max planck institut informatik

SIC Saarland Informatics Campus

High Level Computer Vision

Some Recent Trends: SigLIP, ImageBind, Eyes Wide Shut & GiT

@ July 24, 2024

Bernt Schiele

cms.sic.saarland/hlcvss24/

Max Planck Institute for Informatics & Saarland University,
Saarland Informatics Campus Saarbrücken

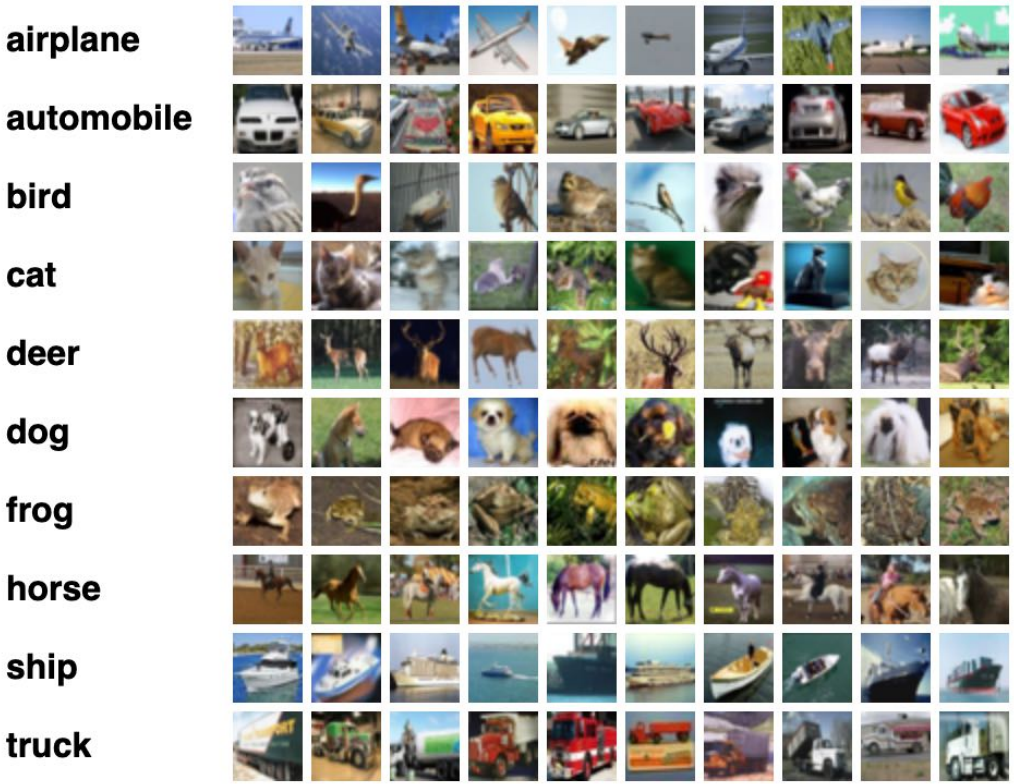
Overview of Today's Lecture

- SigLIP — Sigmoid Loss for Language Image Pre-Training
  - [iccv'23] — https://arxiv.org/abs/2303.15343
  - [arxiv'24] — https://arxiv.org/abs/2405.13777
- Multimodal Learning
  - ImageBind: [cvpr'23] - https://arxiv.org/abs/2305.05665
- Eyes Wide Shut
  - [cvpr'24] — https://arxiv.org/abs/2401.06209
- GiT: Towards Generalist Vision Transformer
  - [eccv'24] — https://arxiv.org/abs/2403.09394

slide credit: Lucas Beyer

supervised data vs

(cifar10, MSCOCO, ImageNet, JFT, ...)



Left image credit: https://www.cs.toronto.edu/~kriz/cifar.html
Right image credit: Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision, https://arxiv.org/abs/2102.05918

slide credit: Lucas Beyer

Contrastive Image-text learning (CLIP)

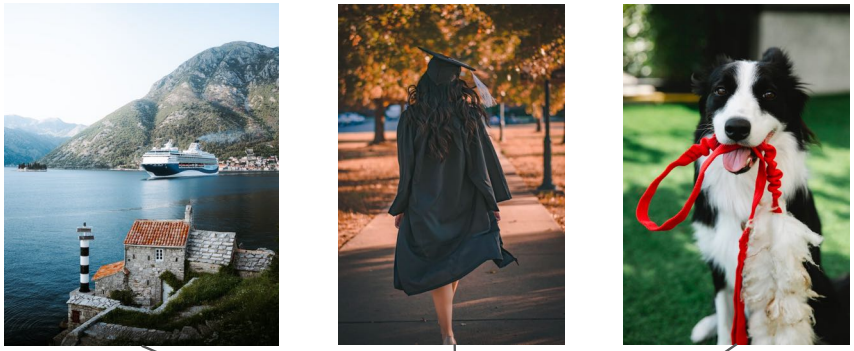
Embed mini-batch of images and text independently, then:

SoftMax: Make emb of matching entries ("positives" +)
more similar than unmatching entries ("negatives" -)

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

Sigmoid: Push emb of matching entries ("positives" +) to 1.0
and push emb of unmatching entries ("negatives" -) to 0.0

$$-\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\underbrace{\log\frac{1}{1+e^{z_{ij}(-t\mathbf{x}_i\cdot\mathbf{y}_j+b)}}}_{\mathcal{L}_{ij}}$$



All images CC0

(CLIP) Learning Transferable Visual Models From Natural Language Supervision - Radford et.al.

(SigLIP) Sigmoid Loss for Language Image Pre-Training - X. Zhai, B. Mustafa, A. Kolesnikov, **L. Beyer**

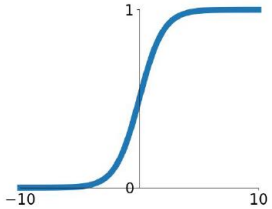
slide credit: Lucas Beyer

SigLIP

Embed mini-batch of images and text independently, then:

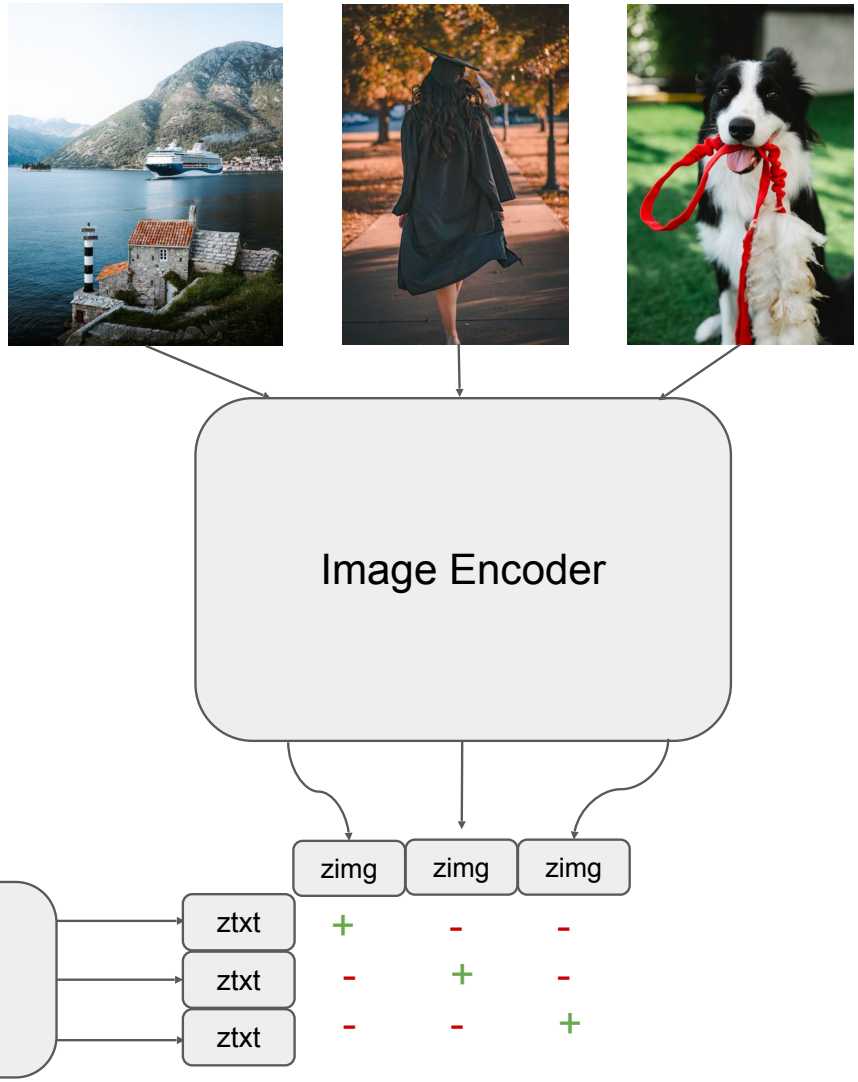
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



Sigmoid: Push emb of matching entries ("positives" +) to 1.0
and push emb of unmatching entries ("negatives" -) to 0.0

$$-\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\underbrace{\log\frac{1}{1+e^{z_{ij}(-t\mathbf{x}_i\cdot\mathbf{y}_j+b)}}}_{\mathcal{L}_{ij}}$$



All images CC0

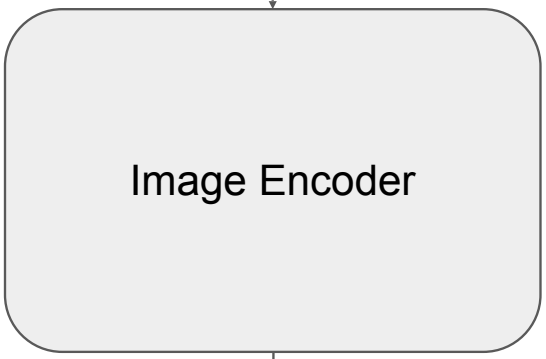
(CLIP) Learning Transferable Visual Models From Natural Language Supervision - Radford et.al.

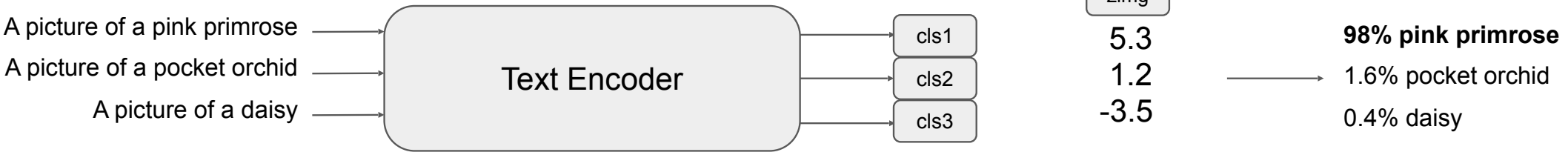
(SigLIP) Sigmoid Loss for Language Image Pre-Training - X. Zhai, B. Mustafa, A. Kolesnikov, **L. Beyer**

slide credit: Lucas Beyer

Using contrastive image-text models for "0shot" transfer

1) "Spell out" your classification task, only once per task.
2) Encode each class's text into the class embedding vector.
3) To classify an image, embed it, compare to class embeddings.





(CLIP) Learning Transferable Visual Models From Natural Language Supervision - Radford et.al.
(SigLIP) Sigmoid Loss for Language Image Pre-Training - X. Zhai, B. Mustafa, A. Kolesnikov, **L. Beyer**

slide credit: Lucas Beyer

SigLIP, demo with new examples:

10%	43%	0%	0%	0%	0%	0%	0%	0%	0%	0%	an apple
43%	69%	0%	0%	0%	0%	0%	0%	0%	0%	0%	a picture of an apple
60%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	an ipod
4%	54%	0%	0%	0%	0%	0%	0%	0%	0%	0%	granny smith
100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	an apple with a note saying "ipod"
0%	0%	26%	0%	0%	0%	0%	0%	0%	0%	0%	a cold drink on a hot day
0%	0%	0%	79%	0%	0%	0%	0%	0%	0%	0%	a hot drink on a cold day
0%	0%	59%	0%	0%	0%	0%	0%	0%	0%	0%	a photo of a cold drink on a hot day
0%	0%	0%	98%	0%	0%	0%	0%	0%	0%	1%	a photo of a hot drink on a cold day
0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	a photo of two guys in need of caffeine
0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	a photo of two guys in need of water
0%	0%	0%	0%	0%	98%	0%	0%	0%	0%	0%	a photo of the SigLIP authors
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	a photo of a rock band
0%	0%	0%	0%	0%	98%	100%	0%	0%	0%	0%	a photo of researchers at Google Brain
0%	0%	0%	0%	0%	20%	35%	0%	0%	0%	0%	a photo of researchers at OpenAI
0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	a robot on a sign
0%	0%	0%	0%	0%	0%	0%	46%	0%	0%	0%	a photo of a robot on a sign
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	an empty street
0%	0%	0%	0%	0%	0%	0%	16%	0%	0%	0%	autumn in Toronto
0%	0%	0%	0%	0%	0%	0%	57%	0%	0%	0%	a photo of autumn in Toronto
0%	0%	0%	0%	0%	0%	0%	34%	0%	0%	0%	a photo of Toronto in autumn
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	a photo of Toronto in summer
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	autumn in Singapore
0%	0%	0%	0%	0%	0%	0%	0%	36%	32%	0%	cow
0%	0%	0%	0%	0%	0%	0%	0%	0%	99%	0%	a cow in a tuxedo
0%	0%	0%	0%	0%	0%	0%	0%	99%	92%	0%	a cow on the beach
0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	a cow in the prairie
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	the real mountain view
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	Zürich
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	San Francisco
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	a picture of a laptop with the lockscreen o window reveals lake Zürich and the Alps ir

SigLIP: Sigmoid Loss for Language Image Pre-Training - X. Zhai, B. Mustafa, A. Kolesnikov, **L. Beyer**

slide credit: Lucas Beyer

First good multi-lingual and multi-cultural CLIP:







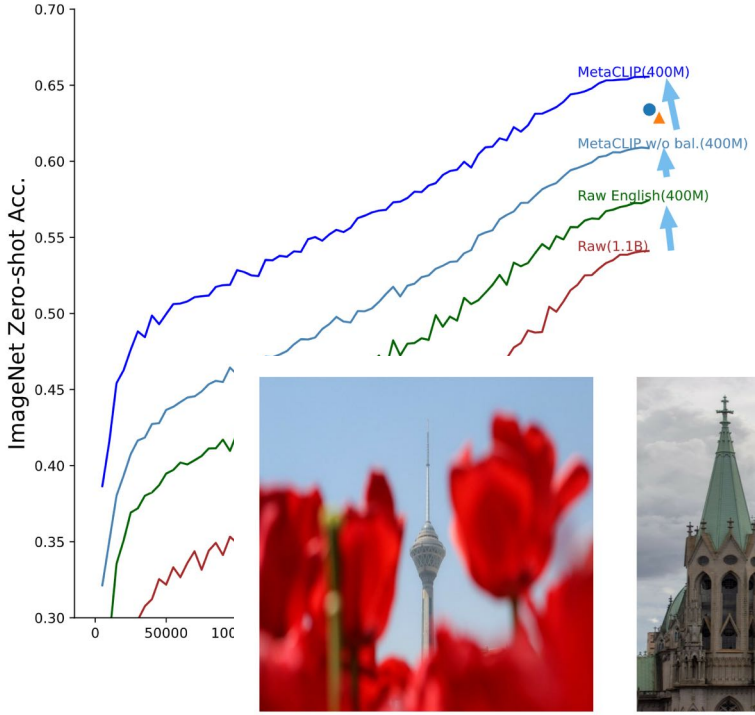


91%	0%	0%	0%	0%	0%	0%	0%	0%	蚂蚁上树
84%	0%	0%	0%	0%	0%	0%	0%	0%	肉末粉丝
0%	93%	0%	0%	0%	0%	0%	0%	0%	ants climbing a tree
2%	0%	0%	0%	0%	0%	0%	0%	0%	minced pork rice noodle
0%	0%	43%	0%	0%	0%	0%	0%	0%	红烧狮子头
0%	0%	0%	11%	0%	0%	0%	0%	0%	red burned lion head
0%	0%	0%	36%	0%	0%	0%	0%	0%	lion head
0%	0%	47%	0%	0%	0%	0%	0%	0%	meat ball with soy sauce
0%	0%	0%	0%	69%	0%	0%	0%	0%	松鼠鳜鱼
0%	0%	0%	0%	0%	0%	20%	0%	0%	squirrel
0%	0%	0%	0%	0%	69%	100%	0%	0%	squirrel and fish
0%	0%	0%	0%	0%	22%	6%	0%	0%	squirrel mandarinfish
0%	0%	0%	0%	0%	41%	12%	0%	0%	squirrel mandarin fish
0%	0%	6%	0%	34%	0%	0%	0%	0%	sweet and sour mandarin fish
0%	0%	0%	0%	0%	0%	0%	35%	20%	cow
0%	0%	0%	0%	0%	0%	0%	0%	100%	a cow in a tuxedo
0%	0%	0%	0%	0%	0%	0%	100%	78%	a cow on the beach
0%	0%	0%	0%	0%	0%	0%	2%	0%	a cow in the prairie
0%	0%	0%	0%	0%	0%	0%	100%	66%	une vache sur la plage
0%	0%	0%	0%	0%	0%	0%	100%	55%	eine Kuh am Strand
0%	0%	0%	0%	0%	0%	0%	78%	9%	วัวอยู่ที่ชายหาด
0%	0%	0%	0%	0%	0%	0%	37%	7%	一只躺在沙滩上的牛
0%	0%	0%	0%	0%	0%	0%	88%	13%	一只沙滩上的牛
0%	0%	0%	0%	0%	0%	0%	99%	90%	корова на пляже
0%	0%	0%	0%	0%	0%	0%	99%	83%	بقرة على الشاطئ

SigLIP: Sigmoid Loss for Language Image Pre-Training - X. Zhai, B. Mustafa, A. Kolesnikov, **L. Beyer**

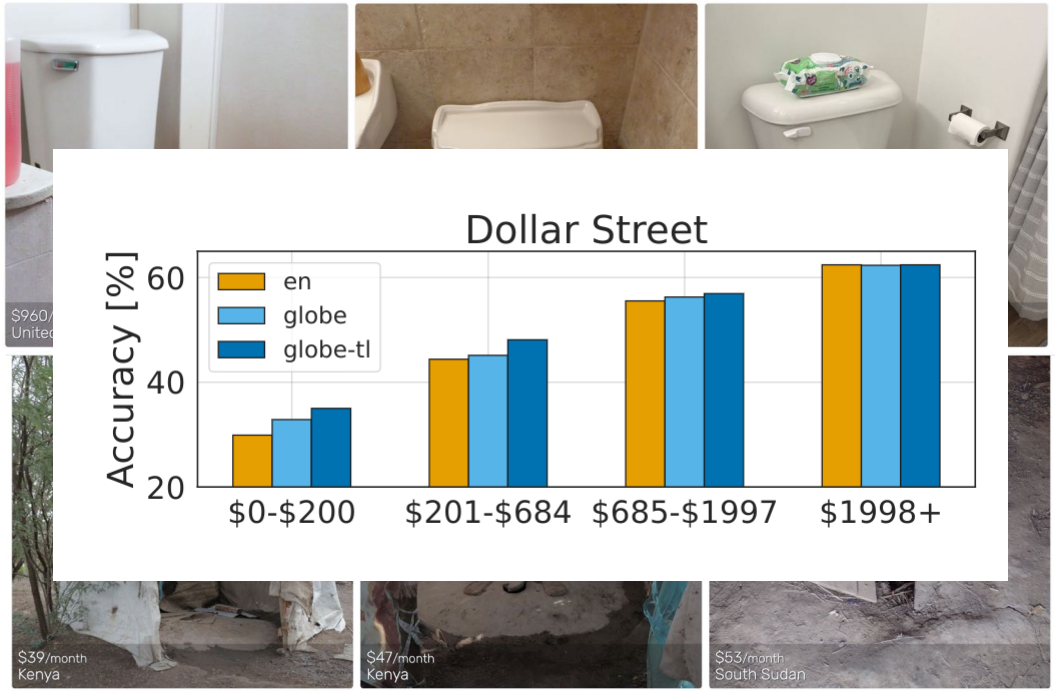
Data filters?

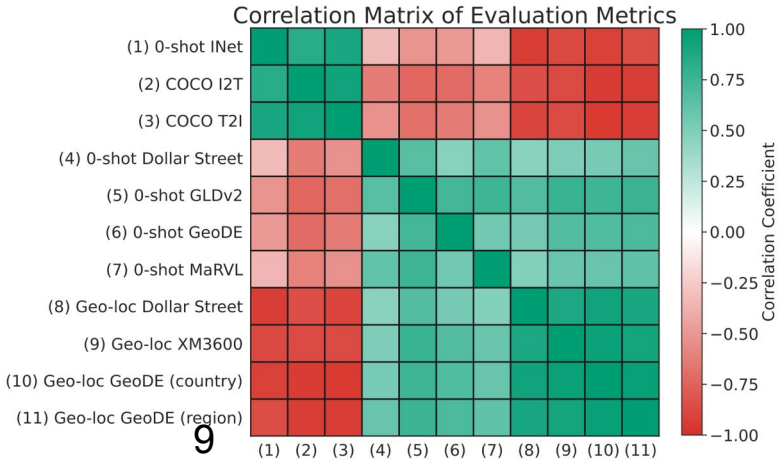
No Filters, please!



(a) **en**: CN Tower (Canada) **globe**: Milad Tower (Iran)

(b) **en**: Notre-Dame Basilica of Montreal (Canada) **globe**: Catedral da Sé de São Paulo (Brazil)





Demystifying CLIP Data — https://arxiv.org/pdf/2309.16671

No Filter: Cultural and Socioeconomic Diversity in Contrastive VLMs - A. Pouget, **L. Beyer**, E. Bugliarello, X. Wang, A. Steiner, X. Zhai

slide credit: Lucas Beyer

Overview of Today's Lecture

- SigLIP — Sigmoid Loss for Language Image Pre-Training
  - [iccv'23] — https://arxiv.org/abs/2303.15343
  - [arxiv'24] — https://arxiv.org/abs/2405.13777
- Multimodal Learning
  - ImageBind: [cvpr'23] - https://arxiv.org/abs/2305.05665
- Eyes Wide Shut
  - [cvpr'24] — https://arxiv.org/abs/2401.06209
- GiT: Towards Generalist Vision Transformer
  - [eccv'24] — https://arxiv.org/abs/2403.09394

Transformers are Modality Guzzlers

Can guzzle all modalities

- Structured input - text, images, speech, audio
- Unordered input - points, graphs

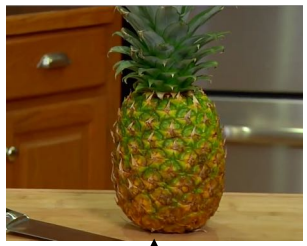
Seem to scale well with data/model size

Great candidate for **multi-modal** learning!

slide credit: Ishan Misra

Recipe for multimodal learning

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



A pineapple sitting on the counter

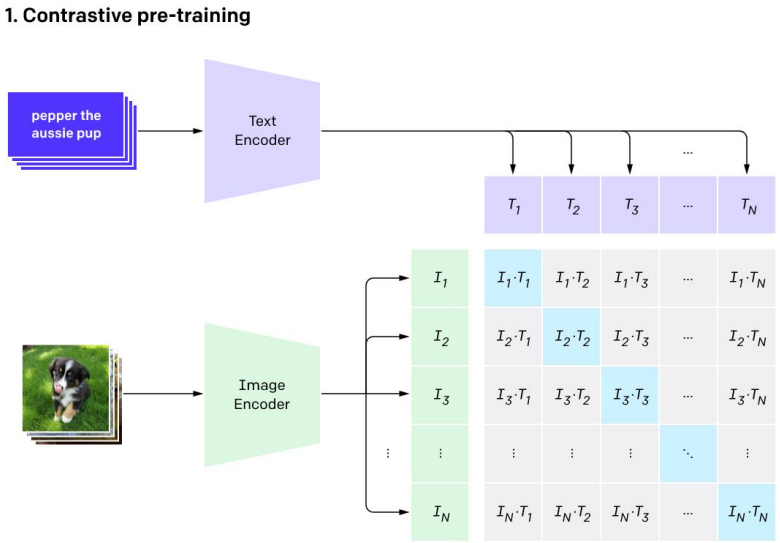


Image source: CLIP - Radford et al., 2021

slide credit: Ishan Misra

Aligned image-text features

- Aligned representations are *really* useful

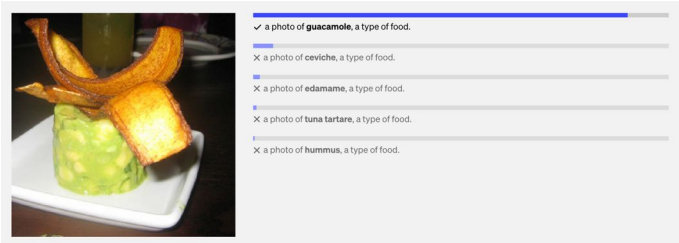
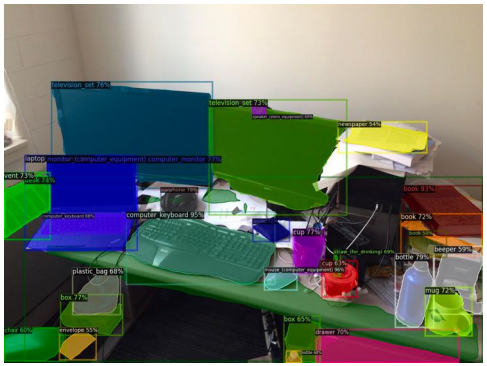
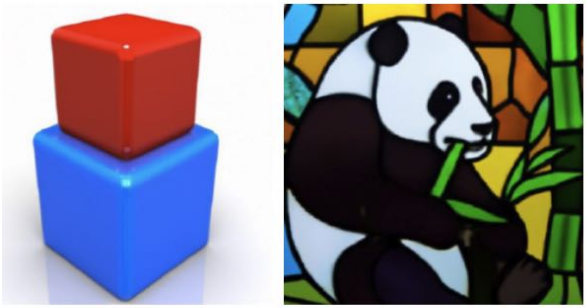


Image-text retrieval
Open-vocabulary classification[1]



Open-vocabulary detection and segmentation[2]



Text to image generation[3]

slide credit: Ishan Misra

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

Aligned image-text features

- Aligned representations are *really* useful

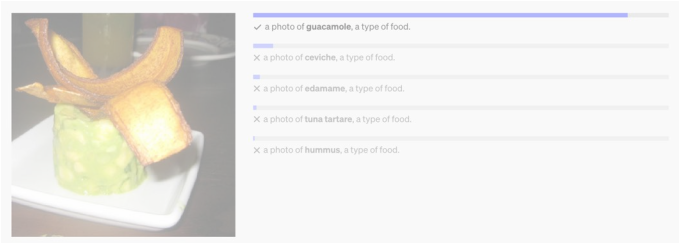
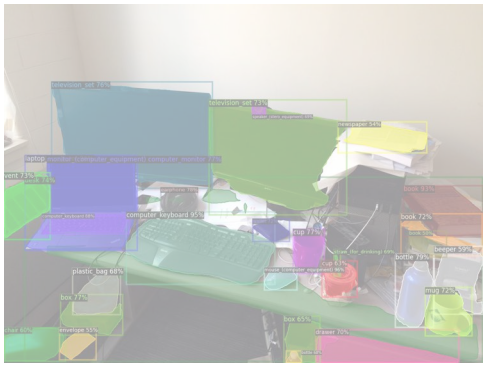
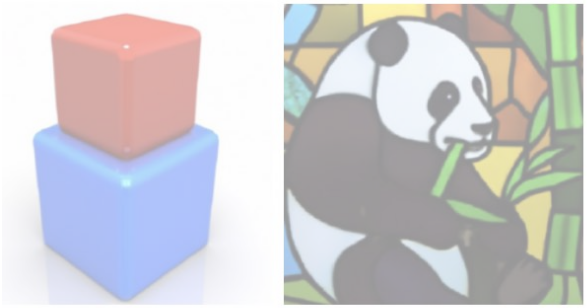


Image-text retrieval
Open-vocabulary classification[1]



Open-vocabulary detection and segmentation[2]



Text to image generation[3]

So have we “solved” multi-modal learning?

slide credit: Ishan Misra

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

Problem 1: **Multi**-modal != **Bi**-modal

There are other modalities …

Image source: Rawpixel, The Rijksmuseum

slide credit: Ishan Misra

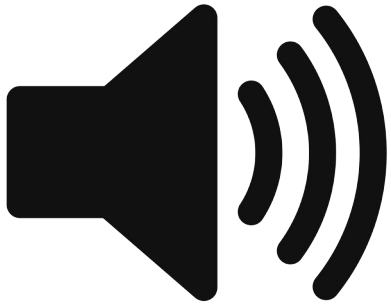
slide credit: Ishan Misra

Problem 2: **Aligned** data is hard to get

Depth

Thermal

Motion (IMU)



Audio

Image source: Rawpixel, The Rijksmuseum

slide credit: Ishan Misra

Images are a universal language

Depth

Thermal

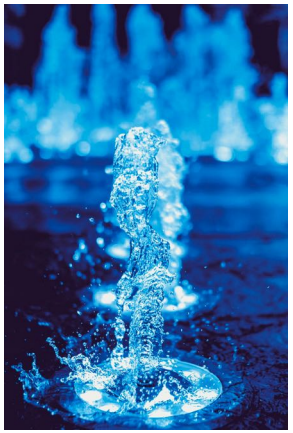
Motion (IMU)

Audio

RGB

RGB

RGB



RGB

slide credit: Ishan Misra

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

Images are a universal language

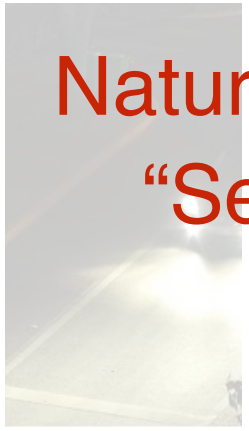
Depth

Thermal

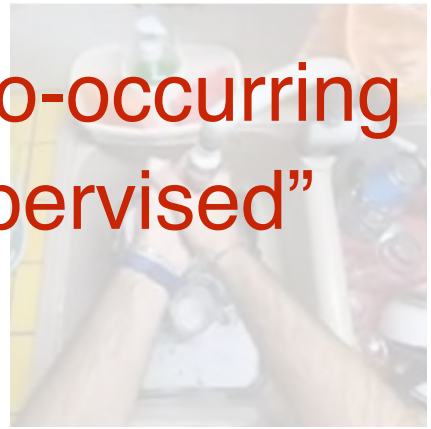
Motion (IMU)

Audio

RGB



RGB



RGB

RGB

Naturally co-occurring
"Self-supervised"

slide credit: Ishan Misra

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

ImageBind: One Embedding to Rule them All

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

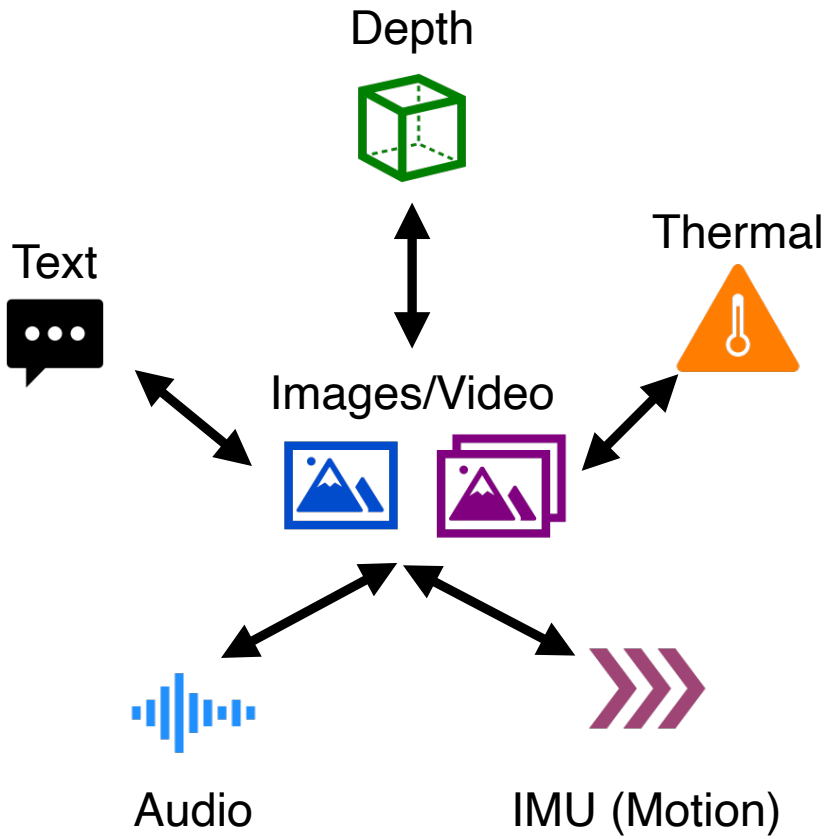
https://github.com/facebookresearch/ImageBind

CVPR 2023

slide credit: Ishan Misra

Key Idea

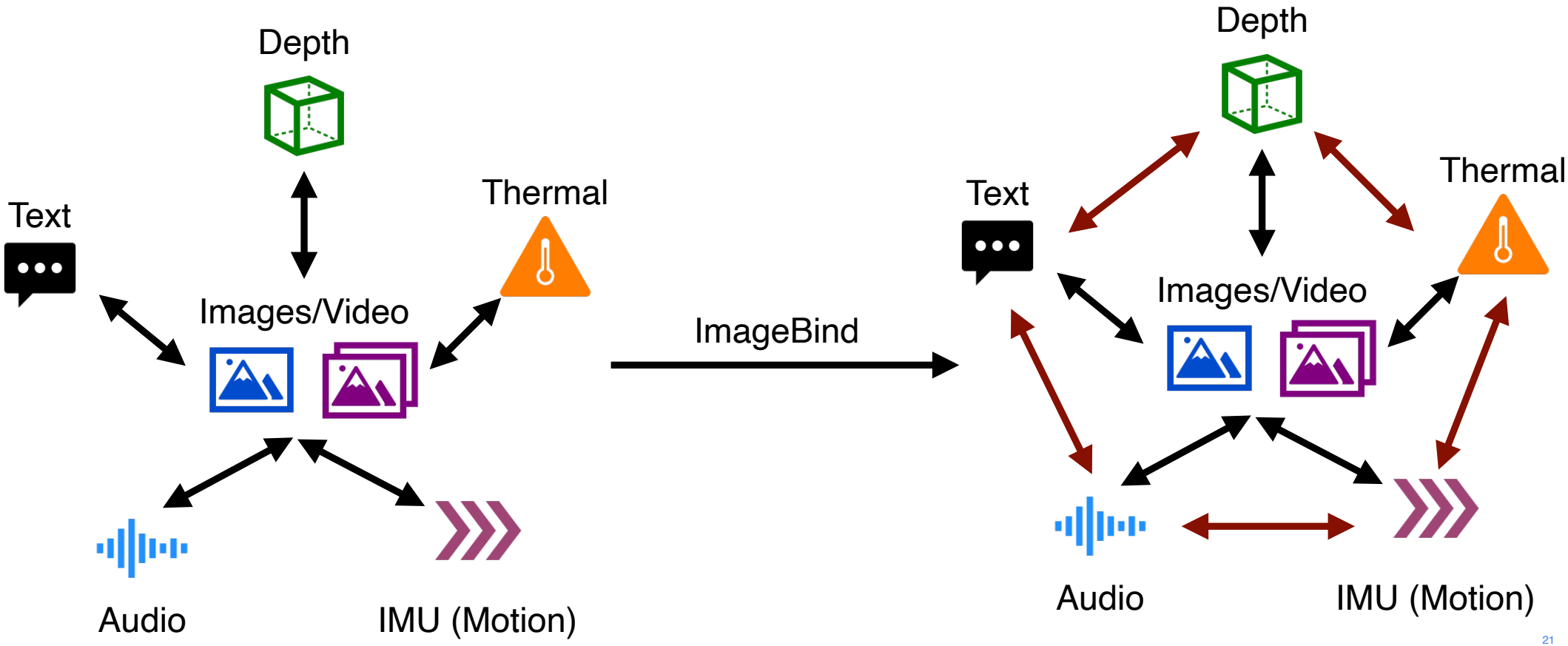
- Images naturally co-occur with different modalities
- Align every modality's representation with images
- Heavily leverage self-supervised learning



slide credit: Ishan Misra

Emergent behavior (Transitive alignment!)

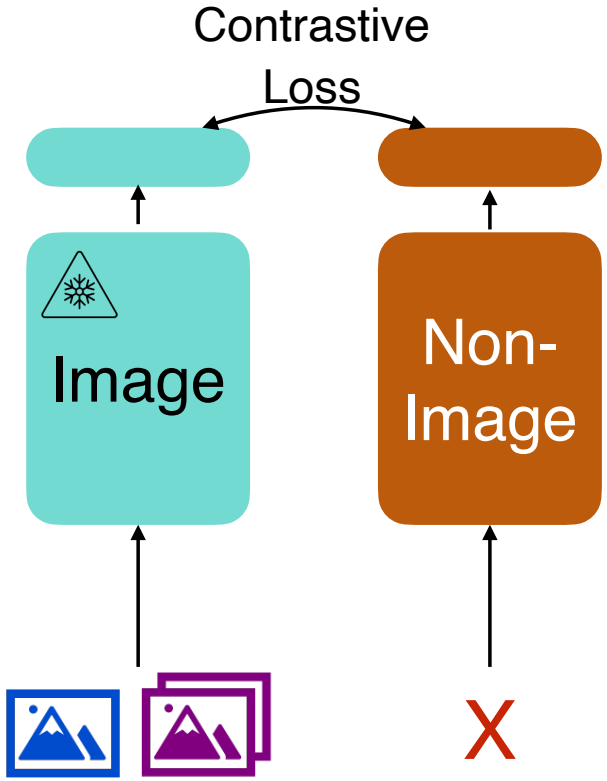
- After training **all** modalities are aligned



slide credit: Ishan Misra

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



slide credit: Ishan Misra

Measuring emergent alignment to text

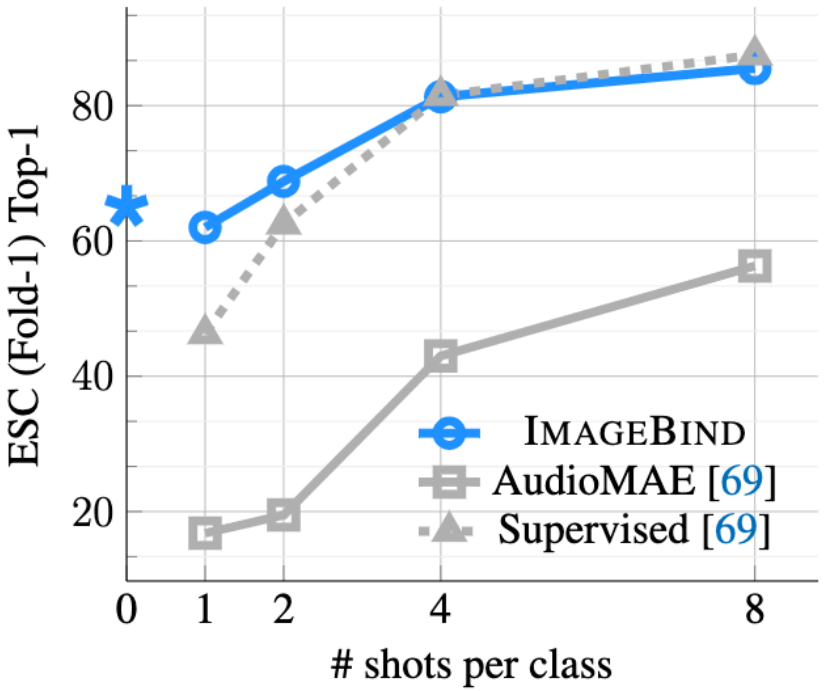
- Train on (Image, X) (Image, Text)
- Test on (X, Text) —> **“Emergent”** zero-shot classification

	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	-	-

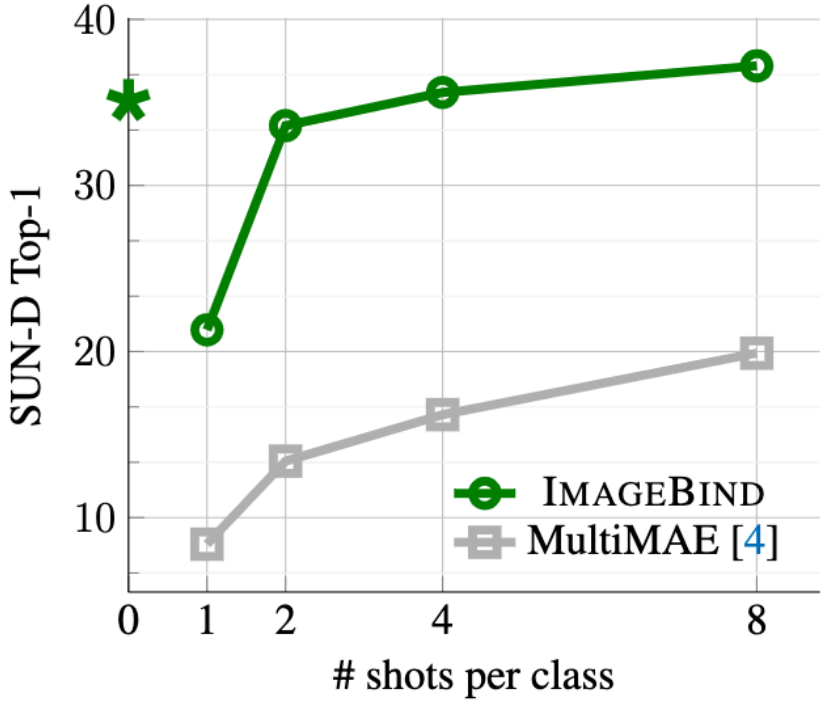
slide credit: Ishan Misra

Measuring performance on few-shot classification

Few-shot audio

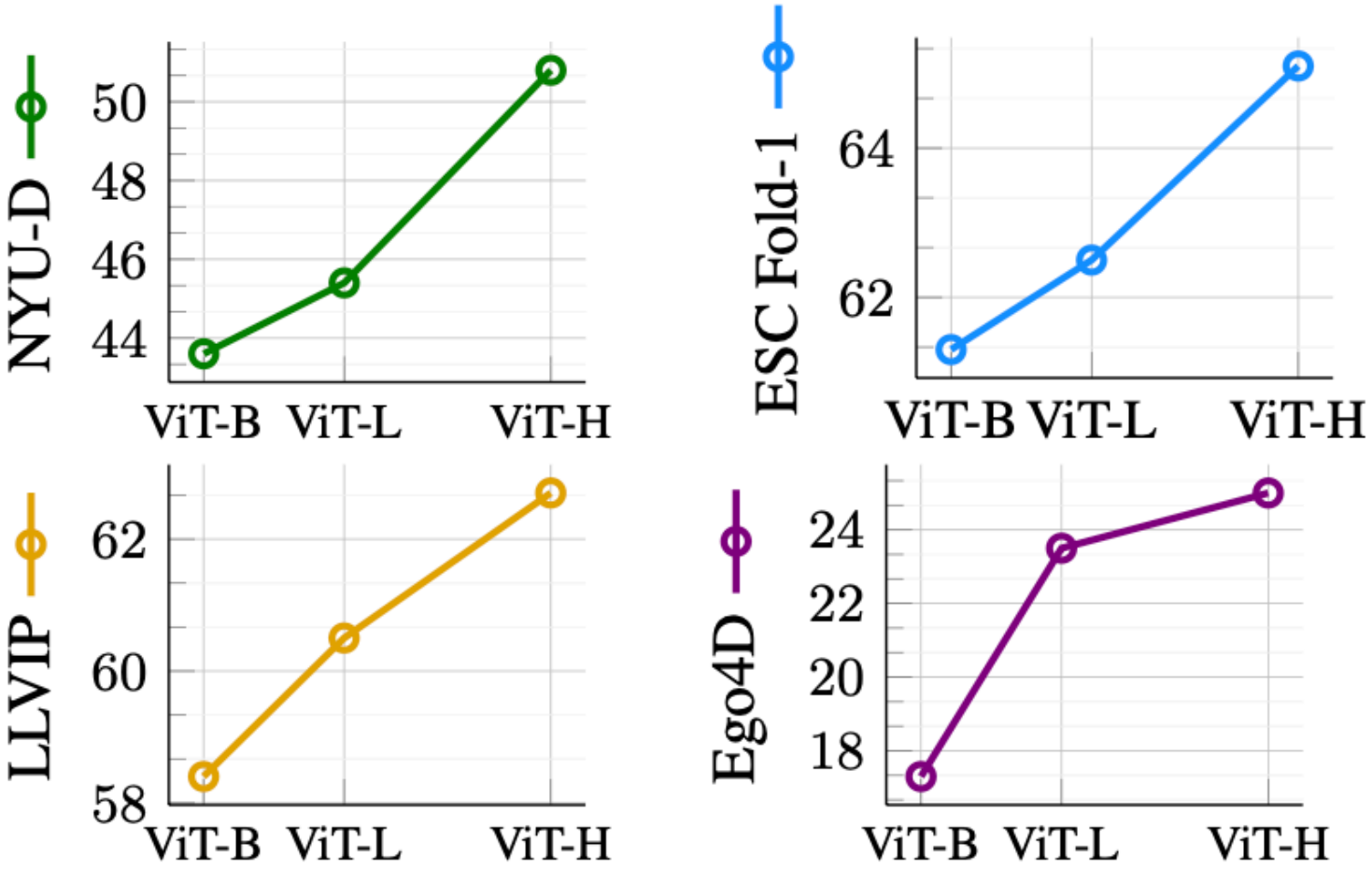


Few-shot depth



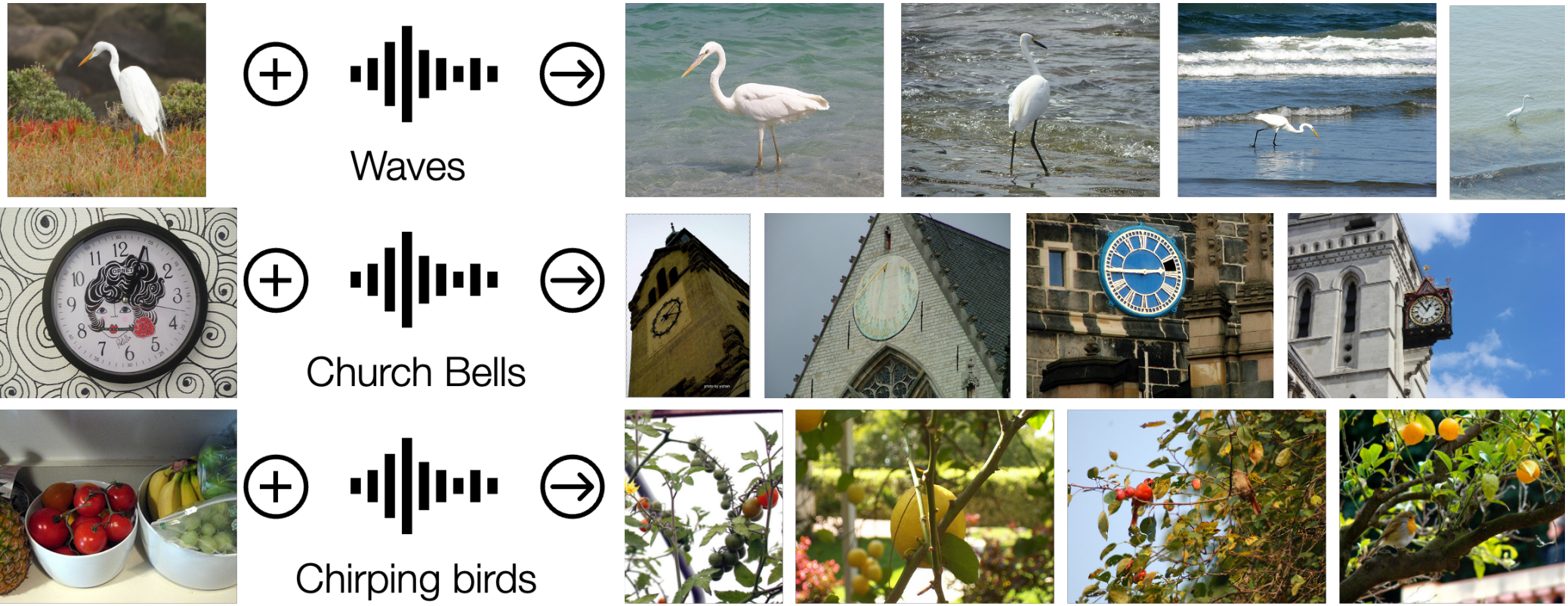
slide credit: Ishan Misra

Binding gets stronger with image-model size



slide credit: Ishan Misra

Aligned embeddings can be “added”



slide credit: Ishan Misra

Overview of Today's Lecture

- SigLIP — Sigmoid Loss for Language Image Pre-Training
  - [iccv'23] — https://arxiv.org/abs/2303.15343
  - [arxiv'24] — https://arxiv.org/abs/2405.13777
- Multimodal Learning
  - ImageBind: [cvpr'23] - https://arxiv.org/abs/2305.05665
- Eyes Wide Shut
  - [cvpr'24] — https://arxiv.org/abs/2401.06209
- GiT: Towards Generalist Vision Transformer
  - [eccv'24] — https://arxiv.org/abs/2403.09394

Eyes Wide Shut?
Exploring the Visual Shortcomings of Multimodal LLMs

Shengbang Tong[1], Zhuang Liu[2], Yuexiang Zhai[3], Yi Ma[3], Yann LeCun[1], Saining Xie[1]

[1]NYU,[2]FAIR, Meta AI, [3]UC Berkeley

slide credit: Shenbang Tong

Preview: Multimodal LLMs are booming

GPT-4V

[OpenAI, et al, 2023]

Gemini

[Google, et al, 2023]

[Anthropic, et al, 2023]

[Liu, Haotian, et al, 2023]

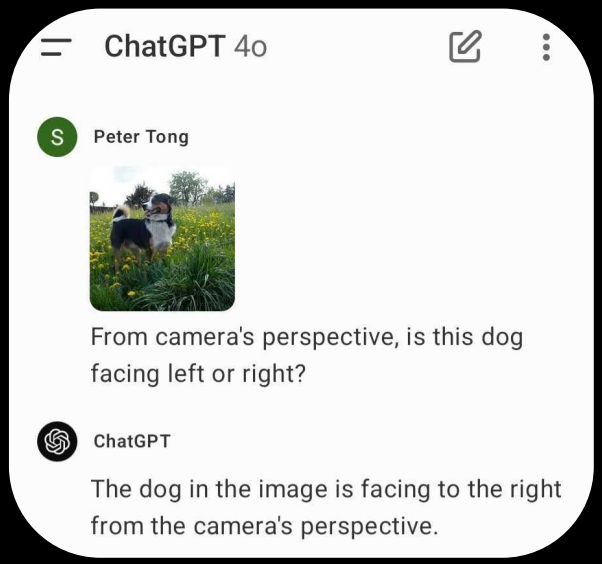
InstructBLIP

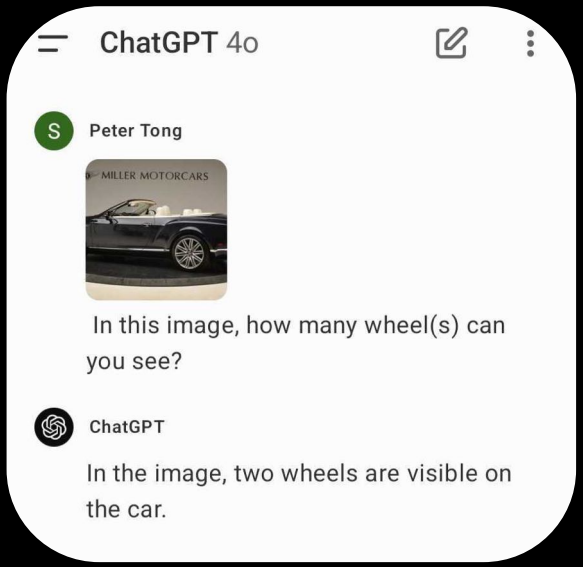
Video Demonstration

[Wenliang, Dai, et al. 2023]

[Hugo Laurençon, et al, 2023]

But MLLMs also make unexpected mistakes





Agenda

- How do we find these mistakes?

- Why do models make these mistakes?

- How do we work towards fixing these mistakes?

slide credit: Shenbang Tong

Recap on the MLLM Architecture

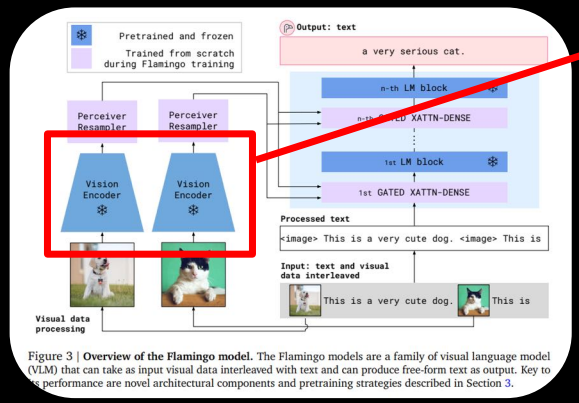
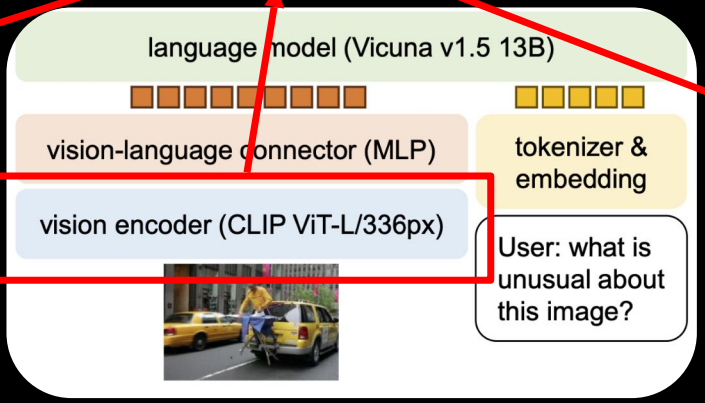


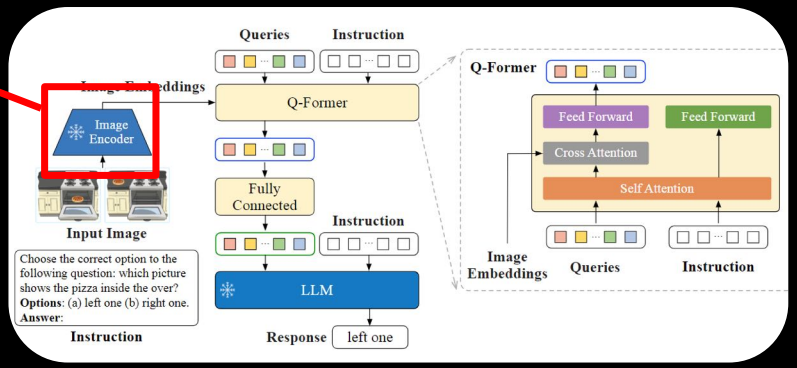
Figure 3 | **Overview of the Flamingo model.** The Flamingo models are a family of visual language model (VLM) that can take as input visual data interleaved with text and can produce free-form text as output. Key to its performance are novel architectural components and pretraining strategies described in Section 3.

Flamingo
[Alayrac, Jean-Baptiste, et al. 2022]

CLIP!



LLaVA
[Liu, Haotian, et al. 2023]

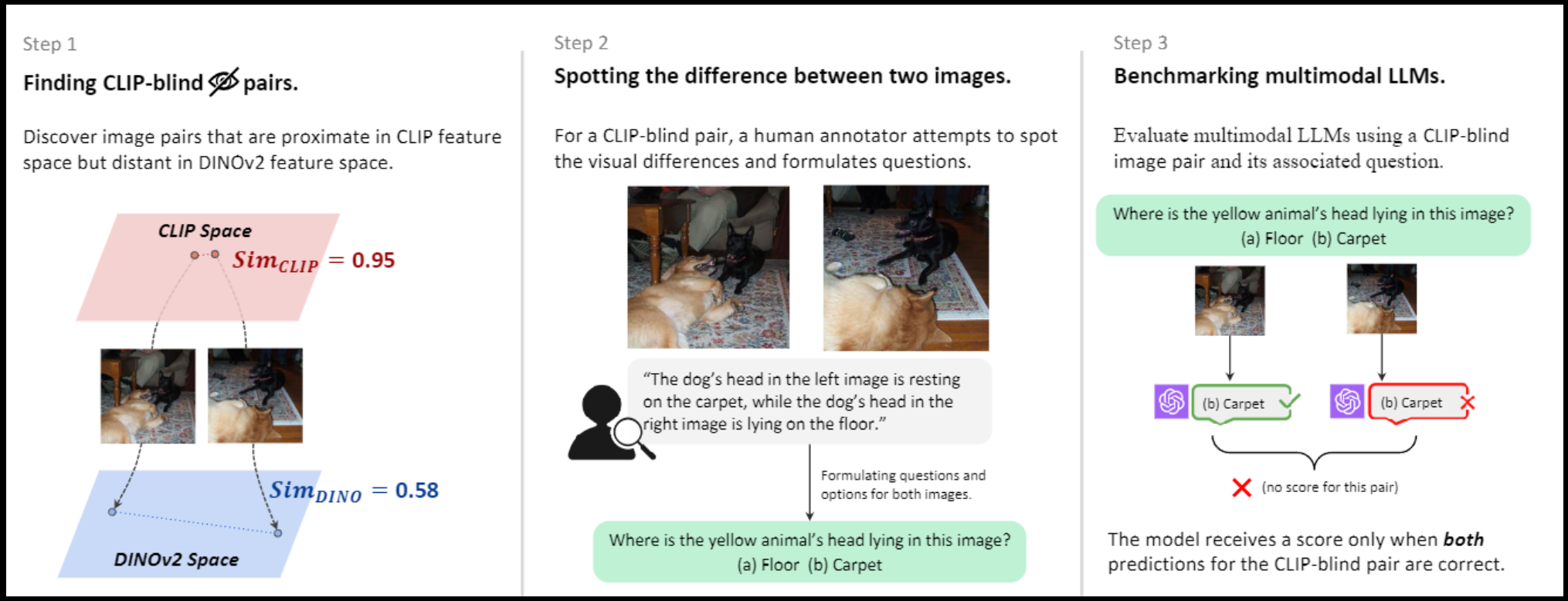


InstructBLIP
[Wenliang, Dai, et al. 2023]

They all use a pretrained Vision Encoder, CLIP!

Finding CLIP-blind pairs

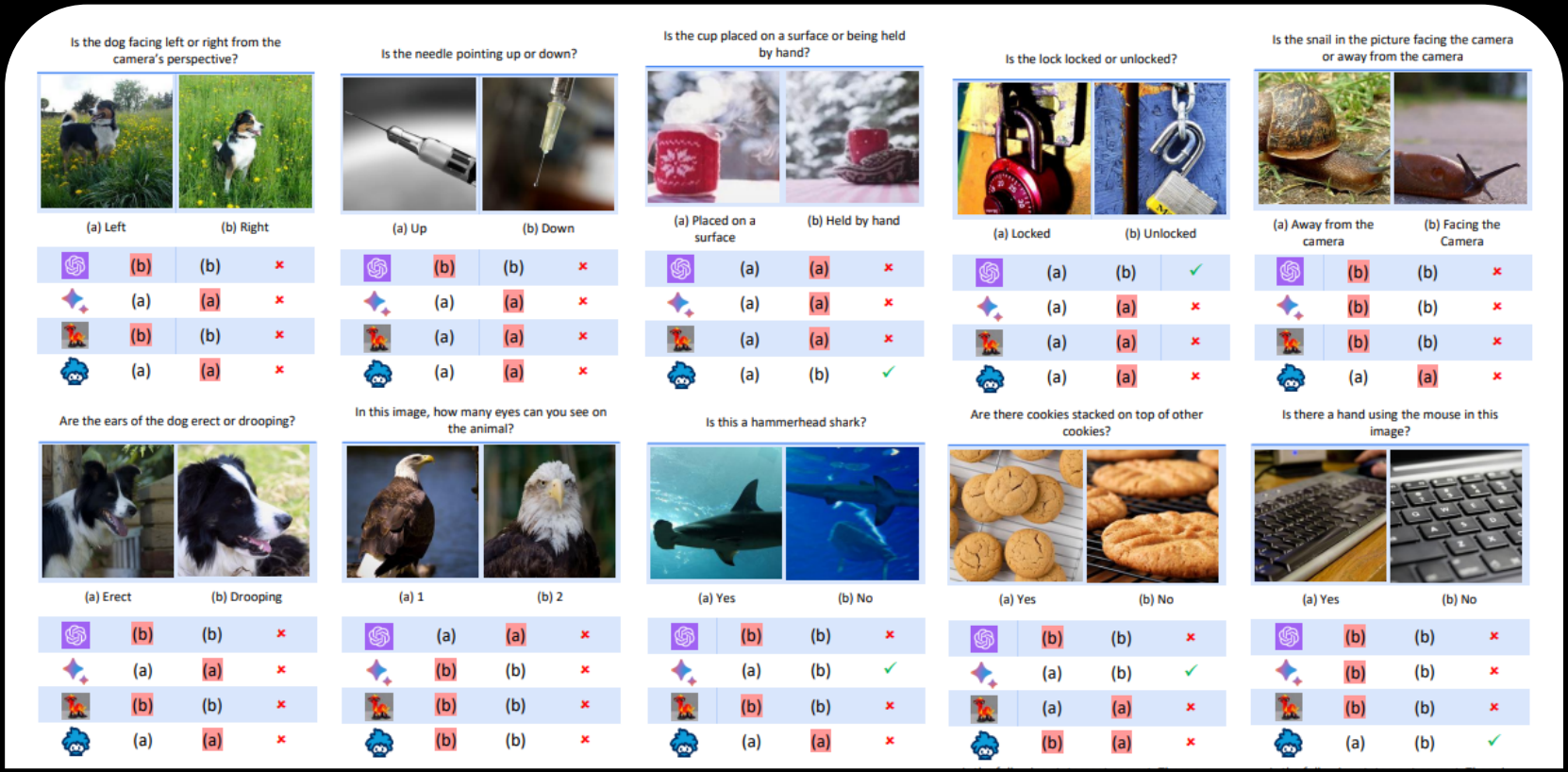
CLIP-blind pairs: If two images are encoded similarly by the CLIP model yet very different in visual appearance, then at least one of them has been inaccurately encoded.



slide credit: Shenbang Tong

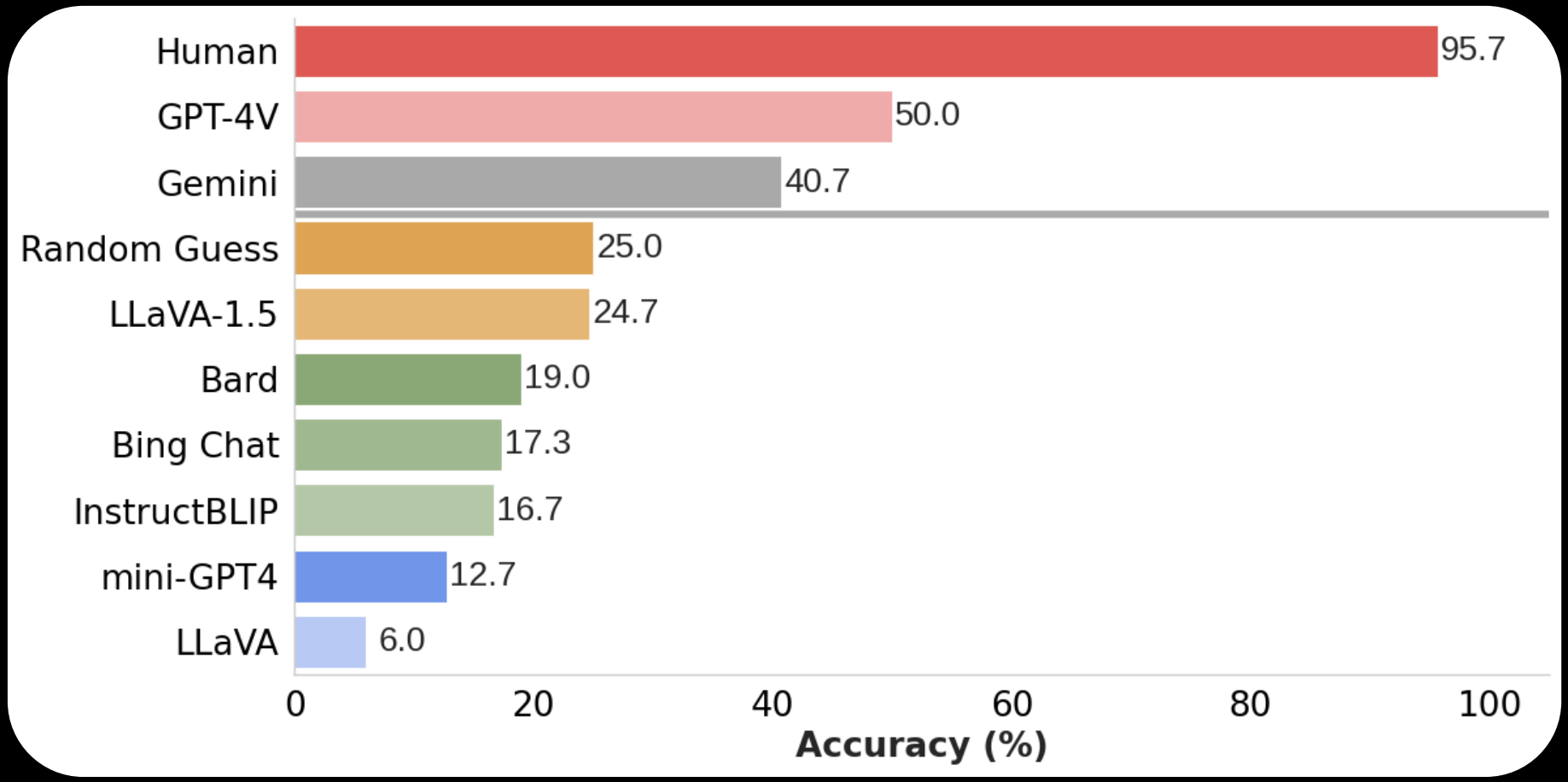
MMVP (MultiModal Visual Patterns) Benchmark

MMVP Benchmark: 150 CLIP-blind pairs & handcrafted questions



slide credit: Shenbang Tong

MMVP Benchmark Results



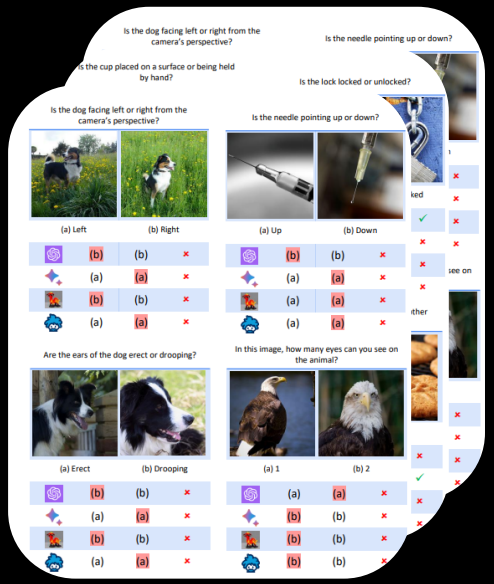
slide credit: Shenbang Tong

Why do models make these mistakes?

slide credit: Shenbang Tong

Finding Patterns in CLIP-blind Pairs

Questions in MMVP:



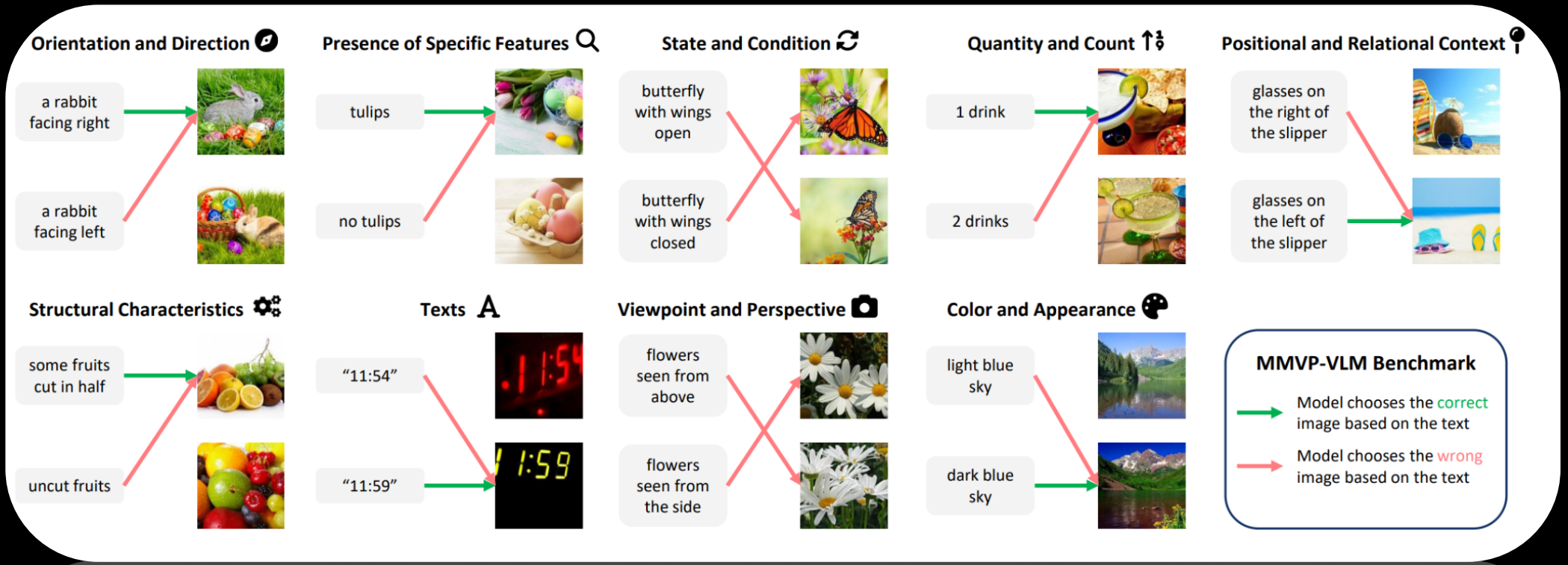
: Summarize Patterns

Visual Patterns:

- Orientation and Direction
- Presence of Specific Features
- State and Condition
- Quantity and Count
- Positional and Relational Context
- Color and Appearance
- Structural and Physical Characteristics
- Texts
- Viewpoint and Perspective

slide credit: Shenbang Tong

MMVP-VLM Benchmark



CLIP models struggle

	Image Size	Params (M)	IN-1k ZeroShot										MMVP Average
OpenAI ViT-L-14 [35]	224^2	427.6	75.5	13.3	13.3	20.0	20.0	13.3	53.3	20.0	6.7	13.3	19.3
OpenAI ViT-L-14 [35]	336^2	427.9	76.6	0.0	20.0	40.0	20.0	6.7	20.0	33.3	6.7	33.3	20.0
SigLIP ViT-SO-14 [53]	224^2	877.4	82.0	26.7	20.0	53.3	40.0	20.0	66.7	40.0	20.0	53.3	37.8
SigLIP ViT-SO-14 [53]	384^2	878.0	83.1	20.0	26.7	60.0	33.3	13.3	66.7	33.3	26.7	53.3	37.0
DFN ViT-H-14 [9]	224^2	986.1	83.4	20.0	26.7	73.3	26.7	26.7	66.7	46.7	13.3	53.3	39.3
DFN ViT-H-14 [9]	378^2	986.7	84.4	13.3	20.0	53.3	33.3	26.7	66.7	40.0	20.0	40.0	34.8
MetaCLIP ViT-L-14 [49]	224^2	427.6	79.2	13.3	6.7	66.7	6.7	33.3	46.7	20.0	6.7	13.3	23.7
MetaCLIP ViT-H-14 [49]	224^2	986.1	80.6	6.7	13.3	60.0	13.3	6.7	53.3	26.7	13.3	33.3	25.2
EVA01 ViT-g-14 [43]	224^2	1136.4	78.5	6.7	26.7	40.0	6.7	13.3	66.7	13.3	13.3	20.0	23.0
EVA02 ViT-bigE-14+ [43]	224^2	5044.9	82.0	13.3	20.0	66.7	26.7	26.7	66.7	26.7	20.0	33.3	33.3

slide credit: Shenbang Tong

CLIP models struggle

#1: Scaling up resolution **does not help**

	Image Size	Params (M)	IN-1k ZeroShot										MMVP Average
OpenAI ViT-L-14 [35]	224^2	427.6	75.5	13.3	13.3	20.0	20.0	13.3	53.3	20.0	6.7	13.3	19.3
OpenAI ViT-L-14 [35]	336^2	427.9	76.6	0.0	20.0	40.0	20.0	6.7	20.0	33.3	6.7	33.3	20.0
SigLIP ViT-SO-14 [53]	224^2	877.4	82.0	26.7	20.0	53.3	40.0	20.0	66.7	40.0	20.0	53.3	37.8
SigLIP ViT-SO-14 [53]	384^2	878.0	83.1	20.0	26.7	60.0	33.3	13.3	66.7	33.3	26.7	53.3	37.0
DFN ViT-H-14 [9]	224^2	986.1	83.4	20.0	26.7	73.3	26.7	26.7	66.7	46.7	13.3	53.3	39.3
DFN ViT-H-14 [9]	378^2	986.7	84.4	13.3	20.0	53.3	33.3	26.7	66.7	40.0	20.0	40.0	34.8
MetaCLIP ViT-L-14 [49]	224^2	427.6	79.2	13.3	6.7	66.7	6.7	33.3	46.7	20.0	6.7	13.3	23.7
MetaCLIP ViT-H-14 [49]	224^2	986.1	80.6	6.7	13.3	60.0	13.3	6.7	53.3	26.7	13.3	33.3	25.2
EVA01 ViT-g-14 [43]	224^2	1136.4	78.5	6.7	26.7	40.0	6.7	13.3	66.7	13.3	13.3	20.0	23.0
EVA02 ViT-bigE-14+ [43]	224^2	5044.9	82.0	13.3	20.0	66.7	26.7	26.7	66.7	26.7	20.0	33.3	33.3

slide credit: Shenbang Tong

CLIP models struggle

#1: Scaling up resolution **does not help**

#2: Scaling up network **helps a little**

	Image Size	Params (M)	IN-1k ZeroShot										MMVP Average
OpenAI ViT-L-14 [35]	224^2	427.6	75.5	13.3	13.3	20.0	20.0	13.3	53.3	20.0	6.7	13.3	19.3
OpenAI ViT-L-14 [35]	336^2	427.9	76.6	0.0	20.0	40.0	20.0	6.7	20.0	33.3	6.7	33.3	20.0
SigLIP ViT-SO-14 [53]	224^2	877.4	82.0	26.7	20.0	53.3	40.0	20.0	66.7	40.0	20.0	53.3	37.8
SigLIP ViT-SO-14 [53]	384^2	878.0	83.1	20.0	26.7	60.0	33.3	13.3	66.7	33.3	26.7	53.3	37.0
DFN ViT-H-14 [9]	224^2	986.1	83.4	20.0	26.7	73.3	26.7	26.7	66.7	46.7	13.3	53.3	39.3
DFN ViT-H-14 [9]	378^2	986.7	84.4	13.3	20.0	53.3	33.3	26.7	66.7	40.0	20.0	40.0	34.8
MetaCLIP ViT-L-14 [49]	224^2	427.6	79.2	13.3	6.7	66.7	6.7	33.3	46.7	20.0	6.7	13.3	23.7
MetaCLIP ViT-H-14 [49]	224^2	986.1	80.6	6.7	13.3	60.0	13.3	6.7	53.3	26.7	13.3	33.3	25.2
EVA01 ViT-g-14 [43]	224^2	1136.4	78.5	6.7	26.7	40.0	6.7	13.3	66.7	13.3	13.3	20.0	23.0
EVA02 ViT-bigE-14+ [43]	224^2	5044.9	82.0	13.3	20.0	66.7	26.7	26.7	66.7	26.7	20.0	33.3	33.3

CLIP models struggle

#1: Scaling up resolution **does not help**

#2: Scaling up network **helps a little**

#3: Scaling up data **helps a little**

	Image Size	Params (M)	IN-1k ZeroShot										MMVP Average
OpenAI ViT-L-14 [35]	224^2	427.6	75.5	13.3	13.3	20.0	20.0	13.3	53.3	20.0	6.7	13.3	19.3
OpenAI ViT-L-14 [35]	336^2	427.9	76.6	0.0	20.0	40.0	20.0	6.7	20.0	33.3	6.7	33.3	20.0
SigLIP ViT-SO-14 [53]	224^2	877.4	82.0	26.7	20.0	53.3	40.0	20.0	66.7	40.0	20.0	53.3	37.8
SigLIP ViT-SO-14 [53]	384^2	878.0	83.1	20.0	26.7	60.0	33.3	13.3	66.7	33.3	26.7	53.3	37.0
DFN ViT-H-14 [9]	224^2	986.1	83.4	20.0	26.7	73.3	26.7	26.7	66.7	46.7	13.3	53.3	39.3
DFN ViT-H-14 [9]	378^2	986.7	84.4	13.3	20.0	53.3	33.3	26.7	66.7	40.0	20.0	40.0	34.8
MetaCLIP ViT-L-14 [49]	224^2	427.6	79.2	13.3	6.7	66.7	6.7	33.3	46.7	20.0	6.7	13.3	23.7
MetaCLIP ViT-H-14 [49]	224^2	986.1	80.6	6.7	13.3	60.0	13.3	6.7	53.3	26.7	13.3	33.3	25.2
EVA01 ViT-g-14 [43]	224^2	1136.4	78.5	6.7	26.7	40.0	6.7	13.3	66.7	13.3	13.3	20.0	23.0
EVA02 ViT-bigE-14+ [43]	224^2	5044.9	82.0	13.3	20.0	66.7	26.7	26.7	66.7	26.7	20.0	33.3	33.3

CLIP models struggle

#1: Scaling up resolution **does not help**

#2: Scaling up network **helps a little**

#3: Scaling up data **helps a little**

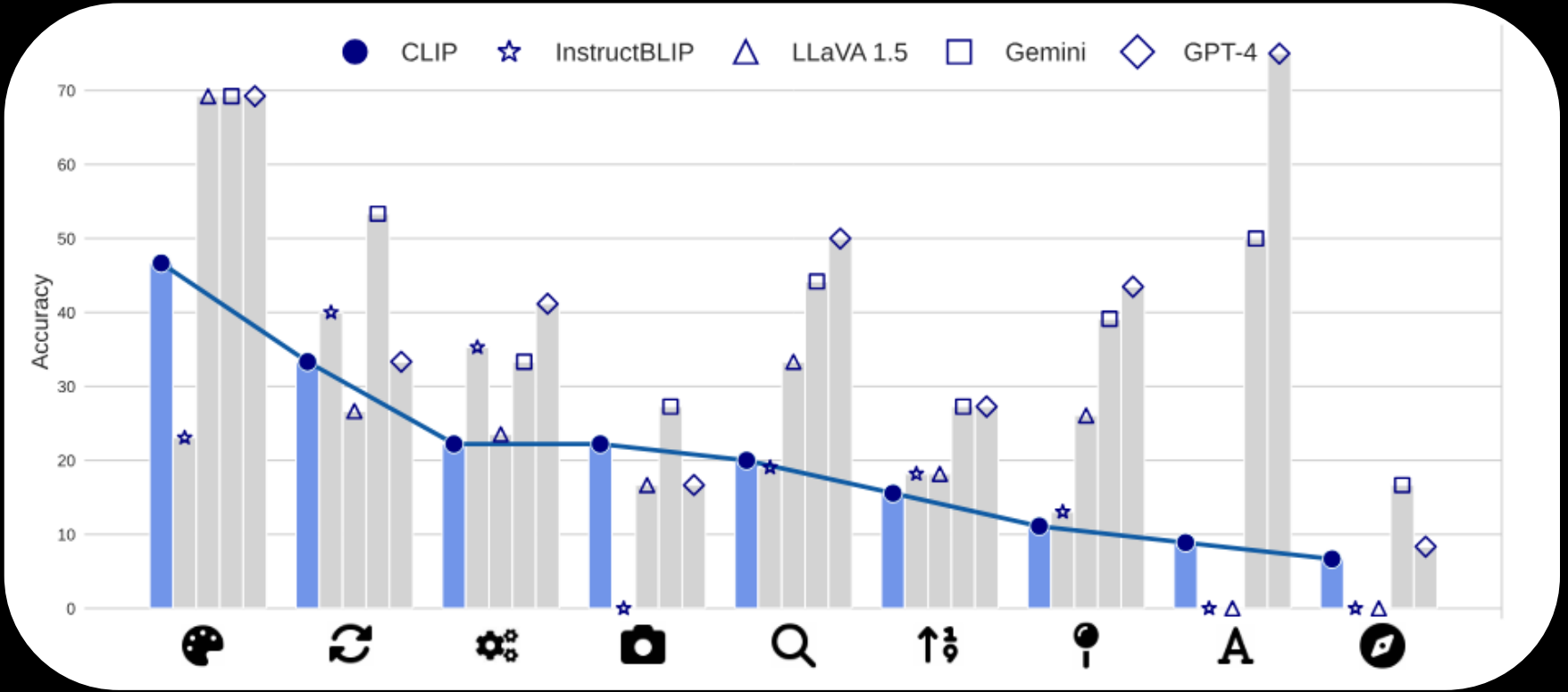
#4: All CLIP-variants **struggle**

	Image Size	Params (M)	IN-1k ZeroShot										MMVP Average
OpenAI ViT-L-14 [35]	224^2	427.6	75.5	13.3	13.3	20.0	20.0	13.3	53.3	20.0	6.7	13.3	19.3
OpenAI ViT-L-14 [35]	336^2	427.9	76.6	0.0	20.0	40.0	20.0	6.7	20.0	33.3	6.7	33.3	20.0
SigLIP ViT-SO-14 [53]	224^2	877.4	82.0	26.7	20.0	53.3	40.0	20.0	66.7	40.0	20.0	53.3	37.8
SigLIP ViT-SO-14 [53]	384^2	878.0	83.1	20.0	26.7	60.0	33.3	13.3	66.7	33.3	26.7	53.3	37.0
DFN ViT-H-14 [9]	224^2	986.1	83.4	20.0	26.7	73.3	26.7	26.7	66.7	46.7	13.3	53.3	39.3
DFN ViT-H-14 [9]	378^2	986.7	84.4	13.3	20.0	53.3	33.3	26.7	66.7	40.0	20.0	40.0	34.8
MetaCLIP ViT-L-14 [49]	224^2	427.6	79.2	13.3	6.7	66.7	6.7	33.3	46.7	20.0	6.7	13.3	23.7
MetaCLIP ViT-H-14 [49]	224^2	986.1	80.6	6.7	13.3	60.0	13.3	6.7	53.3	26.7	13.3	33.3	25.2
EVA01 ViT-g-14 [43]	224^2	1136.4	78.5	6.7	26.7	40.0	6.7	13.3	66.7	13.3	13.3	20.0	23.0
EVA02 ViT-bigE-14+ [43]	224^2	5044.9	82.0	13.3	20.0	66.7	26.7	26.7	66.7	26.7	20.0	33.3	33.3

slide credit: Shenbang Tong

Mistakes in CLIP and MLLM are correlated

The worse CLIP models are, the worse MLLMs are.



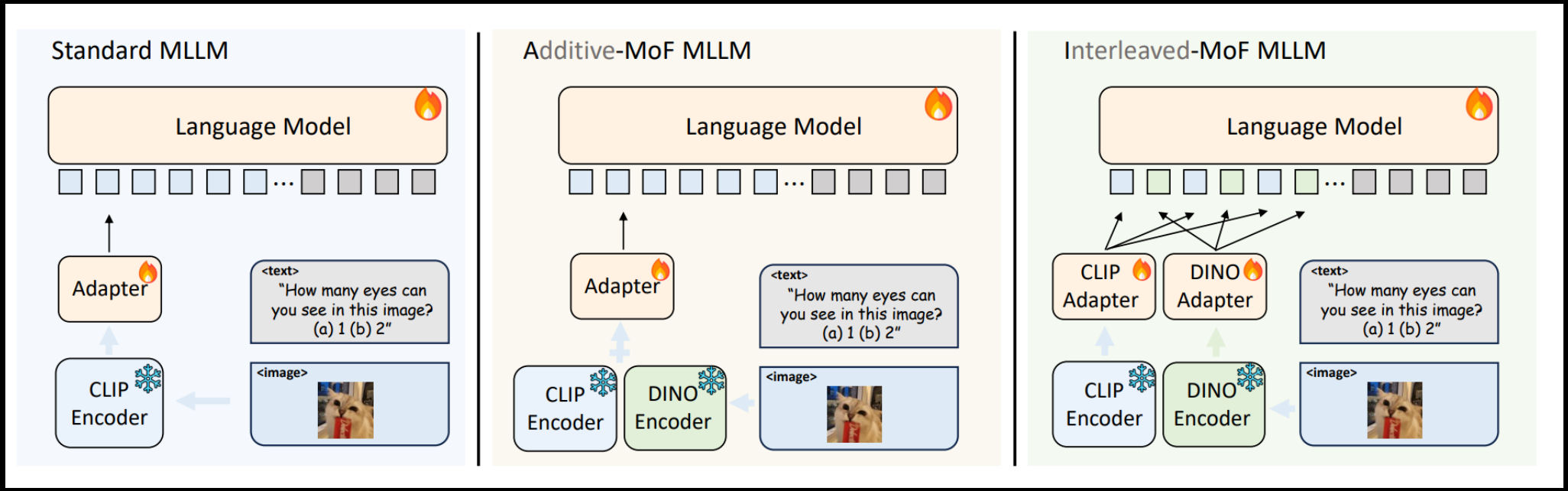
slide credit: Shenbang Tong

How do we work towards fixing these mistakes?

slide credit: Shenbang Tong

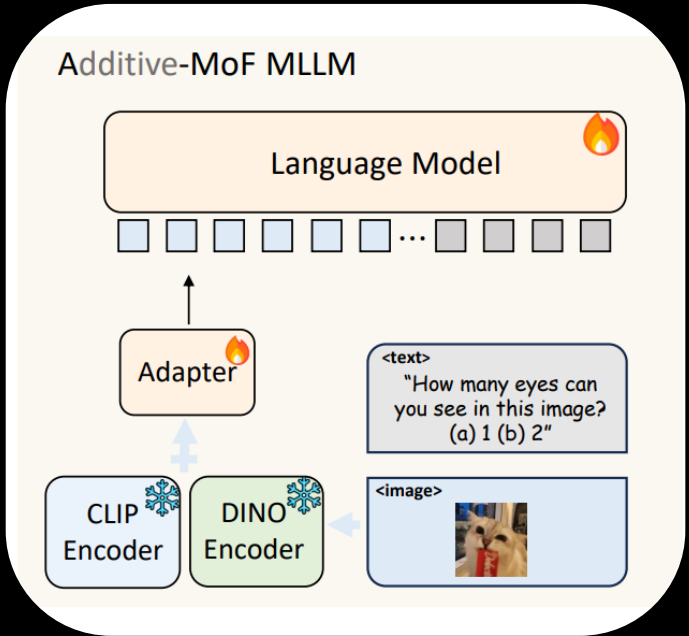
Mix-of-Features (MoF)

Incorporating Vision-Only Features



Additive-MoF MLLM

Vision-SSL model gives you **better visual grounding**, but **worse instruction-following**

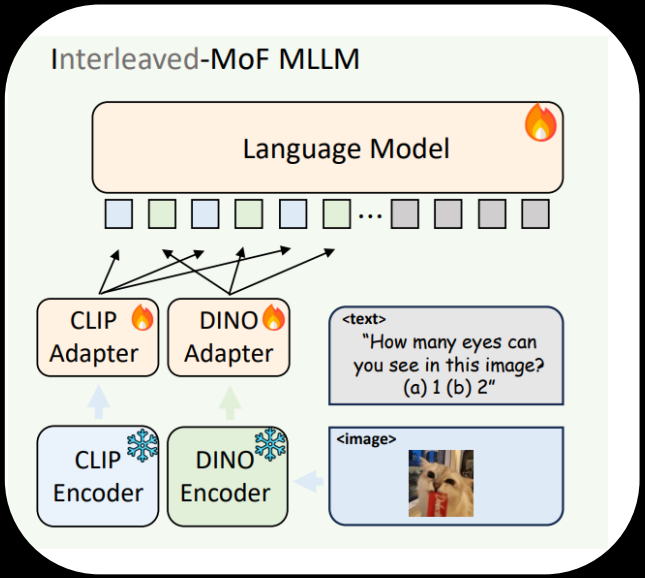


method	SSL ratio	MMVP	LLaVA
LLaVA	0.0	5.5	**81.8**
LLaVA + A-MoF	0.25	7.9 (+2.4)	79.4 (-2.4)
	0.5	12.0 (+6.5)	78.6 (-3.2)
	0.625	15.0 (+9.5)	76.4 (-5.4)
	0.75	**18.7** (+13.2)	75.8 (-6.0)
	0.875	16.5 (+11.0)	69.3 (-12.5)
	1.0	13.4 (+7.9)	68.5 (-13.3)

Table 2. **Empirical Results of Additive MoF.** We use DINOv2 as the image SSL model in our work. With more DINOv2 features added, there is an improvement in visual grounding, while a decline in instruction following ability.

Interleaved-MoF MLLM

Carefully designed mixing between CLIP and Vision-SSL give you both:



method	res	#tokens	MMVP	LLaVA	POPE
LLaVA	224^2	256	5.5	81.8	50.0
LLaVA	336^2	576	6.0	81.4	50.1
LLaVA + I-MoF	224^2	512	16.7 (+10.7)	82.8	51.0
$\text{LLaVA}^{1.5}$	336^2	576	24.7	84.7	85.9
$\text{LLaVA}^{1.5}$ + I-MoF	224^2	512	28.0 (+3.3)	82.7	86.3

Table 3. **Empirical Results of Interleaved MoF.** Interleaved MoF improves visual grounding while maintaining same level of instruction following ability.

slide credit: Shenbang Tong

Other SSL Models Works too

method	SSL Model	res	#tokens	MMVP		POPE
LLaVA$^{1.5}$	None	336^2	576	24.7		85.9
LLaVA$^{1.5}$ + I-MoF	MoCov3	224^2	512	26.7	(+2.0)	86.1
LLaVA$^{1.5}$ + I-MoF	MAE	224^2	512	27.3	(+2.6)	86.1
LLaVA$^{1.5}$ + I-MoF	DINOv2	224^2	512	28.0	(+3.3)	86.3

Takeaway

- MLLMs can be ignorant of unexpected mistakes.

- Scaling up CLIP models does not resolve the issue.

- Adding a Vision-Only SSL Feature is a step towards the cure.

- Vision is *not* ready for Language yet.

slide credit: Shenbang Tong

Overview of Today's Lecture

- SigLIP — Sigmoid Loss for Language Image Pre-Training
  - [iccv'23] — https://arxiv.org/abs/2303.15343
  - [arxiv'24] — https://arxiv.org/abs/2405.13777
- Multimodal Learning
  - ImageBind: [cvpr'23] - https://arxiv.org/abs/2305.05665
- Eyes Wide Shut
  - [cvpr'24] — https://arxiv.org/abs/2401.06209
- GiT: Towards Generalist Vision Transformer
  - [eccv'24] — https://arxiv.org/abs/2403.09394

max planck institut informatik

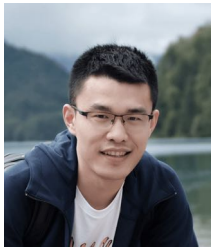
SIC Saarland Informatics Campus

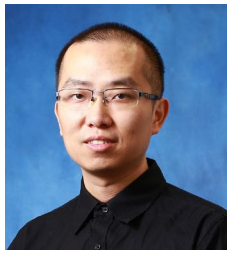
GiT - Generalist Vision Transformer through Universal Language Interface

@ECCV'24 @ github

Haiyang Wang, Hao Tang, Li Jlang, Shaoshuai Shi, Muhammad Ferjad Naeem, Hongsheng Li, Bernt Schiele, Liwei Wang

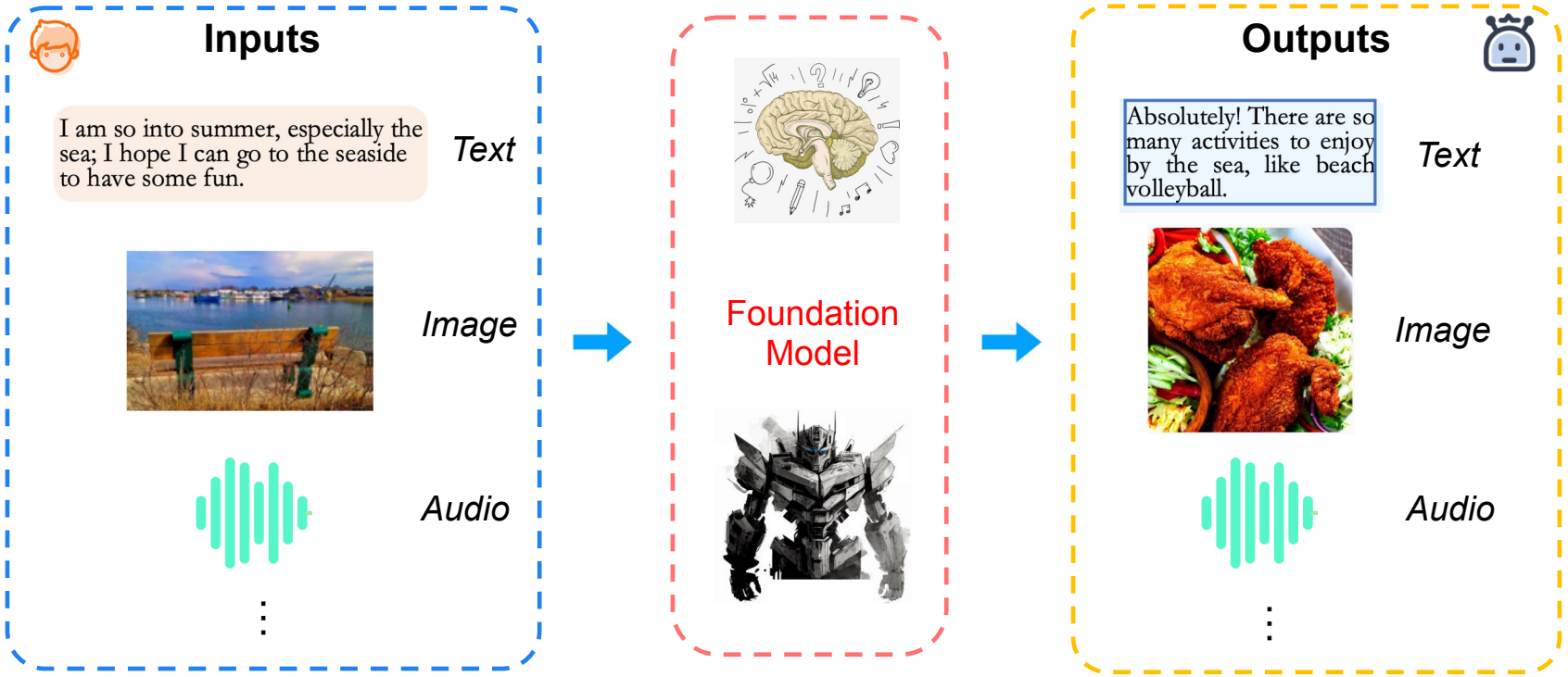






Motivation: Universal Computational Model

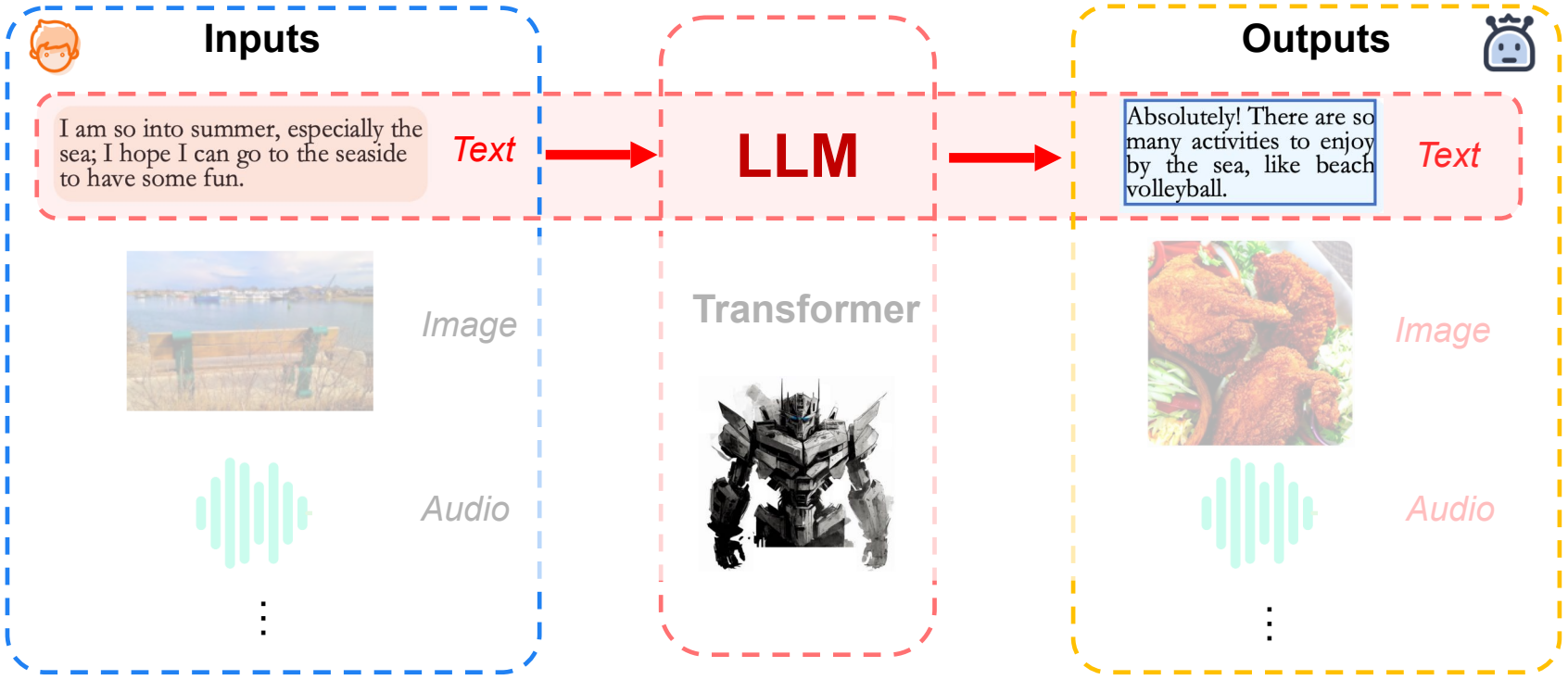
Any-to-Any Foundational Framework



Universal Computational Model for *NLP*

Transformer

Any-to-Any Foundational Framework

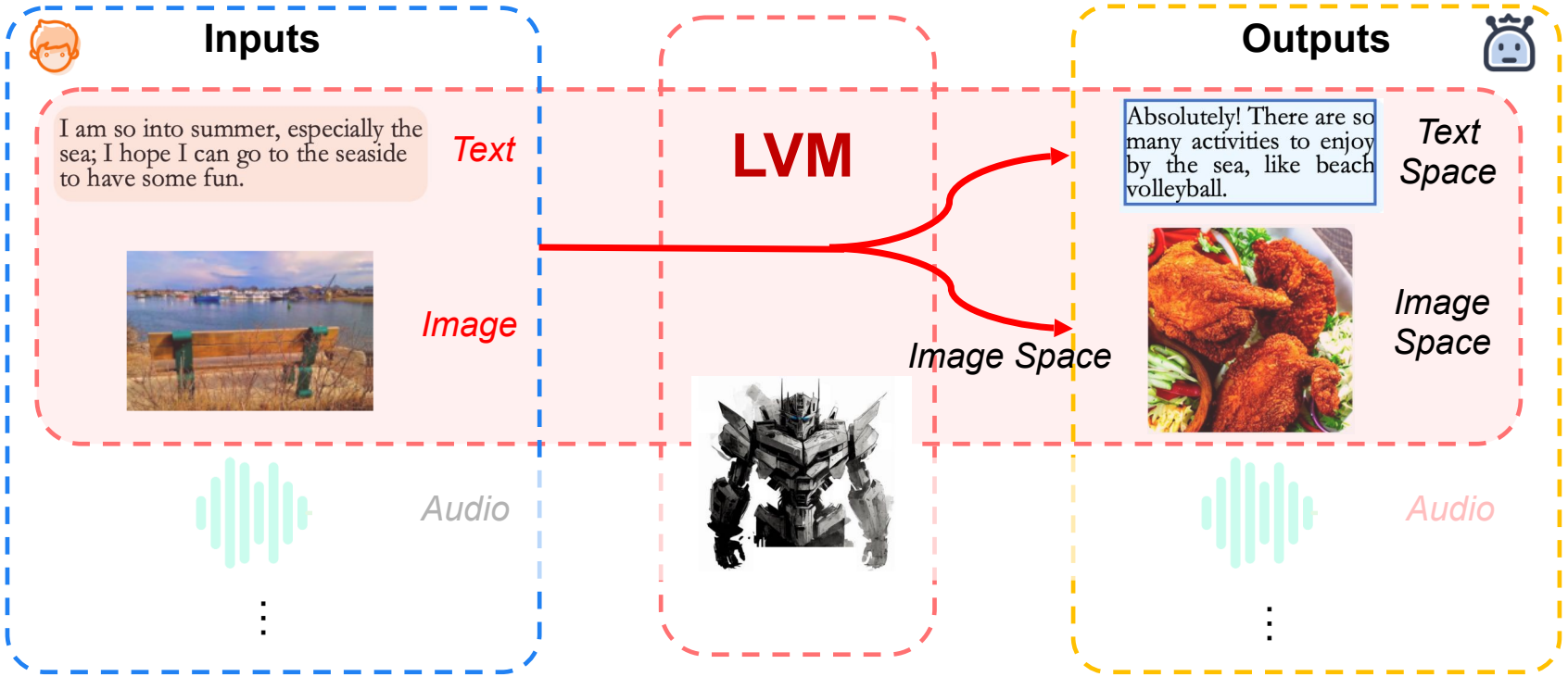


Language-to-Language

Universal Computational Model for *Computer Vision*

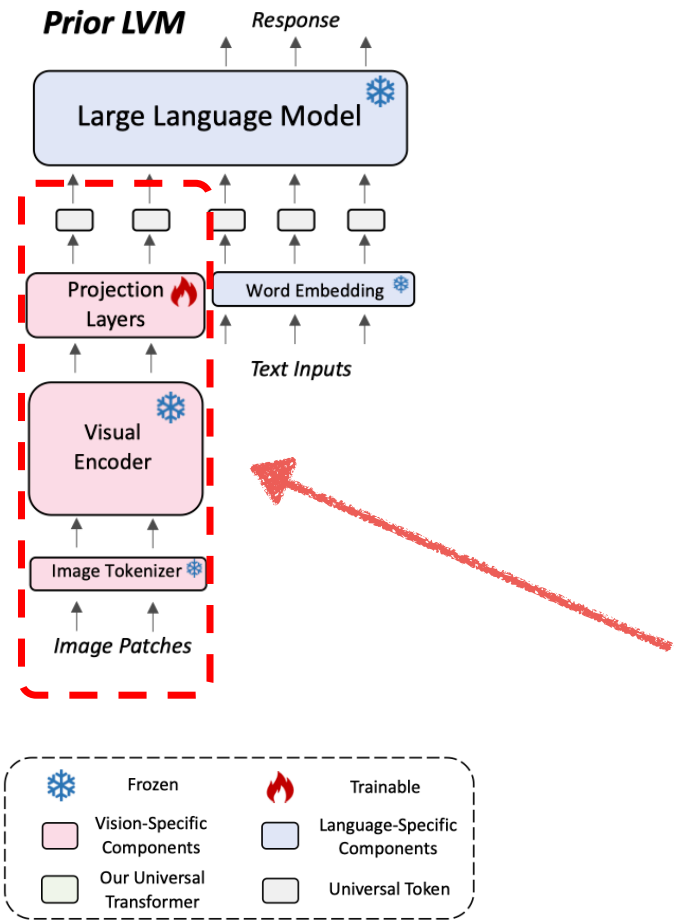
Also Transformer?

Any-to-Any Foundational Framework



Language & Image -to- Language & Image

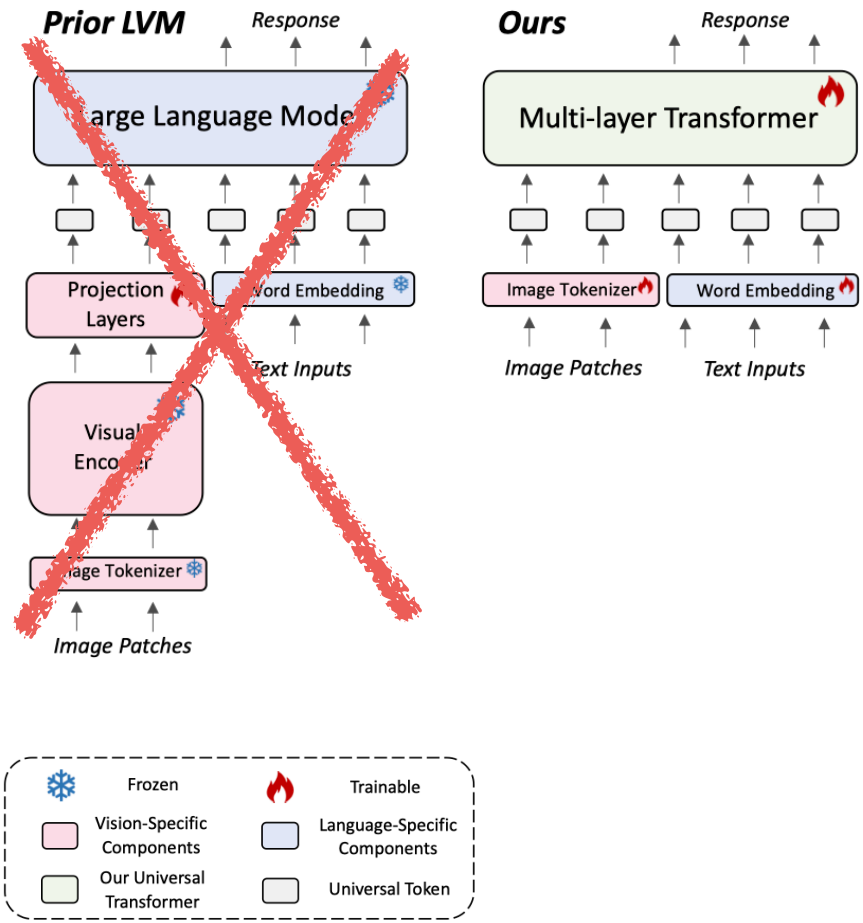
LVM: Large Vision Models



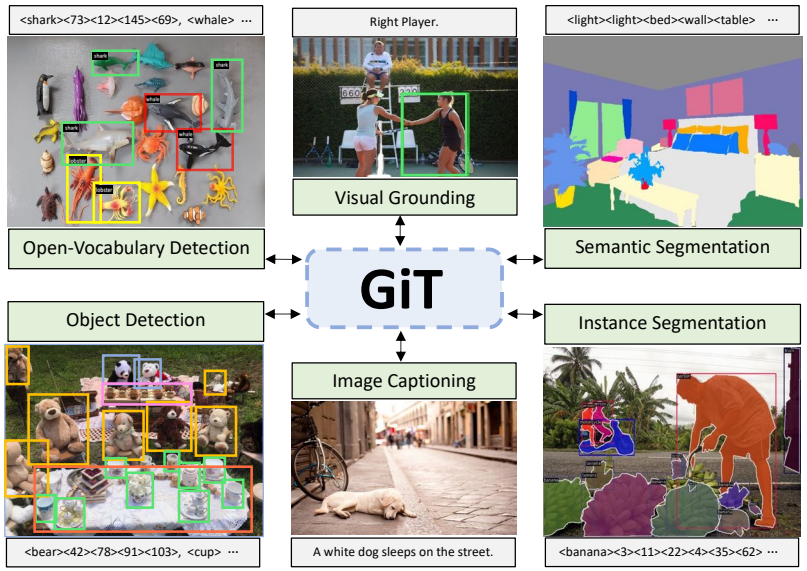
- **Examples include:**
  - **MiniGPT-4** & **LLaVA**: Visual Encoder (ViT)
  - **Unified-IO**: VQ-VAE and Visual Encoder
  - **Uni-Perciever v2**: RPN, task-specific heads
  - **VisionLLM**: Visual Encoder, Deformable Att
  - **NextGPT**: Visual Encoder, Audio Encoder
  - …

Treat vision features as foreign language in LLMs

GiT: Generalist Vision Transformer

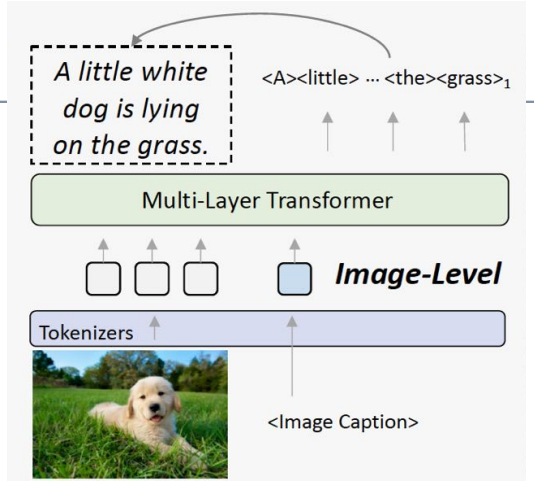


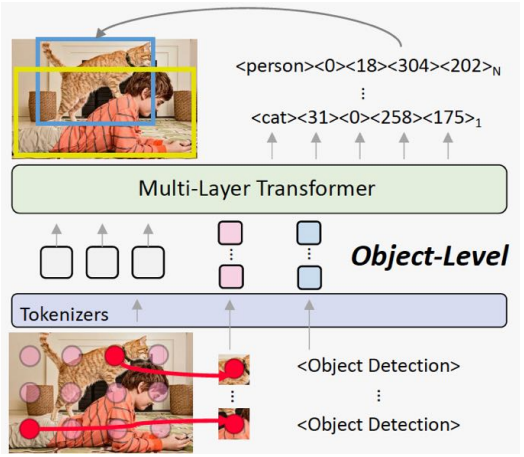
- **A Simple Approach:**
  - single & shared multi-layer transformer
  - multi-task learning for multiple vision tasks

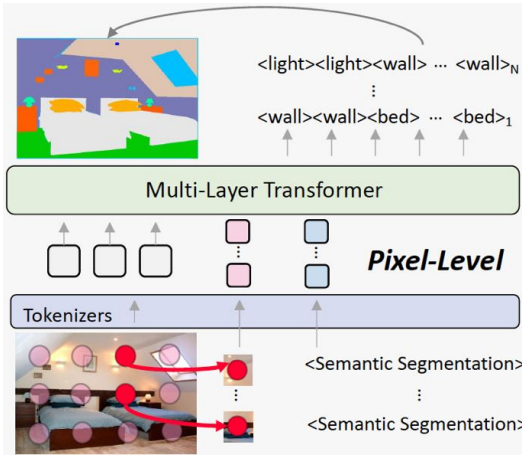


Universal Language Interface for Different Types of Tasks

- Image Level Tasks:
  - image captioning => sentence
  - visual grounding => bounding box
- Region / Object Level Tasks
  - object detection
    (per grid-cell => class + bounding box)
  - instance segmentation
    (per grid-cell => class + polygon)
- Dense Prediction Tasks
  - semantic segmentation
    (per grid-cell => list of classes)

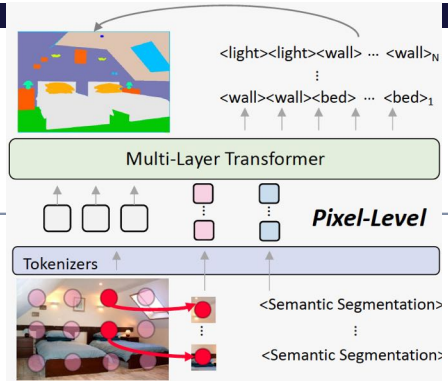


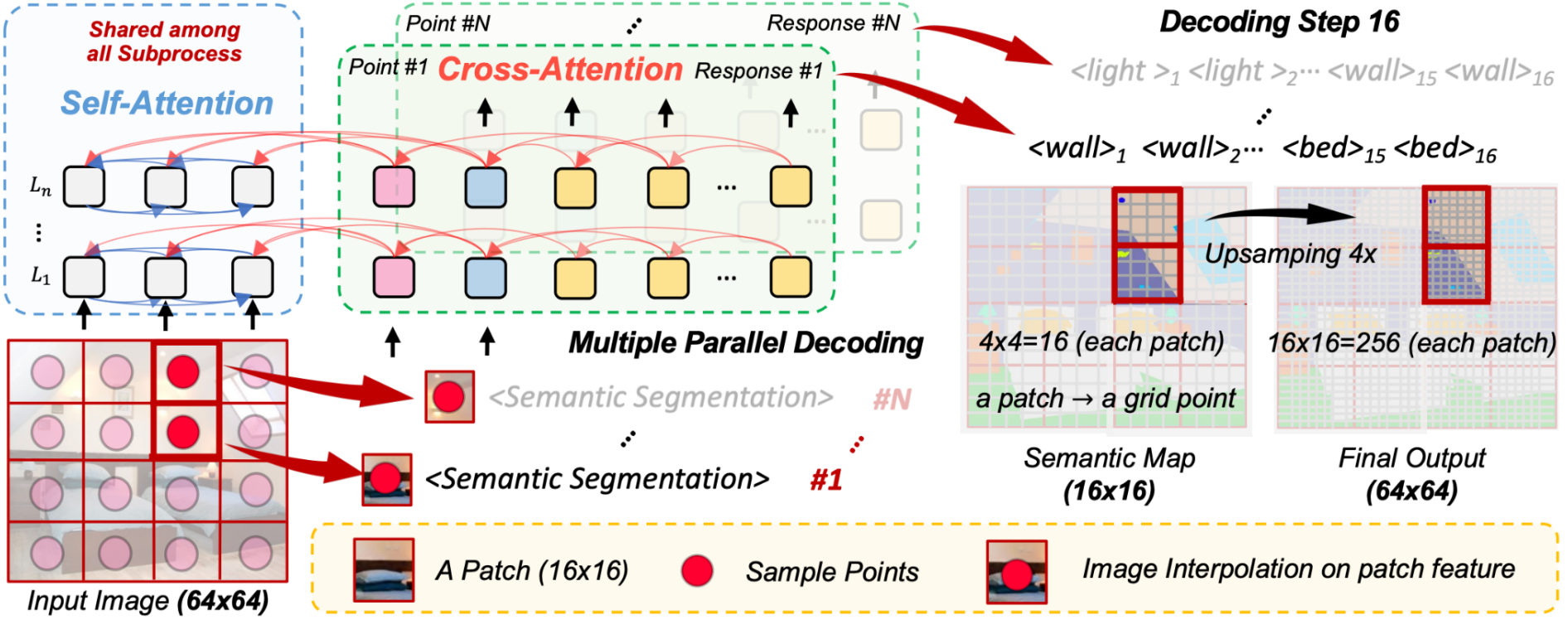




Parallel Processing of all Grid-Points

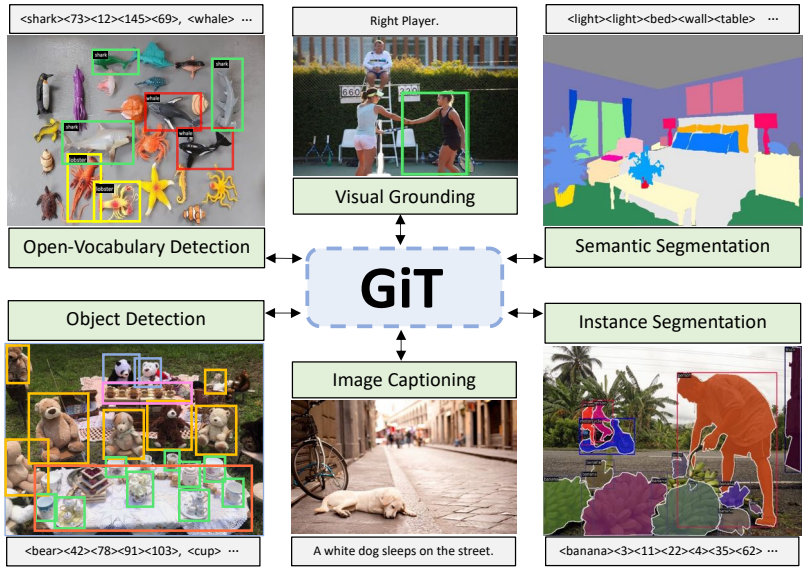
- Example of Semantic Segmentation





GiT - Generalist Vision Transformer

- Experimental setting
  - single multi-layer transformer architecture
    - GiT-B (131M), GiT-L (387M), GiT-H (756M)
    - 98-99% shared parameters across tasks
  - SAM-encoder architecture (used as initialization) + 6 encoder layers (random initialization)
  - trained for multiple vision tasks, across multiple datasets



Training Settings: Multi-Task and Universal Training

- In the following:
  - Multi-Task Setting:

Task	Dataset	Size	Sample ratio
Object detection	COCO 2017	164K	0.2
Instance segmentation	COCO 2017	164K	0.2
Semantic segmentation	ADE20K	20K	0.2
Image captioning	COCO Caption	164K	0.2
Visual grounding	RefCOCO	20K	0.2

  - Universal Setting (training across 27 datasets):

	Example Sources	Size				Input Modalities		Output Modalities		
		Dataset	Size	Percent	Weight	Text	Image	Text	Sparse	Dense
Image-Level		**10**	**11.4m**	**67.1**	**40**	✓	✓	✓	✓	-
Image Captioning	*CC12M [14], VG [46], SBU [66]*	5	11.3m	66.6	30	-	✓	✓	-	-
Visual Grounding	*RefCOCO [100], Flickr30k [68]*	5	115k	0.7	10	✓	✓	-	✓	-
Object-Level		**11**	**5.2m**	**30.9**	**40**	-	✓	-	✓	✓
Object Detection	*Objects365 [75], COCO [54]*	8	3.8m	22.6	20	-	✓	-	✓	-
Instance Segmentation	*OpenImages [48], LVIS [35]*	4	1.4m	7.9	20	-	✓	-	✓	✓
Pixel-Level		**6**	**322k**	**2.0**	**20**	-	✓	-	-	✓
Semantic Segmentation	*COCOStuff [12], ADE20K [103]*	6	322k	2.0	20	-	✓	-	-	✓
All Tasks		**27**	**17m**	**100**	**100**	✓	✓	✓	✓	✓

State-of-the-Art Generalist Performance

- Multi-Task Setting vs. Single-Task Training:

Methods	Specific Modules Examples	Num	#Params	Object Detection AP	AP_{50}	AP_{75}	Instance Seg AP	AP_{50}	AP_{75}	Semantic Seg mIoU(SS)	Captioning BLEU-4	CIDEr	Grounding Acc@0.5
Specialist Models													
Faster R-CNN FPN [73]	ResNet FPN	5	42M	40.3	61.0	44.0							

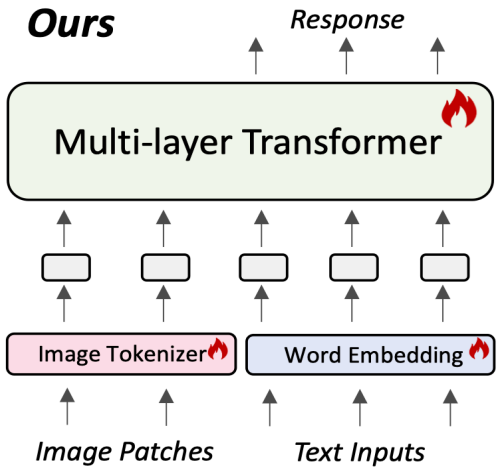
Task	Image	Language	Segment	Localization	*Improve* (single→multi)
Detection	✓	-	-	✓	+***1.6***@AP
InsSeg	✓	-	✓†	✓	+***1.6***@AP_{50}, +***0.2***@AP_{75}
Grounding	✓	✓	-	✓	+***2.5***@Acc
Caption	✓	✓	-	-	+***4.7***@CIDEr
SemSeg	✓	-	✓	-	+***0.1***@mIoU

Generalist Models (MultiTask-Training)													
Uni-Perceiver [107]	None	1	124M	-	-	-	-	-	-	-	32.0	⋆	⋆
Uni-Perceiver-MoE [105]	None	1	167M	-	-	-	-	-	-	-	33.2	⋆	⋆
Uni-Perceiver-V2 [49]	Mask DINO,Swin	8	308M	58.6†	⋆	⋆	50.6†	⋆	⋆	-	35.4	116.9	⋆
VisionLLM-R50 [89]	Deform-DETR	6	7B	44.6	64.0	48.1	25.1	50.0	22.4	-	31.0	112.5	80.6
$GiT\text{-}B_{single\text{-}task}$	None	1	131M	45.1	62.7	49.1	31.4	54.8	31.2	47.7	33.7	107.9	83.3
$GiT\text{-}B_{multi\text{-}task}$	None	1	131M	46.7	64.2	50.7	31.9	56.4	31.4	47.8	35.4	112.6	85.8
Improvement (single→multi)				***+1.6***	***+1.5***	***+1.6***	***+0.5***	***+1.6***	***+0.2***	***+0.1***	***+1.7***	***+4.7***	***+2.5***
$GiT\text{-}L_{multi\text{-}task}$	None	1	387M	51.3	69.2	55.9	35.1	61.4	34.7	50.6	35.7	116.0	88.4
$GiT\text{-}H_{multi\text{-}task}$	None	1	756M	**52.9**	**71.0**	**57.8**	**35.8**	**62.6**	**35.6**	**52.4**	**36.2**	**118.2**	**89.2**

State-of-the-art Zero-shot Transfer Performance

Methods	Specific Modules Examples	Num	#Params	Object Detection Cityscapes [27]	Instance Seg Cityscapes [27]	Semantic Seg Cityscapes [27]	Semantic Seg SUN RGB-D [78]	Captioning nocaps [2]
Supervised								
Faster R-CNN-FPN [73]	ResNet,RPN	5	42M	40.3	-	-	-	-
Mask R-CNN [36]	ResNet,RPN	6	46M	40.9	36.4	-	-	-
DeepLabV3+ [19]	ResNet,Decoder	3	63M	-	-	80.9	⋆	-
Mask2Former [25]	ResNet,Decoder	5	44M	-	-	80.4	⋆	-
TokenFusion [91]	Segformer,YOLOS	4	-	-	-	⋆	48.1	-
Zero-Shot Transfer								
GLIP-T [52]	Swin,Dy-Head	5	156M	28.1†	-	-	-	-
Grounding DINO-T [56]	Swin,DINO	6	174M	31.5†	-	-	-	-
BLIP-2 (129M) [50]	ViT-G,Qformer	4	12.1B	-	-	-	-	**15.8**
ReCo+ [77]	DeiT-SIN	4	46M	-	-	24.2	⋆	-
XDecoder-T [108]	FocalNet,Encoder	4	165M	-	16.0	47.3	34.5	⋆
GiT-$B_{\text{multi-task}}$	None	1	131M	21.8	14.3	34.4	30.9	9.2
GiT-$B_{\text{universal}}$	None	1	131M	29.1	17.9	56.2	37.5	10.6
GiT-$L_{\text{universal}}$	None	1	387M	32.3	**20.3**	58.0	39.9	11.6
GiT-$H_{\text{universal}}$	None	1	756M	**34.1**	18.7	**61.8**	**42.5**	12.6

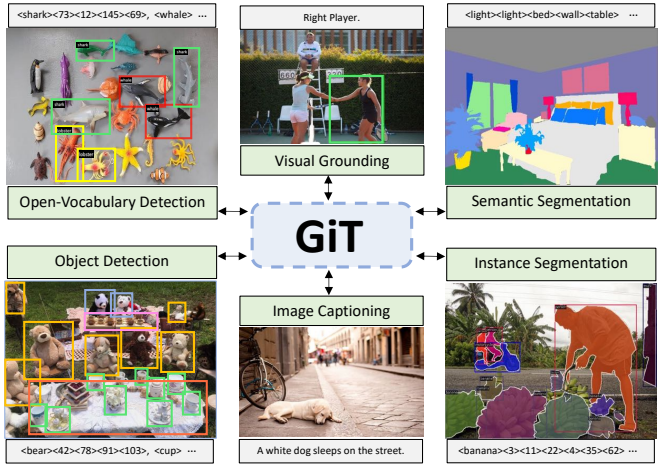
GiT: Generalist Vision Transformer



- Foundational framework for unified visual modeling
  - based one **single** & **shared multi-layer transformer**
    - GiT-B (98% shared parameters) — GiT-H (99% shared parameters)
  - for **Image**-, **Object**-, and **Pixel-Level** tasks
  - **universal "language" interface**

- Multi-task ability like LLMs
  - strong generalizability (within multi-task-setting & out-of-distribution — zero/few-shot transfer)
  - also good open-set detection performance

Task	Image	Language	Segment	Localization	*Improve* (single→multi)
Detection	✓	-	-	✓	***+1.6***@AP
InsSeg	✓	-	✓†	✓	***+1.6***@AP_{75}, ***+0.2***@AP_{50}
Grounding	✓	✓	-	✓	***+2.5***@Acc
Caption	✓	✓	-	-	***+4.7***@CIDEr
SemSeg	✓	-	✓	-	***+0.1***@mIoU



Overview of Today's Lecture

- SigLIP — Sigmoid Loss for Language Image Pre-Training
  - [iccv'23] — https://arxiv.org/abs/2303.15343
  - [arxiv'24] — https://arxiv.org/abs/2405.13777
- Multimodal Learning
  - ImageBind: [cvpr'23] - https://arxiv.org/abs/2305.05665
- Eyes Wide Shut
  - [cvpr'24] — https://arxiv.org/abs/2401.06209
- GiT: Towards Generalist Vision Transformer
  - [eccv'24] — https://arxiv.org/abs/2403.09394
-