Self-supervised Learning for scaling to more modalities and data

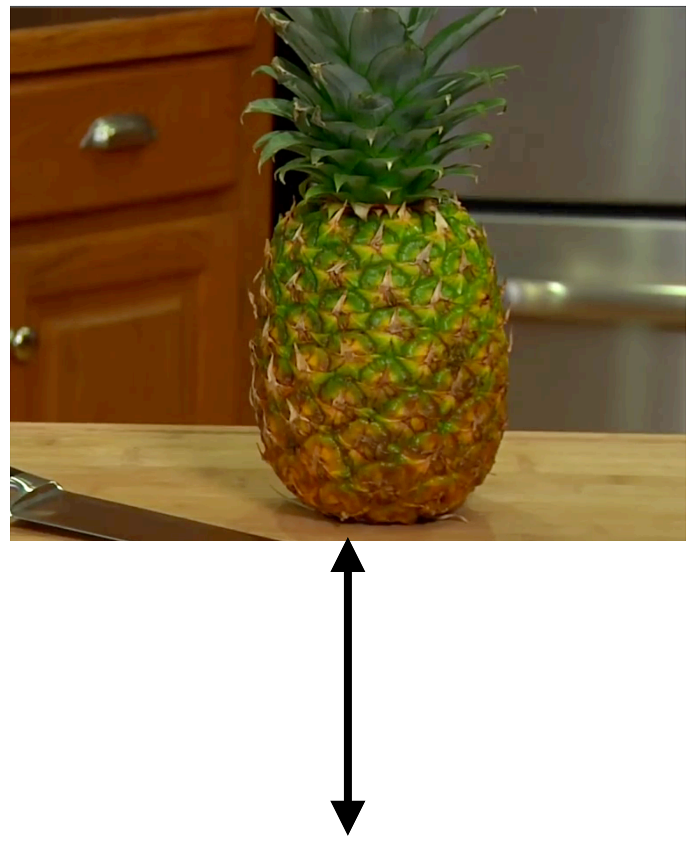
Ishan Misra

GenAI@Meta

BigMAC Workshop, ICCV 2023

The era of multimodal learning

- Get billions of (image, text) pairs
- Learn representations that "align" images with text



A pineapple sitting on the counter

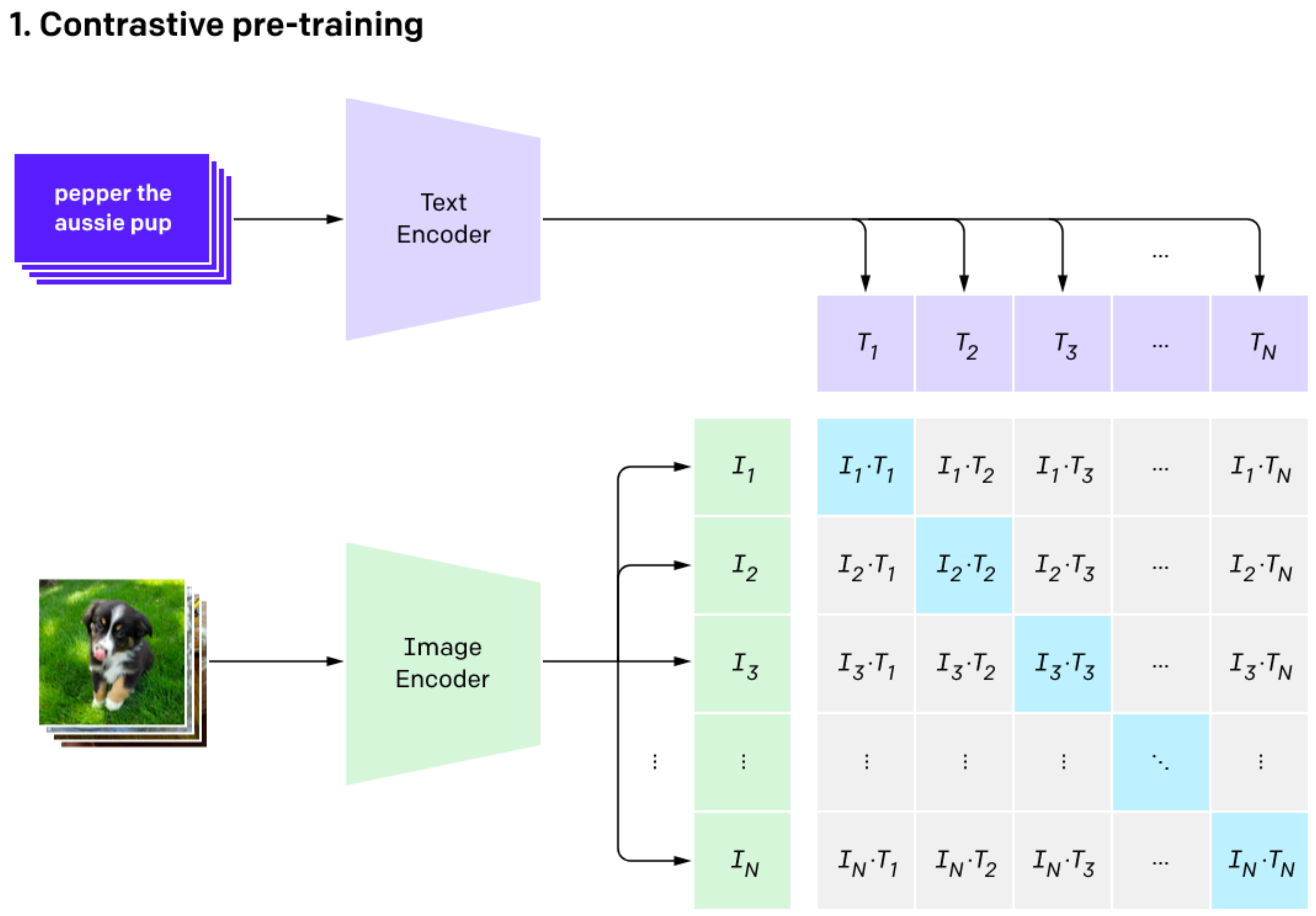


Image source: CLIP - Radford et al., 2021

Aligned image-text features

- Aligned representations are *really* useful

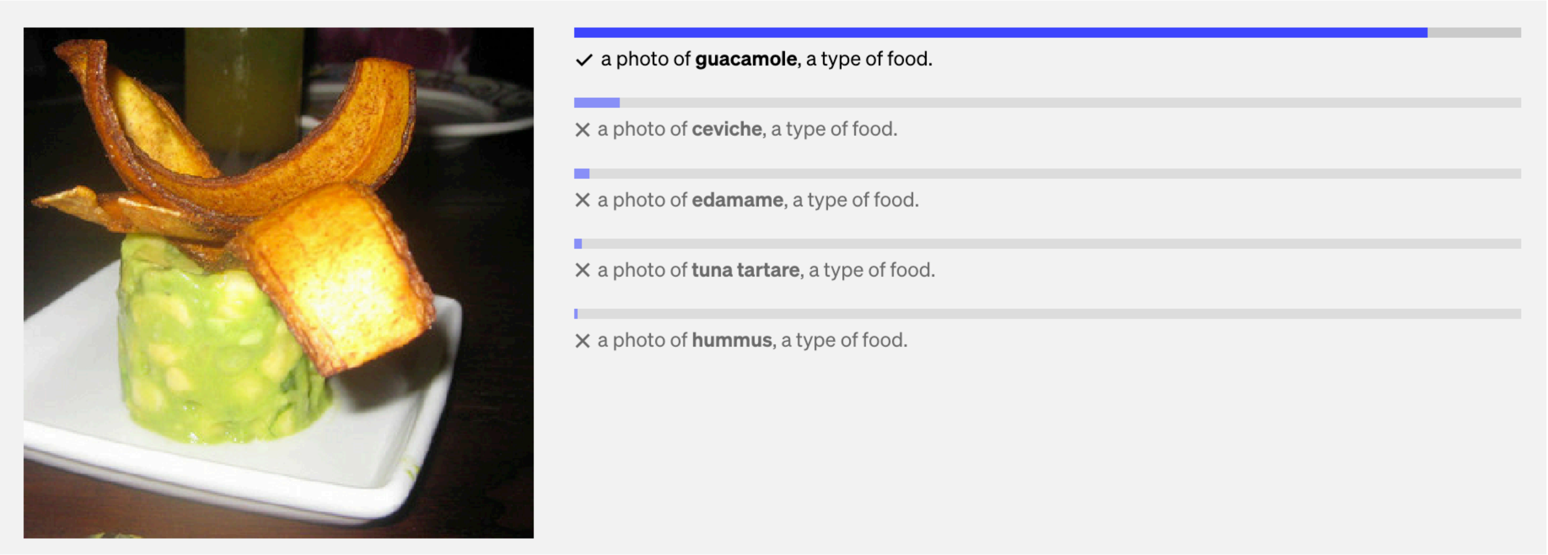
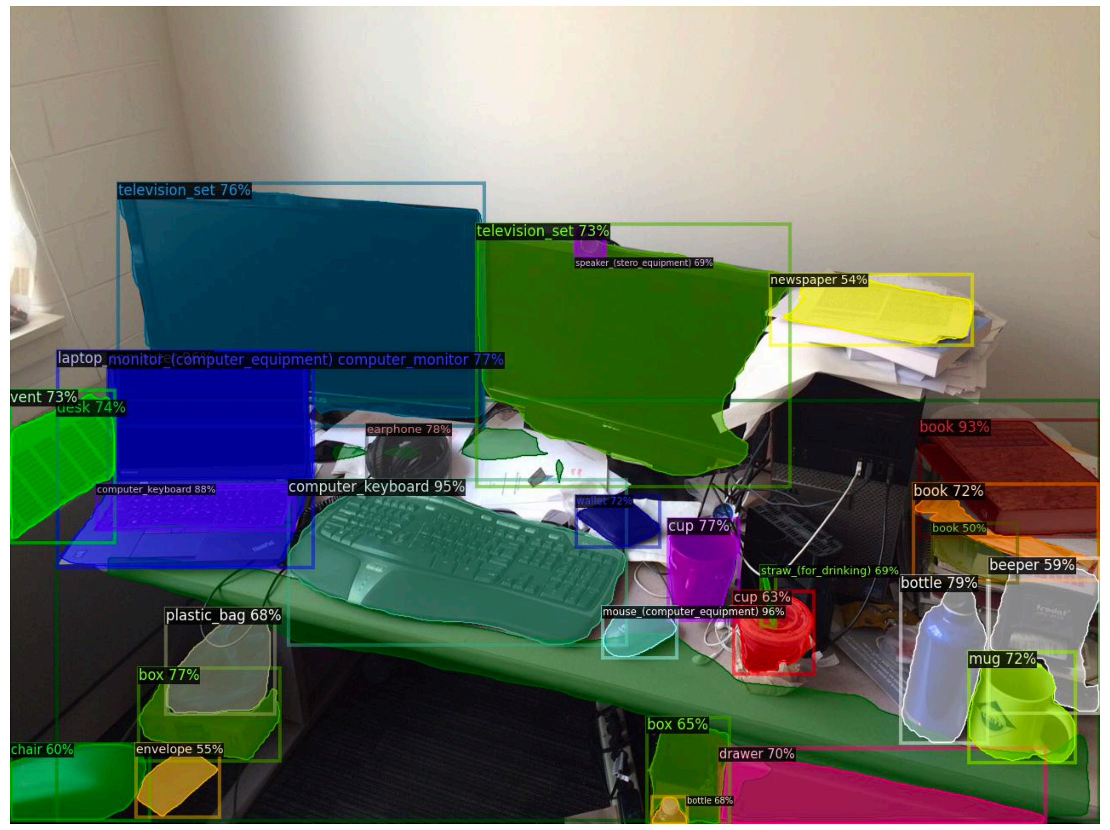
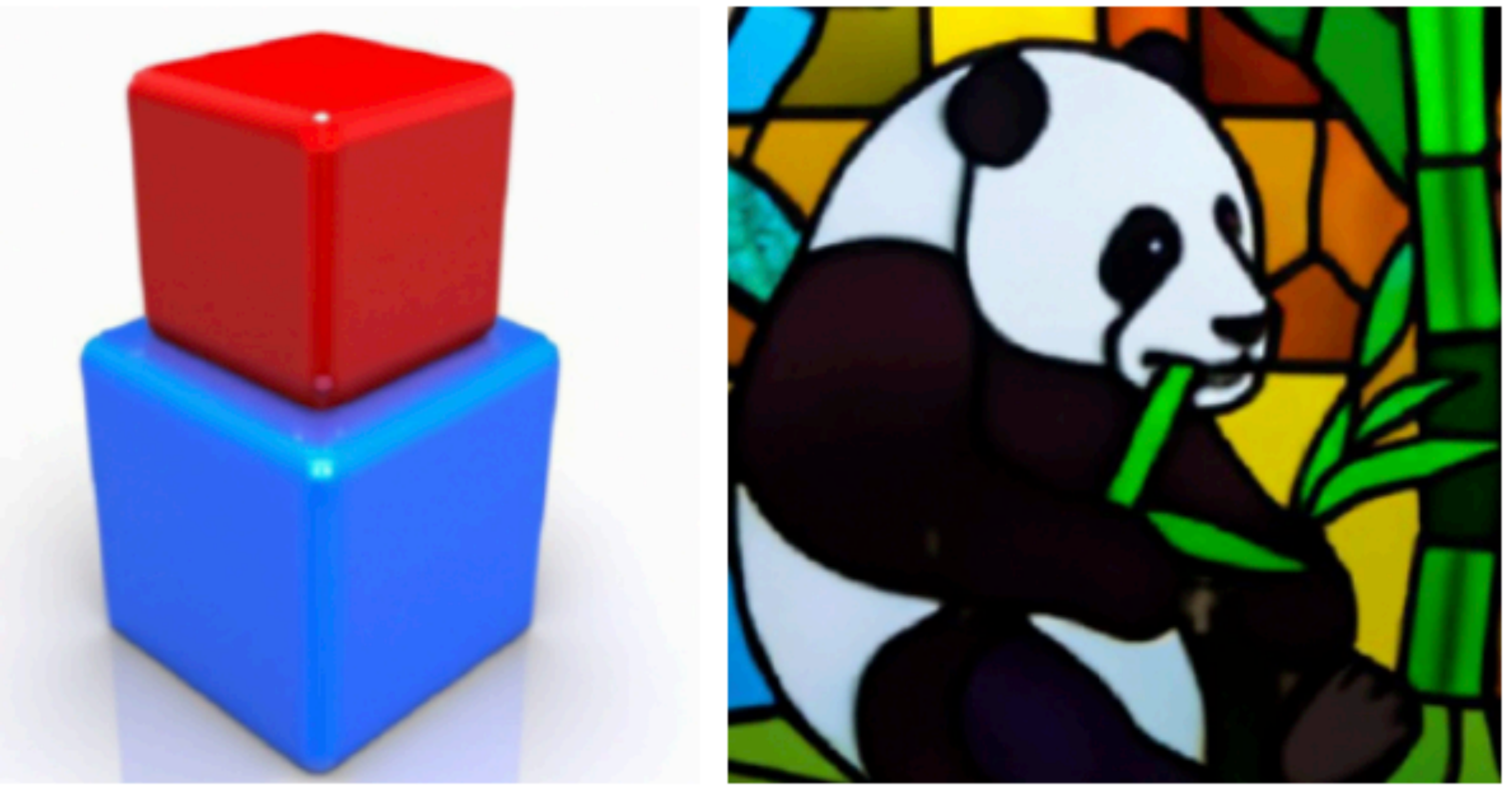


Image-text retrieval
Open-vocabulary classification[1]



Open-vocabulary detection and segmentation[2]



“a red cube on top of a blue cube”

“a stained glass window of a panda eating bamboo”

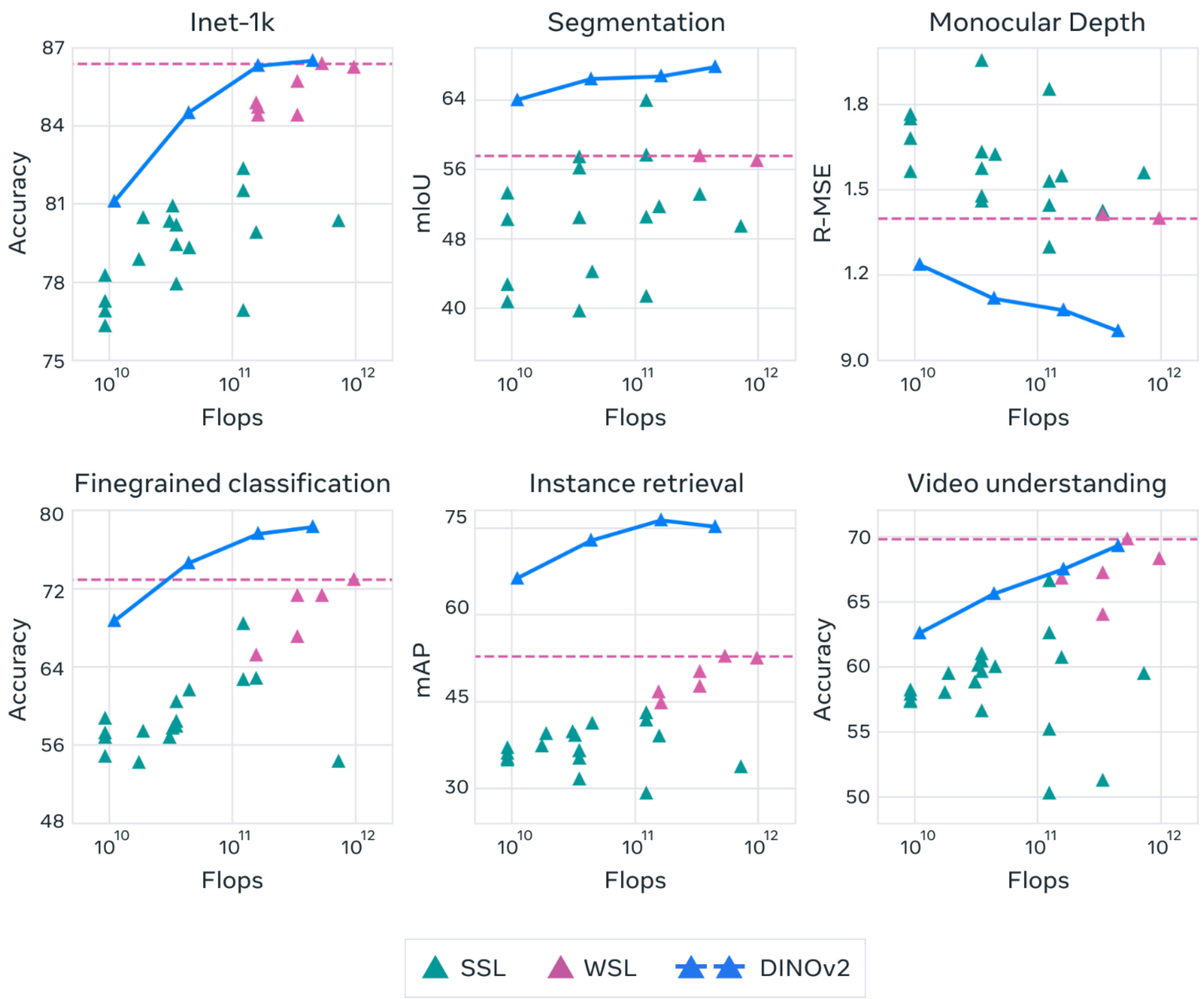
Text to image generation[3]

[1] CLIP - Radford et al., 2021
[2] Detic - Zhou et al., 2022
[3] GLIDE - Nichol et al., 2022, LAFITE - Zhou et al., 2022

Does SSL Matter?!

- Especially in this era of strong image features from (image, text)?
- Scaling (image, text) data is the way forward?

Standalone SSL is scaling well



DINOv2 - Oquab et al., 2023

SSL vs. Weakly supervised Debate

Image credit - Wikimedia

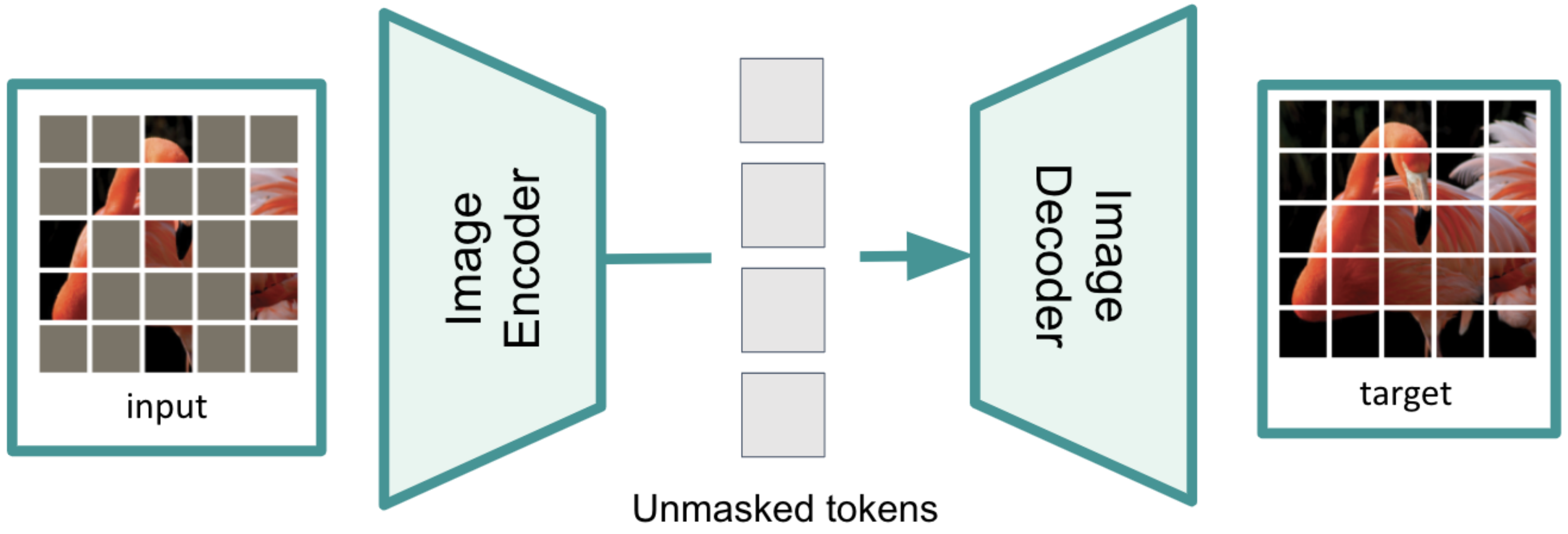
SSL ~~vs.~~ and Weakly supervised ~~Debate~~

Image credit - Wikimedia

Image credit - Flickr

SSL

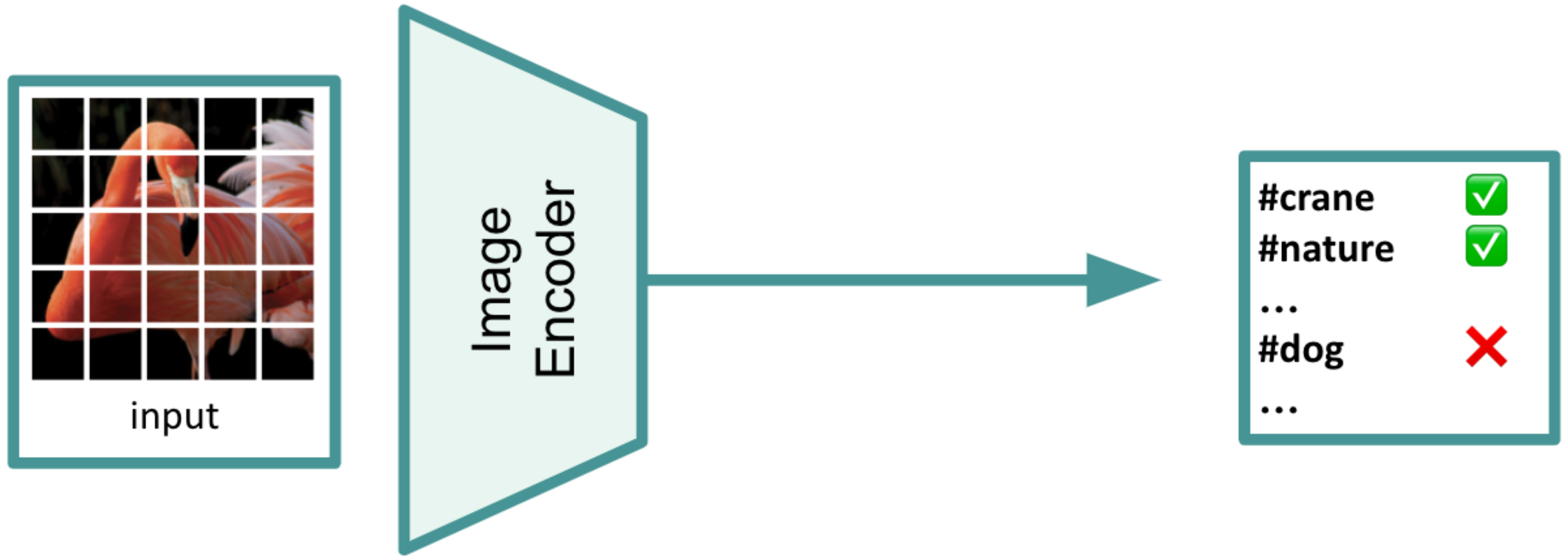
Self-Supervised Learning



Ex: Image Reconstruction (**MAE**)

WSP

Weakly Supervised Pretraining



Ex: Noisy Label Supervision (**SwAG**)

Meta AI

MAE

Great potential on diverse downstream tasks

Great fine-tuning classification performance

Great on dense prediction tasks like detection (ViTDeT)

F

WSP

Basis for SOTA foundational models

SOTA for classification (fine-tuning)

SOTA Zero Shot Capabilities (CLIP, LiT)

The effectiveness of MAE pre-pretraining for billion scale pretraining

Mannat Singh*, Quentin Duval*, Kalyan Vasudev Alwala*, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, Rohit Girdhar, Ishan Misra

ICCV 2023 Poster (Wednesday)

Key idea

- Introduce a “pre” pre-training stage
- Pre-pretraining uses self-supervised learning (no labels)
- Initialize and train as usual

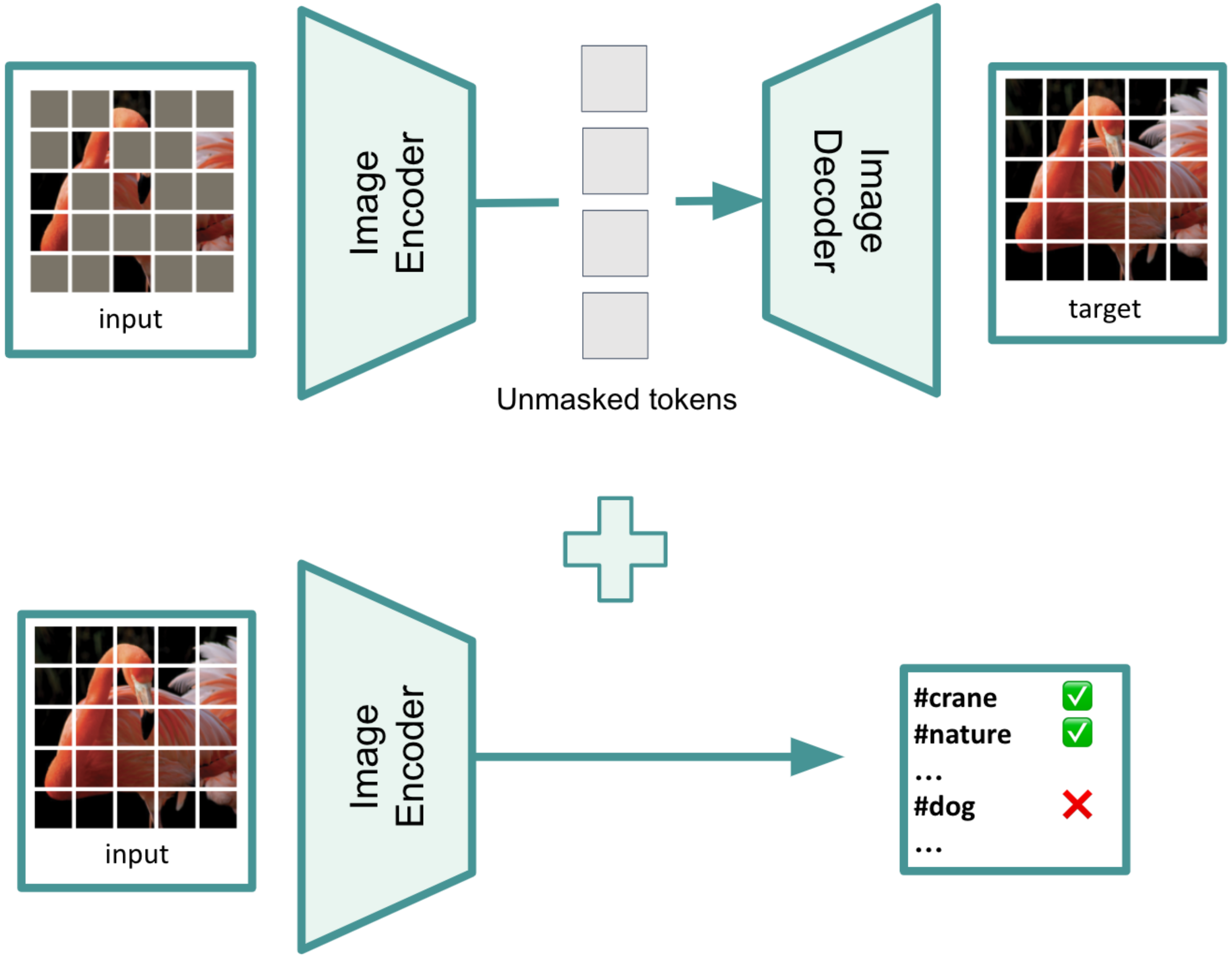
Pre-pretraining

Step 1: Pre-pretraining

- Use Masked AutoEncoders (MAE)
- Low FLOPs (75% masking)

Step 2: Standard weakly supervised training

- Use image labels
- Multi-target prediction (no contrastive learning!)
- Simple yet SOTA



Pre-pretraining at scale

Dataset: Instagram-3B

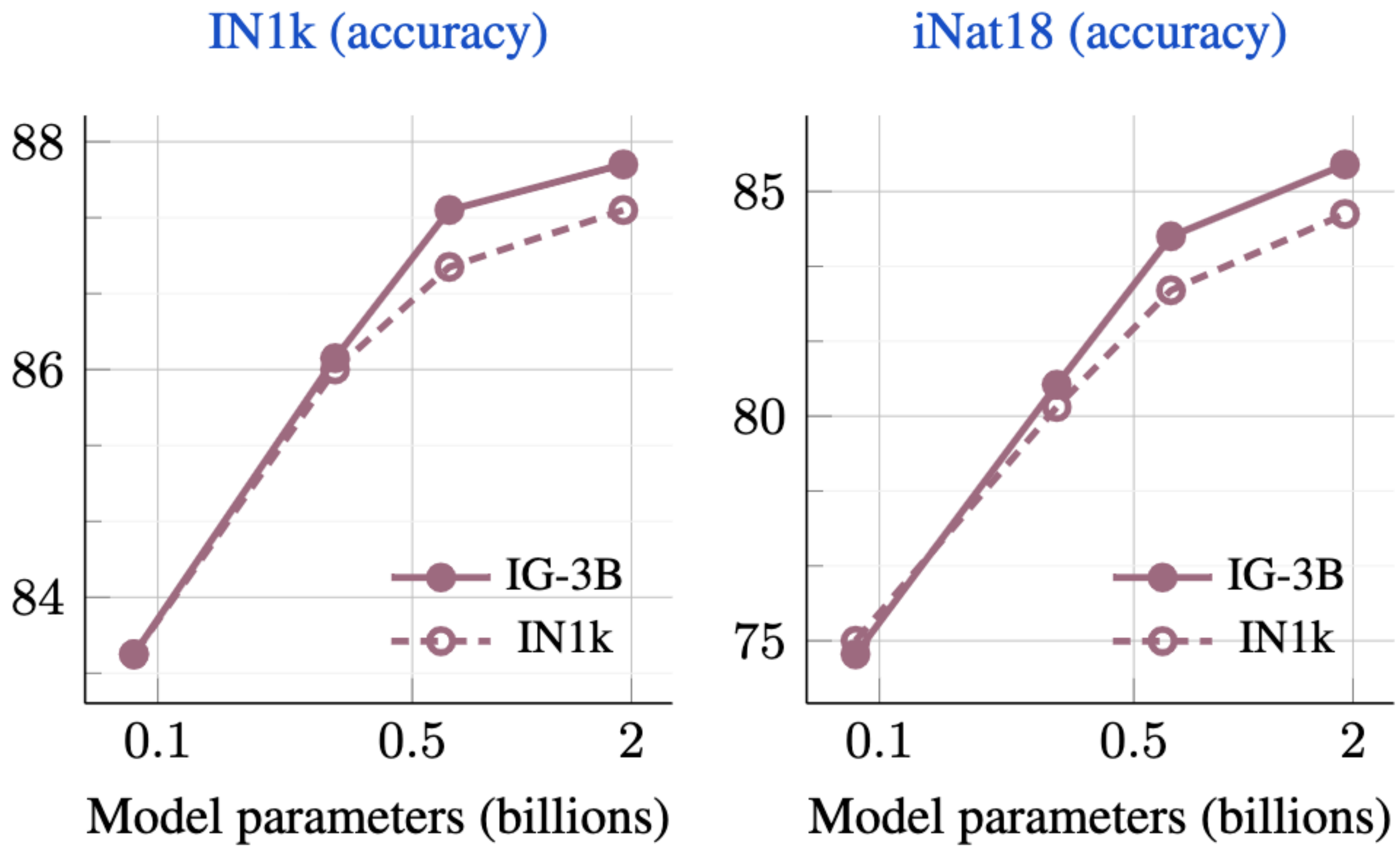
- 3B unique images
- 5B images after resampling

For weakly-supervised

- 28K unique hashtags

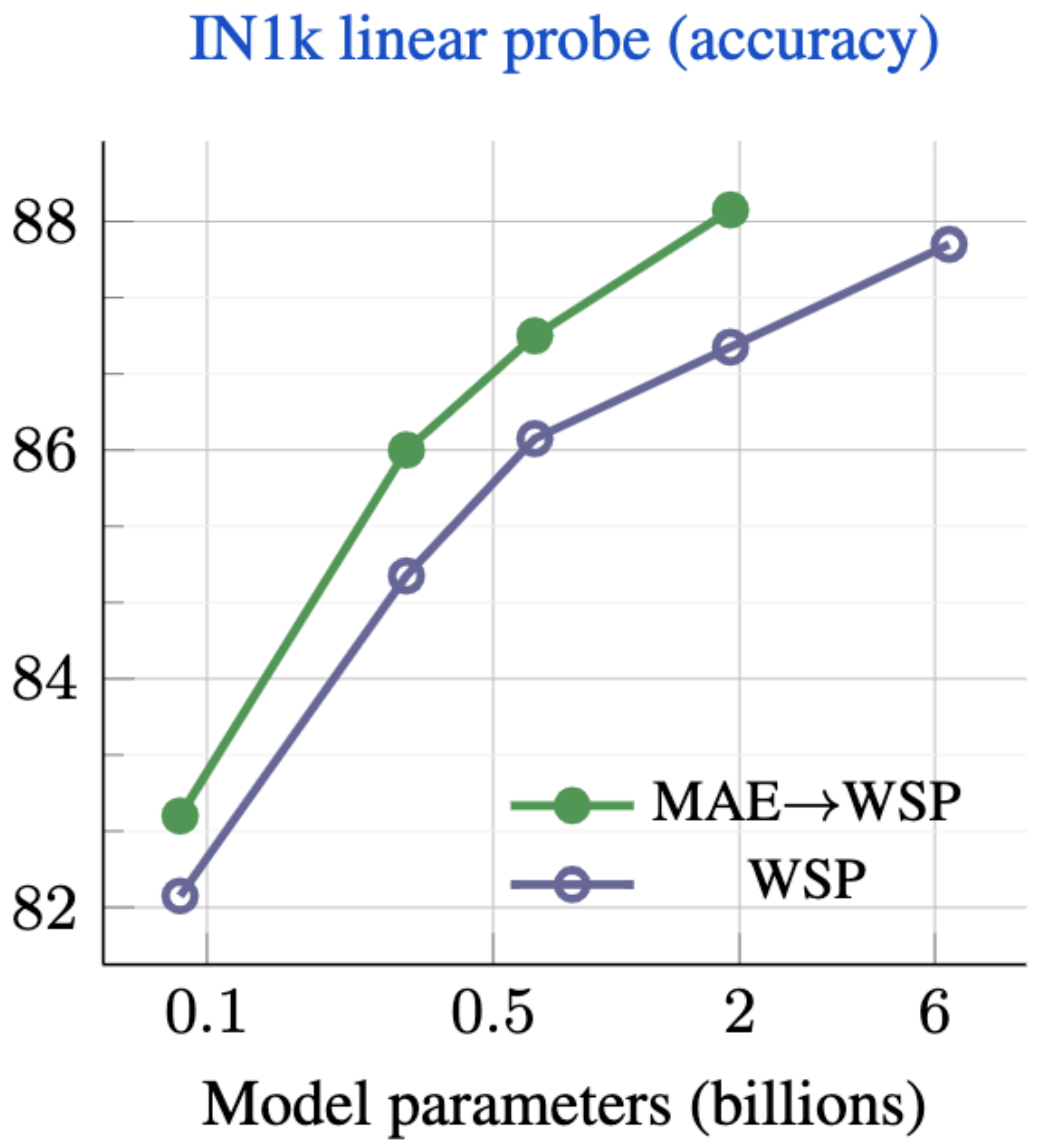
Architecture: ViT up to **6.5B** params

MAE scales with **both** data and model



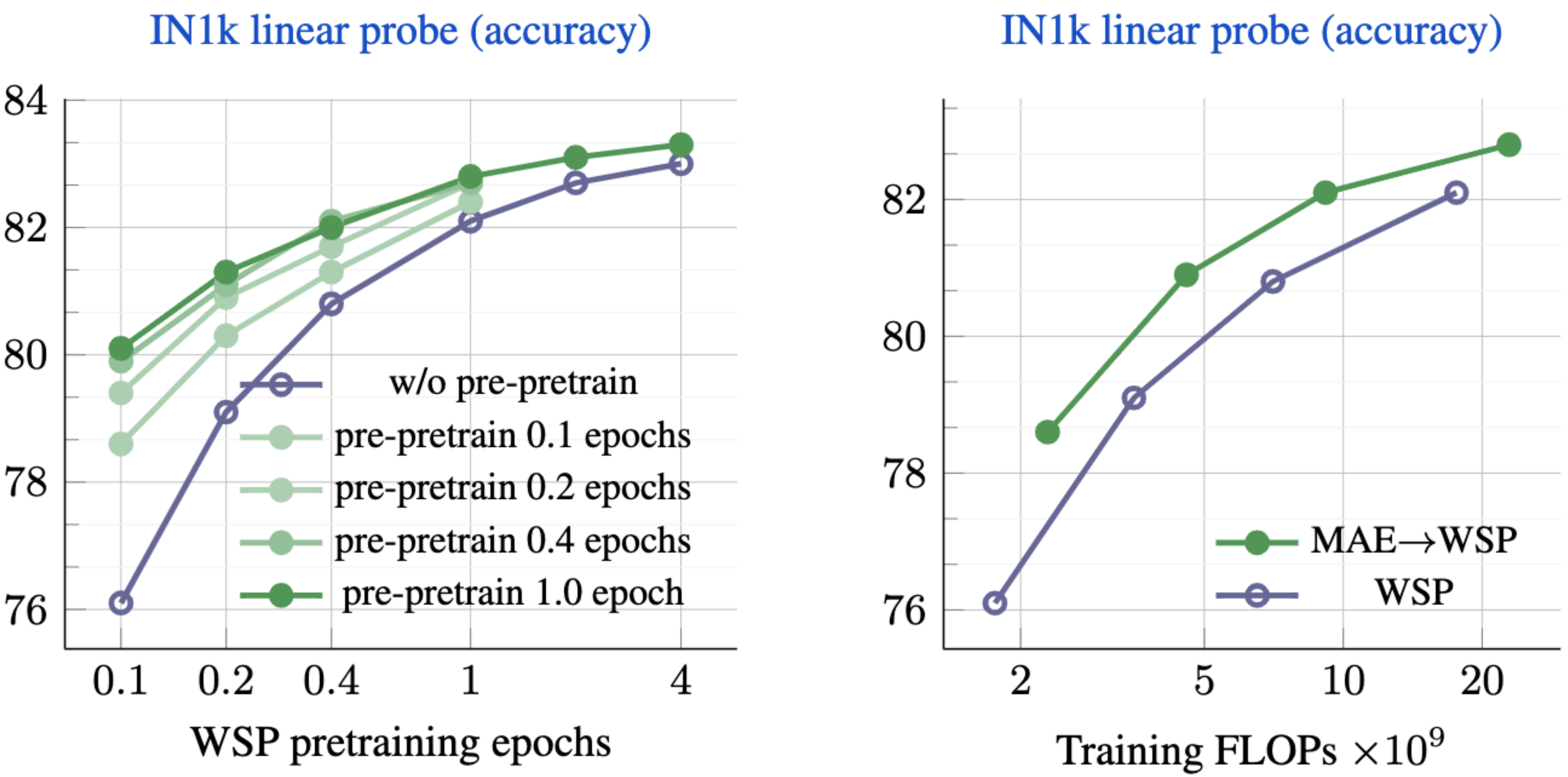
He et al., 2022 showed it scaled only with model size

Pre-pretraining matters at large scale too!



- Improves performance across all model & data sizes

Pre-pretraining matters at large scale too!



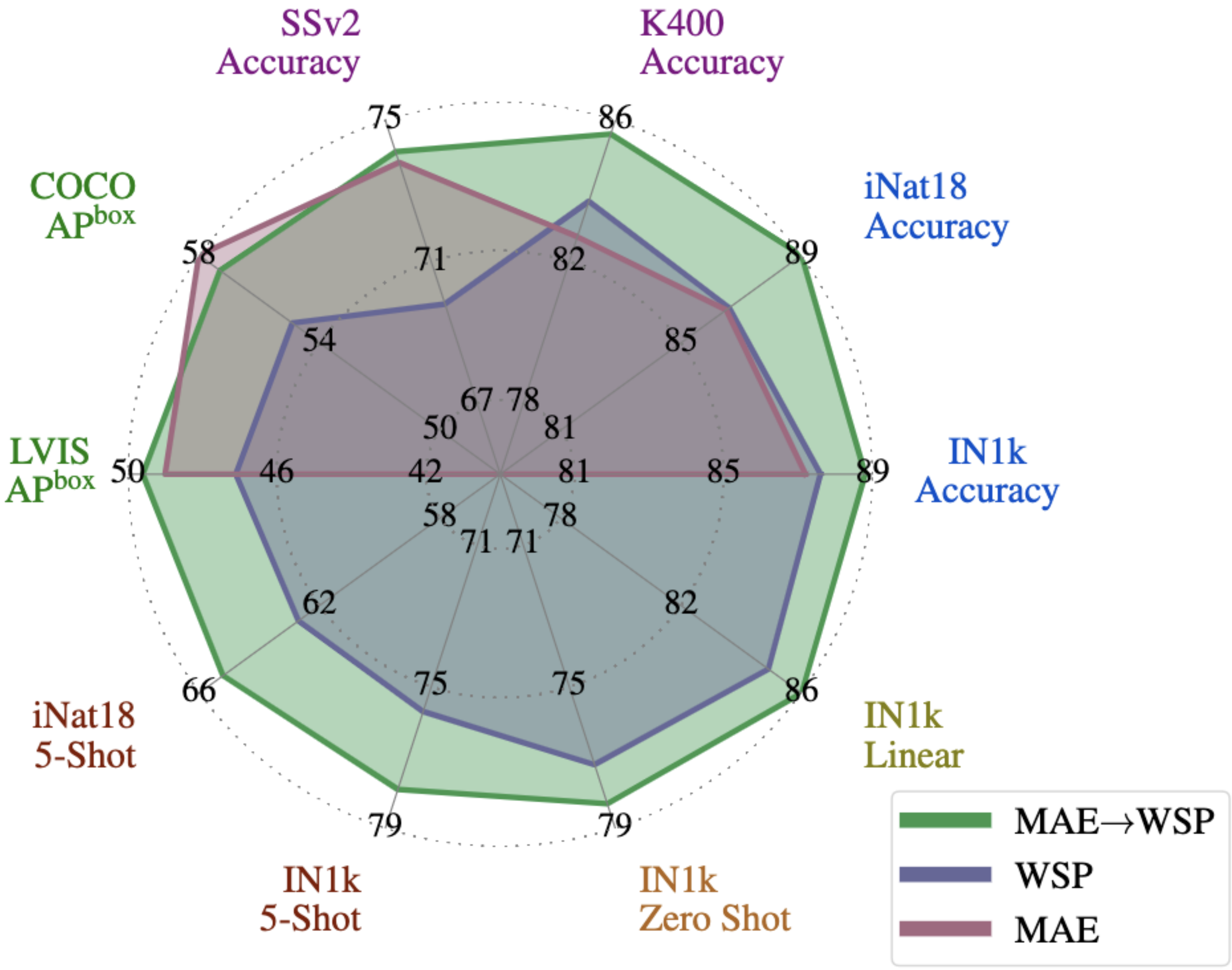
- More efficient! —> Better performance at fewer FLOPs

Best of SSL and WSP

MAE shines on dense prediction tasks

WSP shines on classification tasks

MAE->WSP combines their strengths



Pushing the state-of-the-art

iNaturalist-18 Fine-tuning	ImageNet1k 1-shot	Food101 0-shot	Object Net OOD eval
91.3% top-1 accuracy	**62.1%** top-1 accuracy	**96.2%** top-1 accuracy	**75.8%** top-1 accuracy

Multi-modal != **Bi**-modal

There are other modalities …

Image source: Rawpixel, The Rijksmuseum

Aligned data is hard to get

Depth

Thermal

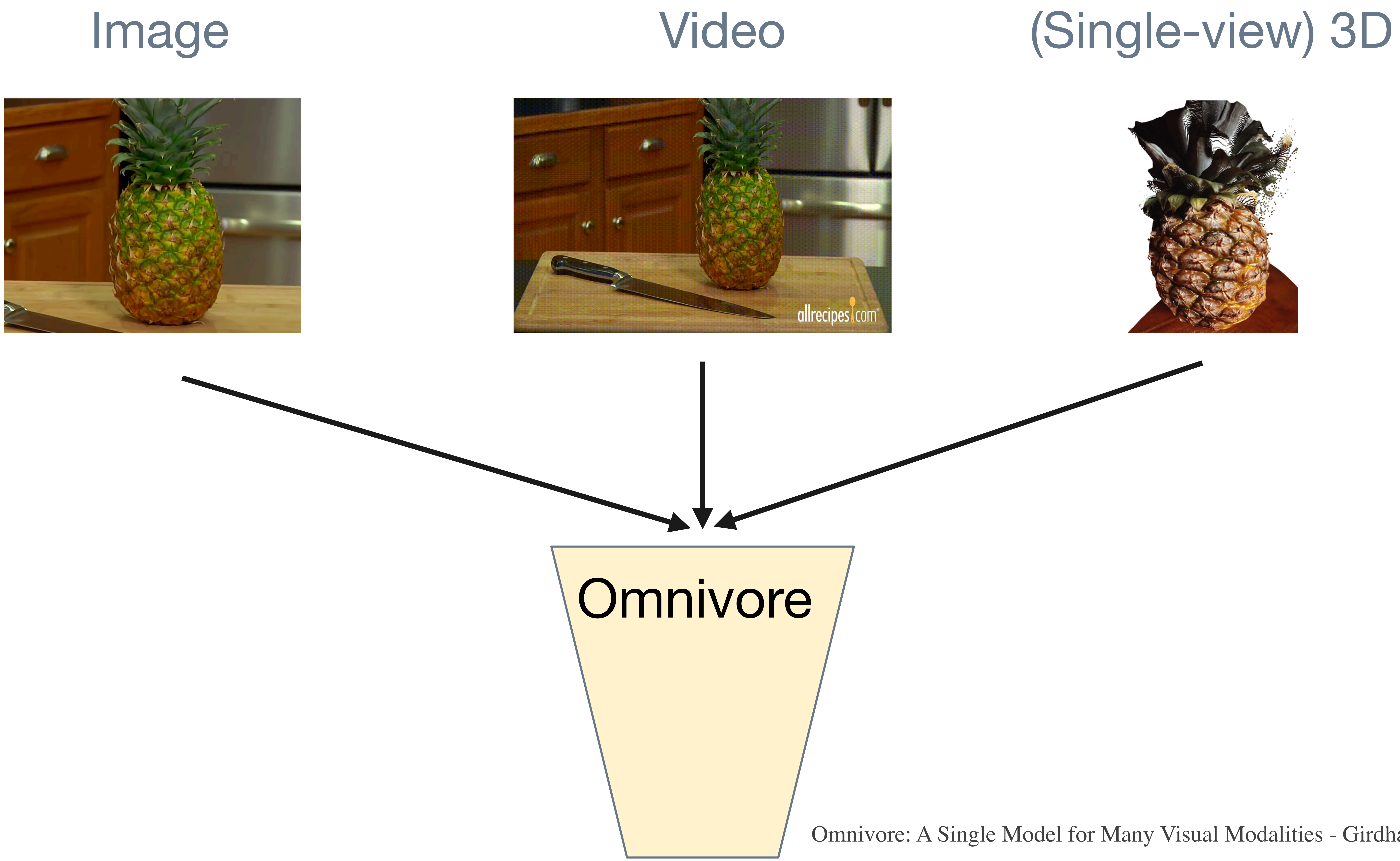
Motion (IMU)

Audio

Image source: Rawpixel, The Rijksmuseum

Solution 1: Single model

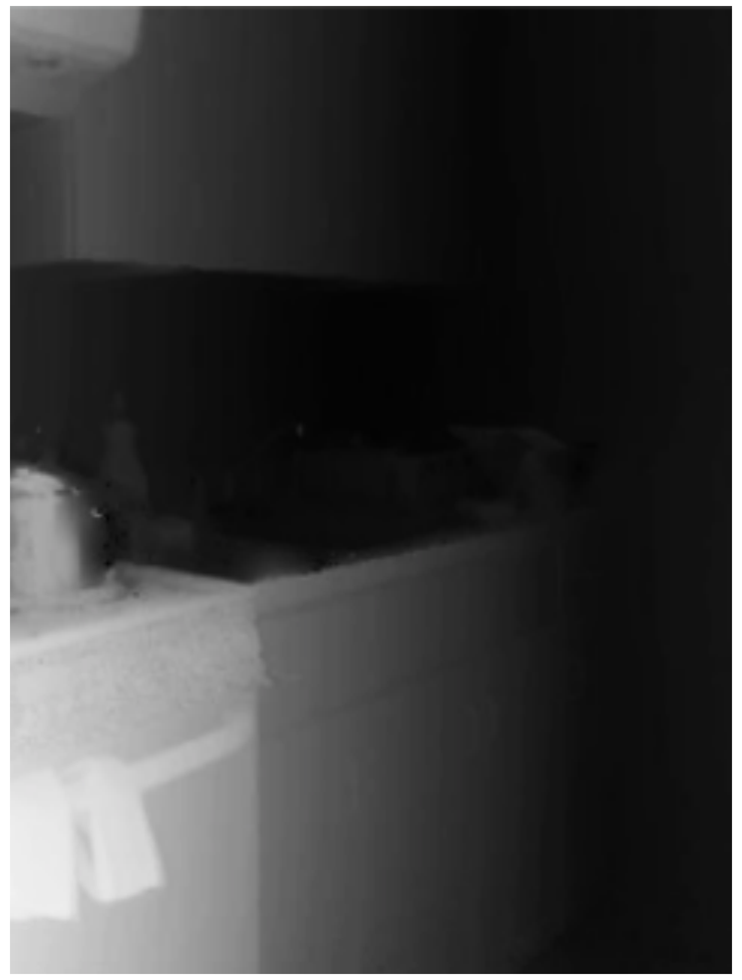
Omnivore: A Single Model for Many Visual Modalities



Omnivore: A Single Model for Many Visual Modalities - Girdhar et al., CVPR 2022

Omnivore: Cross-modal alignment emerges!

Images are a universal language



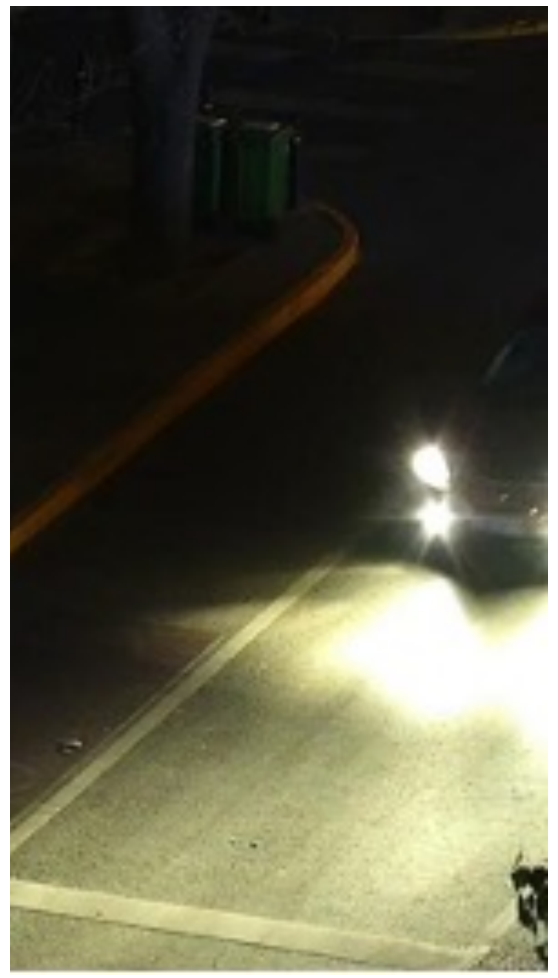
Depth

Thermal

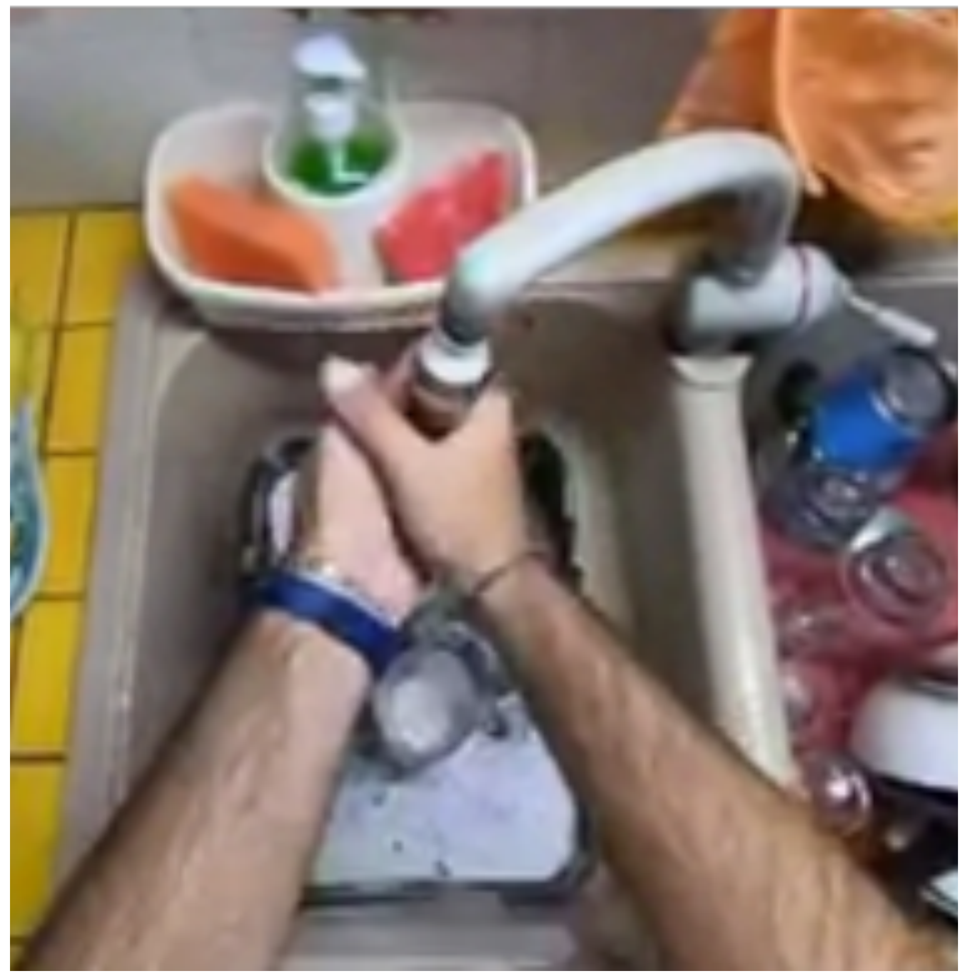
Motion (IMU)

Audio

RGB



RGB



RGB

RGB

Image source (L to R): SUN RGB-D, LLVIP, Isaque Pereira, Ego4D, Wikimedia, Gabriel Peter

Images are a universal language

Depth

Thermal

Motion (IMU)

Audio

RGB

RGB

RGB

RGB

Naturally co-occurring
“Self-supervised”

Image source (L to R): SUN RGB-D, LLVIP, Isaque Pereira, Ego4D, Wikimedia, Gabriel Peter

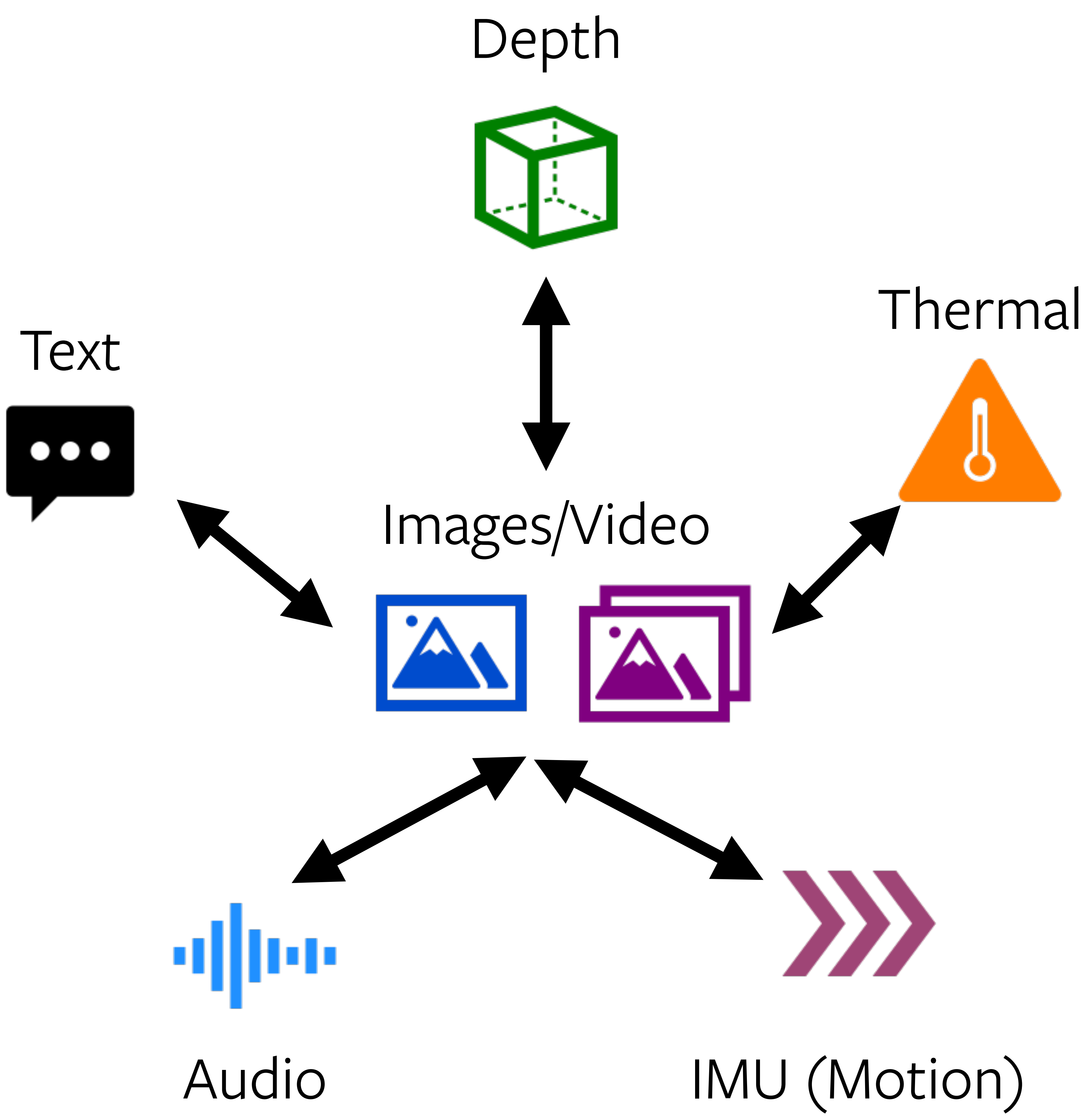
ImageBind: One Embedding to Rule them All

Rohit Girdhar*, Alaaeldin El-Nouby*, Zhuang Liu, Mannat Singh,
Kalyan Vasudev Alwala, Armand Joulin, Ishan Misra*

https://github.com/facebookresearch/ImageBind

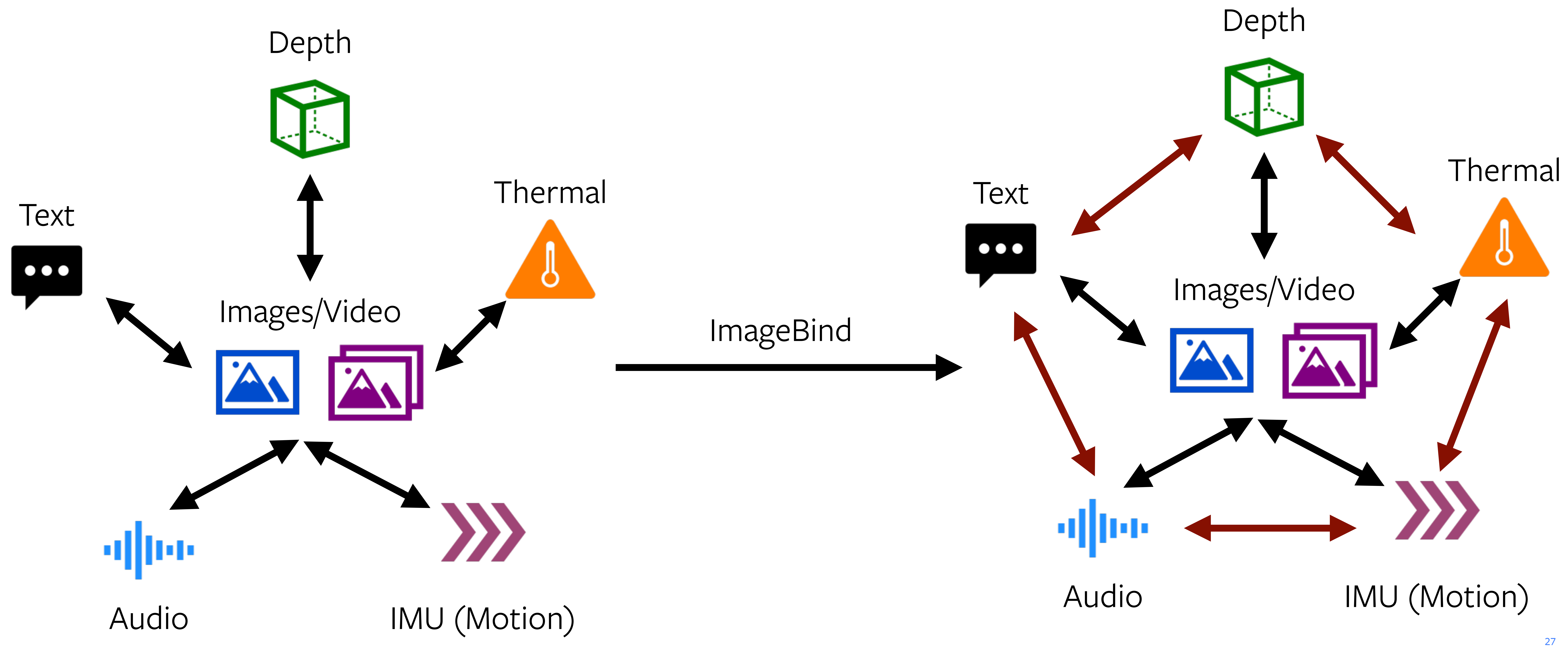
Key Idea

- Images naturally co-occur with different modalities
- Align every modality's representation with images
- Heavily leverage self-supervised learning



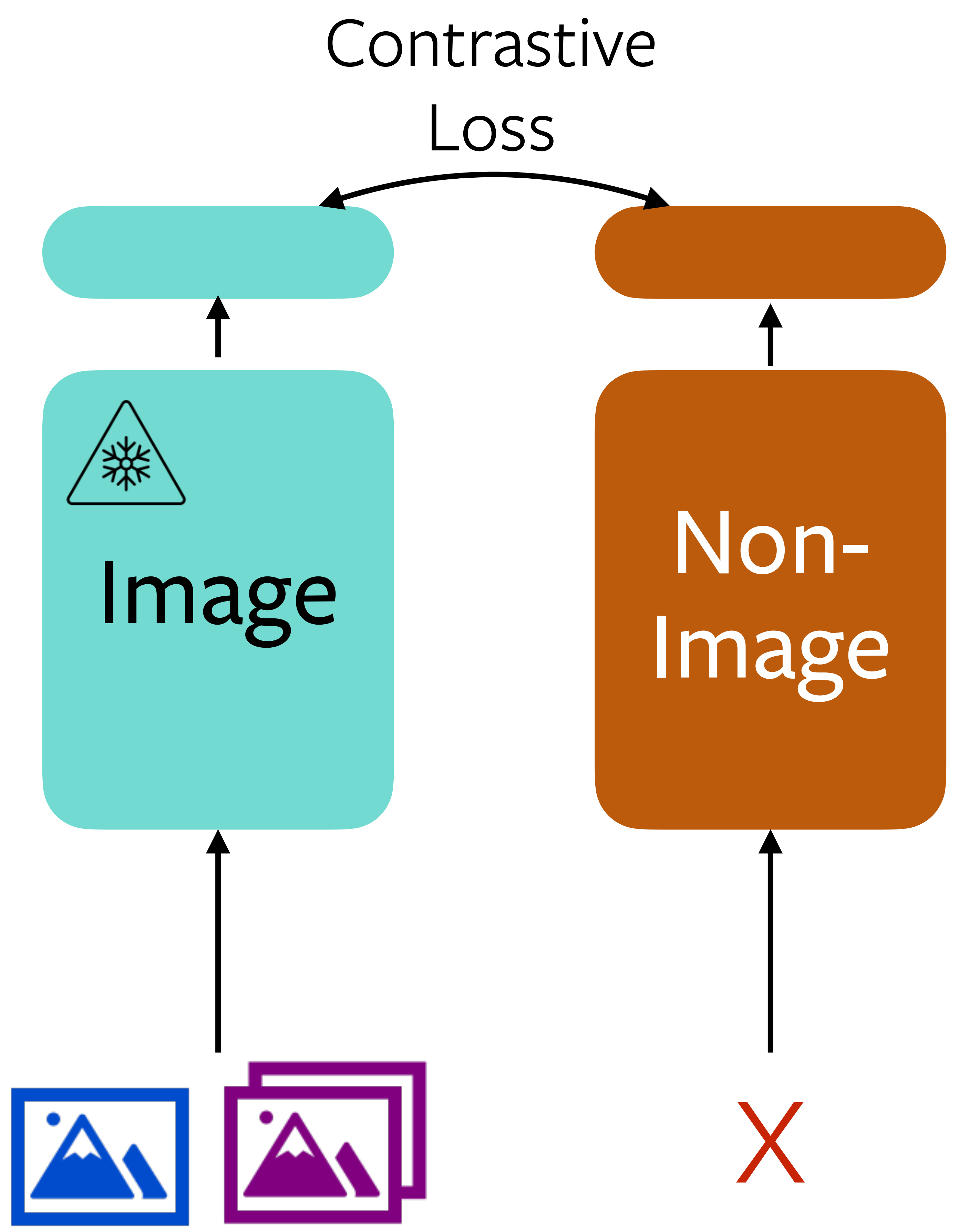
Emergent behavior (Transitive alignment!)

- After training **all** modalities are aligned



Training setup

- 6 modalities — Image/Video, Text, Audio, Depth, IMU, Thermal
- Train only with image-paired data
- Separate encoder per modality
- Initialize image & text encoder from CLIP/OpenCLIP and keep frozen



Measuring emergent alignment to text

- Train on (Image, X) (Image, Text)
- Test on (X, Text) —> **"Emergent"** zero-shot classification

	Image		Video		Depth		Audio			Thermal	IMU
	IN1k	P365	K400	MSVTT	NYU	SUN	AudioSet	VGGS	ESC	LLVIP	Ego4D
Random	0.1	0.27	0.25	0.1	10.0	5.26	0.62	0.32	2.75	50.0	0.9
ImageBind	77.7	45.4	50.0	36.1	**54.0**	**35.1**	17.6	**27.8**	66.9	**63.4**	**25.0**
Text paired	-	-	-	-	41.9	25.4	**28.4**	-	**68.6**	-	-
Absolute SOTA	91.0	60.7	89.9	57.7	76.7	64.9	49.6	52.5	97.0	-	-

ImageBind for “upgrading” existing models

Only takes text inputs

Your Favorite Model

Text

ImageBind for "upgrading" existing models

Only takes text inputs

Your Favorite Model

Text

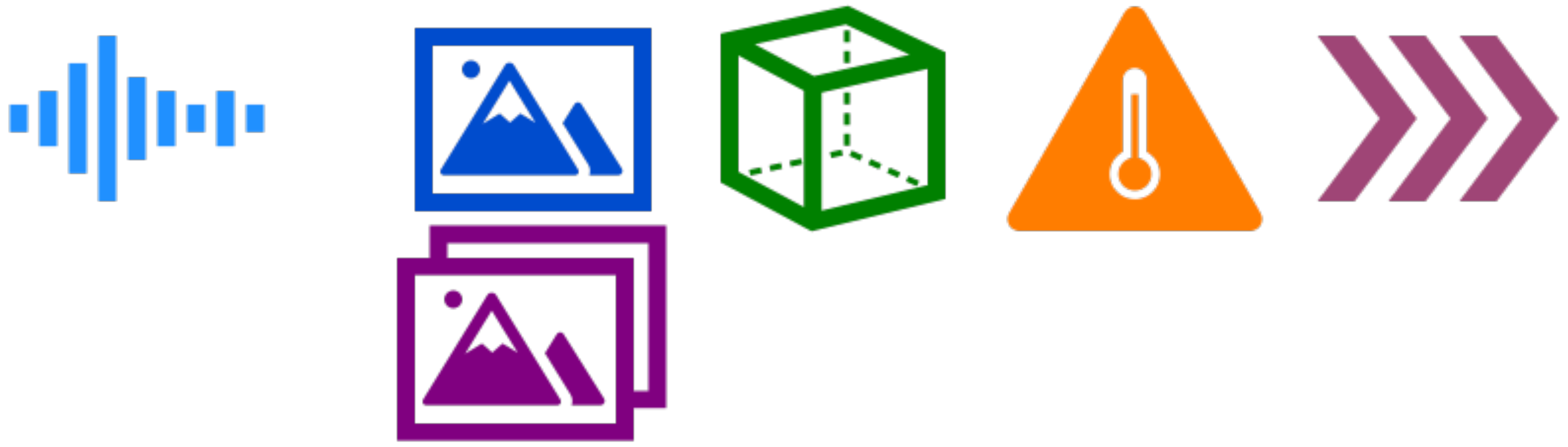
"Upgrade"

No re-training

"Multi" Modal

Your Favorite Model

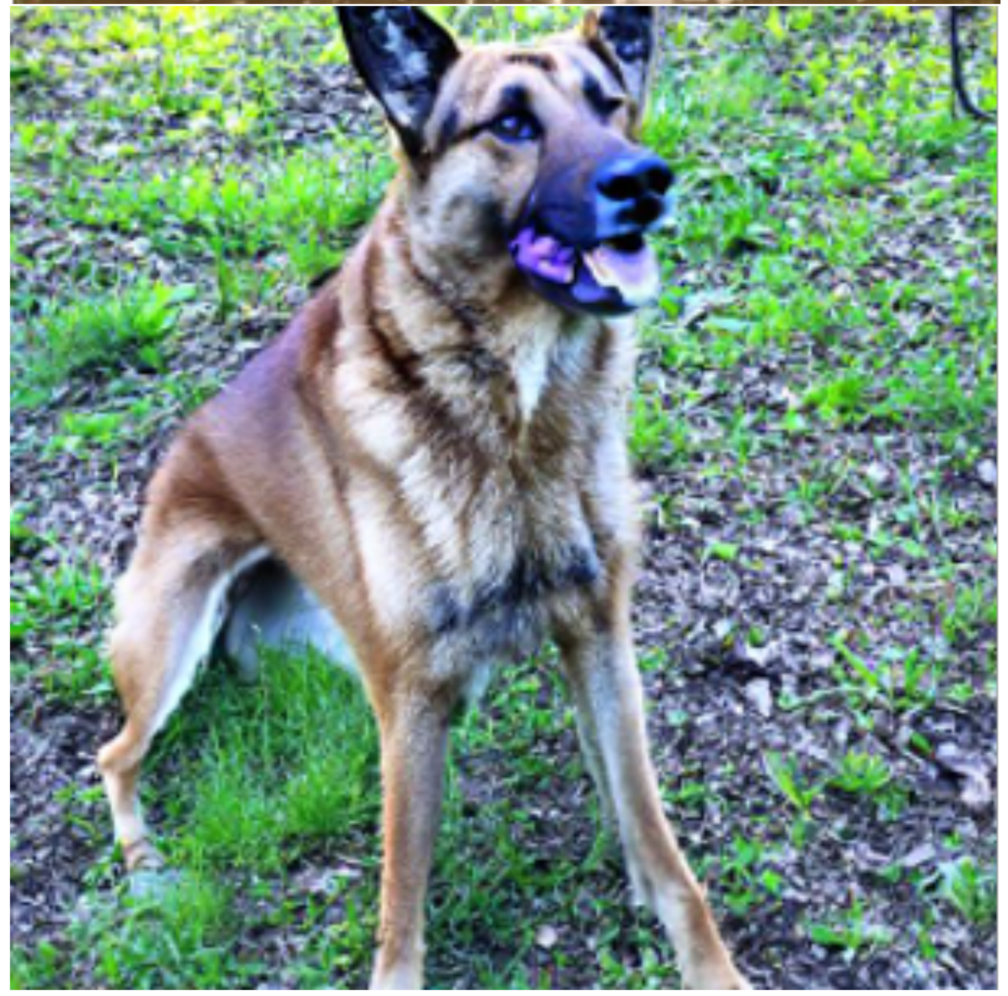
New modality



Audio-based prompting for image generation

Rain

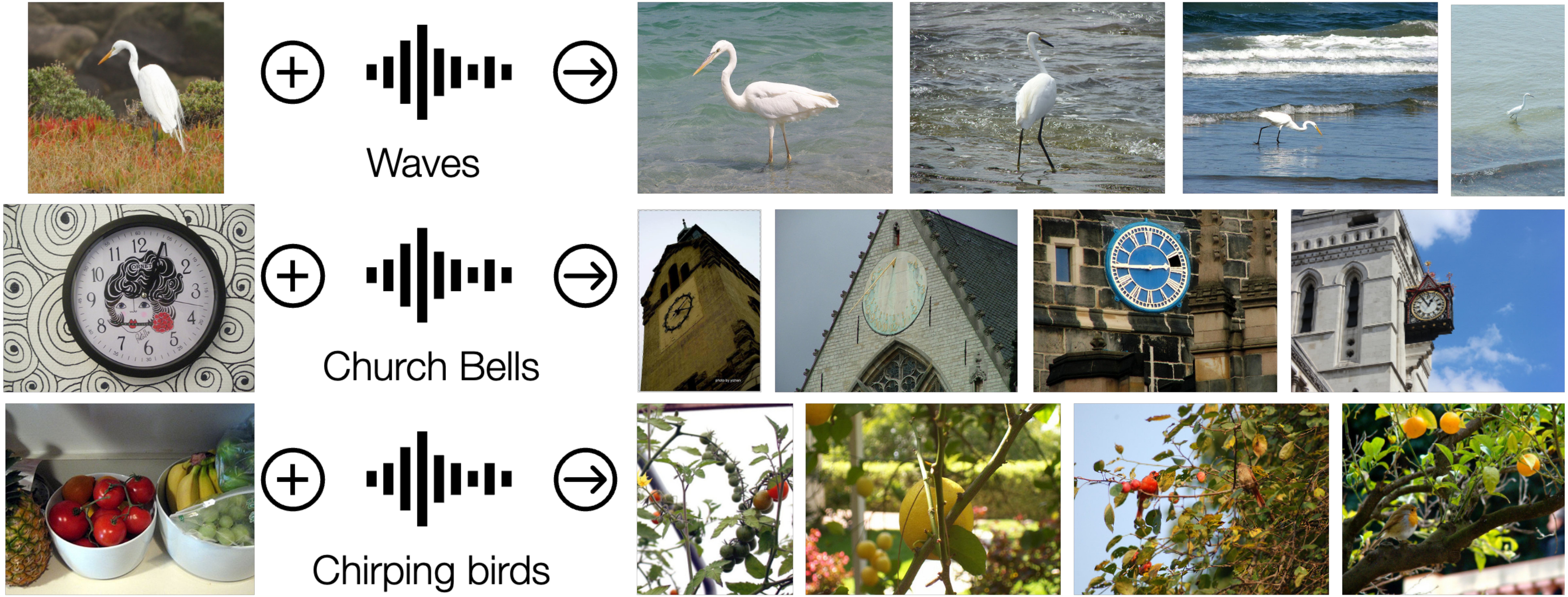
Bark



Fire

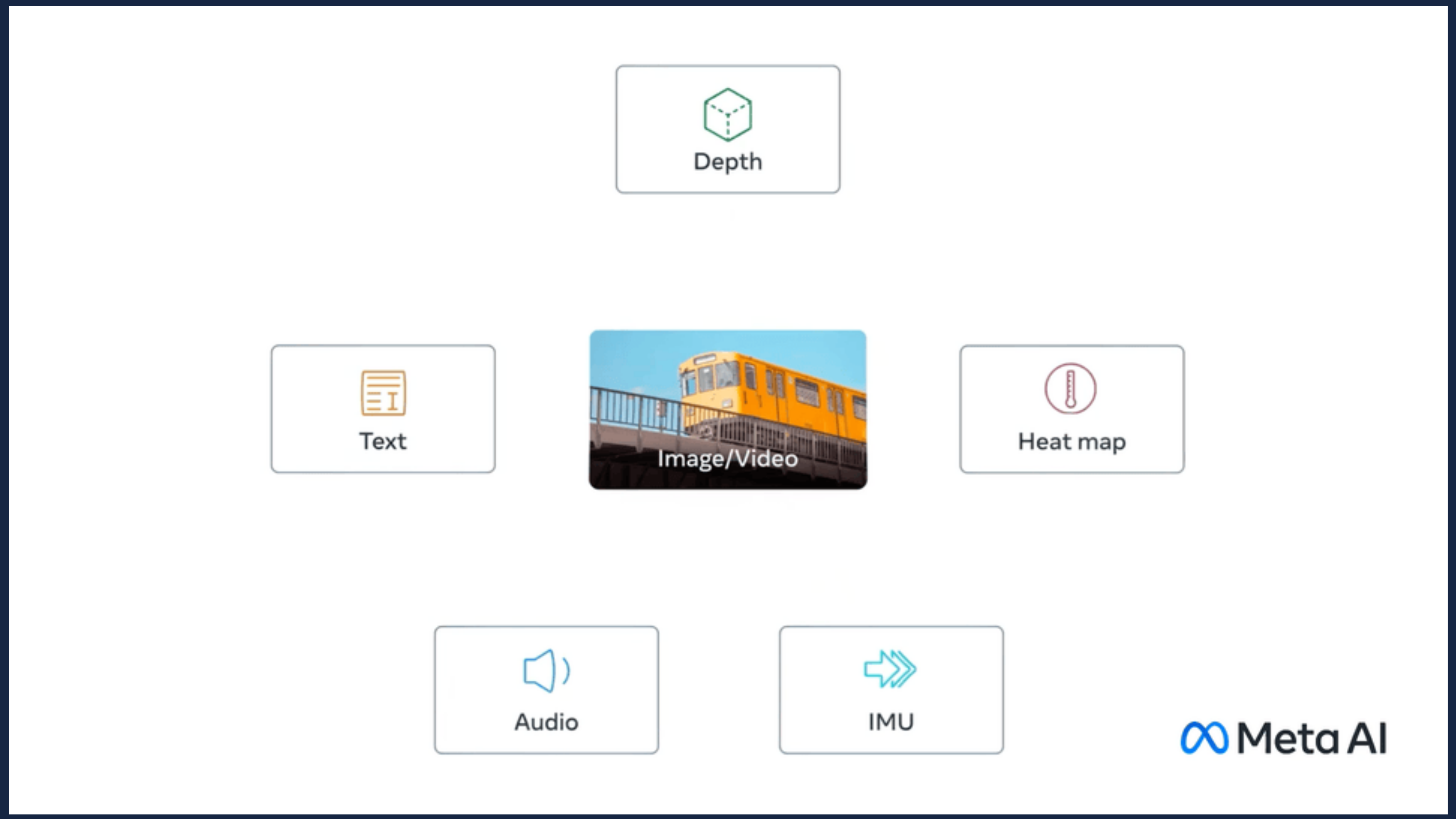
Engine

Aligned embeddings can be “added”



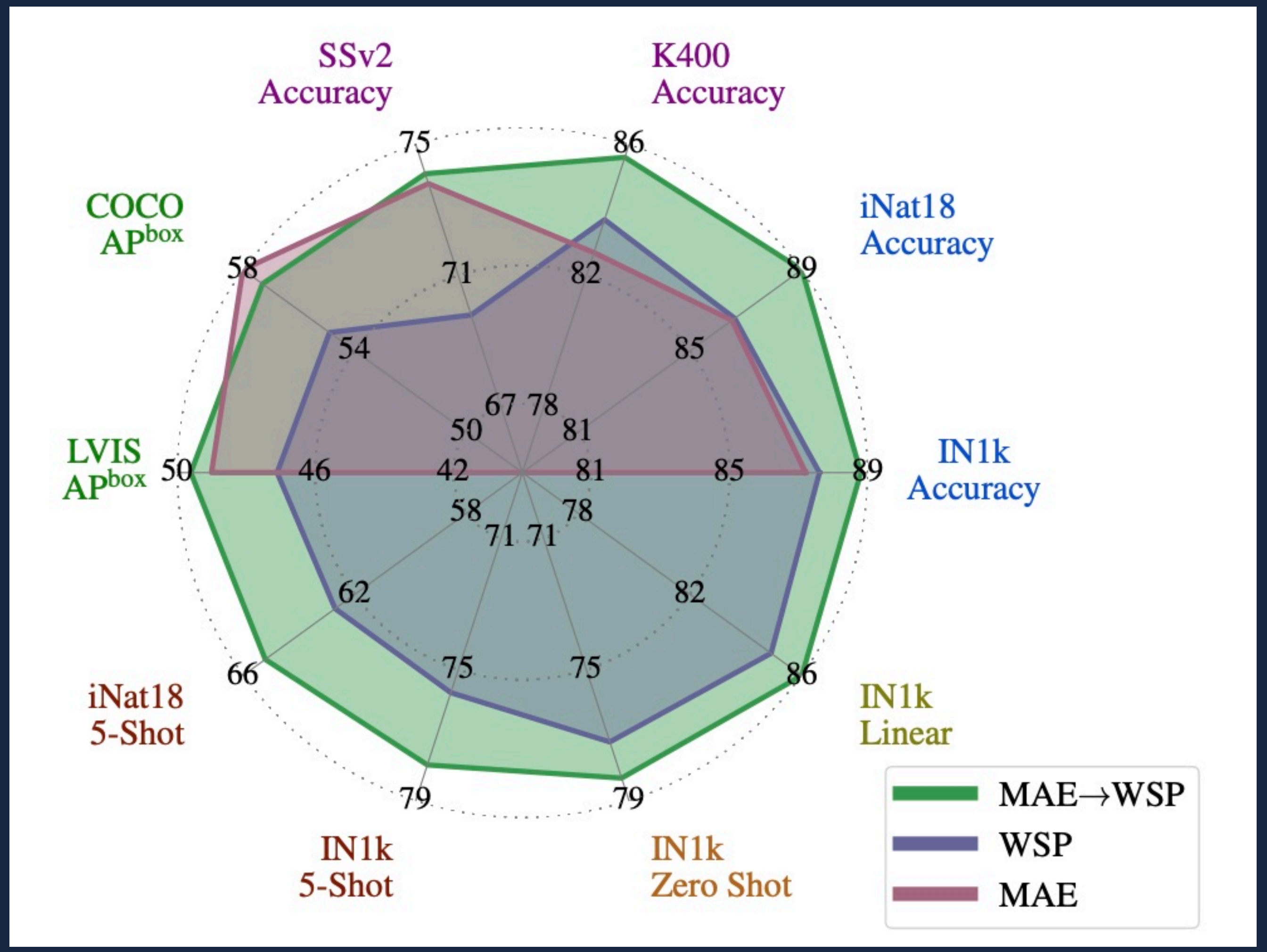
Thanks!

ImageBind



Code & Models released
https://imagebind.metademolab.com/

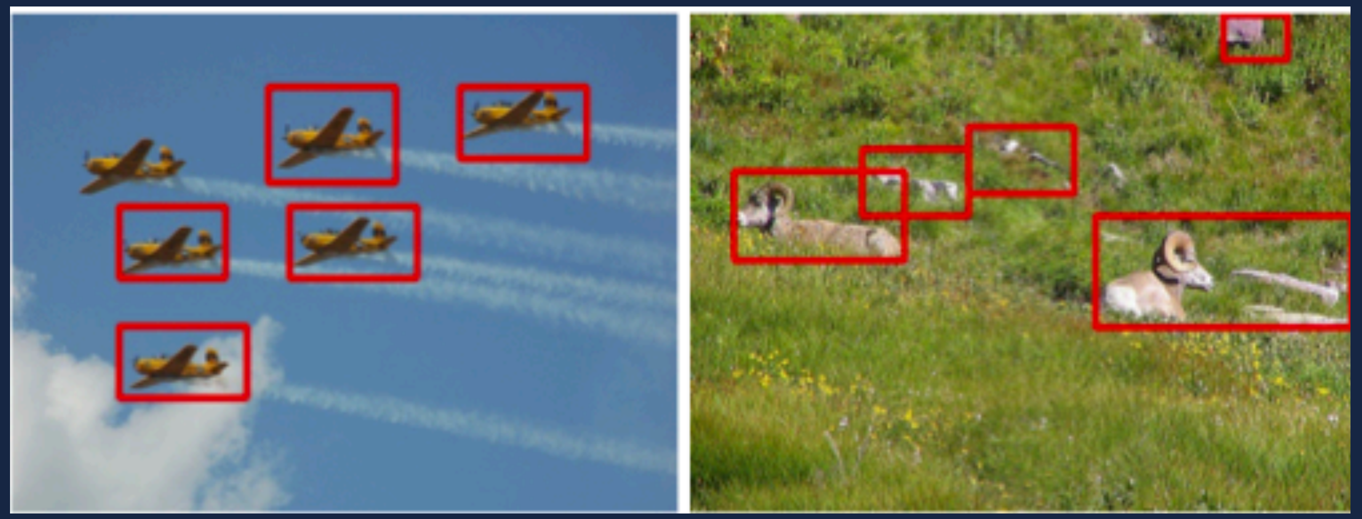
Effectiveness of MAE Pre-pretraining



Poster session (Wednesday)

Code & Models
https://github.com/facebookresearch/maws

MOST: Unsupervised Object Discovery



Poster session (Friday)

Code & Models
https://github.com/rssaketh/MOST/