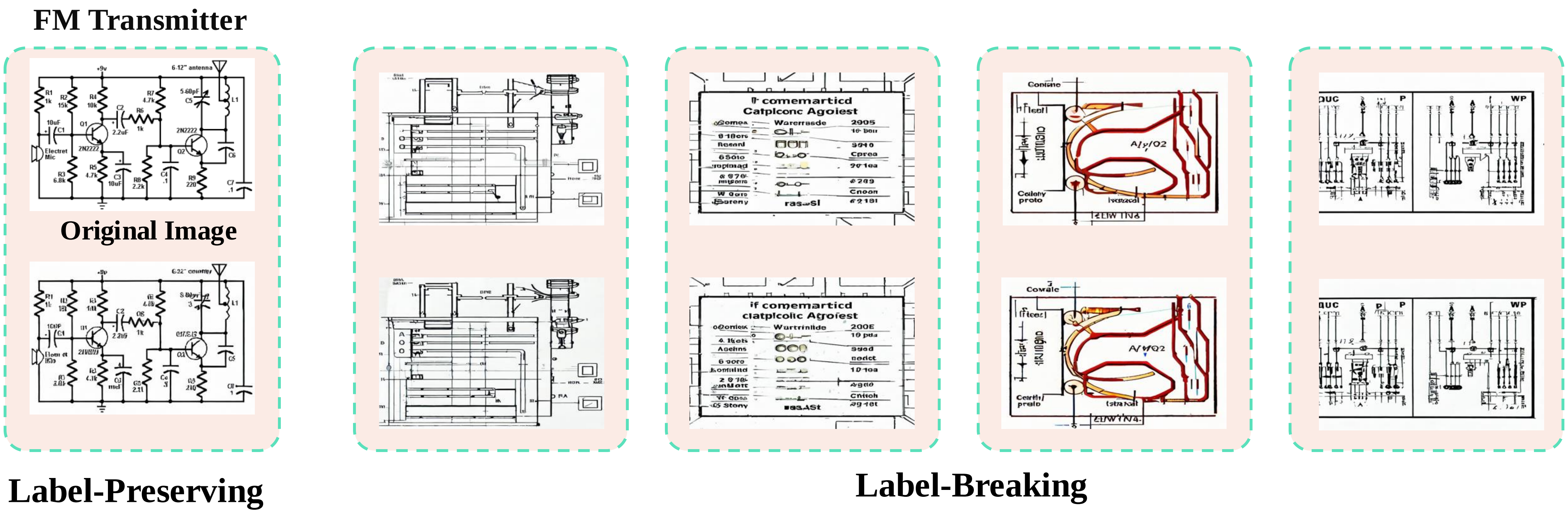
Baseline I: Generated Data for Data Scarcity



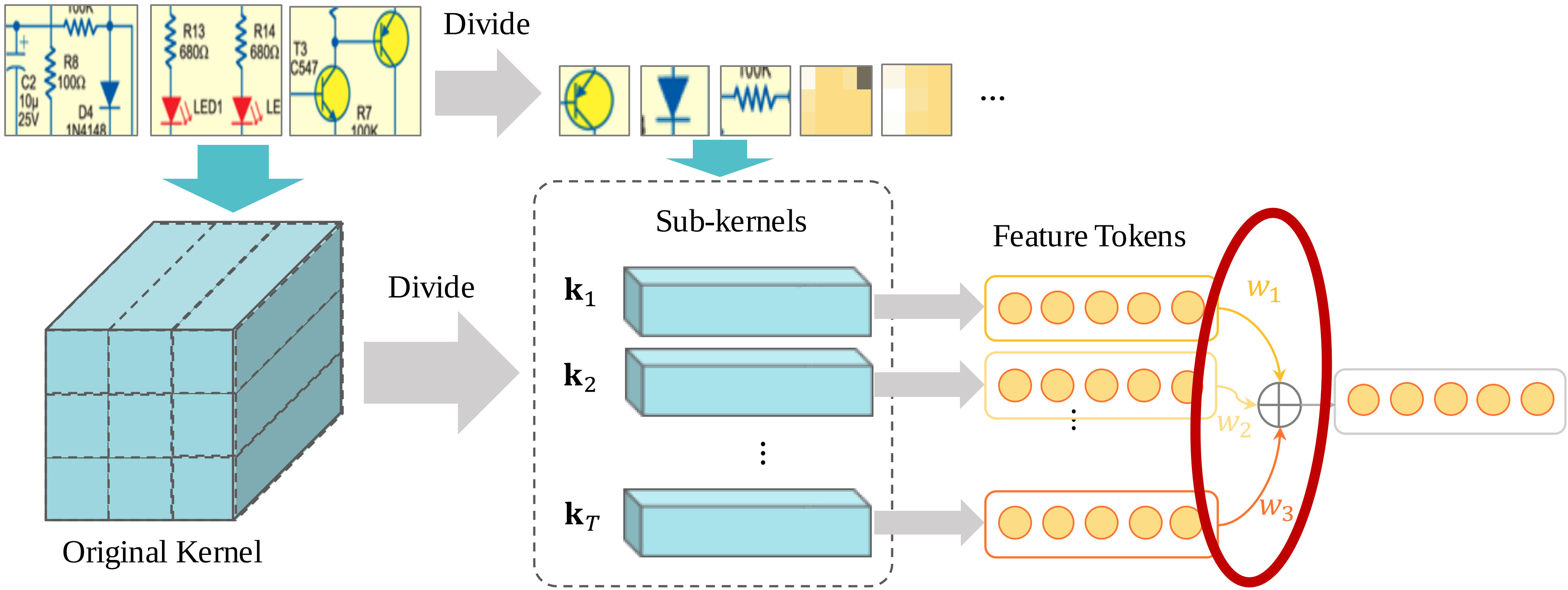
$$L = L_{\text{task}} + \lambda L_{\text{label-breaking}}$$

We generate images close to the input image where the label is preserved as well as more diverse images which break the label.

Circuit diagram examples



Baseline II: Tokenization for Fine-Grained



As we have limited data we cannot train a tokenization layer from scratch

Instead, we divide the linear projection kernel into sub-kernels for image patches.

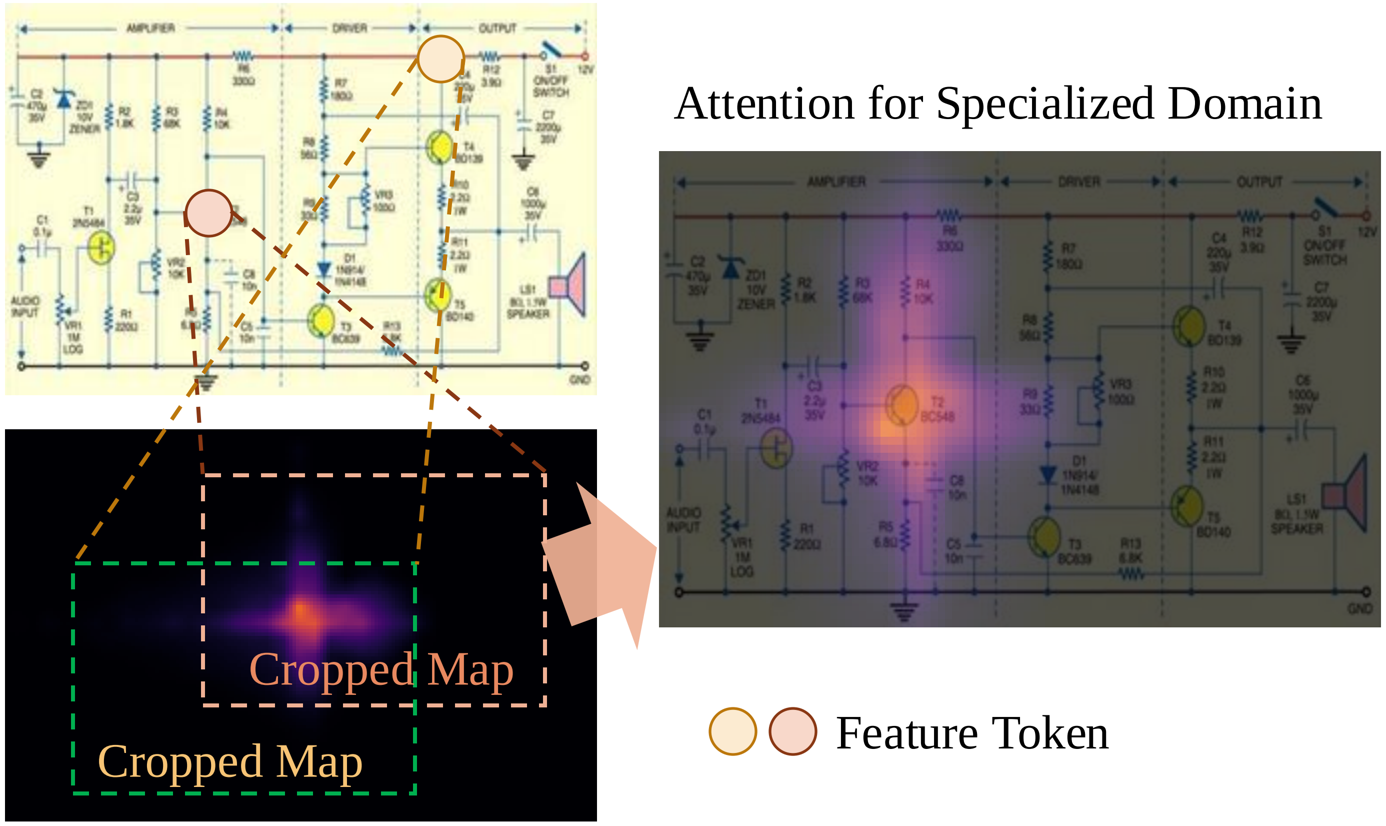
Then create patch-level features with a learned weighting

Baseline III: Attention for Specialized Domains

1. Learn global attention maps with common patterns particular to the specialized domain

2. For each token, crop its region from the global attention map.

3. Combine with multi-head self-attention.



Results of baselines for the three challenges

	Circuit Classification	
	Top-1 (%) ↑	Top-5 (%) ↑
Zero-Shot Transfer	19.3	45.1
Simple Transformation		
Random Crop and Flip	19.8	45.3
Mixup	20.8	46.0
CutMix	20.0	45.5
Random Erasing	20.8	46.2
Generative Models		
DA-Fusion	19.6	45.1
SyntheticData	20.8	46.0
Our Baselines		
Generated Data for Data Scarcity	21.3	46.9
Combination of Baselines	24.1	49.3

Challenge I: Data Scarcity

	Circuit Classification	
	Top-1 (%) ↑	Top-5 (%) ↑
Zero-Shot Transfer	19.3	45.1
Fine-Grained		
Adaptive-FGSBIR	[illegible]	43.2
PLEor	17.1	44.1
PDiscoNet	16.2	43.5
Our Baselines		
Tokenization for Fine-Grained	20.9	45.5
Combination of Baselines	24.1	49.3

Challenge II: Fine-Grained

	Circuit Classification	
	Top-1 (%) ↑	Top-5 (%) ↑
Zero-Shot Transfer	19.3	45.1
Transfer Learning		
Linear Probe	18.7	45.9
TOAST	16.4	43.3
A3	18.2	45.4
VPT	19.4	45.2
LoRA	15.5	42.2
AdaptFormer	19.8	45.5
Our Baselines		
Attention for Specialized Domains	20.6	47.0
Combination of Baselines	24.1	49.3

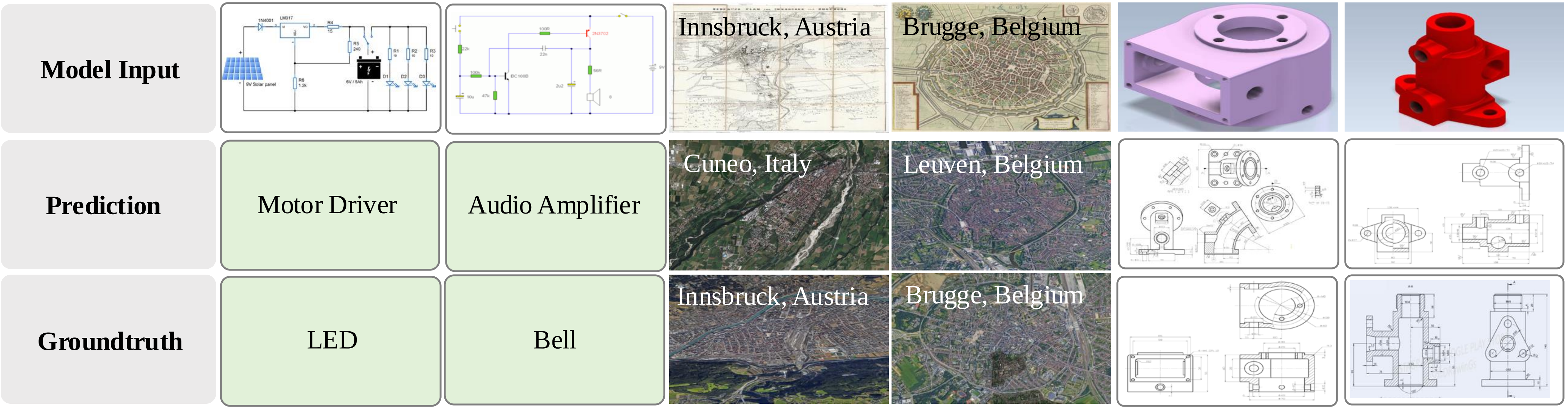
Challenge III: Specialized Domain

Our baselines are effective

Effective adapter for several foundation models

Results for Historic Map Retrieval

Qualitative results: hard samples



Our predictions are overconfident, often basing predictions on one key region such as the presence of the battery in the LED circuit.

We cannot yet generalize to rare image styles such as used for the Innsbruck map

Michael Dorkenwald

Nimrod Barazani

Yuki Asano

2. Space

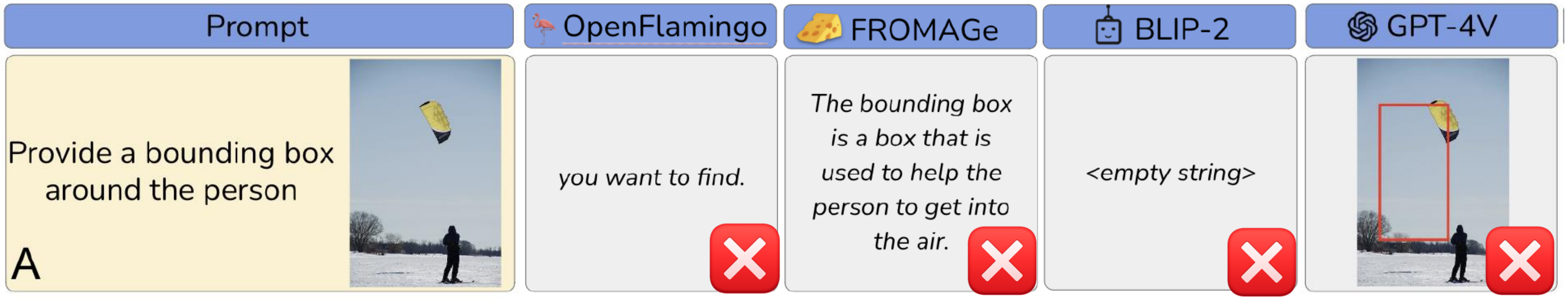
Michael Dorkenwald, Nimrod Barazani, Cees G M Snoek, Yuki M Asano: **PIN: Positional Insert Unlocks Object Localisation Abilities in VLMs**. In: CVPR, 2024.

CVPR SEATTLE, WA JUNE 17-21, 2024

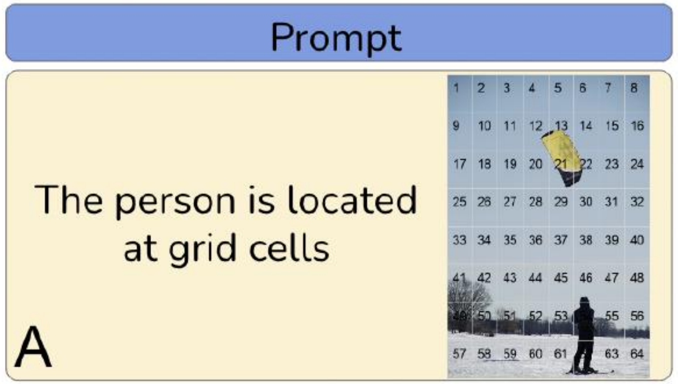
w/ Kien Nguyen et al. CVPR 2022 / ICLR 2024
w/ Aritra Bhowmik et al. ICCV 2023

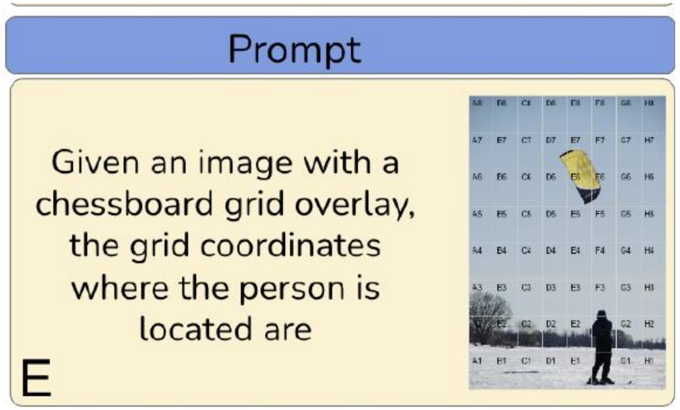
Special purpose object localization is very mature

Can vision-language models localize objects?

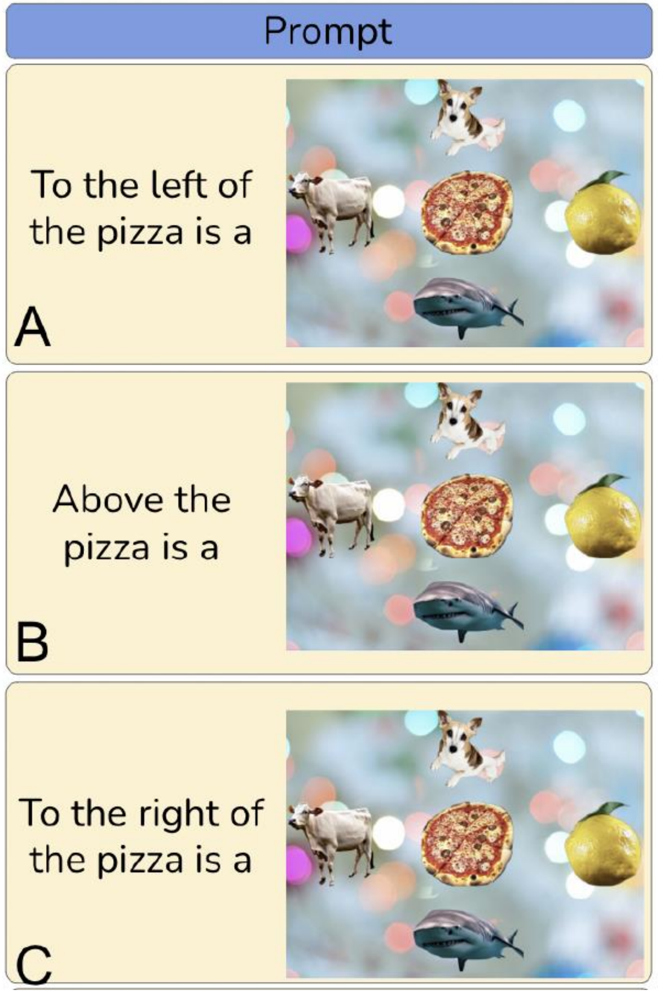


Perhaps we need another type of prompt?





Can vision-language models do spatial reasoning?

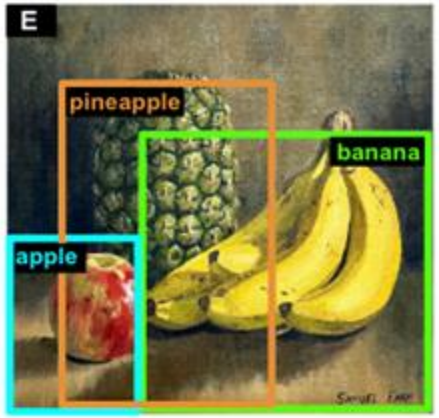


Our proposal

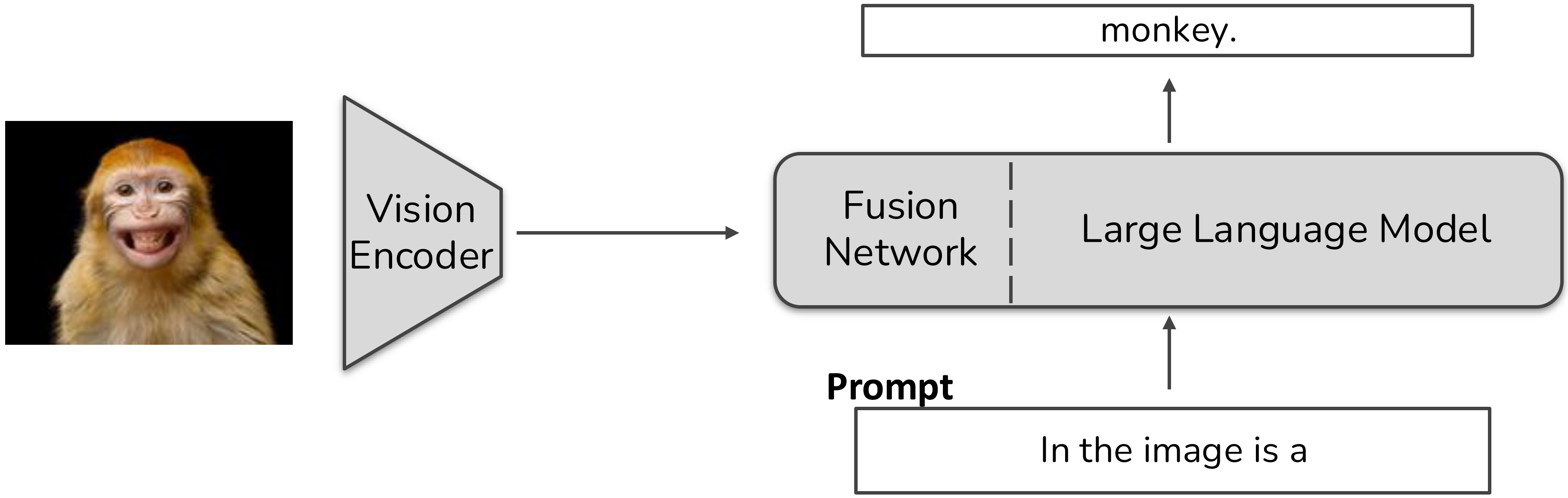
Frozen VLM,
e.g. Flamingo

PIN: positional
learnable prompt

Self-generated
supervision signal



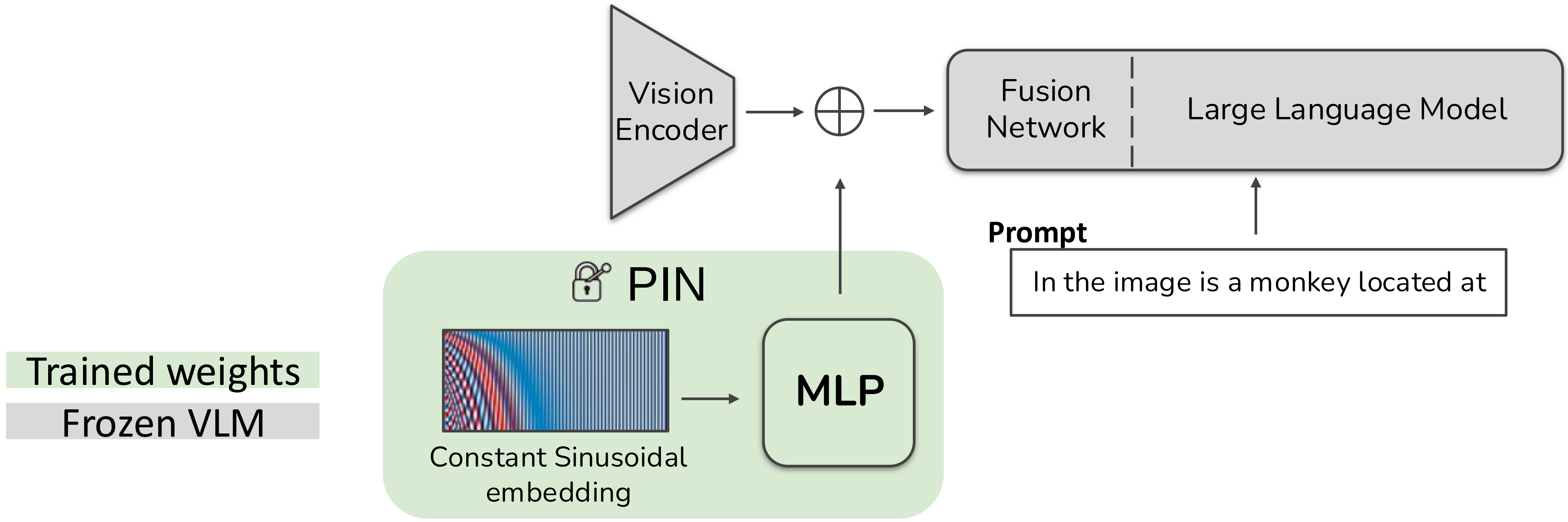
Vanilla Flamingo next token prediction



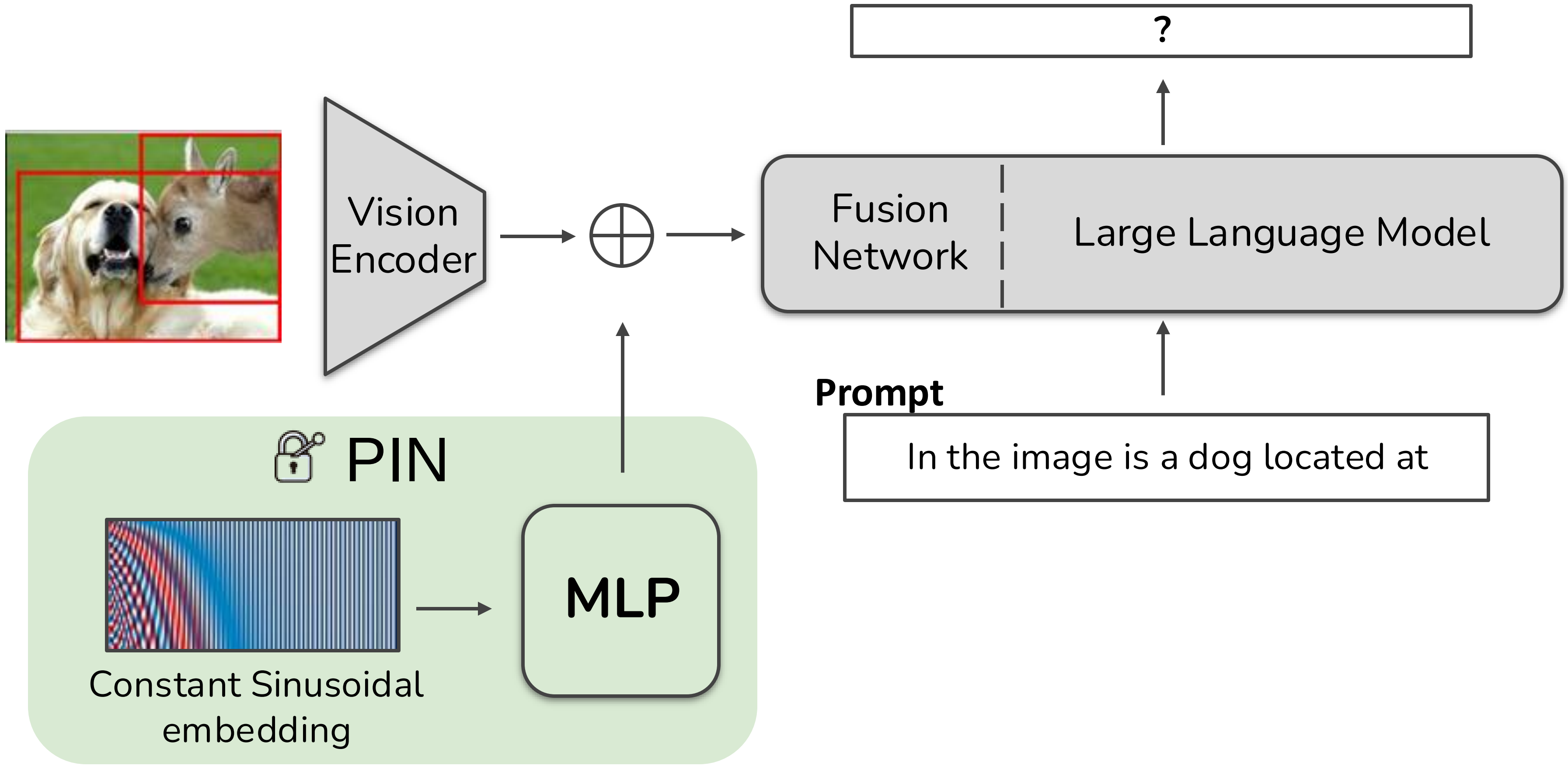
Frozen VLM

Alayrac, et al. Flamingo: a visual language model for few-shot learning. In NeurIPS, 2022.

Positional Insert (PIN)



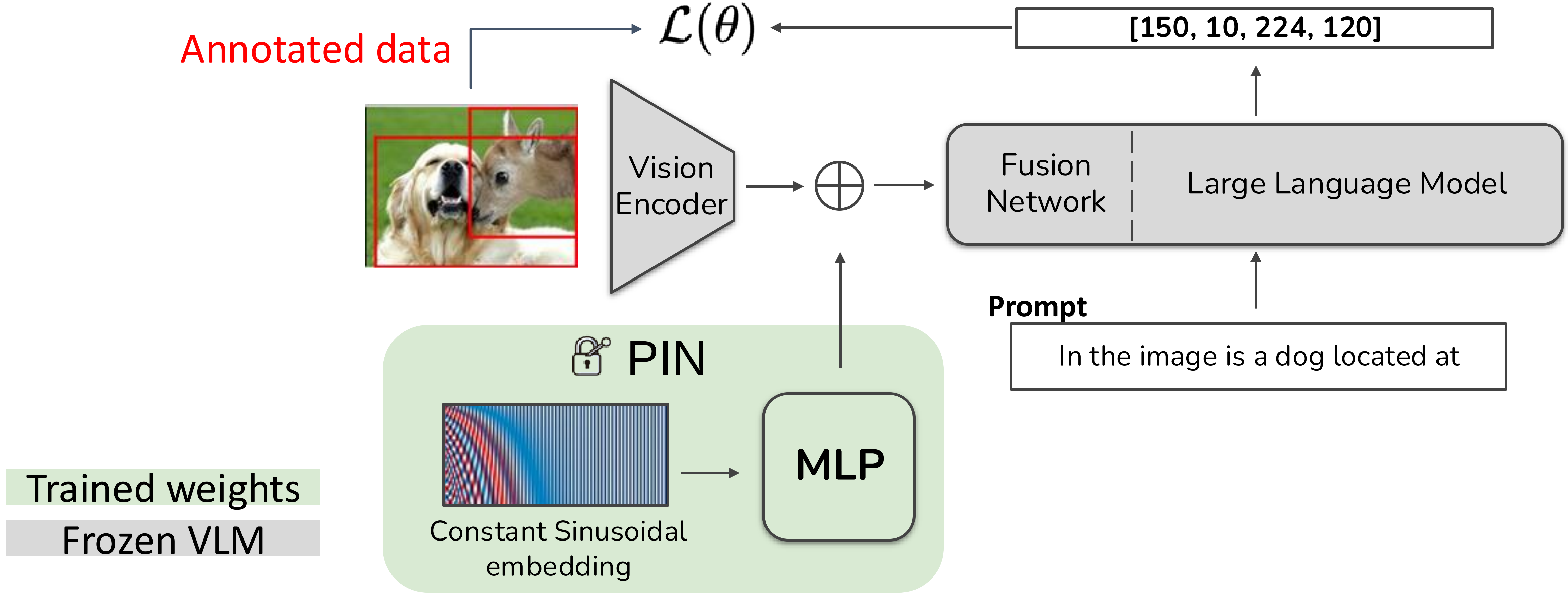
Positional Insert (PIN)



Trained weights

Frozen VLM

Do we need labeled data?



Self-generated supervision signal

Generate objects via Stable Diffusion for 1203 categories from LVIS.

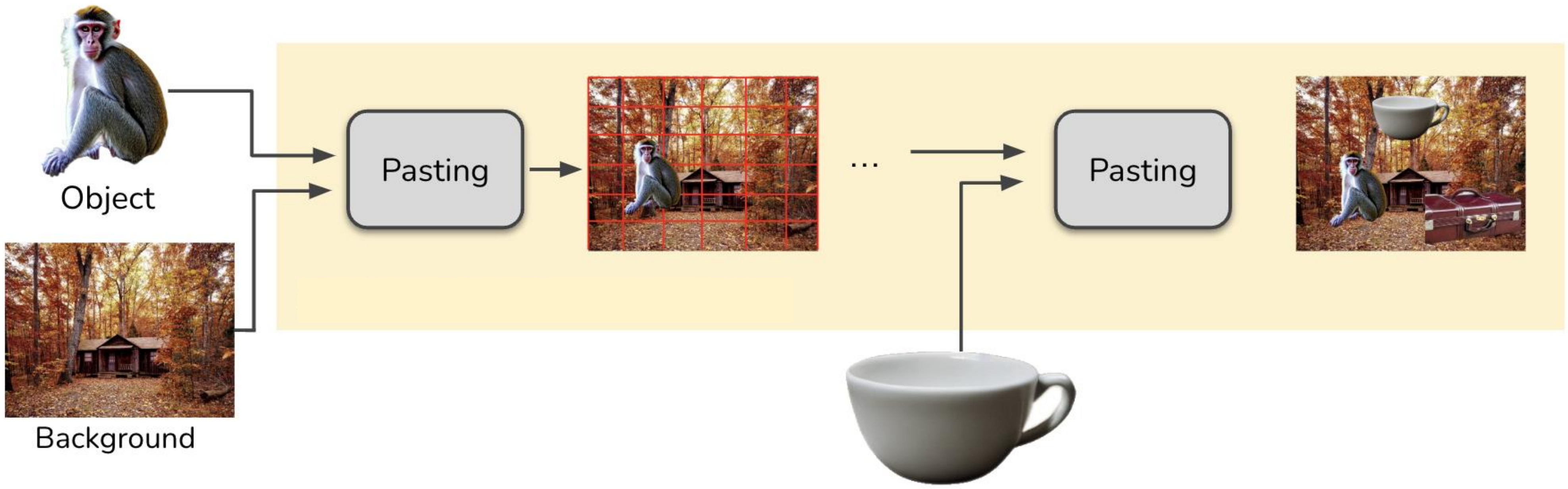
Paste objects into BG20k background dataset



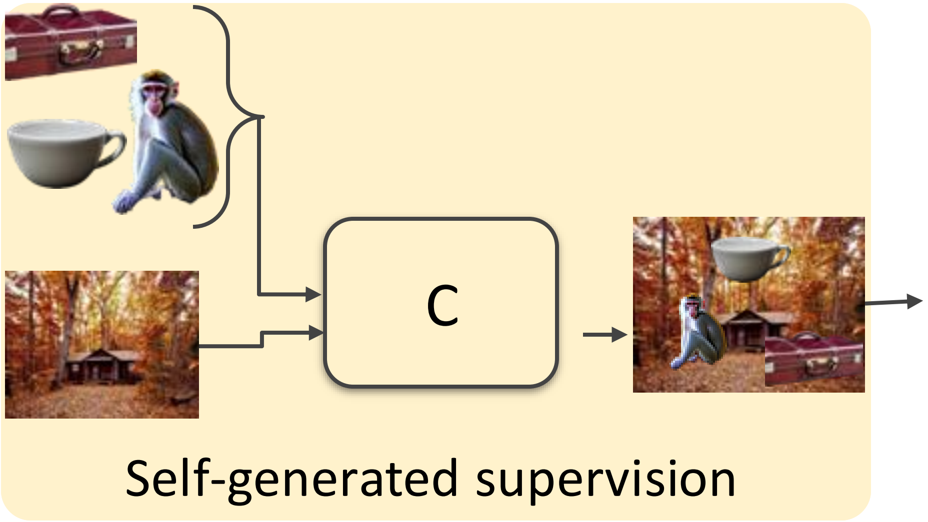


Hanqing Zhao *et al.* X-Paste: Revisiting Scalable Copy-Paste for Instance Segmentation using CLIP and Stable Diffusion. ICML 2023.
Jizhizi Li *et al.* Bridging composite and real: towards end-to-end deep image matting. IJCV 2022.

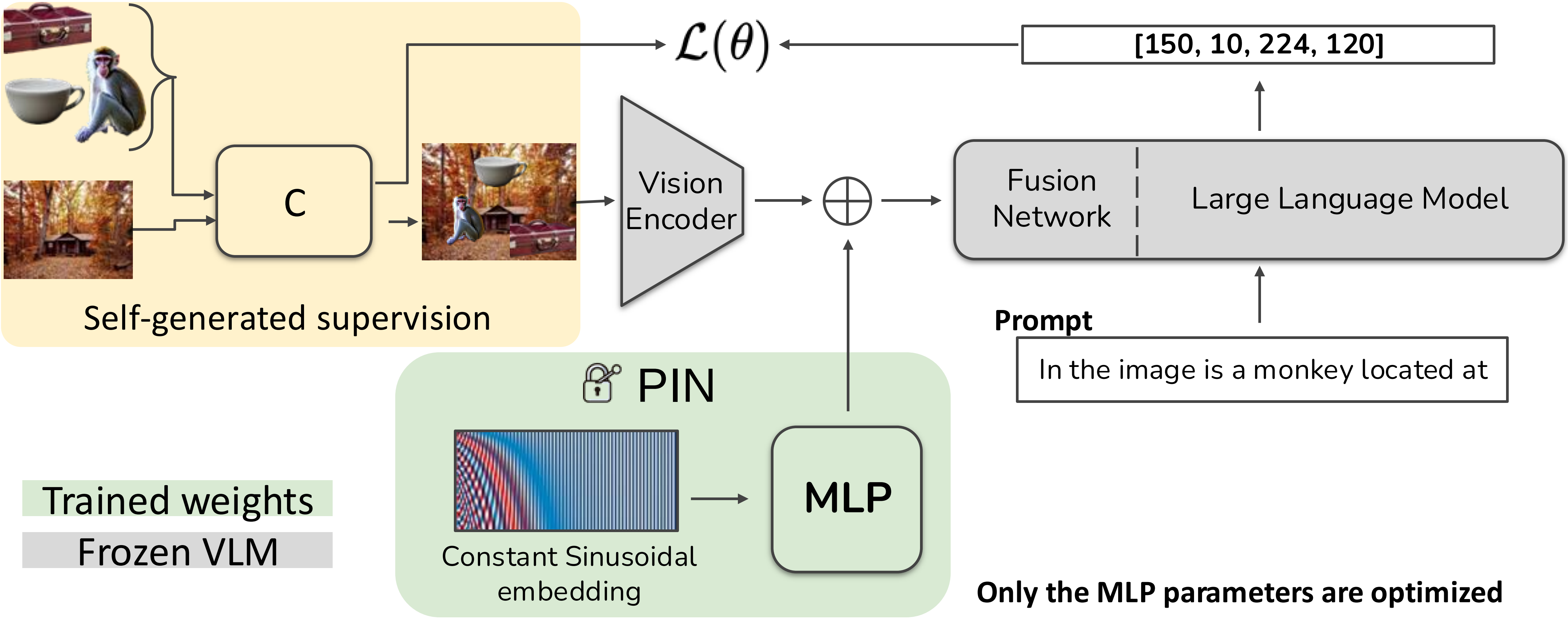
Self-generated supervision signal



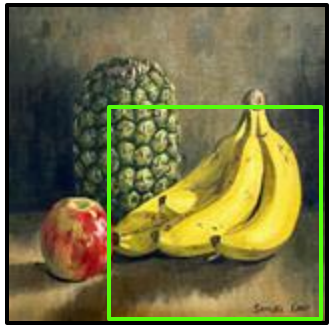
Training

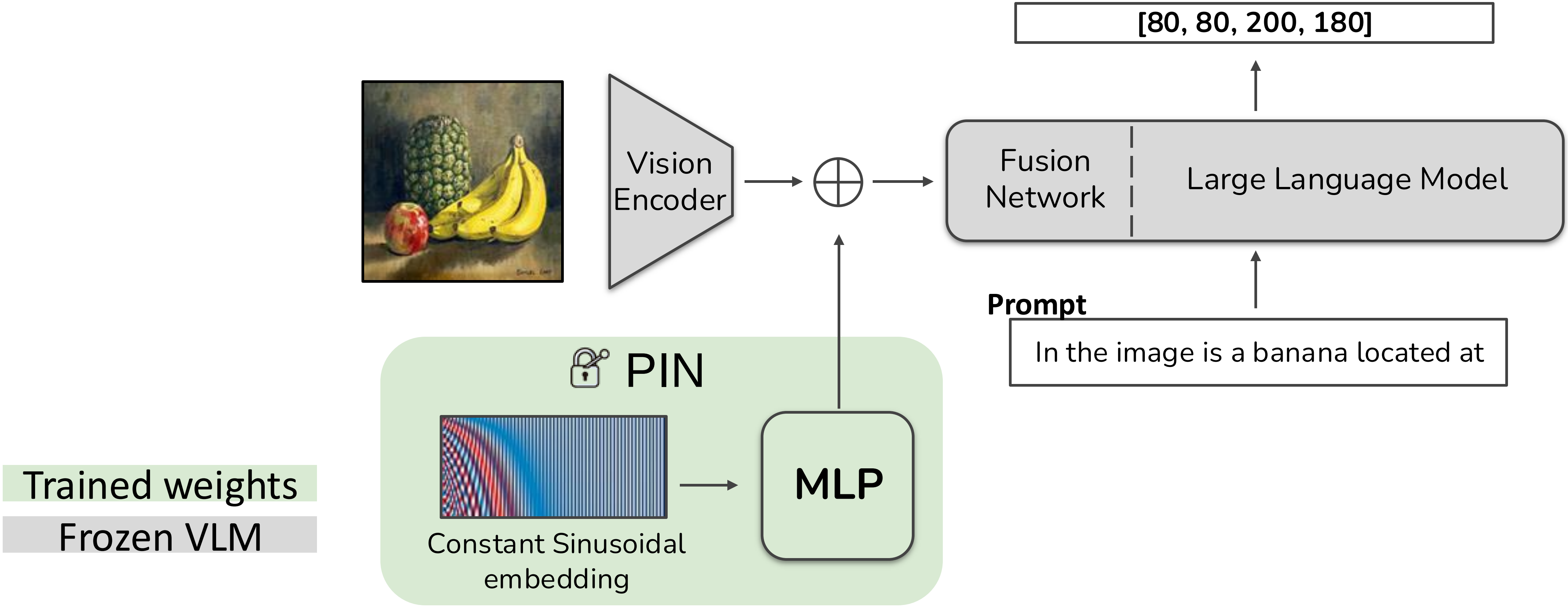


Training: next-token prediction

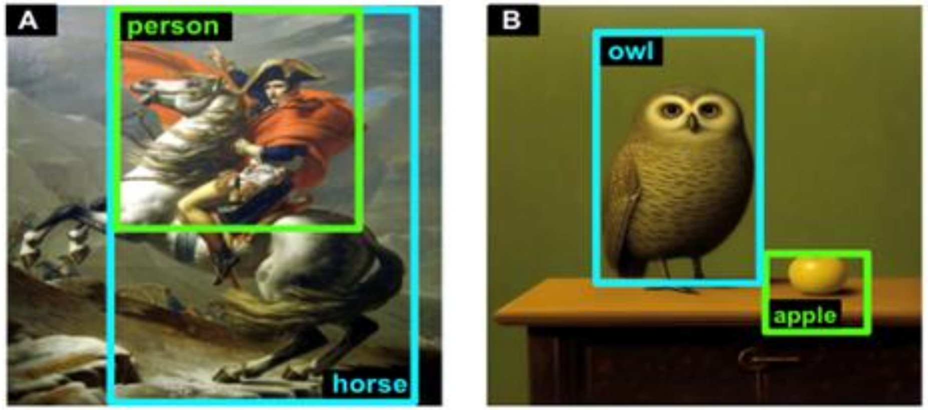


Inference

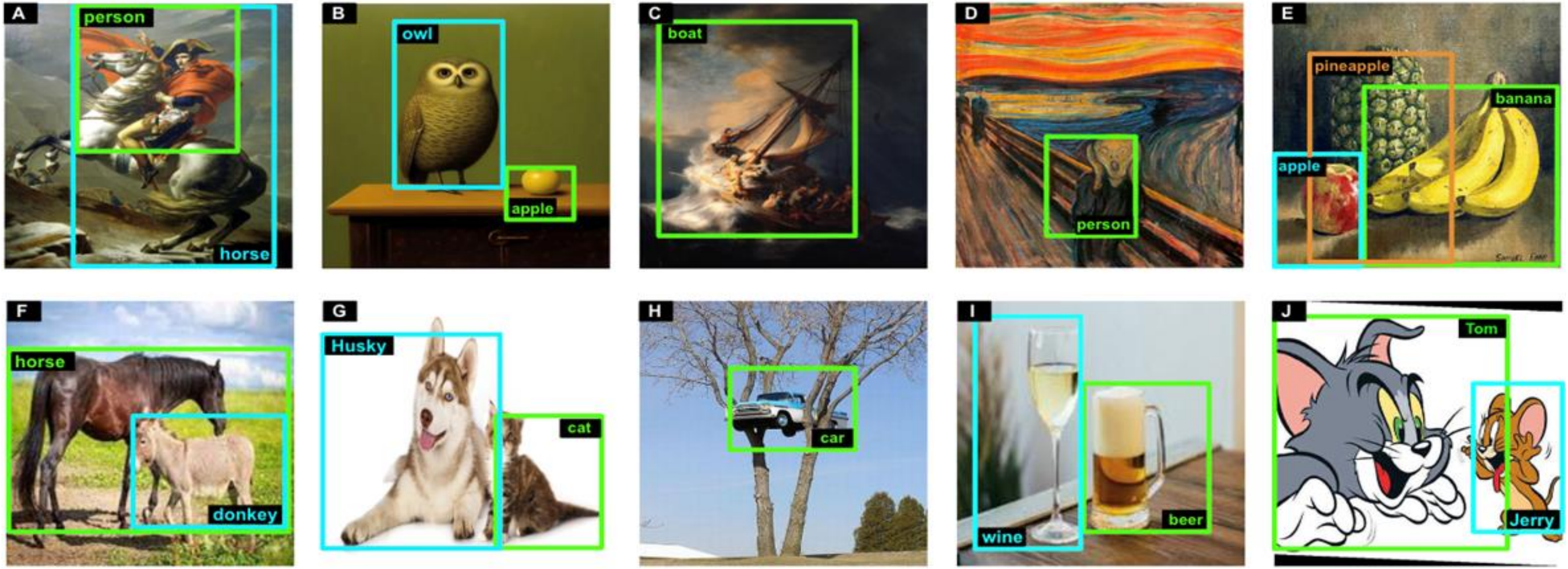



[80, 80, 200, 180]
Vision Encoder
Fusion Network
Large Language Model
Prompt
In the image is a banana located at
PIN
MLP
Constant Sinusoidal embedding
Trained weights
Frozen VLM

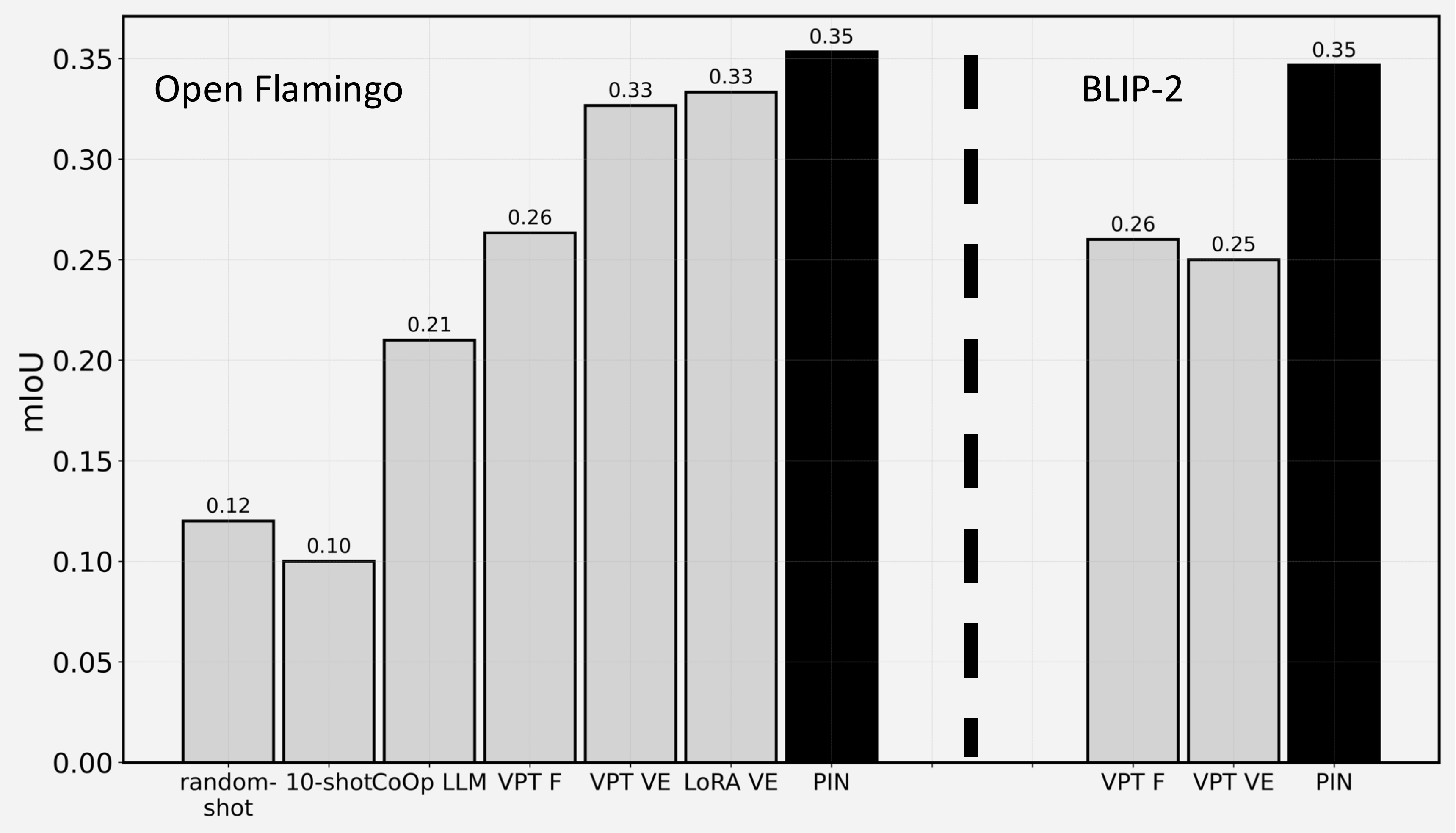
The PIN module unlocks spatial localisation



The PIN module unlocks spatial localisation



PIN outperforms PEFT alternatives



Piyush Bagad

Makarand Tapaswi

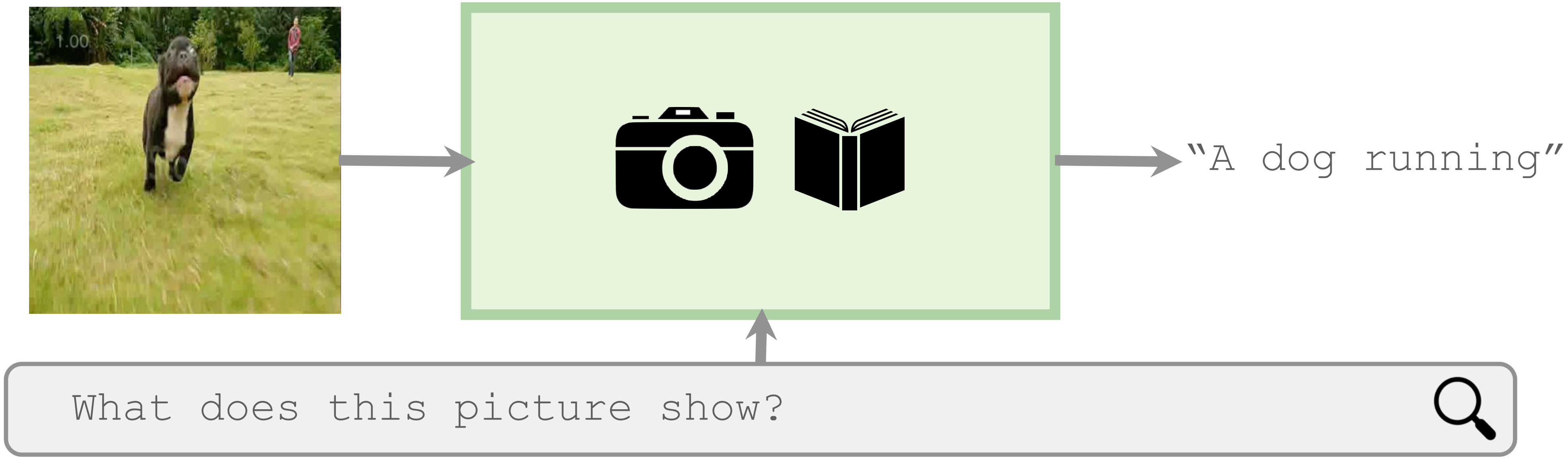
3. Time

Piyush Bagad, Makarand Tapaswi, Cees G M Snoek: **Test of Time: Instilling Video-Language Models with a Sense of Time.** In: CVPR, 2023.

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

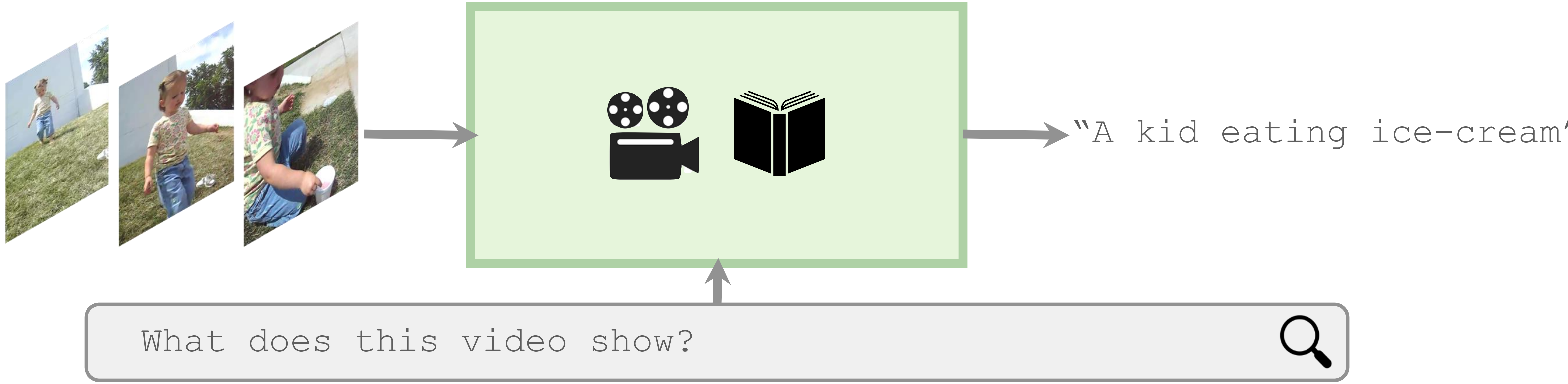
The problem

- Foundation models: Language interface + a few (or no) training samples



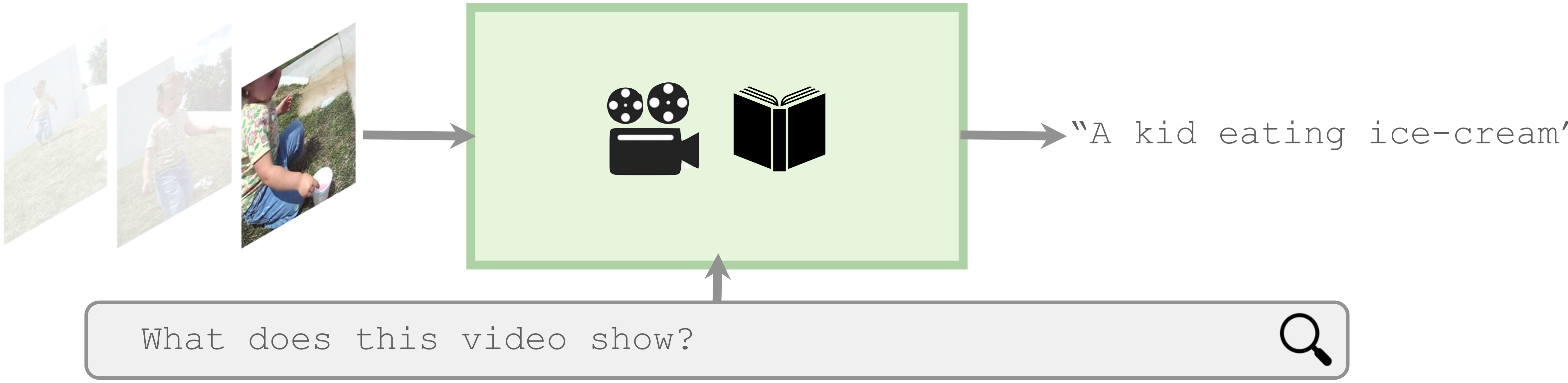
The problem

- Foundation models: Language interface + a few (or no) training samples
- Particularly attractive for videos given high cost



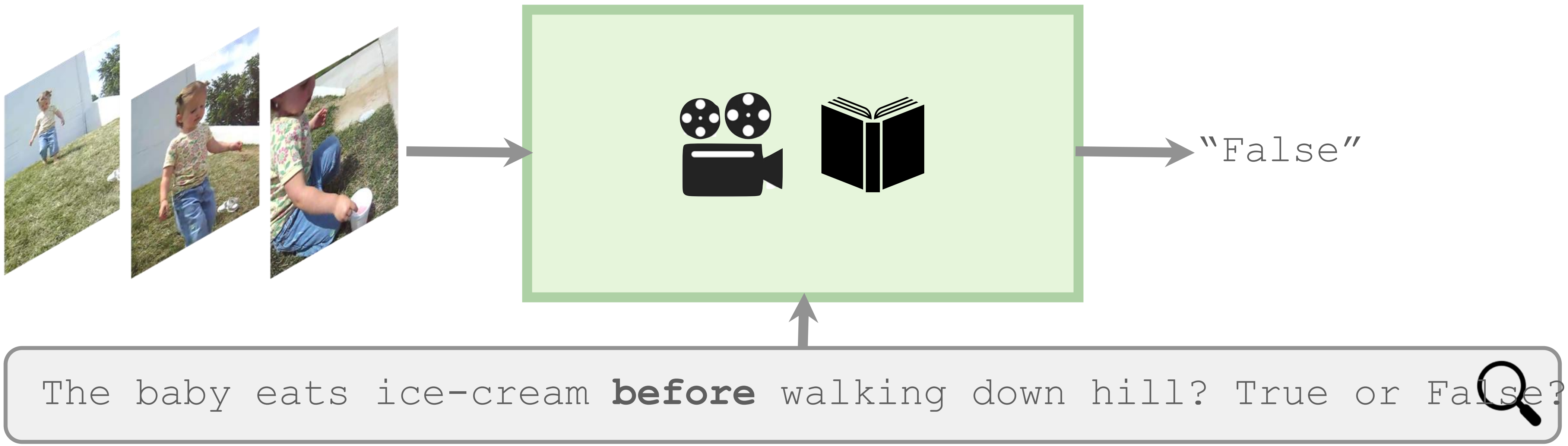
The problem

- Do video foundation models truly understand time?



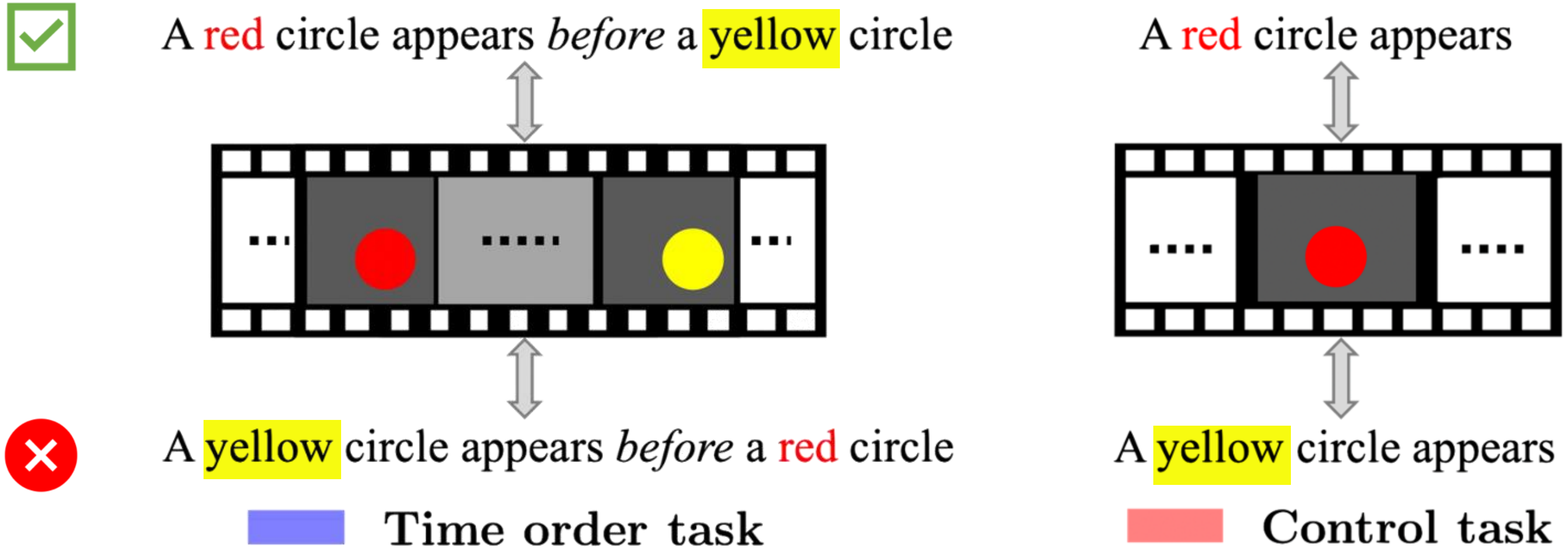
The problem

- Do video foundation models truly understand time?
- Our idea for a "test of time": ask questions that have temporal relations



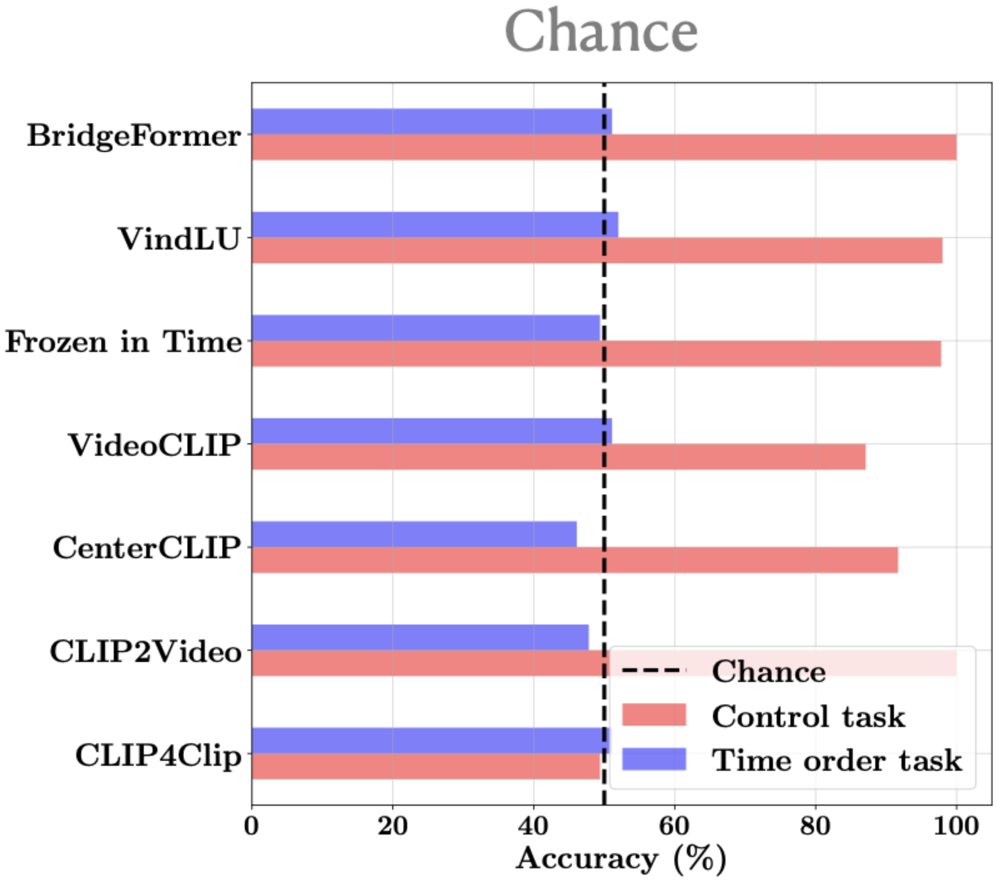
The test of time

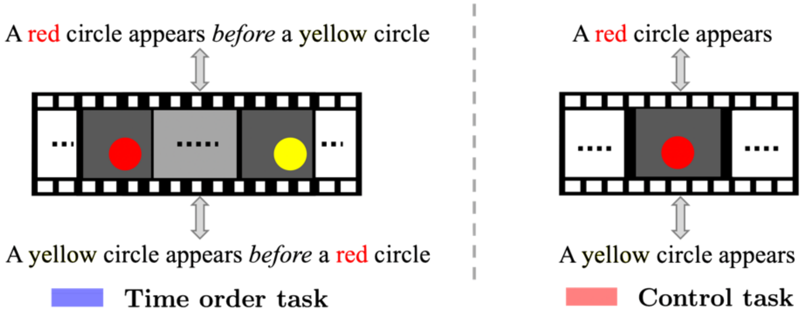
- Synthetic benchmark
- Simple 'true' or 'false' predictions



Existing models fail this test of time

- We pick a suite of seven openly available video-language models
- While excelling at the control task, they all fail at the time-order task



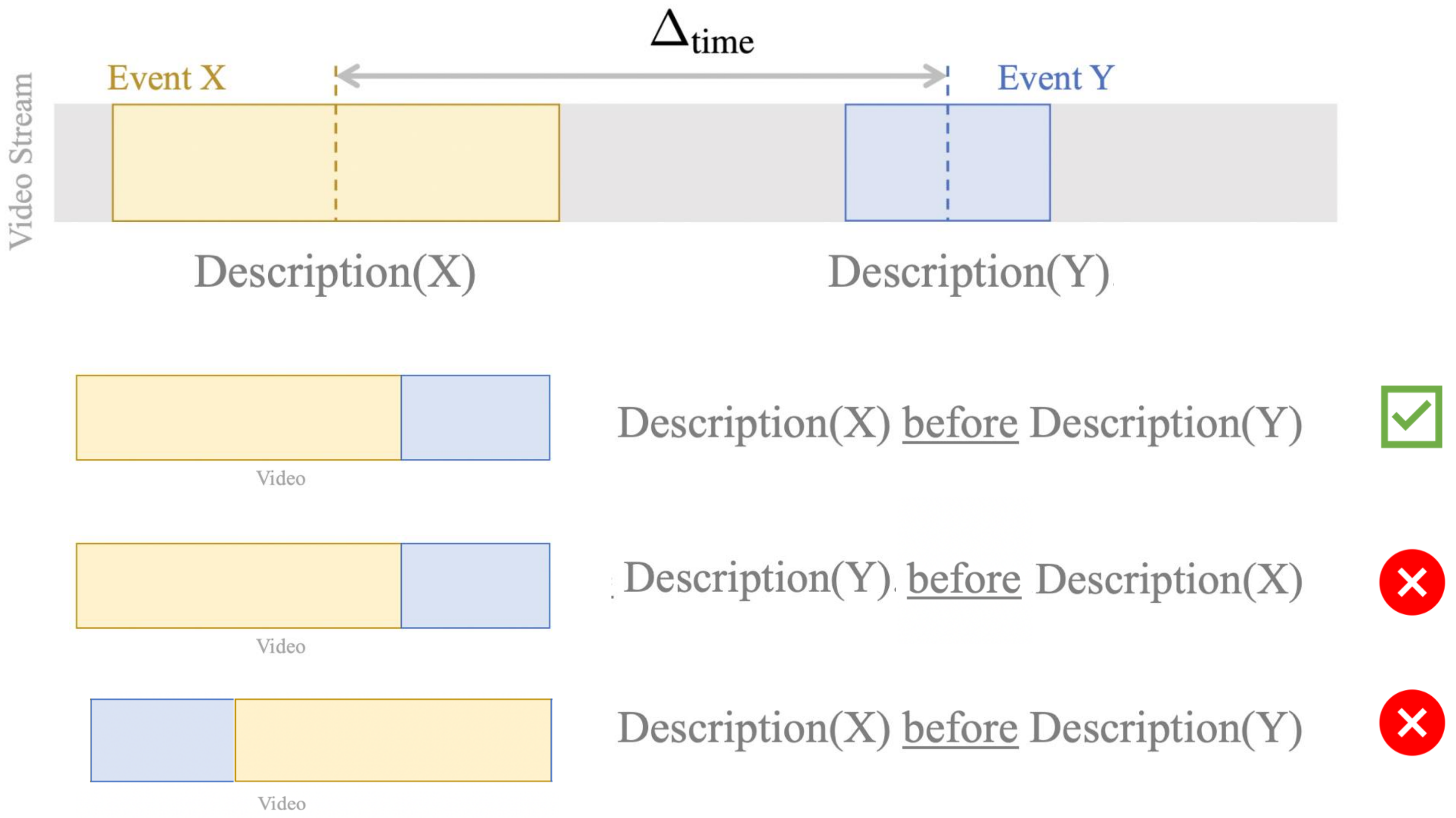


How to instil this sense of time?

- Post-pretraining: instead of training from scratch, we run another round of pre-training

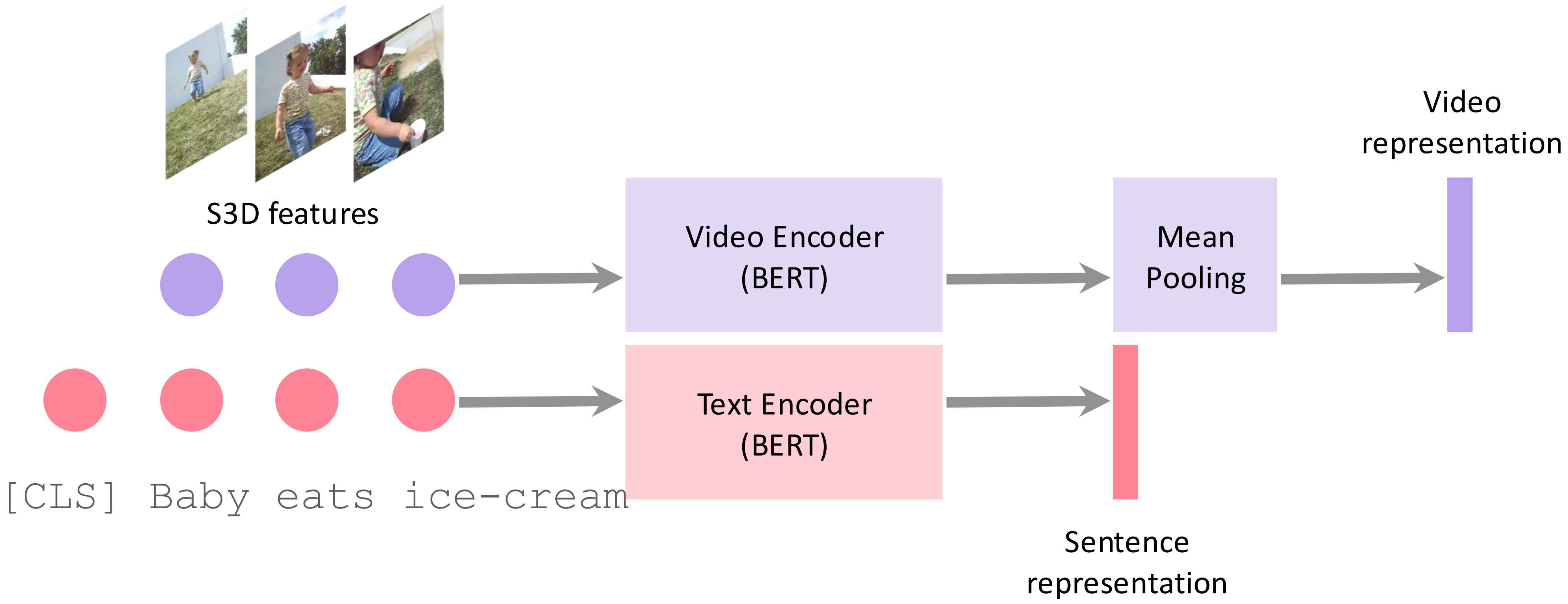
How to instil this sense of time?

- Data: any dense video-captioning dataset!



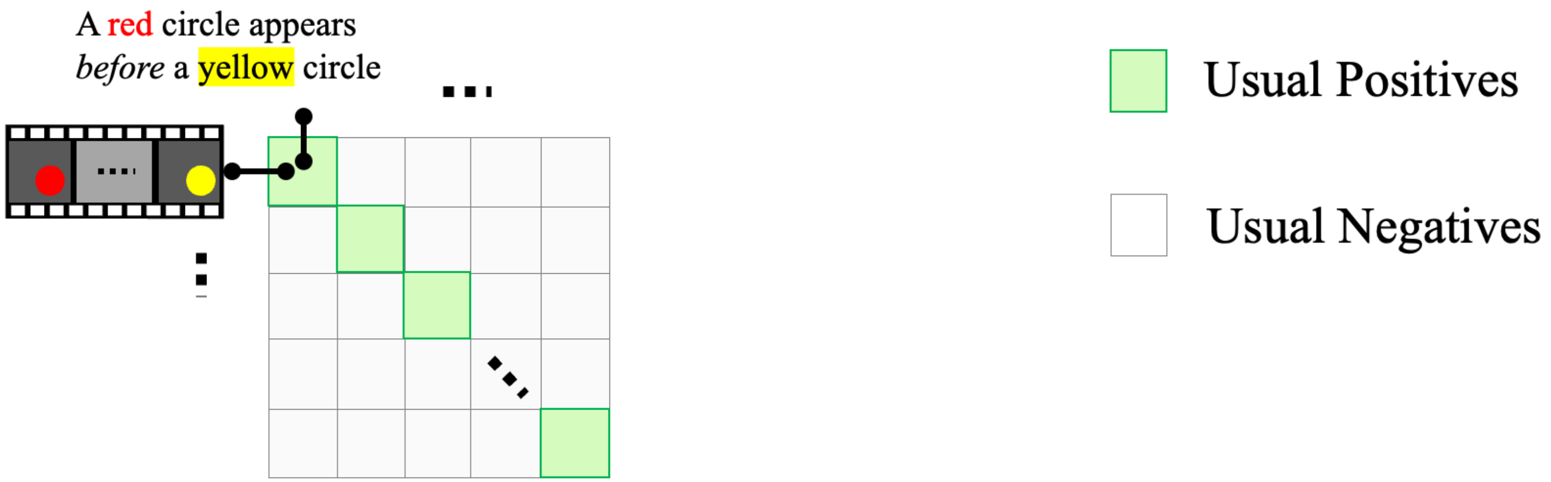
How to instil this sense of time?

- Base model: We start with a pre-trained model: VideoCLIP

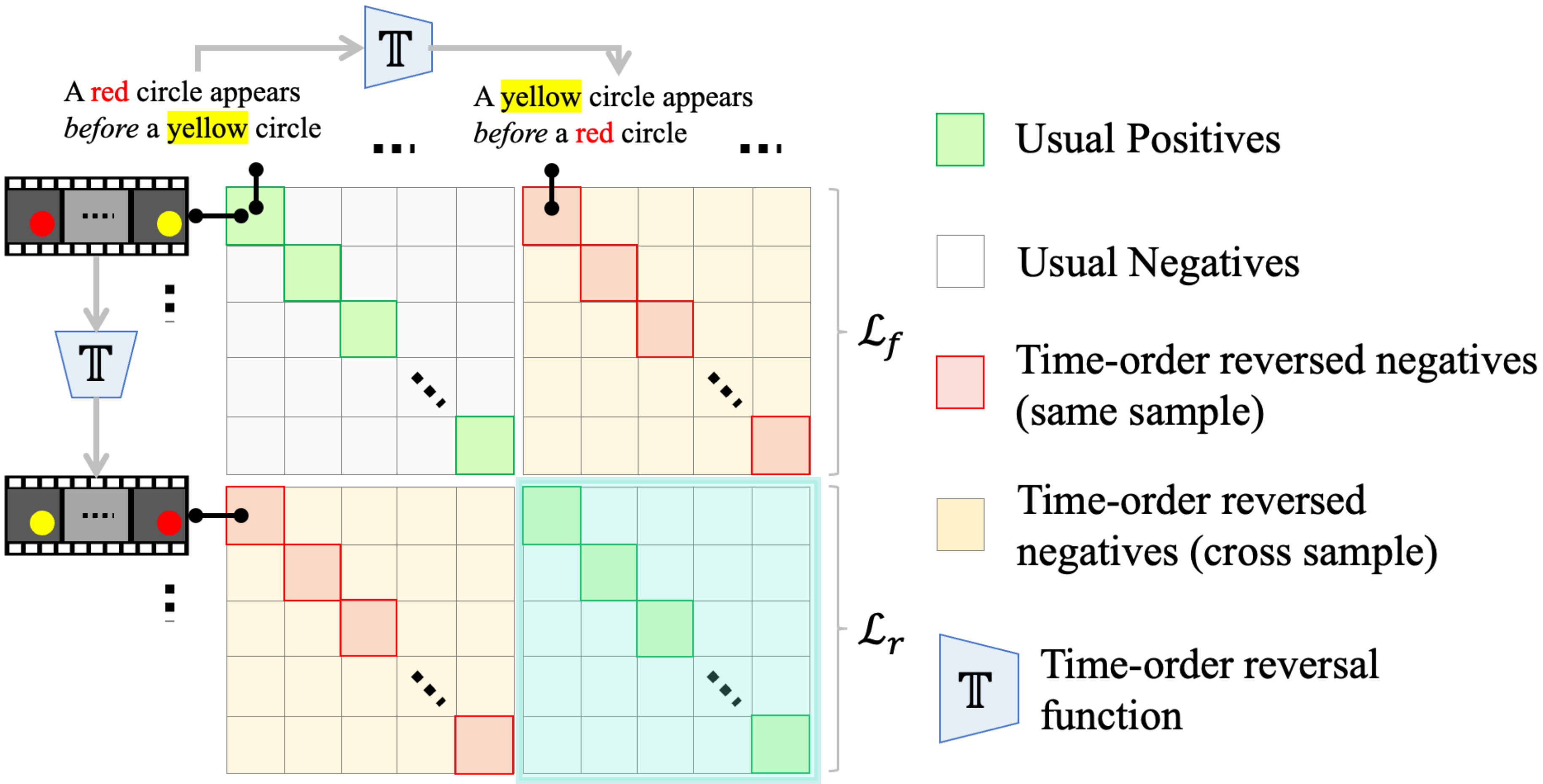


Xu et al, VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding, EMNLP 2021.

How to instill this sense of time?



How to instill this sense of time?



Experiments

Little girl eats from cup after the child walks downhill

(a) TEMPO

A woman is standing in a room holding a hula hoop before she begins to use the hula hoop

The team shakes hands with the opposing team after a team groups together holding a trophy

(b) ActivityNet

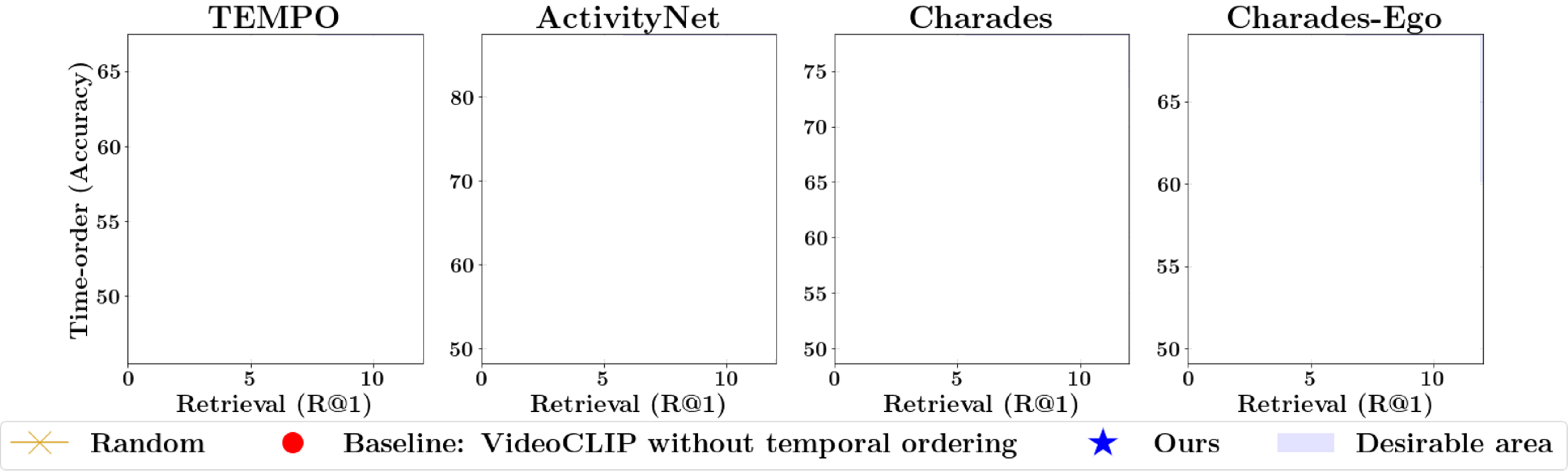
Putting on shoe/shoes before holding a mirror

(c) Charades

Taking a broom from somewhere before holding a dish

(d) Charades-Ego

Experiments



4. Human values

Work in progress with the UvA Data Science Center HAVA-Lab.


Heineken
PREMIUM QUALITY
CAFE MOOY
KOLKSTEEG
(CENTRUM)
Heineken Bieren Café
Café mooy
Poffertjes
Soep v/d Dag
Tosti's
Div. belegde
Pistolets
Hot Dog
Pizza
Ham/Cheese
Burger
Bal gehakt
Koffie/Thee
Irish Coffee
Warme Choco
met Slagroom
POOLTABLE
Café
Mooy


世界人民大团结万岁

HAVA-Lab

What defines **human-aligned video-AI**, how can it be made computable, and what determines its societal acceptance?

How can we **embed laws, societal values, and ethics** in video AI's algorithm lifecycle?

Is there one solution for all, or do we need specialized **algorithms for each domain?**

Cees Snoek

Pascal Mettes

Iris Groen

Heleen Janssen

Tobias Blanke

Paula Helm

Marie Lindegaard

Erwin Berkhout

Stevan Rudinac

Marlies Schijven

Conclusions

Foundation models are amazing.

But have perceptual difficulty with **scarcity**, **space**, **time** and **human values**.

Small-capacity adapters and **synthetic data generation** may help.

Bonus: both sustainable and responsible.

Thank you

Contact info

VIS
LAB
VIDEO & IMAGE SENSE LAB

Prof. dr. Cees Snoek
https://ivi.fnwi.uva.nl/vislab/
@cgmsnoek {x, bsky.social}