CS 839: Foundation Models
Multimodal Models

Fred Sala

University of Wisconsin-Madison

Oct. 29, 2024

Announcements

- **Logistics:**
  - HW3 out tonight.
  - Class roadmap:

Tuesday Oct. 29	Multimodal Foundation Models
Thursday Oct. 31	Diffusion Models
Tuesday Nov. 5	Scaling & Scaling Laws
Thursday Nov. 7	Security, Privacy, Toxicity + Future Areas

Outline

- **Multimodal Models Intro + One-Encoder Models**
  - Short history, adapting models to incorporate multiple modalities, BERT-like vision-language models, ViTs
- **VLM Variations and Types**
  - Multi-encoder setups, contrastive training, CLIP, joint training, few-shot models, visual instructions
- **Other Modalities and Domains**
  - Audio, video, code generation, RL

Outline

- **Multimodal Models Intro + One-Encoder Models**
  - Short history, adapting models to incorporate multiple modalities, BERT-like vision-language models, ViTs
- **VLM Variations and Types**
  - Multi-encoder setups, contrastive training, CLIP, joint training, few-shot models, visual instructions
- **Other Modalities and Domains**
  - Audio, video, code generation, RL

Short History of Multimodal Models

Multimodal models pre-date foundation models

- Image-captioning models, VQA models, esc…
  - But it has become more popular

- Ex: **joint embedding spaces**

(Weston, Bengio, Usunier ’11)

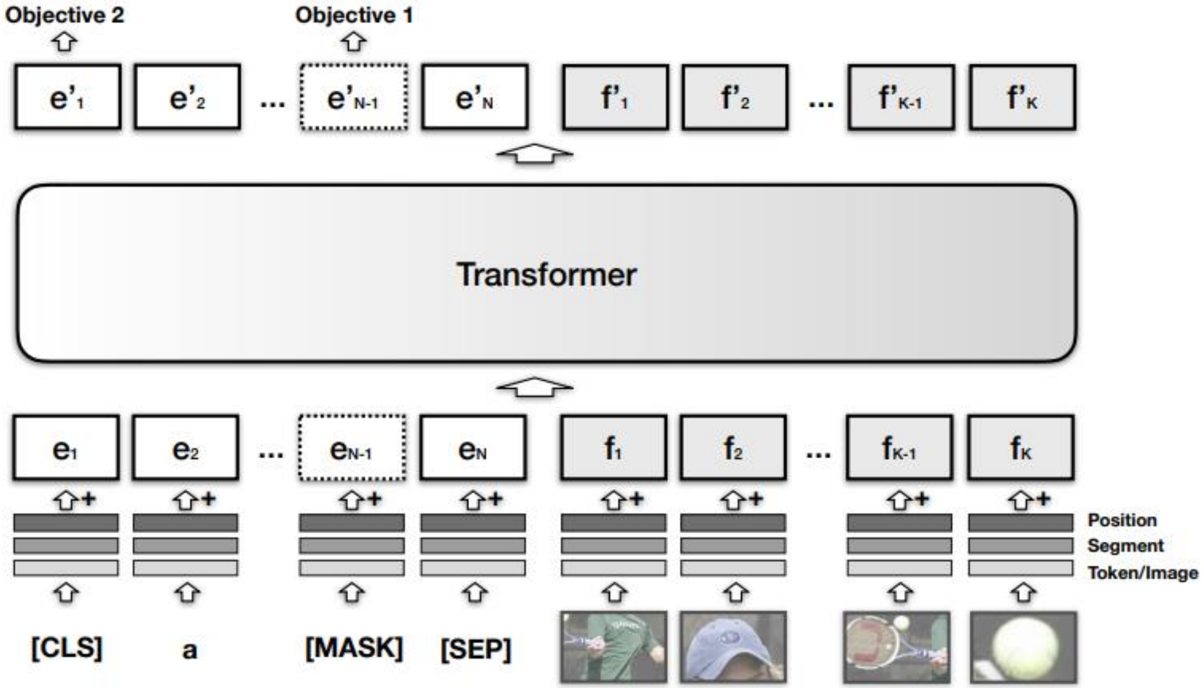
Image	One-vs-Rest	WSABIE
	surf, bora, belize, sea world, balena, wale, tahiti, delfini, surfing, mahi mahi	delfini, orca, **dolphin**, mar, delfin, dauphin, whale, cancun, killer whale, sea world
	**eiffel tower**, tour eiffel, snowboard, blue sky, empire state building, luxor, eiffel, lighthouse, jump, adventure	**eiffel tower**, statue, eiffel, mole antoneliana, la tour eiffel, londra, cctv tower, big ben, calatrava, tokyo tower
	falco, barack, daniel craig, **obama**, barack obama, kanye west, pharrell williams, 50 cent, barrack obama, bono	barrack obama, barack obama, barack hussein obama, barack obama, james marsden, jay z, **obama**, nelly, falco, barack

Making LLMs Multimodal

How do we use a language architecture for multiple modalities?

VisualBERT: take all the ideas from BERT, add images

- Use bounding boxes from image detector + image embedder



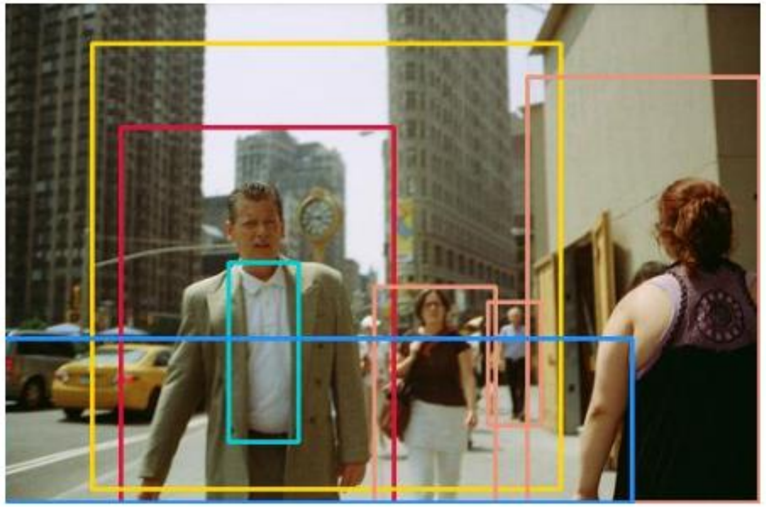
A person hits a ball with a tennis racket

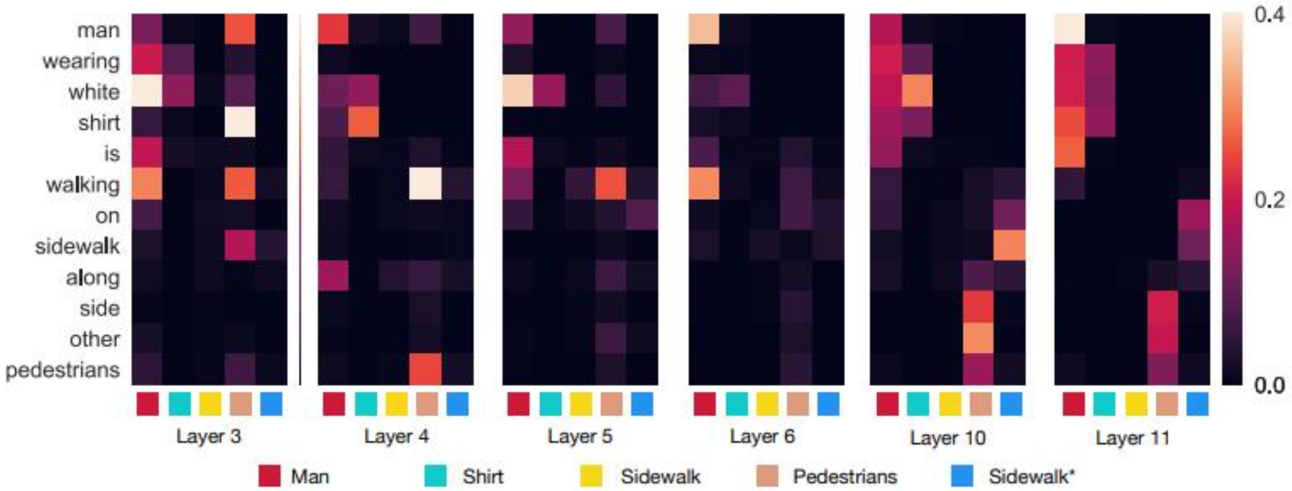
Li et al '19

Making LLMs Multimodal: **VisualBERT**

VisualBERT: take all the ideas from BERT, add images

- What about training? Recall BERT training…
  - Masked language modeling + image (text is masked, image same)
  - Sentence-image prediction
- Results (Li et al, '19)



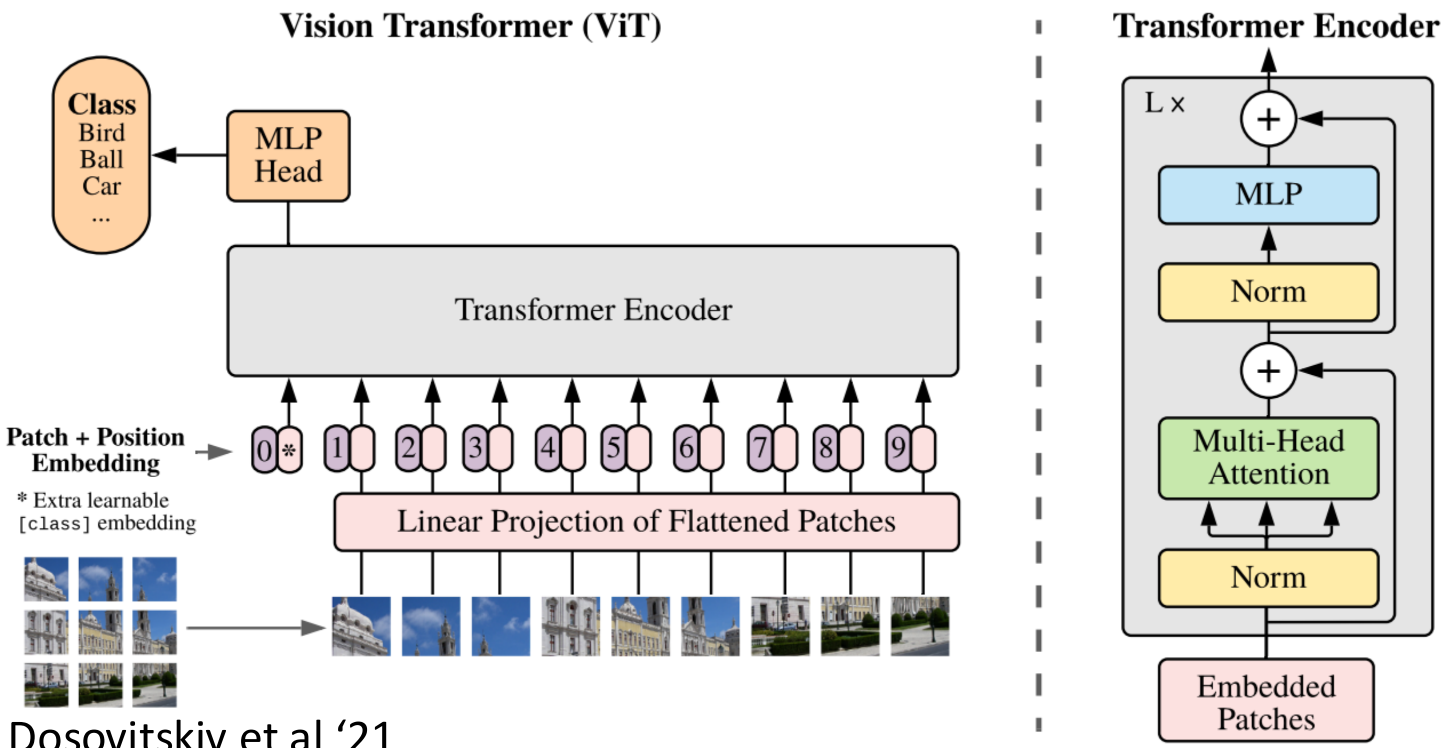


How Do We Get Image Embeddings?

Could always user resnets, etc., but…

An Image is Worth 16x16 Words: Transformers for ...
by A Dosovitskiy · 2020 · Cited by 46708 — A pure transformer applied directly to sequences of image patches can perform very well on image classification tasks.

- Didn't Transformers make a big difference for text?
- Can also use for vision: **ViT.** Just use patches!

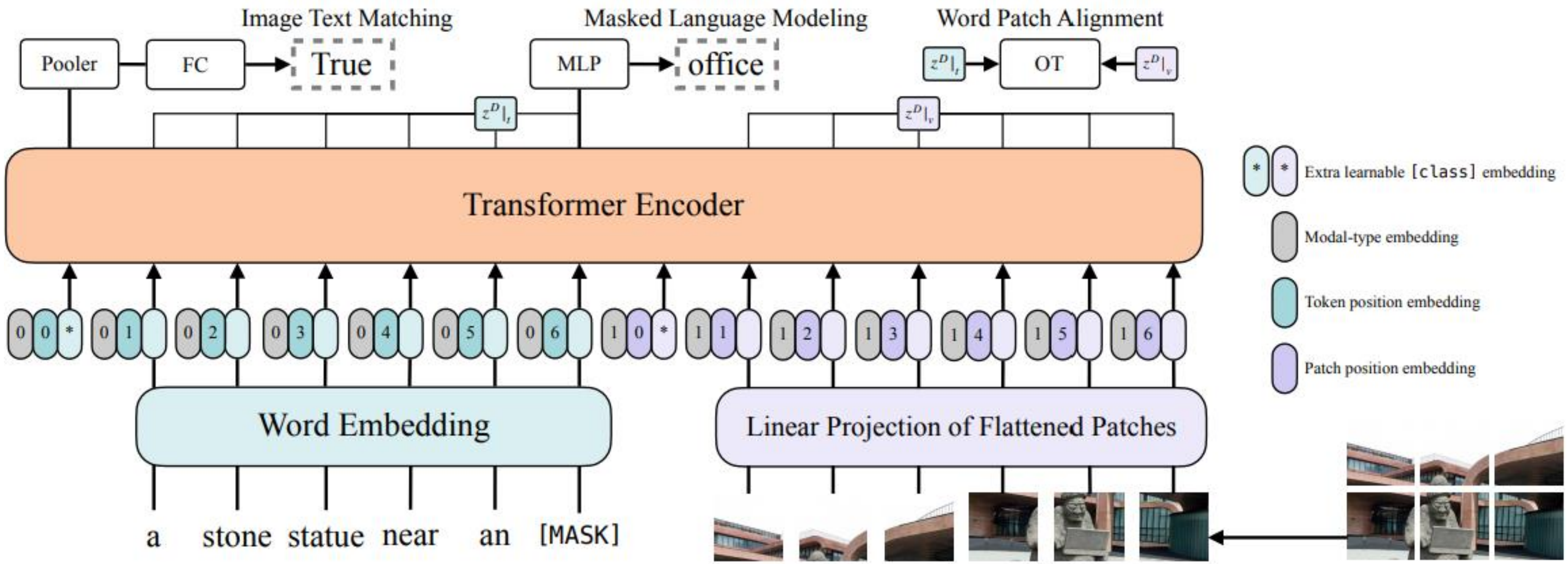


Dosovitskiy et al '21

Put It Together

Multimodal with language and vision transformers: **ViLT**

- Kim et al '21



Variations…

Lots of different approaches!

- Du et al ‘22, “A Survey of Vision-Language Pre-Trained Models”

VL-PTM	Text encoder	Vision encoder	Fusion scheme	Pre-training tasks	Multimodal datasets for pre-training
Fusion Encoder					
VisualBERT [2019]	BERT	Faster R-CNN	Single stream	MLM+ITM	COCO
Uniter [2020]	BERT	Faster R-CNN	Single stream	MLM+ITM+WRA+MRFR+MRC	CC+COCO+VG+SBU
OSCAR [2020c]	BERT	Faster R-CNN	Single stream	MLM+ITM	CC+COCO+SBU+Flickr30k+VQA
InterBert [2020]	BERT	Faster R-CNN	Single stream	MLM+MRC+ITM	CC+COCO+SBU
ViLBERT [2019]	BERT	Faster R-CNN	Dual stream	MLM+MRC+ITM	CC
LXMERT [2019]	BERT	Faster R-CNN	Dual stream	MLM+ITM+MRC+MRFR+VQA	COCO+VG+VQA
VL-BERT [2019]	BERT	Faster R-CNN+ ResNet	Single stream	MLM+MRC	CC
Pixel-BERT [2020]	BERT	ResNet	Single stream	MLM+ITM	COCO+VG
Unified VLP [2020]	UniLM	Faster R-CNN	Single stream	MLM+seq2seq LM	CC
UNIMO [2020b]	BERT, RoBERTa	Faster R-CNN	Single stream	MLM+seq2seq LM+MRC+MRFR+CMCL	COCO+CC+VG+SBU
SOHO [2021]	BERT	ResNet + Visual Dictionary	Single stream	MLM+MVM+ITM	COCO+VG
VL-T5 [2021]	T5, BART	Faster R-CNN	Single stream	MLM+VQA+ITM+VG+GC	COCO+VG
XGPT [2021]	transformer	Faster R-CNN	Single stream	IC+MLM+DAE+MRFR	CC
Visual Parsing [2021]	BERT	Faster R-CNN + Swin transformer	Dual stream	MLM+ITM+MFR	COCO+VG
ALBEF [2021a]	BERT	ViT	Dual stream	MLM+ITM+CMCL	CC+COCO+VG+SBU
SimVLM [2021b]	ViT	ViT	Single stream	PrefixLM	C4+ALIGN
WenLan [2021]	RoBERTa	Faster R-CNN + EffcientNet	Dual stream	CMCL	RUC-CAS-WenLan
ViLT [2021]	ViT	Linear Projection	Single stream	MLM+ITM	CC+COCO+VG+SBU
Dual Encoder					
CLIP [2021]	GPT2	ViT, ResNet		CMCL	self-collected
ALIGN [2021]	BERT	EffcientNet		CMCL	self-collected
DeCLIP [2021b]	GPT2, BERT	ViT, ResNet, RegNetY-64GF		CMCL+MLM+CL	CC+self-collected
Fusion Encoder+ Dual Encoder					
VLMo [2021a]	BERT	ViT	Single stream	MLM+ITM+CMCL	CC+COCO+VG+SBU
FLAVA [2021]	ViT	ViT	Single stream	MMM+ITM+CMCL	CC+COCO+VG+SBU+RedCaps

Datasets

Trained on? Datasets with image-text pairs

Dataset	Year	Num. of Image-Text Pairs	Language	Public
SBU Caption [92] [link]	2011	1M	English	✓
COCO Caption [93] [link]	2016	1.5M	English	✓
Yahoo Flickr Creative Commons 100 Million (YFCC100M) [94] [link]	2016	100M	English	✓
Visual Genome (VG) [95] [link]	2017	5.4 M	English	✓
Conceptual Captions (CC3M) [96] [link]	2018	3.3M	English	✓
Localized Narratives (LN) [97] [link]	2020	0.87M	English	✓
Conceptual 12M (CC12M) [98] [link]	2021	12M	English	✓
Wikipedia-based Image Tex (WIT) [99] [link]	2021	37.6M	108 Languages	✓
Red Caps (RC) [100] [link]	2021	12M	English	✓
LAION400M [28] [link]	2021	400M	English	✓
LAION5B [27] [link]	2022	5B	Over 100 Languages	✓
WuKong [101] [link]	2022	100M	Chinese	✓
CLIP [14]	2021	400M	English	✗
ALIGN [24]	2021	1.8B	English	✗
FILIP [25]	2021	300M	English	✗
WebLI [102]	2022	12B	109 Languages	✗

Zhang et al ‘23

Break & Questions

Outline

- **Multimodal Models Intro + One-Encoder Models**
  - Short history, adapting models to incorporate multiple modalities, BERT-like vision-language models, ViTs
- **VLM Variations and Types**
  - Multi-encoder setups, contrastive training, CLIP, joint training, few-shot models, visual instructions
- **Other Modalities and Domains**
  - Audio, video, code generation, RL

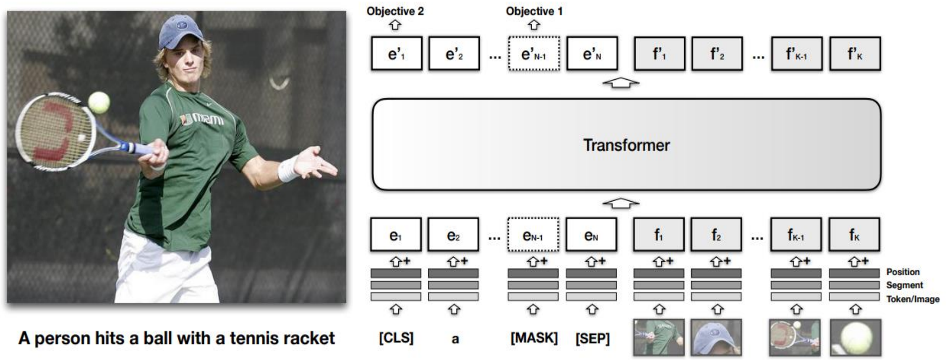
Contrastive Vision-Language Models

So far, trained the modalities together

- I.e., text and images were both inputs to a transformer
- This is "fusion", but we could do it **later**…

- I.e., produce two representations separately, then produce some means of connecting/tying them together

- **Contrastive** approach



Li et al '19

VLMs: **Constrastive Training**

Training approach: contrastive

- Loss example: InfoNCE (noise contrastive estimation) loss:

$$\mathcal{L}_I^{\text{InfoNCE}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(z_i^I \cdot z_+^I/\tau\right)}{\sum_{j=1,j\neq i}^{B+1}\exp(z_i^I \cdot z_j^I/\tau)}$$

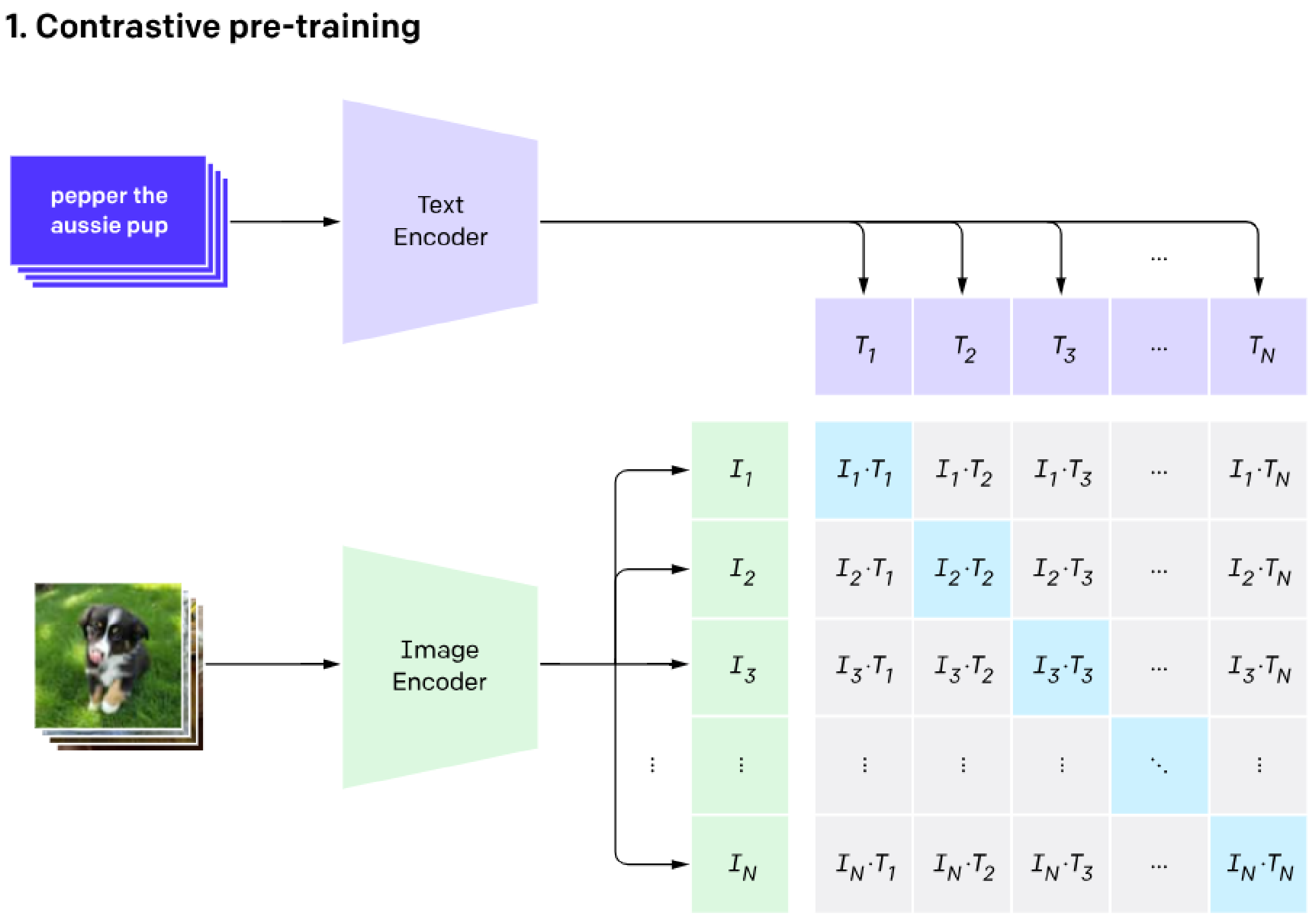
- To train a text and image encoder simultaneously, symmetrize:

$$\mathcal{L}_{I\to T} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(z_i^I \cdot z_i^T/\tau\right)}{\sum_{j=1}^{B}\exp(z_i^I \cdot z_j^T/\tau)}$$

$$\mathcal{L}_{T\to I} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(z_i^T \cdot z_i^I/\tau\right)}{\sum_{j=1}^{B}\exp(z_i^T \cdot z_j^I/\tau)}$$

VLMs: **CLIP**

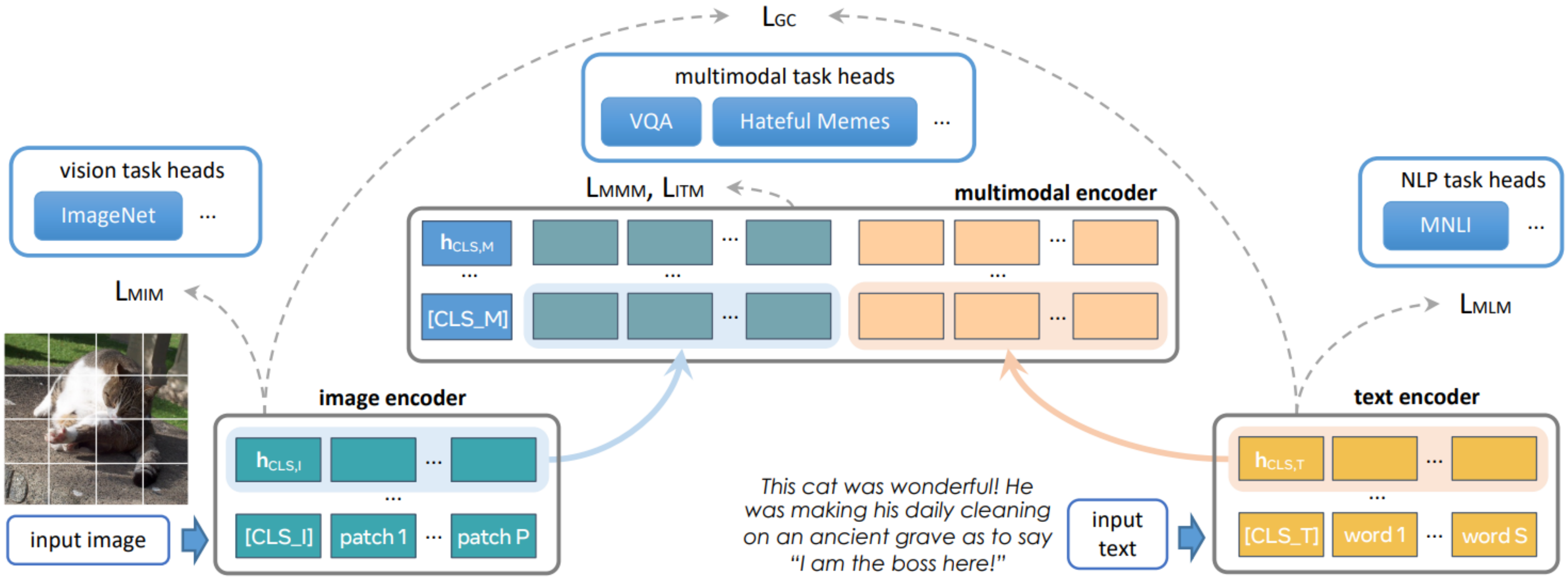
A simple but easily scalable constrastive VLM



VLMs: **FLAVA**

Foundational Language And Vision Alignment Model (FLAVA)

- Combines everything
- Pretrain **separately** and **jointly**



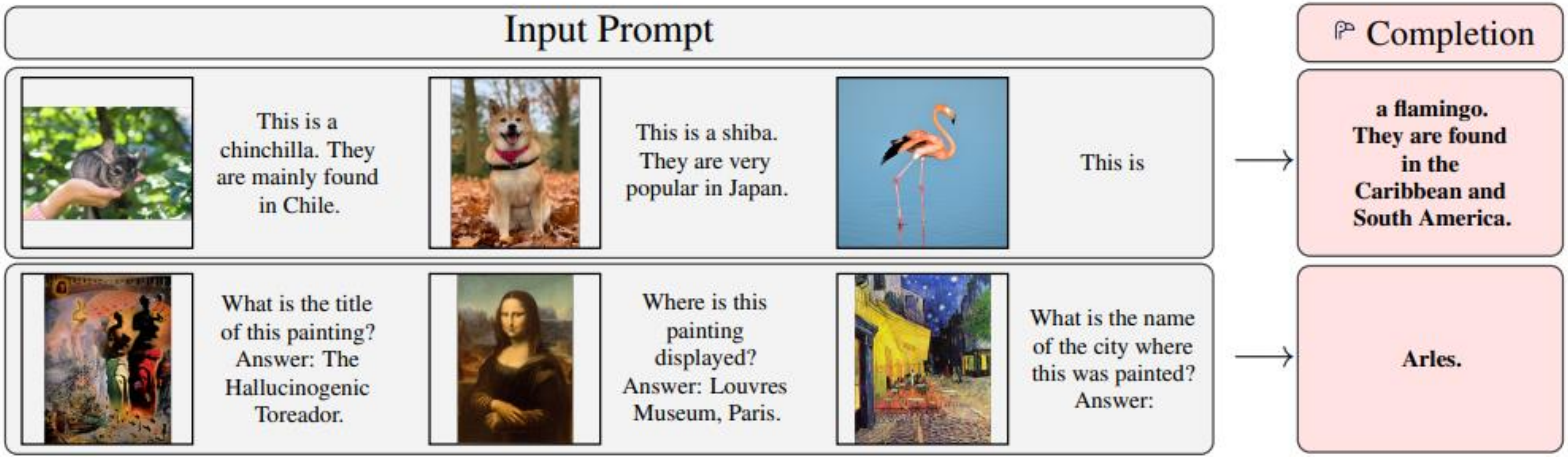
Singh et al '22

Few-Shot VLMs

The models we've talked about are either meant to

- Do zero-shot prediction, OR
- Be fine-tuned for a particular task

- What about **few-shot** (like in LLMs) for VLMs?

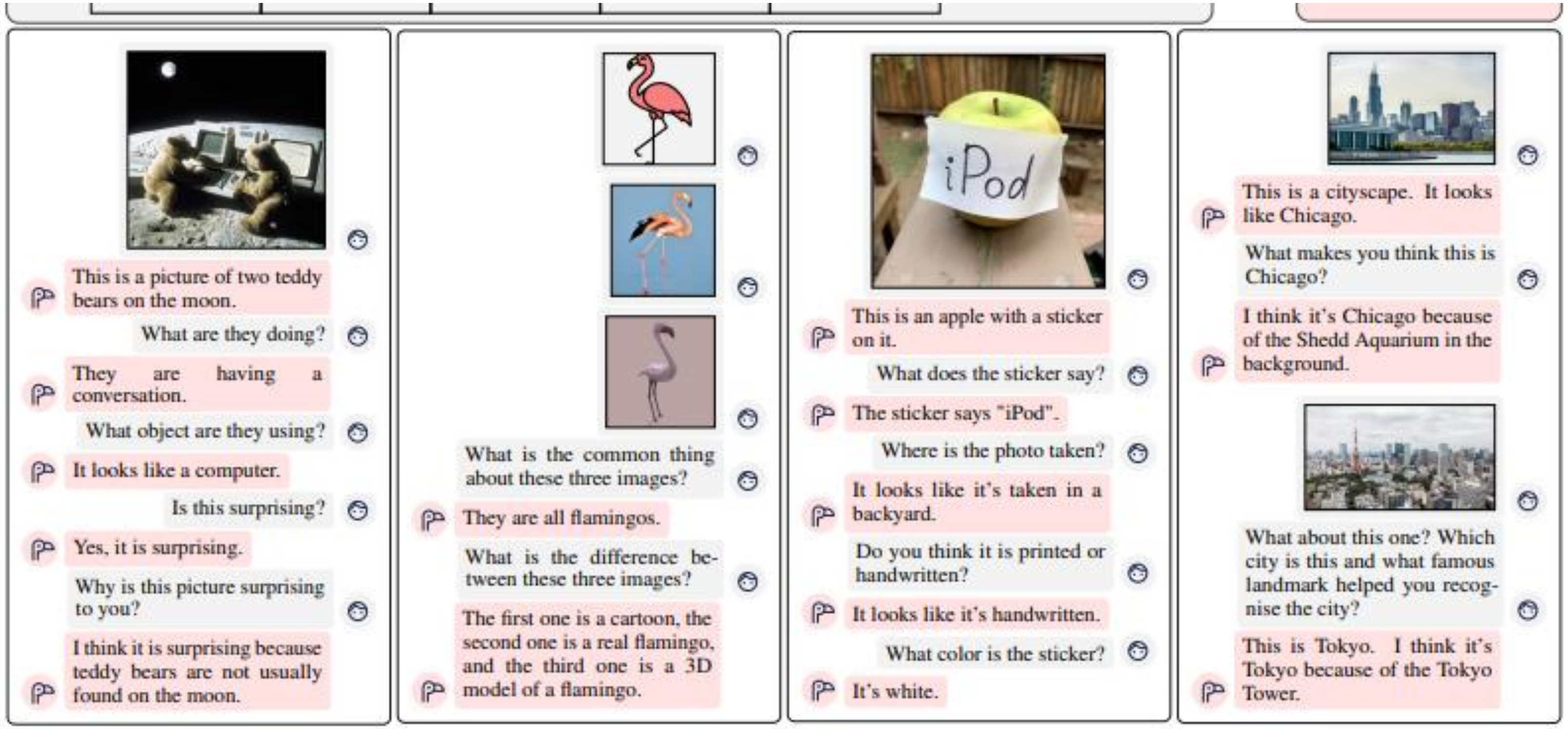


Alayrac et al '22

Few-Shot VLMs: **Flamingo**

Flamingo: 80B parameter model (based on an LLM)

- Multi-image!
- More complex interleaved architecture

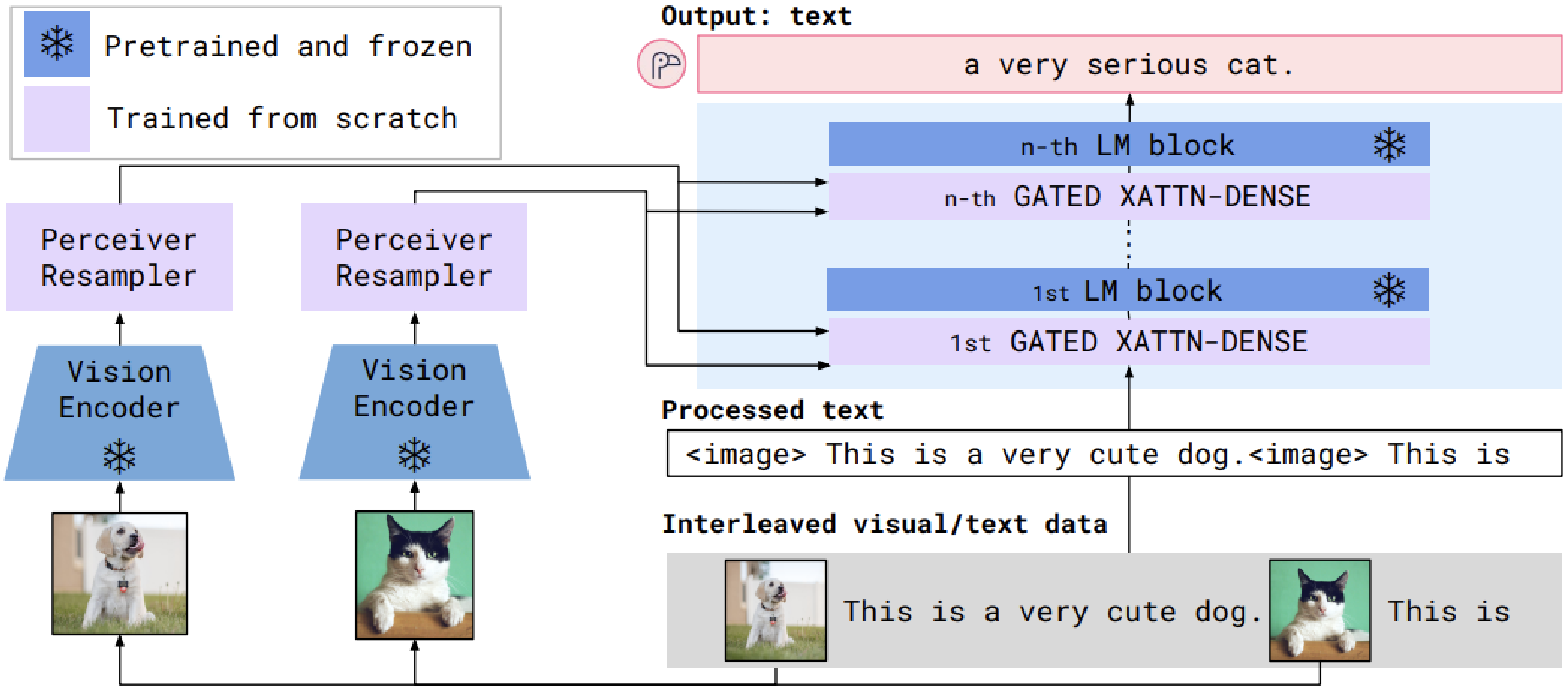


Alayrac et al '22

Few-Shot VLMs: **Flamingo**

Flamingo: 80B parameter model (based on an LLM)

- Multi-image!
- More complex interleaved architecture



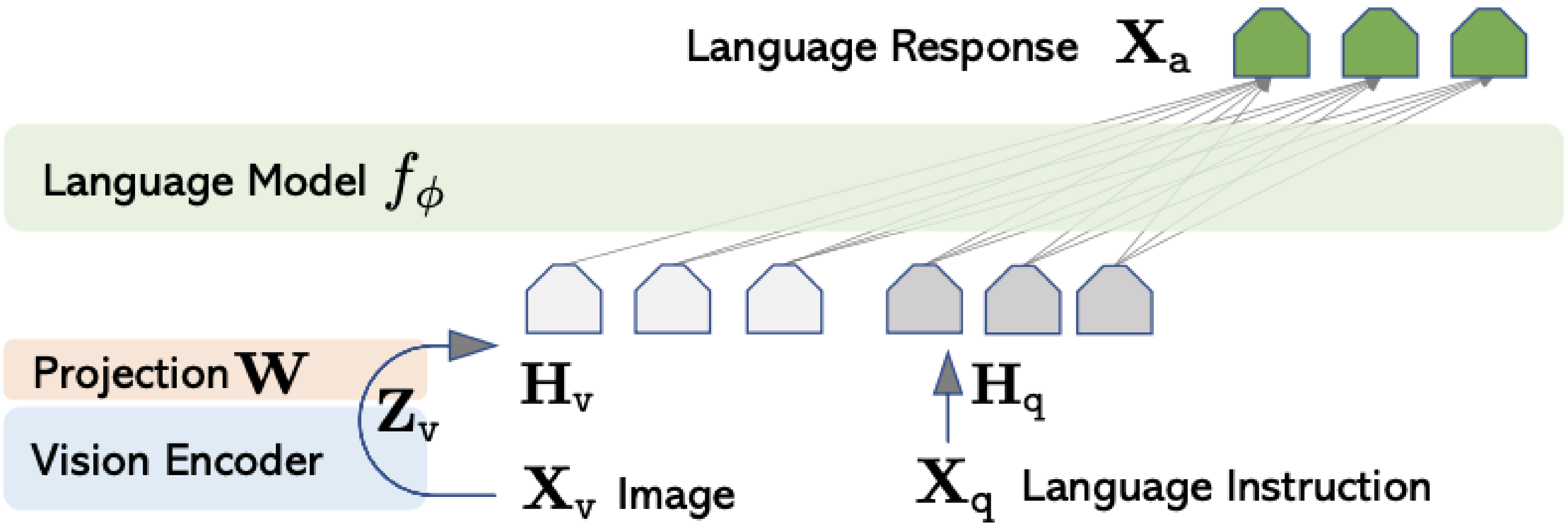
Alayrac et al '22

Visual Instructions: **Llava**

Visual instruction tuning

Goal: achieve close to GPT-4’s multimodal performance

- Simple, scalable architecture:



Liu et al ‘23

Visual Instructions: **Llava**

Example conversation

Note: challenging out-of-distribution example

Model designed to handle multi-round conversation

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

User	What is unusual about this image?
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
	[*Start a new conversation, and clear the history*]
User	What's happening in the scene?
LLaVA	The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.
User	What is unusual about this image?
GPT-4 [36]	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
User	What is unusual about this image?
BLIP-2	a man is sitting on the back of a yellow cab
User	What is unusual about this image?
OpenFlamingo	The man is drying his clothes on the hood of his car.

Liu et al '23

Closed-source: Gemini

Now, back to closed-source models.

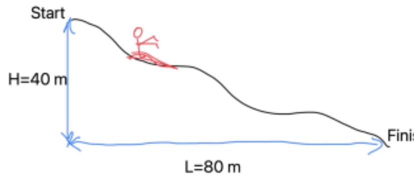
Similar capabilities, and can be more powerful because they are trained on all modalities **jointly**.

Downside: **expensive**!

Prompt

Here's a solution to a Physics problem by a student.

23. A skier slides down the frictionless slope as shown. What is the skier's speed at the bottom?



Solution:

The conservation of energy means that the energy of the skier is same at the beginning and at the end.

Start: $E = mgL$

End: $E = \frac{1}{2}mv^2$

$\Rightarrow mgL = \frac{1}{2}mv^2$

$\Rightarrow v = \sqrt{2gL} = \sqrt{2 \times 9.81 \times 80} = 39.6\ m.s^{-1}$

Try to reason about the question step by step.

Did the student get the correct answer? If the solution is wrong, please explain what is wrong and solve the problem.

Make sure to use LaTeX for math and round off the final answer to two decimal places.

Response

The student did not get the correct answer. The student made a mistake in the calculation of the potential energy at the start of the slope. The potential energy at the start of the slope is given by *mgh*, where *m* is the mass of the skier, *g* is the acceleration due to gravity, and *h* is the height of the slope. The student used the length of the slope, *L*, instead of the height, *H*, in the calculation of the potential energy.

The correct solution is:

The conservation of energy means that the total energy of the skier is the same at the beginning and at the end of the slope.

At the start of the slope, the skier has only potential energy, which is given by *mgh*.

At the end of the slope, the skier has only kinetic energy, which is given by $\frac{1}{2}mv^2$

Therefore, we can write:

$$mgh = \frac{1}{2}mv^2$$

Canceling out the mass *m* on both sides, we get:

$$gh = \frac{1}{2}v^2$$

Solving for *v*, we get:

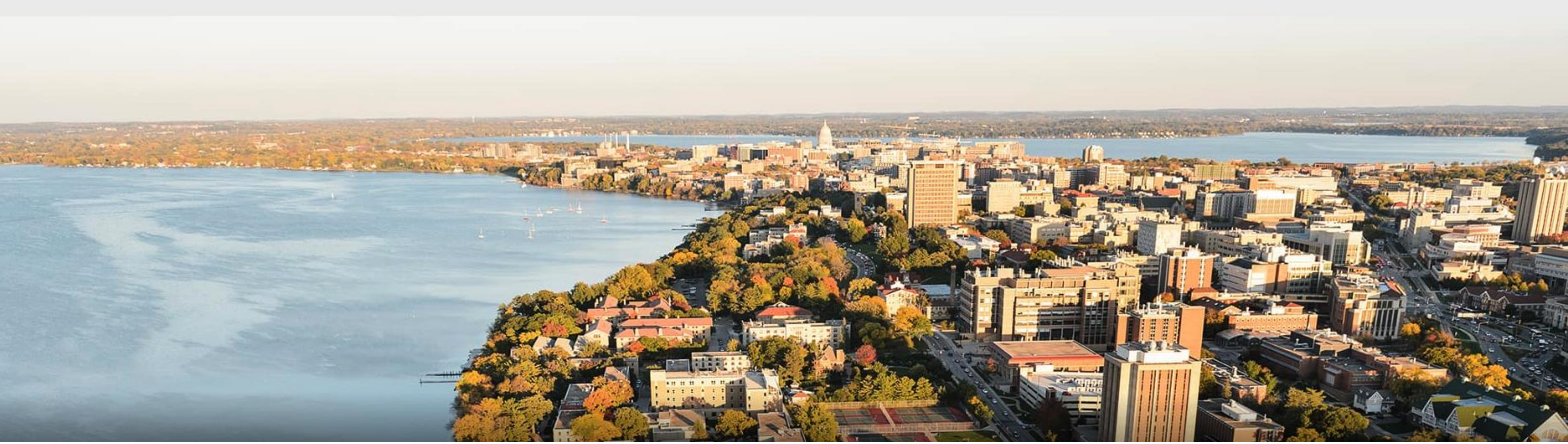
$$v = \sqrt{2gh}$$

Substituting the given values, we get:

$$v = \sqrt{2 \times 9.81 \times 40} = 28.01 \text{ m/s}$$

Therefore, the skier's speed at the bottom of the slope is 28.01 m/s.

Gemini Team Google:



Break & Questions

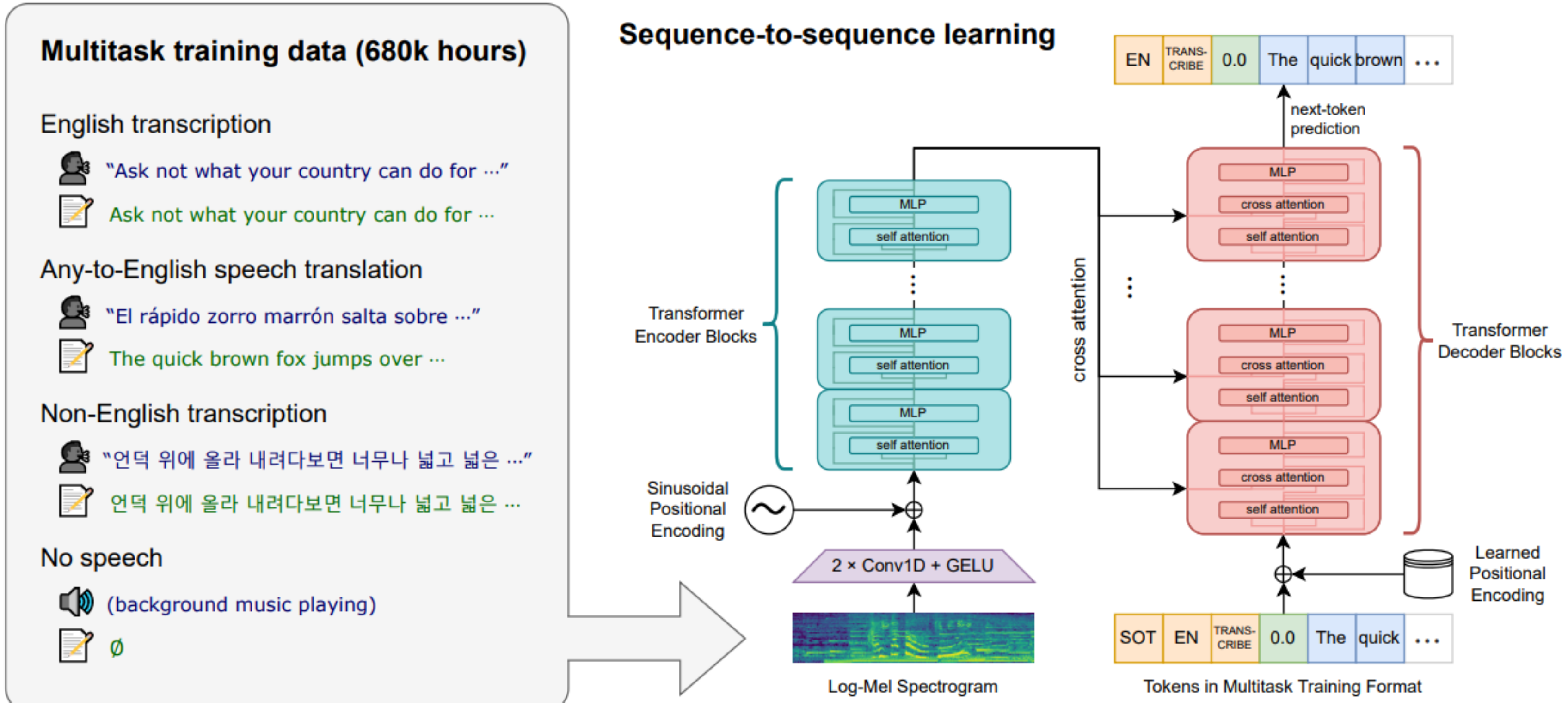
Outline

- **Multimodal Models Intro + One-Encoder Models**
  - Short history, adapting models to incorporate multiple modalities, BERT-like vision-language models, ViTs
- **VLM Variations and Types**
  - Multi-encoder setups, contrastive training, CLIP, joint training, few-shot models, visual instructions
- **Other Modalities and Domains**
  - Audio, video, code generation, RL

Other Modalities: Audio

Can do similar things with all sorts of other modalities

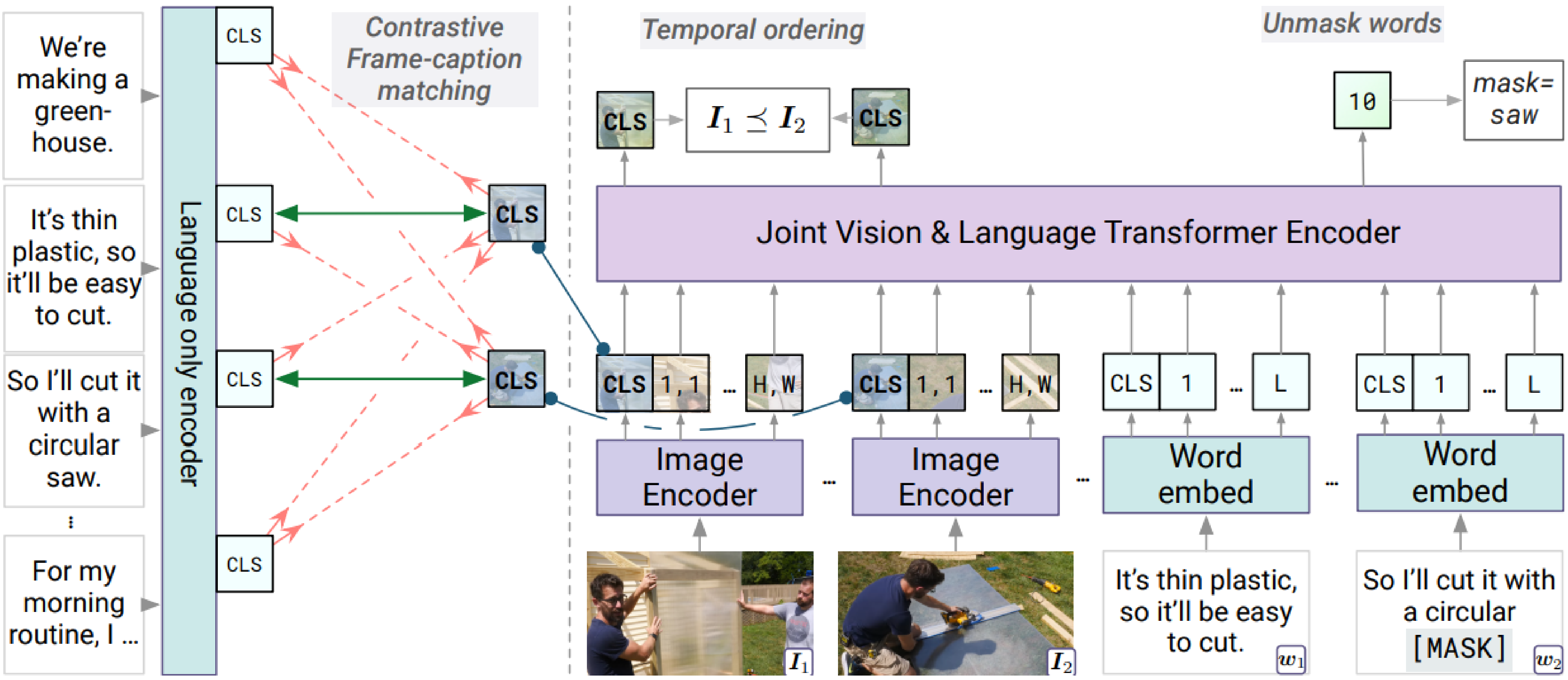
- Audio: can always convert to image and apply directly
- **Ex: Whisper.** 680K hours of audio supervision



Radford et al '22

Other Modalities: **Audio + Video + Text**

Merlot: video + text



Zellers et al '21

Code Models: **Codex**

Warning: for historical reasons. Code out-of-the-box now.

Start with GPT-3 and fine-tune on large-scale code.

- Data: “0 from 54 million public software repositories hosted on GitHub, containing 179 GB of … Python files under 1 MB. “
- Plus pre-processing. Filter out
  - High-chance of autogenerated
  - Long average line length
- ~160GB of data.
- **Eval**: pass @ k
  - k samples per prob, correct if any pass

	PASS@k		
	$k = 1$	$k = 10$	$k = 100$
GPT-NEO 125M	0.75%	1.88%	2.97%
GPT-NEO 1.3B	4.79%	7.47%	16.30%
GPT-NEO 2.7B	6.41%	11.27%	21.37%
GPT-J 6B	11.62%	15.74%	27.74%
TABNINE	2.58%	4.35%	7.59%
CODEX-12M	2.00%	3.62%	8.58%
CODEX-25M	3.21%	7.1%	12.89%
CODEX-42M	5.06%	8.8%	15.55%
CODEX-85M	8.22%	12.81%	22.4%
CODEX-300M	13.17%	20.37%	36.27%
CODEX-679M	16.22%	25.7%	40.95%
CODEX-2.5B	21.36%	35.42%	59.5%
CODEX-12B	28.81%	46.81%	72.31%

Chen et al ‘21

Code Models: **StarCoder**

Codex (and descendants) are not open source.

Lots of open variants. Trained on open dataset: “The Stack”

- “From the 358 programming languages… we selected 86 languages”
- 15B model
- 1T tokens for pretraining
- 35B Python tokens
  for fine-tuning

Model	HumanEval	MBPP
LLaMA-7B	10.5	17.7
LaMDA-137B	14.0	14.8
LLaMA-13B	15.8	22.0
CodeGen-16B-Multi	18.3	20.9
LLaMA-33B	21.7	30.2
CodeGeeX	22.9	24.4
LLaMA-65B	23.7	37.7
PaLM-540B	26.2	36.8
CodeGen-16B-Mono	29.3	35.3
StarCoderBase	30.4	49.0
code-cushman-001	33.5	45.9
StarCoder	33.6	**52.7**
StarCoder-Prompted	**40.8**	49.5

Foundation Models in Robotics

Can use language models for planning/robotics, but

- Not “grounded” since not aware of the environment
- Can mix together with RL concepts

Ahn et al ‘22

Foundation Models in Robotics: **SayCan**

Can use language models for planning/robotics, but

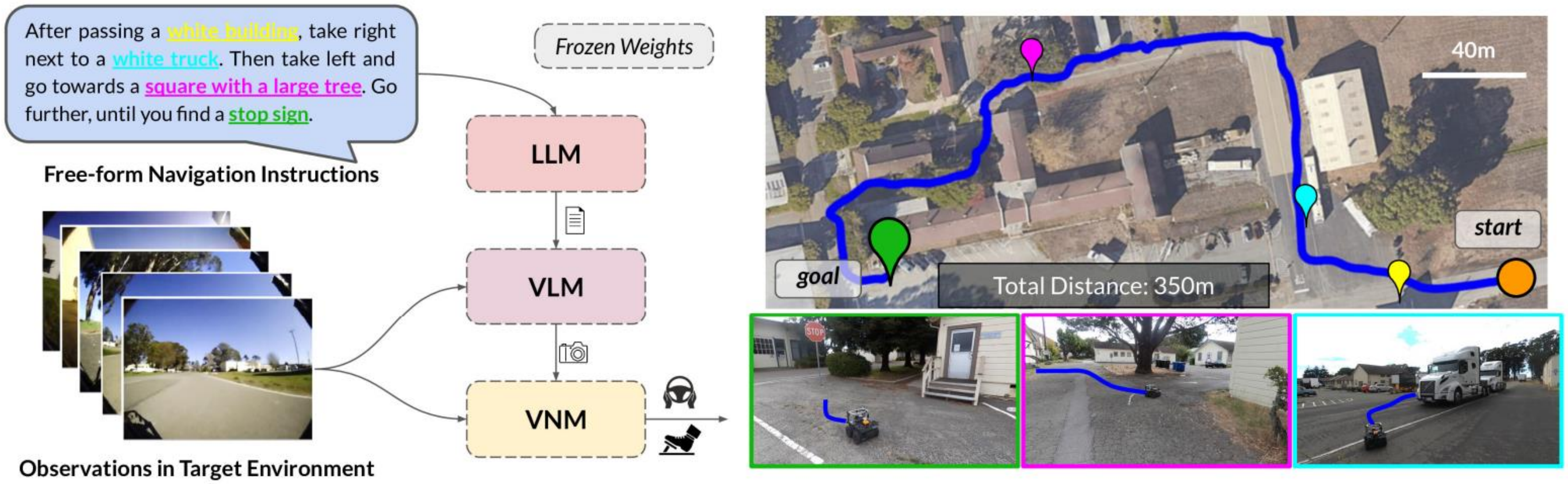
- Not "grounded" since not aware of the environment
- Can mix together with RL concepts
- Basic idea (Ahn et al '22)

$$\pi \;=\; \arg\max_{\pi \in \Pi} \underbrace{p(c_\pi | s, \ell_\pi)}_{\text{Prob. of completing skill/step from state s}} \underbrace{p(\ell_\pi | i)}_{\text{LLM-provided prob of next step being valid}}$$

Foundation Models in Robotics: **Navigation**

For navigation:

- Connect multiple FMs (language, vision, action)
- **Inputs**: observations, instructions
- **Output**: plan

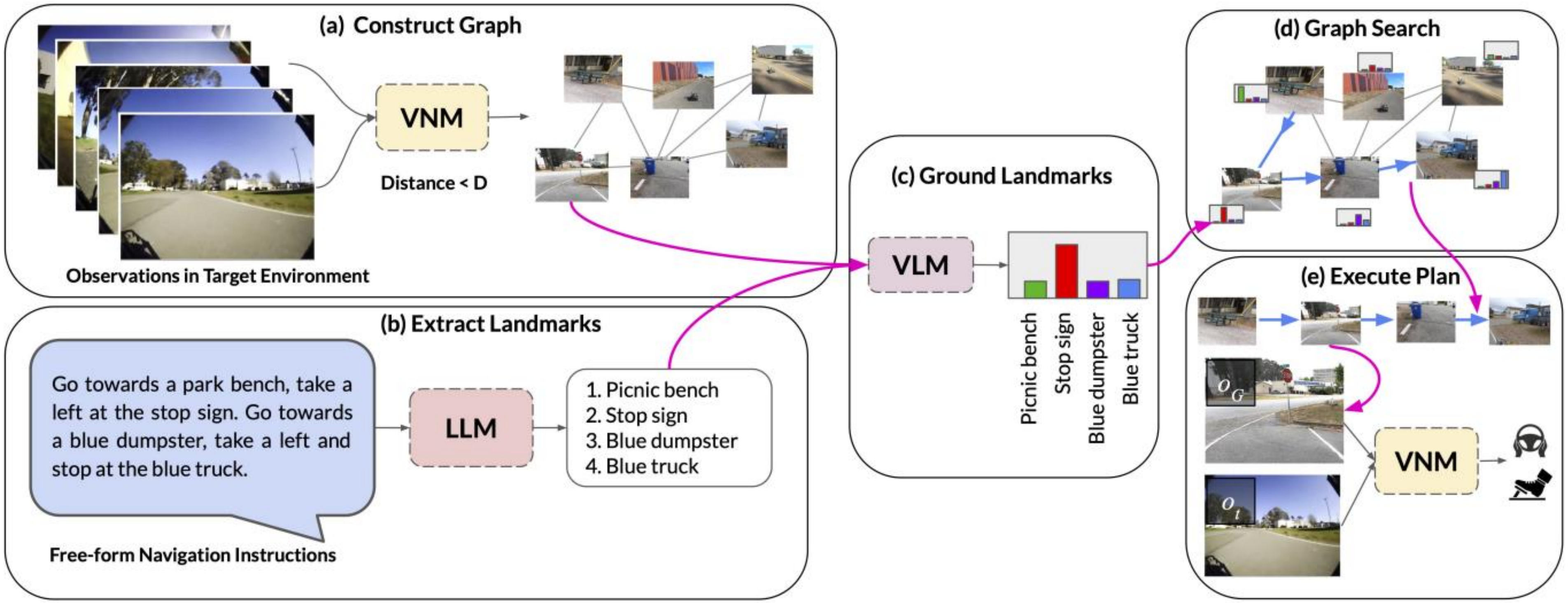


Shah et al ‘22

Foundation Models in Robotics: **Navigation**

For navigation:

- Connect multiple FMs (language, vision, action)



Shah et al '22

Bibliography

- Weston, Bengio, Usunier '11: Jason Weston, Samy Bengio, Nicolas Usunier, "Wsabie: Scaling Up To Large Vocabulary Image Annotation" (https://research.google/pubs/pub37180/)
- Li et al '19: Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, Kai-Wei Chang, "VisualBERT: A Simple and Performant Baseline for Vision and Language" (https://arxiv.org/abs/1908.03557)
- Dosovitskiy et al '20: Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" (https://arxiv.org/abs/2010.11929)
- Kim et al '21: Wonjae Kim, Bokyung Son, Ildoo Kim, "ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision" (https://arxiv.org/abs/2102.03334)
- Du et al '22: Yifan Du, Zikang Liu, Junyi Li, Wayne Xin Zhao, "A Survey of Vision-Language Pre-Trained Models" https://arxiv.org/abs/2202.10936)
- Zhang et al '23: Jingyi Zhang, Jiaxing Huang, Sheng Jin, Shijian Lu, "Vision-Language Models for Vision Tasks: A Survey" (https://arxiv.org/abs/2304.00685)
- Singh et al '22: Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, Douwe Kiela, "FLAVA: A Foundational Language And Vision Alignment Model" (https://arxiv.org/pdf/2112.04482.pdf)
- Alayrac et al '22: Jean-Baptiste Alayrac and others, "Flamingo: a Visual Language Model for Few-Shot Learning" (https://arxiv.org/abs/2204.14198)
- Liu et al ‘23: Haotian Liu, Chunyuan Li, Qingyang Wu, Yong Jae Lee, “Visual Instruction Tuning” (https://arxiv.org/abs/2304.08485)
- Gemini Team Google ‘23, Gemini Team Google, “Gemini: A Family of Highly Capable Multimodal Models” (https://arxiv.org/abs/2312.11805)
- Radford et al '22: Alec Radford and others, "Robust Speech Recognition via Large-Scale Weak Supervision" (https://cdn.openai.com/papers/whisper.pdf)
- Zellers et al '21: Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, Yejin Choi, "MERLOT: Multimodal Neural Script Knowledge Models" (https://arxiv.org/abs/2106.02636)
- Ahn et al '21: Michael Ahn and others, "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances" (https://arxiv.org/pdf/2204.01691.pdf)
- Shah et al ‘22: Dhruv Shah, Błazej Osinski, Brian Ichter, Sergey Levine “LM-Nav: Robotic Navigation with Large Pre-Trained Models of Language, Vision, and Action” (https://proceedings.mlr.press/v205/shah23b/shah23b.pdf)

Thank You!