



# CS 639: Foundation Models **Multimodal Models II**

Fred Sala

University of Wisconsin-Madison

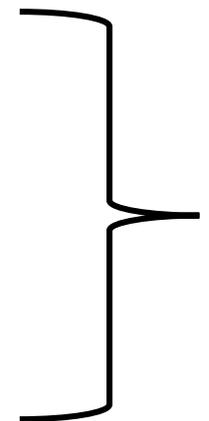
**March 3, 2026**



# Announcements

- **Midterm: March 11, 5:40 pm - 7:20 pm**
  - Sample problems out **tonight**
- **Homework 2: due in a week**
- **Project information: released**  
[https://pages.cs.wisc.edu/~fredsala/cs639/files/project\\_info\\_639.pdf](https://pages.cs.wisc.edu/~fredsala/cs639/files/project_info_639.pdf)
- **Class roadmap:**

Thursday Feb. 26	Multimodal Architectures I
Tuesday March 3	Multimodal Architectures II
Thursday March 5	Prompting, ICL, and Others
Tuesday March 10	Specialization I



# Outline

- **Multimodal Models Intro + One-Encoder Models**
  - 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, RL

# Outline

- **Multimodal Models Intro (from last time)**

- 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, RL

# Making LLMs Multimodal

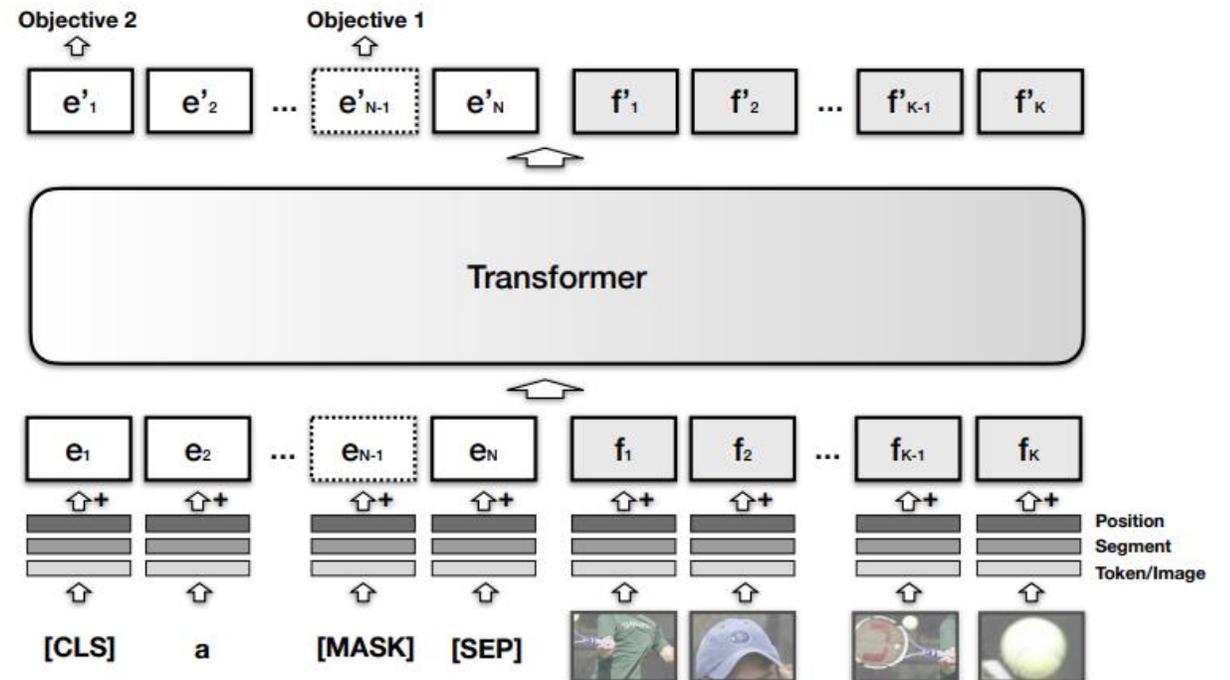
How do we use a language architecture for multiple modalities?

**VisualBERT:** take all the ideas from BERT, add images

- Use bounding boxes from image detector + image embedder



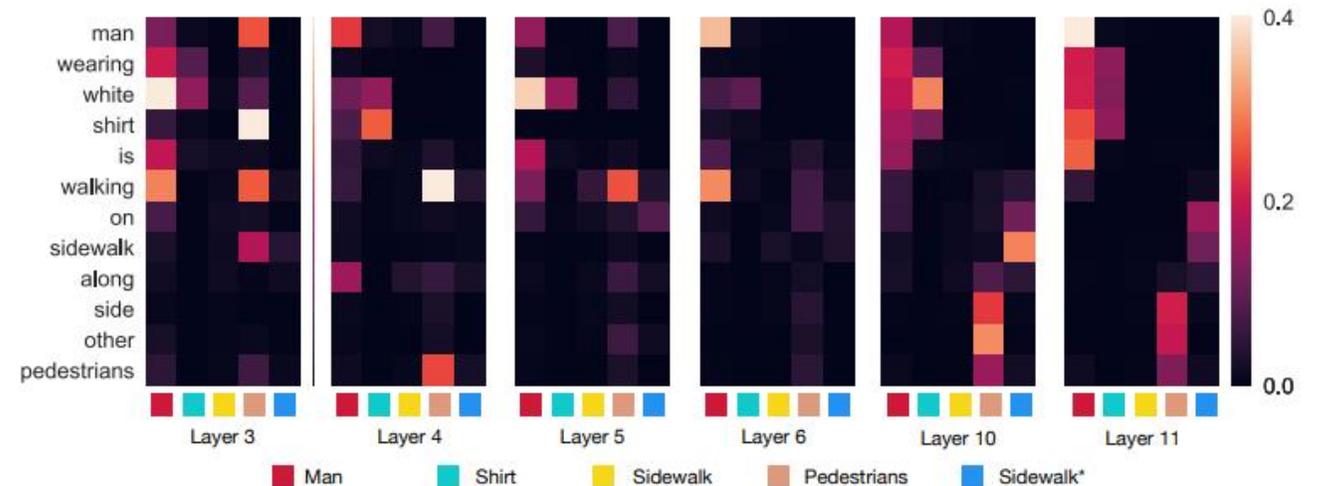
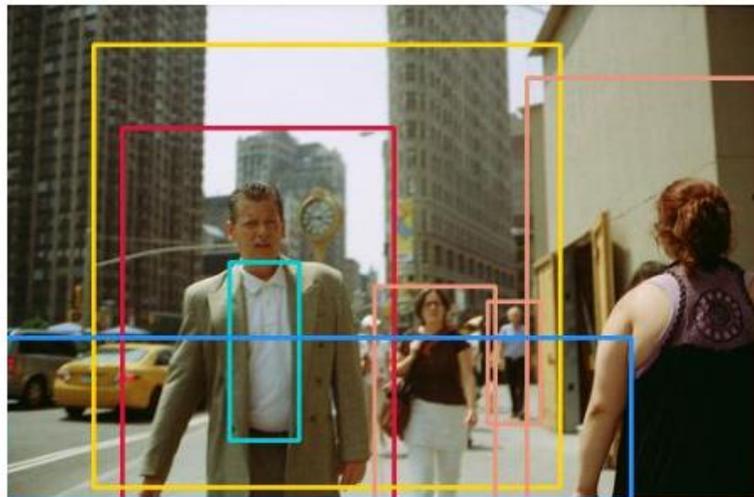
A person hits a ball with a tennis racket  
Li et al '19



# 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)



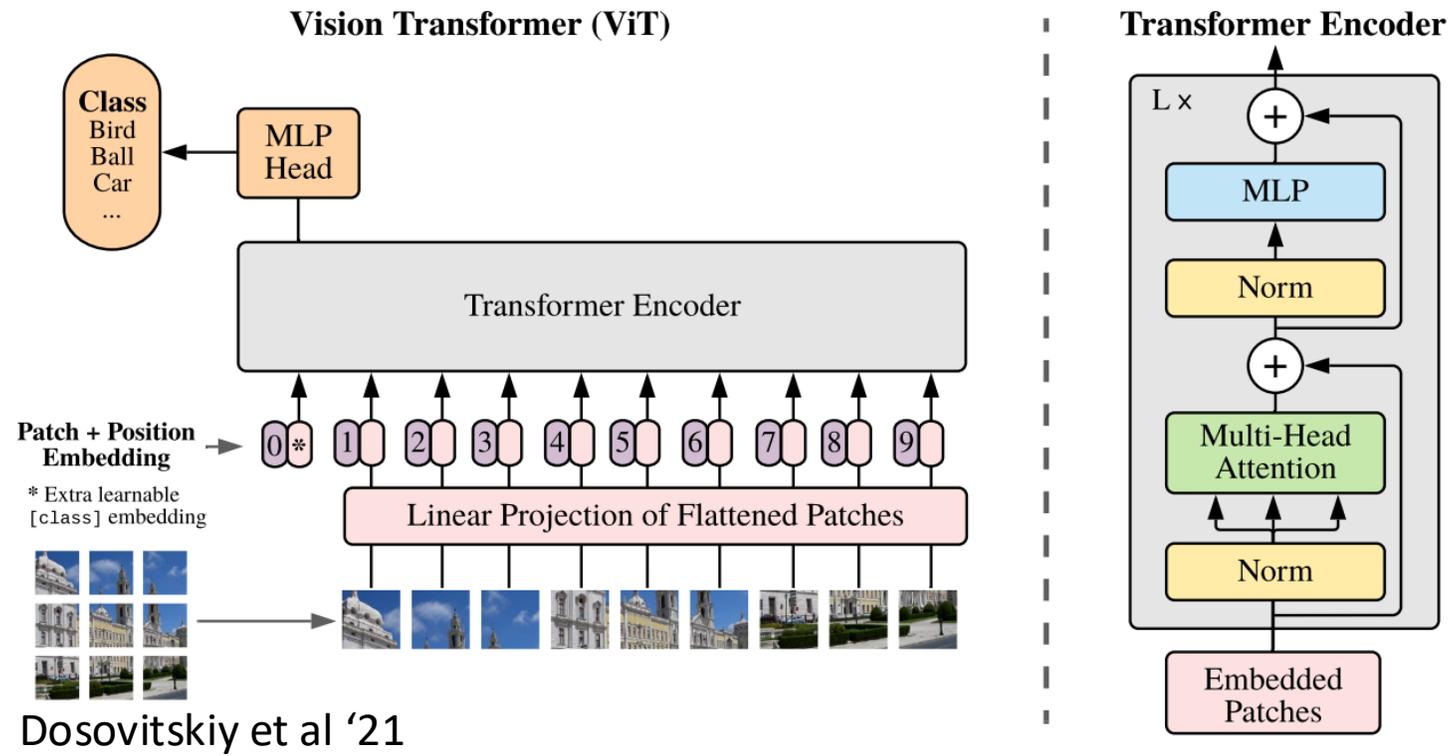
# How Do We Get Image Embeddings?

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.

Could always use Resnets, etc., but...

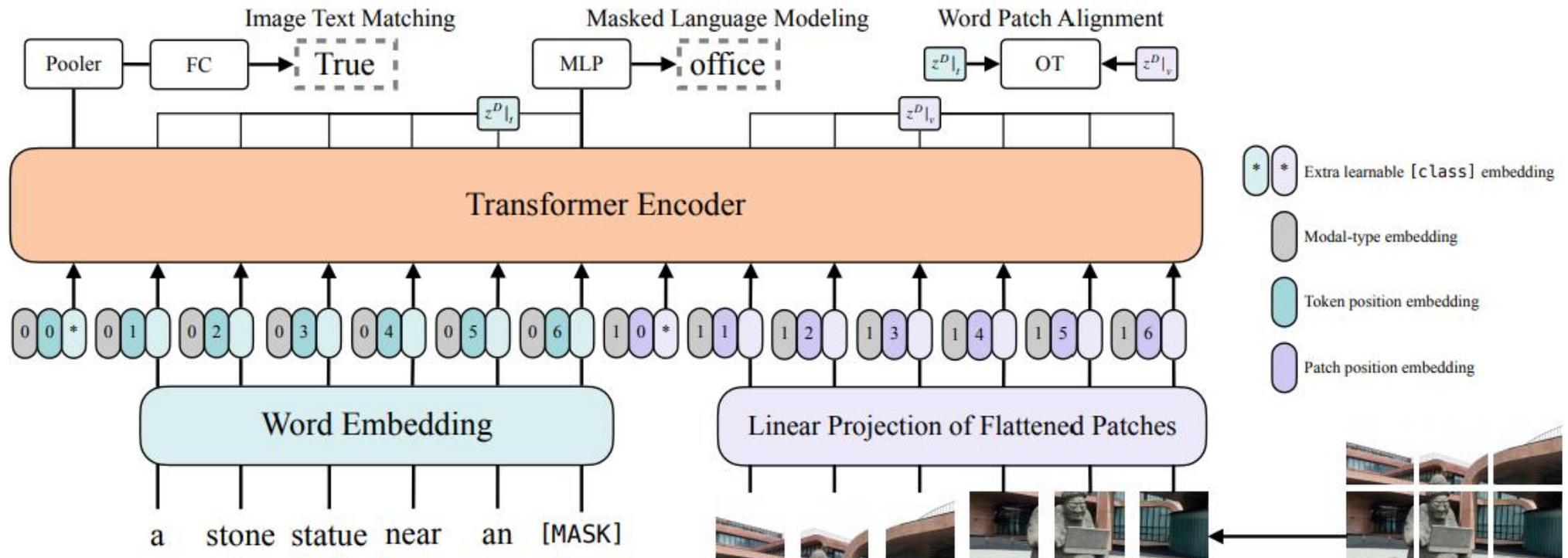
- Didn't Transformers make a big difference for text?
- Can also use for vision: **ViT**. Just use patches!



# 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
<b>Fusion Encoder</b>					
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 + EfficientNet	Dual stream	CMCL	RUC-CAS-WenLan
ViLT [2021]	ViT	Linear Projection	Single stream	MLM+ITM	CC+COCO+VG+SBU
<b>Dual Encoder</b>					
CLIP [2021]	GPT2	ViT, ResNet		CMCL	self-collected
ALIGN [2021]	BERT	EfficientNet		CMCL	self-collected
DeCLIP [2021b]	GPT2, BERT	ViT, ResNet, RegNetY-64GF		CMCL+MLM+CL	CC+self-collected
<b>Fusion Encoder+ Dual Encoder</b>					
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	✗



# Break & Questions

# Outline

- **Multimodal Models Intro + One-Encoder Models**
  - 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, 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**...

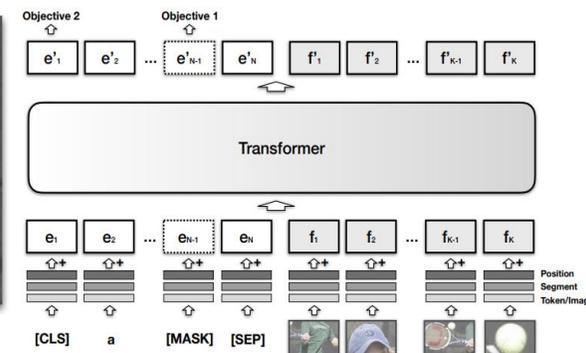
- I.e., produce two representations separately, then produce some means of connecting/tying them together

- **Contrastive** approach



A person hits a ball with a tennis racket

Li et al '19



# VLMs: Contrastive Training

Training approach: contrastive

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

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

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

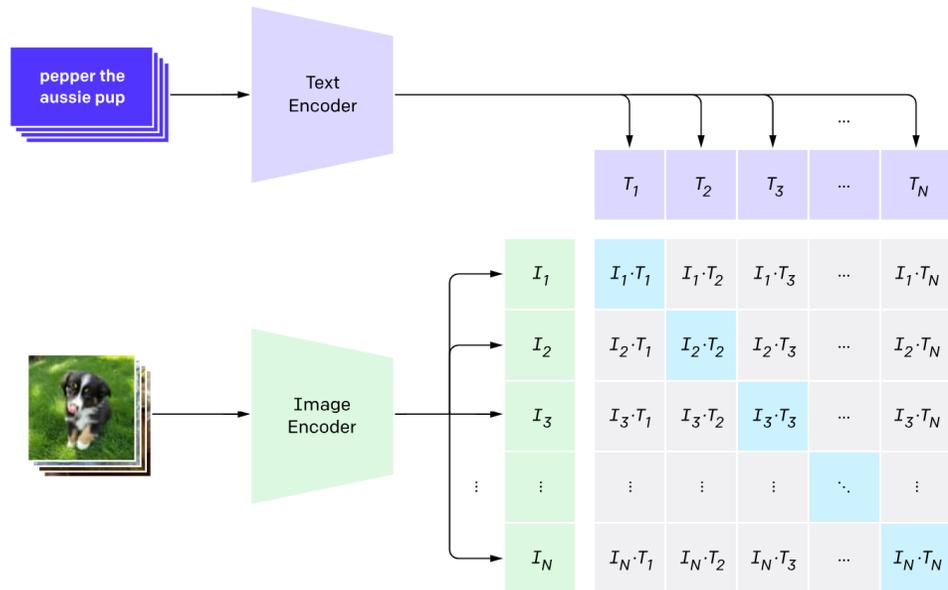
$$\mathcal{L}_{I \rightarrow T} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(z_i^I \cdot z_i^T / \tau)}{\sum_{j=1}^B \exp(z_i^I \cdot z_j^T / \tau)}$$

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

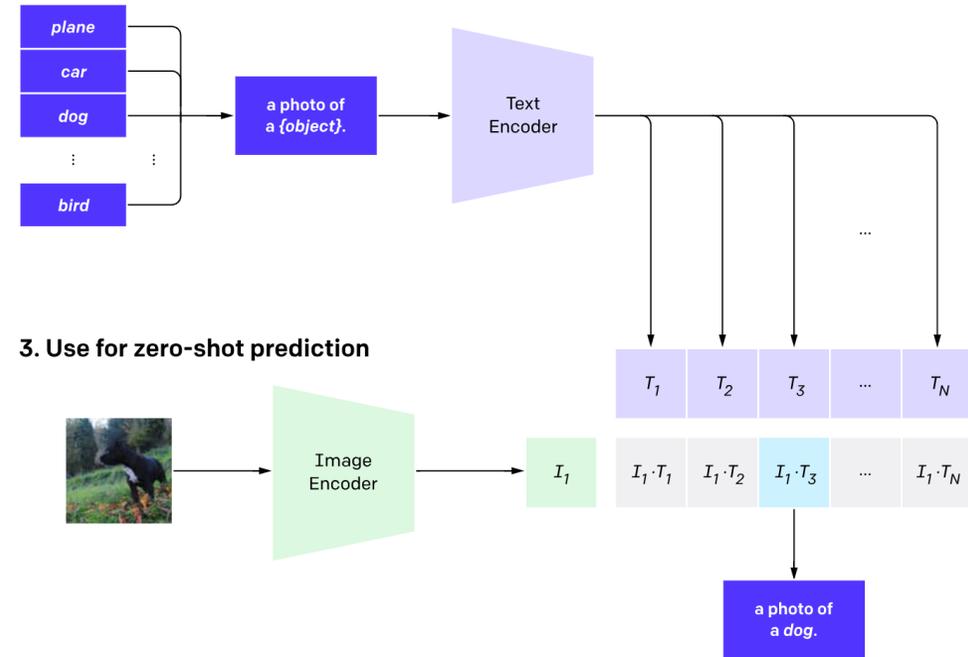
# VLMs: CLIP

## A simple but easily scalable contrastive VLM

### 1. Contrastive pre-training



### 2. Create dataset classifier from label text



### 3. Use for zero-shot prediction

# How to use CLIP?

Standard way: use pre-defined templates

- E.g., “a photo of a [X]”

**SUN397**  
television studio (90.2%) Ranked 1 out of 397 labels

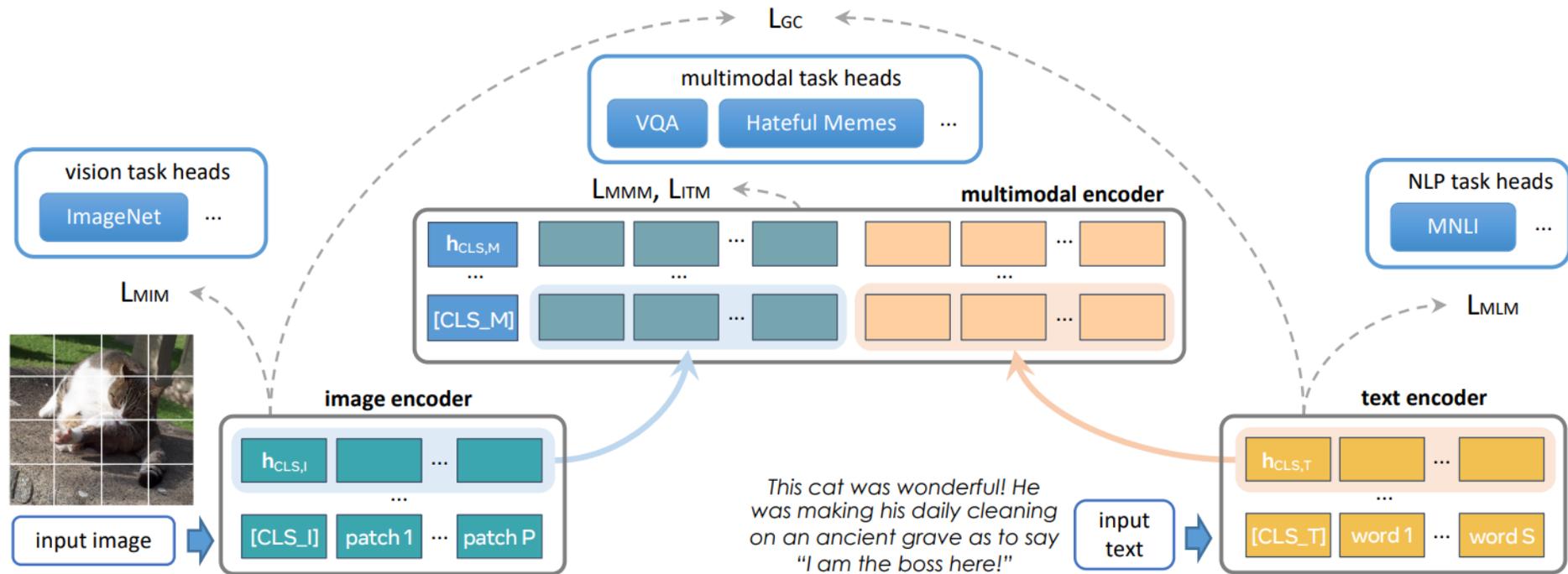


- ✓ a photo of a **television studio**.
- ✗ a photo of a **podium indoor**.
- ✗ a photo of a **conference room**.
- ✗ a photo of a **lecture room**.
- ✗ a photo of a **control room**.

# VLMs: FLAVA

## Foundational Language And Vision Alignment Model (FLAVA)

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



# 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?



# Few-Shot VLMs: Flamingo

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

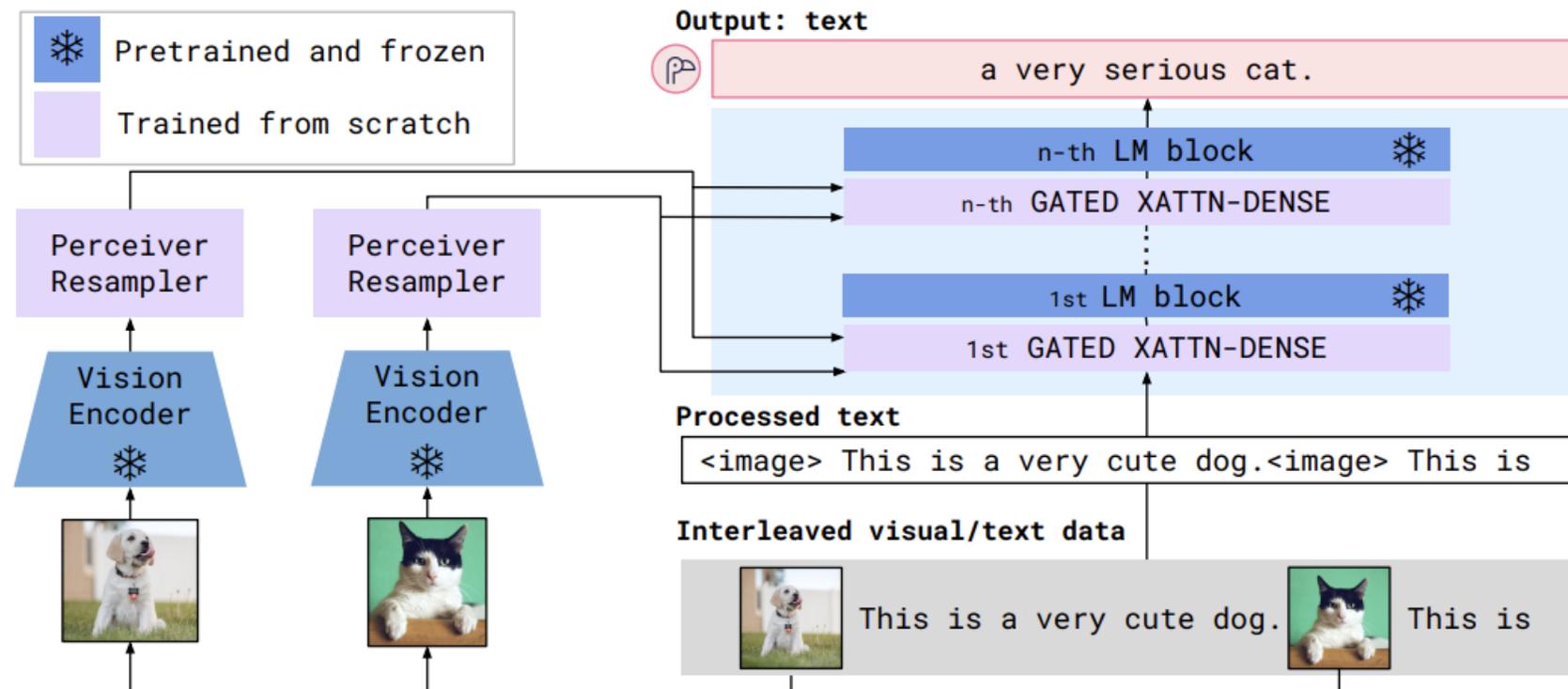
- Multi-image!
- More complex interleaved architecture

 <p>This is a picture of two teddy bears on the moon.</p> <p>What are they doing?</p> <p>They are having a conversation.</p> <p>What object are they using?</p> <p>It looks like a computer.</p> <p>Is this surprising?</p> <p>Yes, it is surprising.</p> <p>Why is this picture surprising to you?</p> <p>I think it is surprising because teddy bears are not usually found on the moon.</p>	 <p>What is the common thing about these three images?</p> <p>They are all flamingos.</p> <p>What is the difference between these three images?</p> <p>The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.</p>	 <p>This is an apple with a sticker on it.</p> <p>What does the sticker say?</p> <p>The sticker says "iPod".</p> <p>Where is the photo taken?</p> <p>It looks like it's taken in a backyard.</p> <p>Do you think it is printed or handwritten?</p> <p>It looks like it's handwritten.</p> <p>What color is the sticker?</p> <p>It's white.</p>	 <p>This is a cityscape. It looks like Chicago.</p> <p>What makes you think this is Chicago?</p> <p>I think it's Chicago because of the Shedd Aquarium in the background.</p>  <p>What about this one? Which city is this and what famous landmark helped you recognise the city?</p> <p>This is Tokyo. I think it's Tokyo because of the Tokyo Tower.</p>
---	--	---	--

# Few-Shot VLMs: Flamingo

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

- Multi-image!
- More complex interleaved architecture

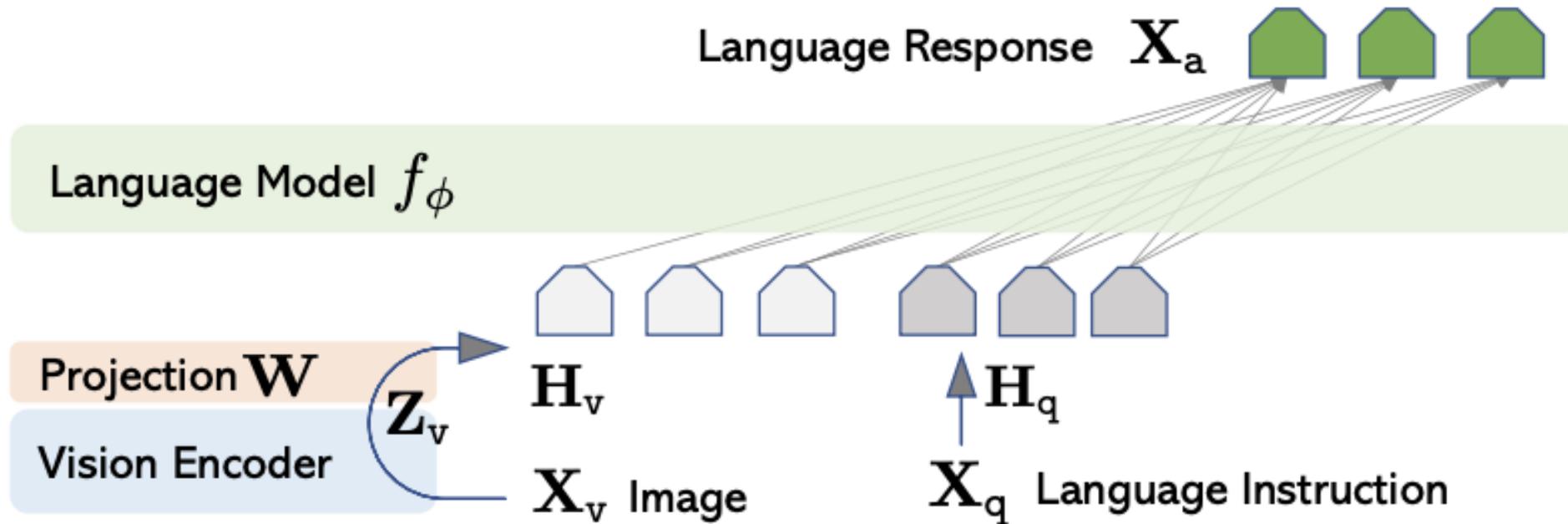


# Visual Instructions: Llava

Visual instruction tuning

**Goal:** achieve close to GPT-4's multimodal performance

- Simple, scalable architecture:



# 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  
LLaVA

What is unusual about this image?

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  
LLaVA

What's happening in the scene?

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  
GPT-4 [36]

What is unusual about this image?

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  
BLIP-2

What is unusual about this image?

a man is sitting on the back of a yellow cab

User  
OpenFlamingo

What is unusual about this image?

The man is drying his clothes on the hood of his car.

# Closed-source Multimodal Models

Now, back to closed-source models.

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

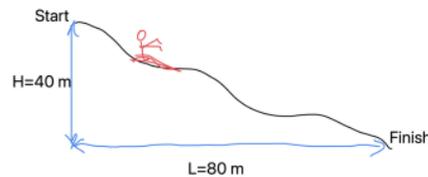
Downside: **expensive!**



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

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

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

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

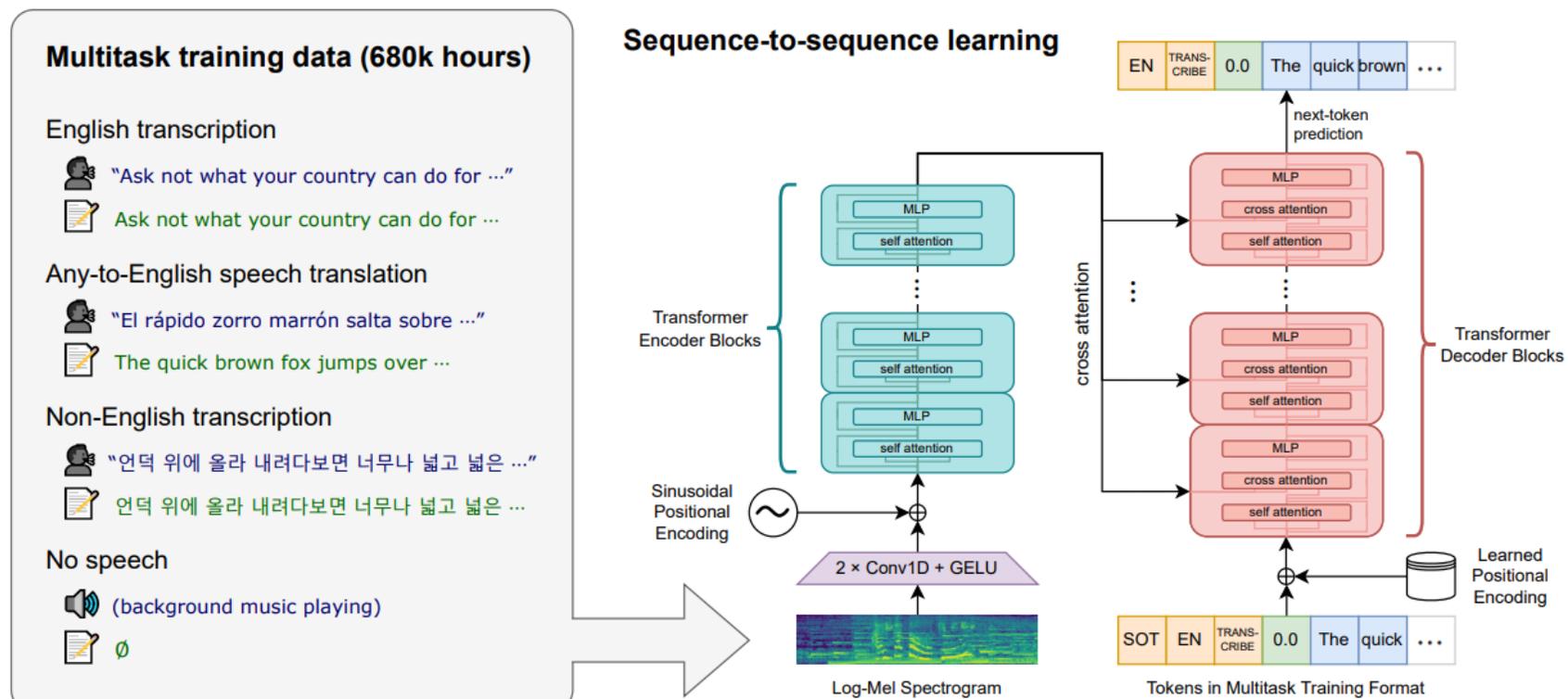
# Outline

- **Multimodal Models Intro + One-Encoder Models**
  - 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, RL

# Other Modalities: Audio

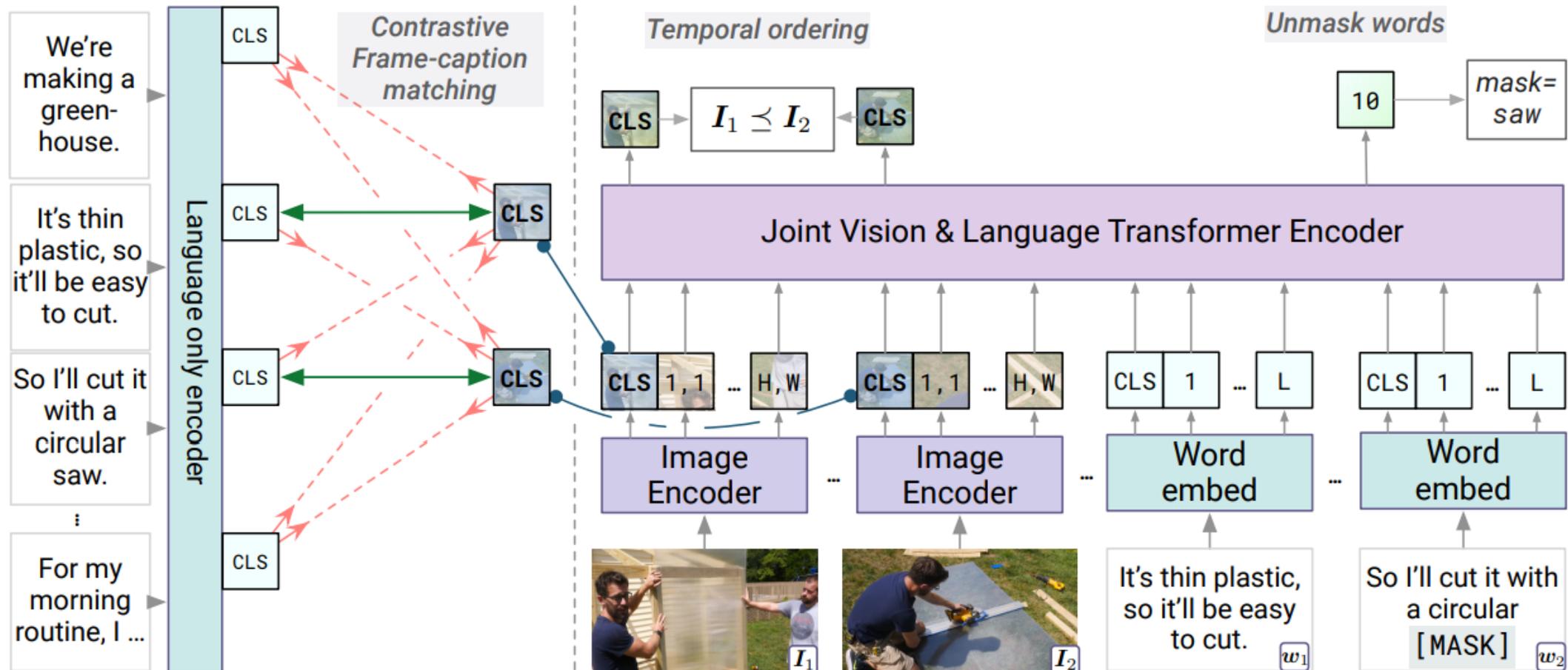
Can do similar things with all sorts of other modalities

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



# Other Modalities: Audio + Video + Text

## Merlot: video + text



# 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



# 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)

$$\pi = \arg \max_{\pi \in \Pi} p(c_{\pi} | s, l_{\pi}) p(l_{\pi} | i)$$



Prob. of completing  
skill/step from state  $s$



LLM-provided  
prob of next  
step being valid

# Foundation Models in Robotics: Navigation

For navigation:

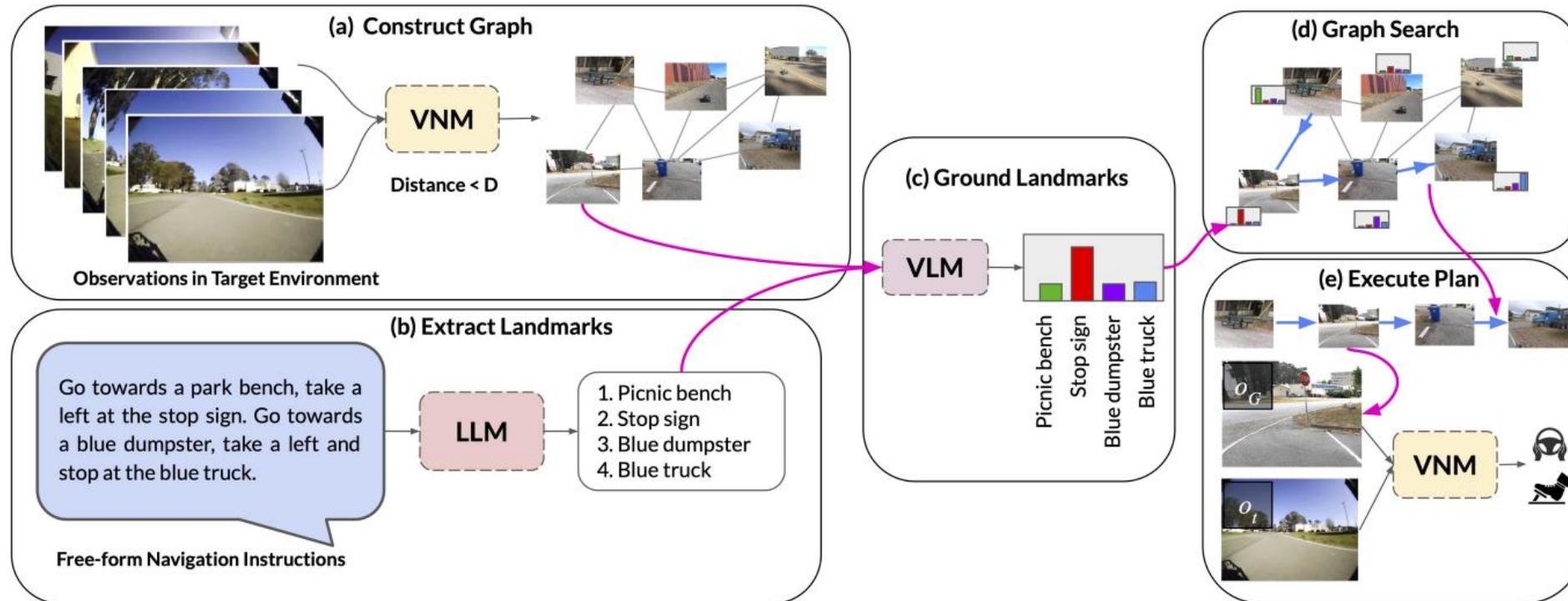
- Connect multiple FMs (language, vision, action)
- **Inputs:** observations, instructions
- **Output:** plan



# Foundation Models in Robotics: Navigation

For navigation:

- Connect multiple FMs (language, vision, action)





**Thank You!**