

# Multimodal Learning

对比学习、多模态生成式模型（语言生成）

高鑫 2023.6.6

# 目录

## CONTENT

01

### ImageBind

Learn a single joint embedding space for all modalities

02

### 华西

Well-designed medical prompts are the key to elicit knowledge from pre-trained VLMs

03

### MMCoT

The first to study CoT reasoning in different modalities

04

### LLaVA

The first attempt to multimodal instruction-following data

# IMAGEBIND: One Embedding Space To Bind Them

2023,5,9

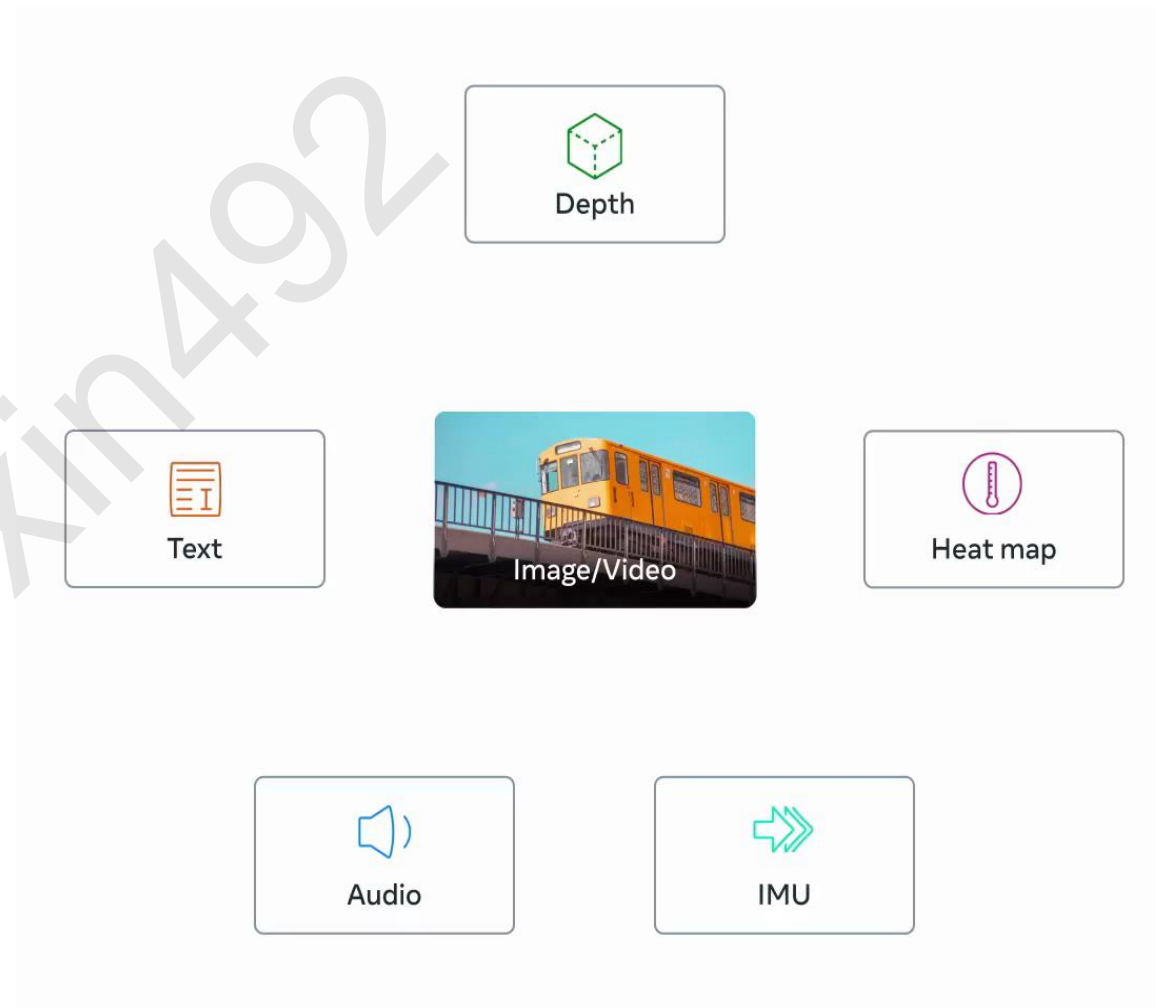
All Rohit Girdhar\*   Alaaeldin El-Nouby\*

Zhuang Liu   Mannat Singh   Kalyan Vasudev Alwala

Armand Joulin   Ishan Misra\*

FAIR, Meta AI

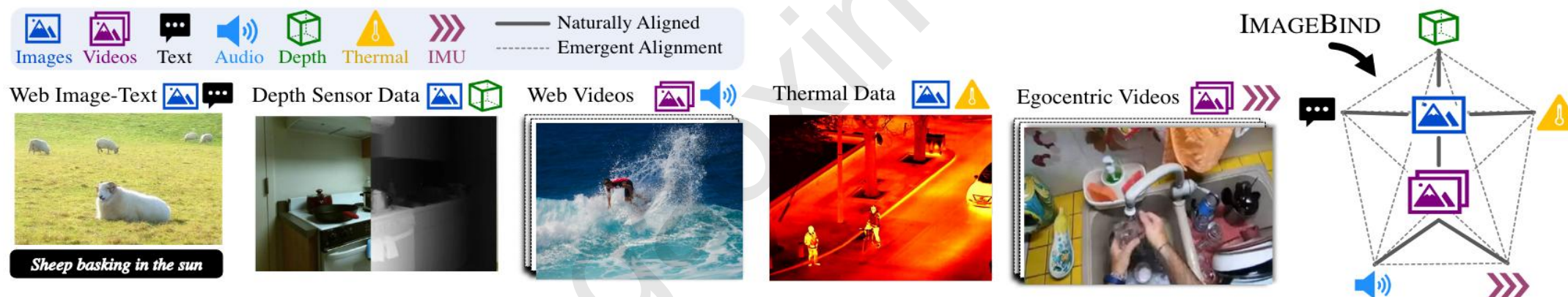
<https://facebookresearch.github.io/ImageBind>



- The 'binding' property of images offers many sources of supervision to learn visual features

**!! the absence of large quantities of multimodal data where all modalities are present together !!**

- IMAGEBIND**: multiple types of image-paired data ==> a single shared representation space



**an emergent alignment across all of the modalities**

IMAGEBIND's emergent zero-shot classification matches or outperforms specialist models trained with direct audio-text supervision on benchmarks

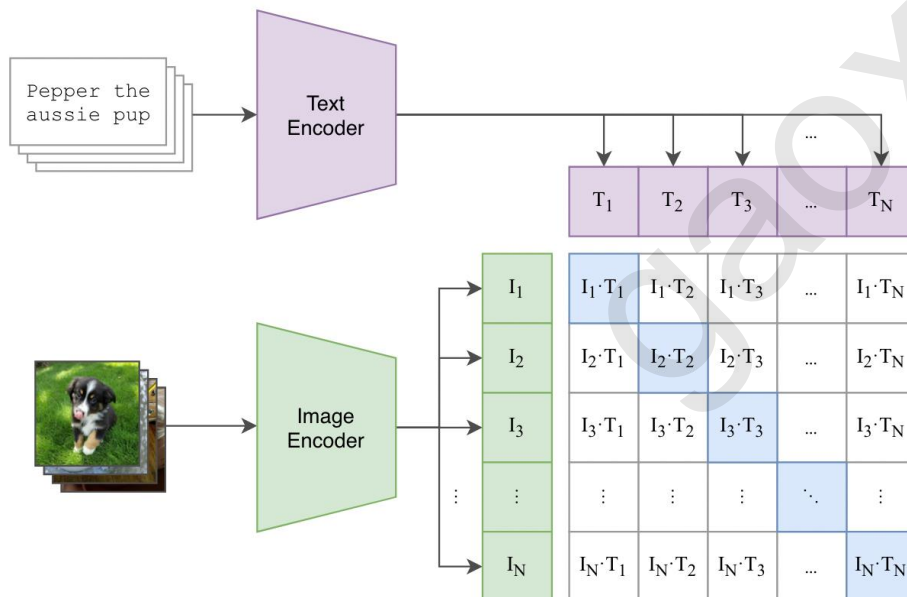
an image  $I_i$ , its corresponding observation in the other modality  $M_i$

encode them:  $q_i = f(I_i)$  and  $k_i = g(M_i)$  ( $f, g$  are deep networks)

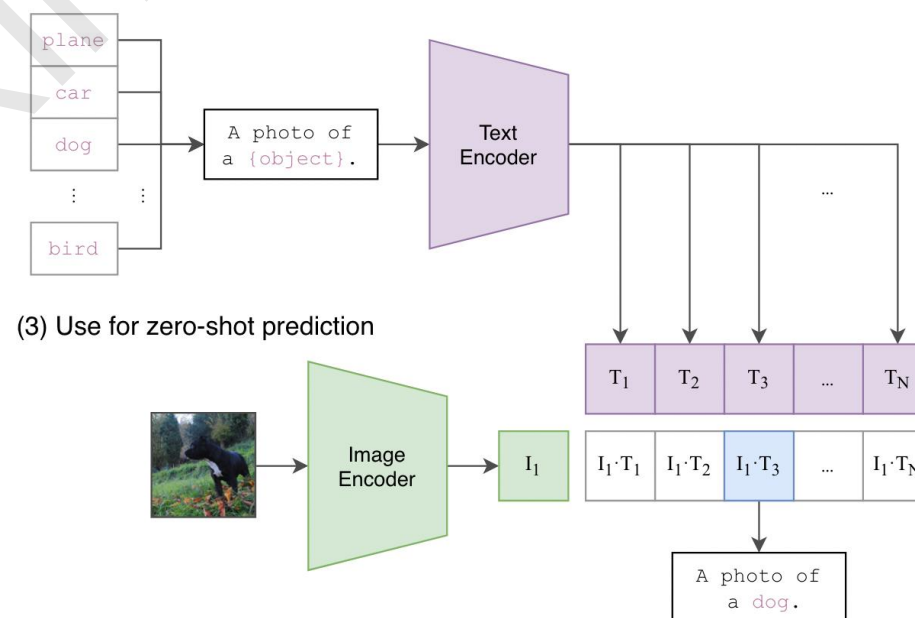
InfoNCE loss:  $L_{\mathcal{I}, \mathcal{M}} = -\log \frac{\exp(\mathbf{q}_i^\top \mathbf{k}_i / \tau)}{\exp(\mathbf{q}_i^\top \mathbf{k}_i / \tau) + \sum_{j \neq i} \exp(\mathbf{q}_i^\top \mathbf{k}_j / \tau)}$ ,  $L_{\mathcal{I}, \mathcal{M}} + L_{\mathcal{M}, \mathcal{I}}$

**an emergent behavior in the embedding space that aligns two pairs of modalities ( $M_1, M_2$ ) (only train using the pairs ( $I, M_1$ ) and ( $I, M_2$ ))**

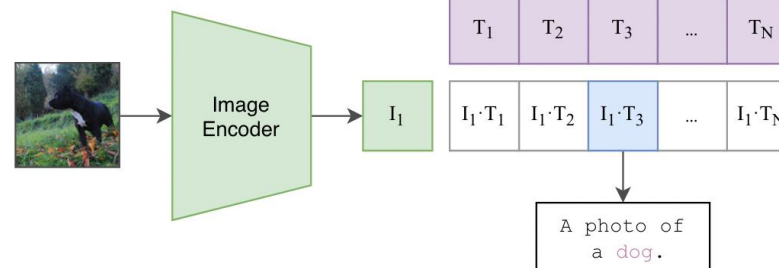
(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction



**Architecture:**

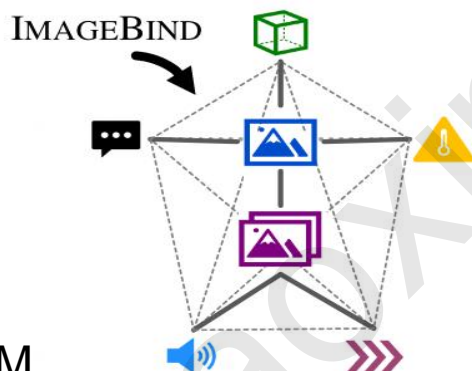
images(videos), text encoder: pretrained OpenCLIP

audio, depth, thermal, IMU: “2D images”, ViT

32GB V100 or 40GB A100

**Training dataset:**

|                |          |               |       |
|----------------|----------|---------------|-------|
| videos&audio   | AudioSet | 2M            | } ×50 |
| images&depth   | SUN      | ~5K           |       |
| images&thermal | LLVIP    | 12025, 3463   |       |
| videos&IMU     | Ego4D    | 510142, 68865 |       |



| Config               | AS                              | SUN        | LLVIP      | Ego4D |
|----------------------|---------------------------------|------------|------------|-------|
| Vision encoder       | ViT-Huge                        |            |            |       |
| embedding dim.       | 768                             | 384        | 768        | 512   |
| number of heads      | 12                              | 8          | 12         | 8     |
| number of layers     | 12                              | 12         | 12         | 6     |
| Optimizer            | AdamW                           |            |            |       |
| Optimizer Momentum   | $\beta_1 = 0.9, \beta_2 = 0.95$ |            |            |       |
| Peak learning rate   | 1.6e-3                          | 1.6e-3     | 5e-4       | 5e-4  |
| Weight decay         | 0.2                             | 0.2        | 0.05       | 0.5   |
| Batch size           | 2048                            | 512        | 512        | 512   |
| Gradient clipping    | 1.0                             | 1.0        | 5.0        | 1.0   |
| Warmup epochs        | 2                               |            |            |       |
| Sample replication   | 1.25                            | 50         | 25         | 1.0   |
| Total epochs         | 64                              | 64         | 64         | 8     |
| Stoch. Depth [28]    | 0.1                             | 0.0        | 0.0        | 0.7   |
| Temperature          | 0.05                            | 0.2        | 0.1        | 0.2   |
| Augmentations:       |                                 |            |            |       |
| RandomResizedCrop    |                                 |            |            |       |
| size                 | -                               | 224px      |            | -     |
| interpolation        | -                               | Bilinear   | Bilinear   | -     |
| RandomHorizontalFlip | -                               | $p = 0.5$  | $p = 0.5$  | -     |
| RandomErase          | -                               | $p = 0.25$ | $p = 0.25$ | -     |
| RandAugment          | -                               | 9/0.5      | 9/0.5      | -     |
| Color Jitter         | -                               | 0.4        | 0.4        | -     |
| Frequency masking    | 12                              | -          | -          | -     |



## IMAGEBIND's joint embedding space enables novel multimodal capabilities

### 1) Cross-Modal Retrieval

#### Audio



Crackle of a Fire

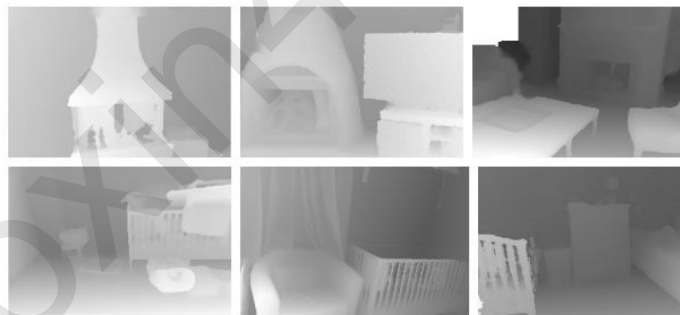


Baby Cooing

#### Images & Videos



#### Depth



#### Text

"A fire crackles while a pan of food is frying on the fire."

"Fire is crackling then wind starts blowing."

"Firewood crackles then music..."

"A baby is crying while a toddler is laughing."

"A baby is laughing while an adult is laughing."

"A baby laughs and something..."

### 2) Embedding-Space Arithmetic



Waves



### 3) Audio to Image Generation



Dog



Engine



Fire









Rain



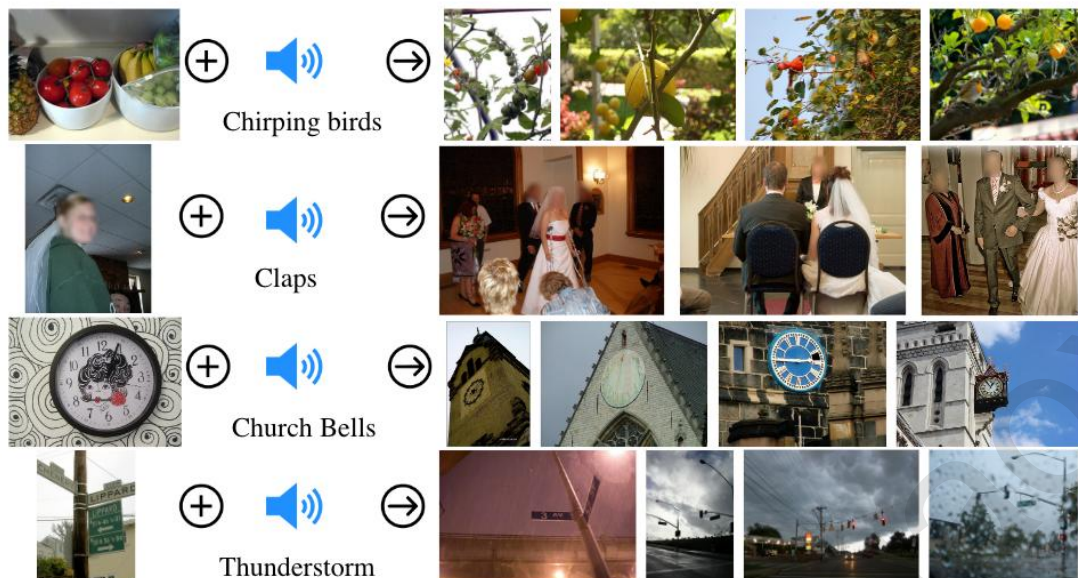
## Emergent zero-shot classification and retrieval

| Dataset                         | Task          | #cls | Metric | #test | Emergent                        | Clotho | AudioCaps | ESC  |      |       |      |
|---------------------------------|---------------|------|--------|-------|---------------------------------|--------|-----------|------|------|-------|------|
|                                 |               |      |        |       |                                 | R@1    | R@10      | R@1  | R@10 | Top-1 |      |
| Audioset Audio-only (AS-A) [18] | Audio cls.    | 527  | mAP    | 19048 | Uses audio and text supervision |        |           |      |      |       |      |
| ESC 5-folds (ESC) [58]          | Audio cls.    | 50   | Acc    | 400   | AudioCLIP [26]                  | ✗      | -         | -    | -    | -     | 68.6 |
| Clotho (Clotho) [16]            | Retrieval     | -    | Recall | 1045  | Uses audio and text loss        |        |           |      |      |       |      |
| AudioCaps (AudioCaps) [36]      | Retrieval     | -    | Recall | 796   | AVFIC [50]                      | ✗      | 3.0       | 17.5 | 8.7  | 37.7  | -    |
| VGGSound (VGGs) [8]             | Audio cls.    | 309  | Acc    | 14073 | No audio and text supervision   |        |           |      |      |       |      |
| SUN Depth-only (SUN-D) [67]     | Scene cls.    | 19   | Acc    | 4660  | IMAGEBIND                       | ✓      | 6.0       | 28.4 | 9.3  | 42.3  | 66.9 |
| NYU-v2 Depth-only (NYU-D) [64]  | Scene cls.    | 10   | Acc    | 653   | Supervised                      |        |           |      |      |       |      |
| LLVIP (LLVIP) [31]              | Person cls.   | 2    | Acc    | 15809 | AVFIC finetuned [50]            | ✗      | 8.4       | 38.6 | -    | -     | -    |
| Ego4D (Ego4D) [22]              | Scenario cls. | 108  | Acc    | 68865 | ARNLQ [52]                      | ✗      | 12.6      | 45.4 | 24.3 | 72.1  | -    |

|               |  |           |  |           |  |           |  |           |  |       |  |
|---------------|---|-----------|--|-----------|---|-----------|---|-----------|---|-------|---|
|               | IN1K  | P365      | K400   | MSR-VTT   | NYU-D   | SUN-D     | AS-A  | 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 <sup>†</sup> [26]  | -         | 68.6 <sup>†</sup> [26]  | -     | -   |
| Absolute SOTA | 91.0 [80]   | 60.7 [65] | 89.9 [78]  | 57.7 [77] | 76.7 [20]   | 64.9 [20] | 49.6 [38]   | 52.5 [35] | 97.0 [9]  | -     | -   |



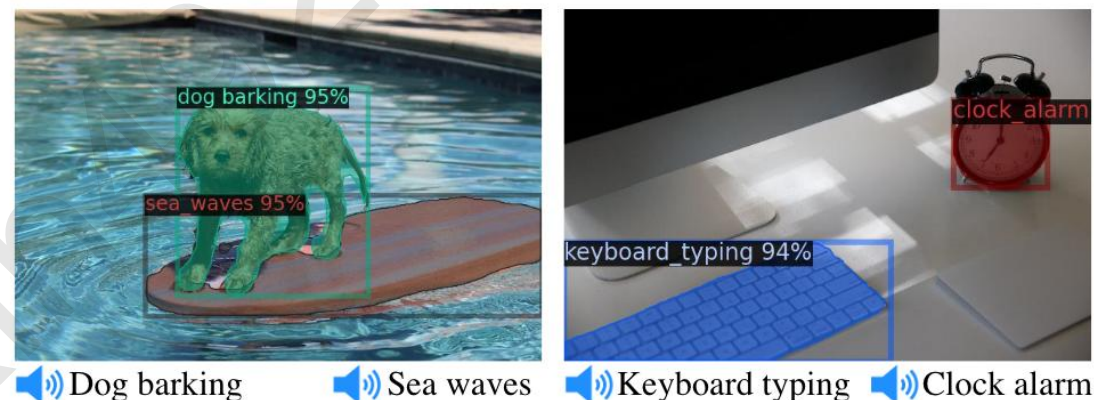
**Embedding space arithmetic:** image retrievals obtained by adding together image and audio embeddings.



### 3) Audio to Image Generation



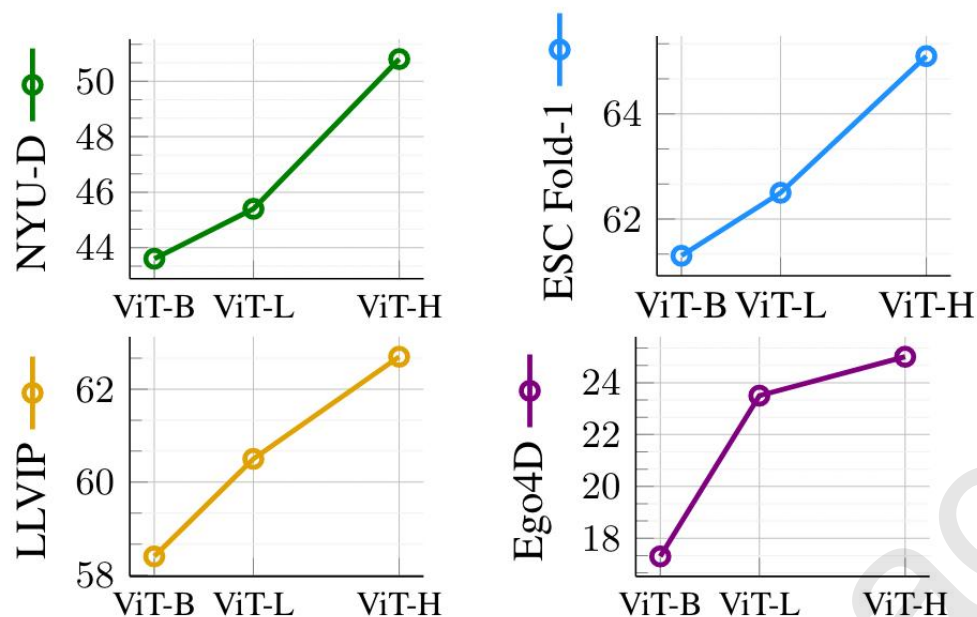
**Without training  
haven't seen (audio, text) pairs**



a pretrained **text-based detection model, Detic**, and simply replace its CLIP-based 'class' (text) embeddings with IMAGEBIND's audio embeddings

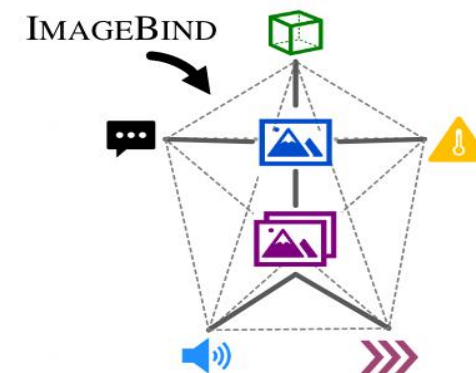
a pretrained **DALLE-2 diffusion model** (private reimplementation) and replace its prompt embeddings by audio embeddings

## Scaling the Image Encoder



**Figure 6. Scaling the image encoder** size while keeping the other modality encoders' size fixed. We measure the performance on the emergent zero-shot classification of depth, audio, thermal, and IMU modalities. Scaling the image encoder significantly improves the zero-shot classification results suggesting that a stronger visual representation improves the 'binding' of modalities.

- No improvement in the ability to extract visual features;
- Excessive reliance on the ability of image encoders;
- Imbalances between different modalities, the dataset sizes vary greatly
- How to further train and enhance the representation ability with other modalities in pairs



# MEDICAL IMAGE UNDERSTANDING WITH PRETRAINED VISION LANGUAGE MODELS: A COMPREHENSIVE STUDY

**Ziyuan Qin<sup>1\*</sup>, Huahui Yi<sup>1\*</sup>, Qicheng Lao<sup>2,3\*†</sup>, Kang Li<sup>1,3†</sup>**

ICLR 2023

<sup>1</sup>West China Biomedical Big Data Center, West China Hospital, Sichuan University

<sup>2</sup>School of Artificial Intelligence, BUPT, Beijing, China

<sup>3</sup>Shanghai Artificial Intelligence Laboratory, Shanghai, China

qicheng.lao@bupt.edu.cn

**Can pre-trained VLMs learned from a large number of natural text-image pairs help with the understanding of medical images?**

**Well-designed medical prompts are the key to elicit knowledge from pre-trained VLMs**



## domain gap between medical images and natural images

With the help of well-designed text prompts, the model can be equipped with **high-level semantics** describing the characteristic of target objects instead of only providing **object names**.

expressive attributes that are shared between domains: **Color**, **Shape**, **Location**

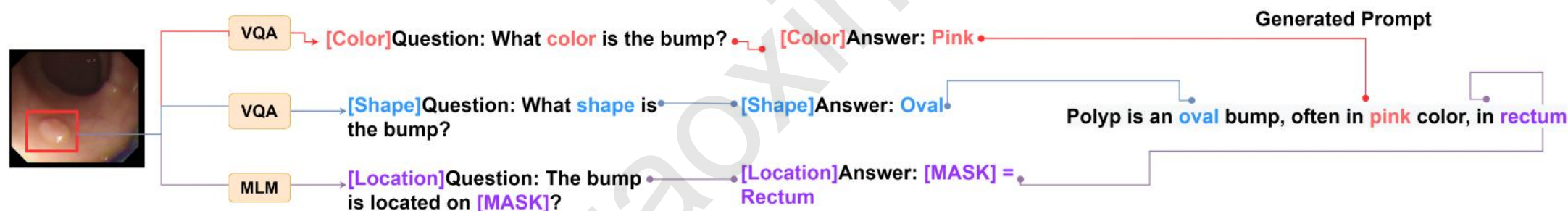


Figure 4: Auto-prompt generation show case.

### Overall goal:

Transfer pre-trained VLMs to the medical field, and use them for downstream object detection.

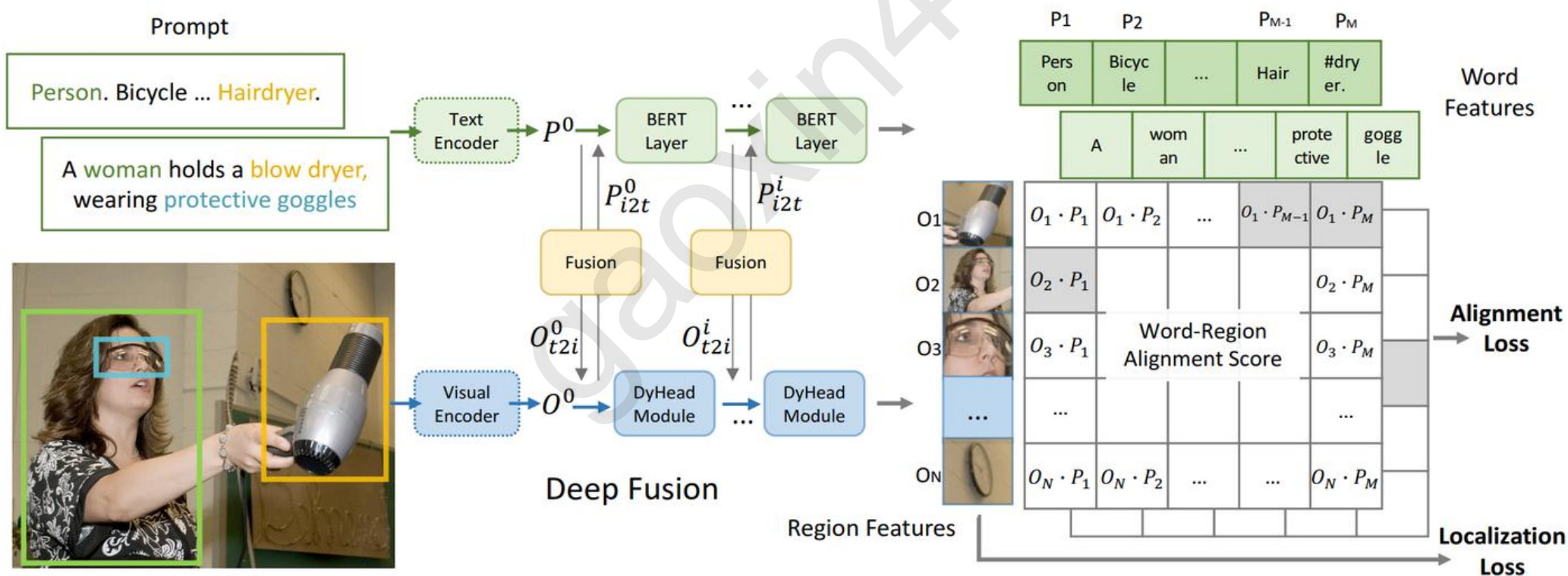
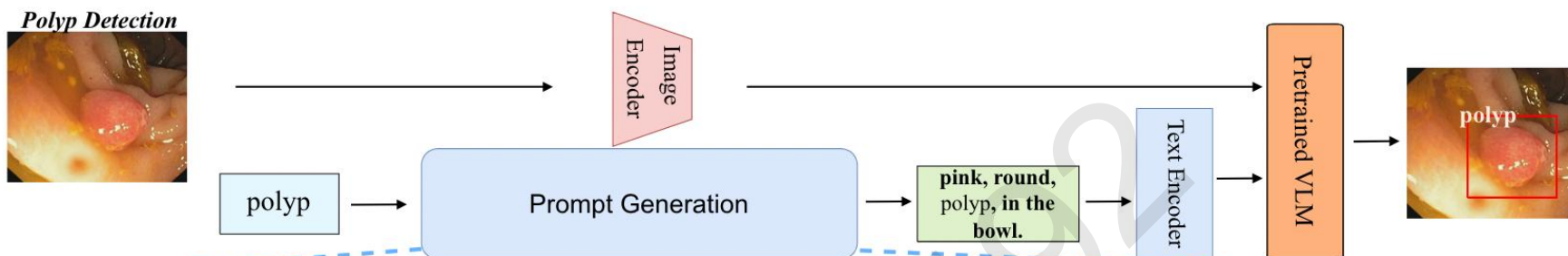
GLIP

IMAGEBIND

华西

MMCoT

LLaVA





- manually designing an effective prompt requires expert-level knowledge and a lot of effort;
- in the current vision-language models, the prompts are normally fixed for all samples during inference, i.e., not image-specific, which is not ideal for grounding novel objects that may have varying appearances.

### generate knowledge-rich and image-specific prompts

**【PubMedBERT】** Masked Language Model  
Driven Auto-Prompt Generation  
‘The [Attr] of an [Object] is [MASK]’  
top-k predicted words for the [MASK] token

**【OFA】** Image Specific Auto-Prompt Generation  
“What color is this wound?”.

hybrid prompts

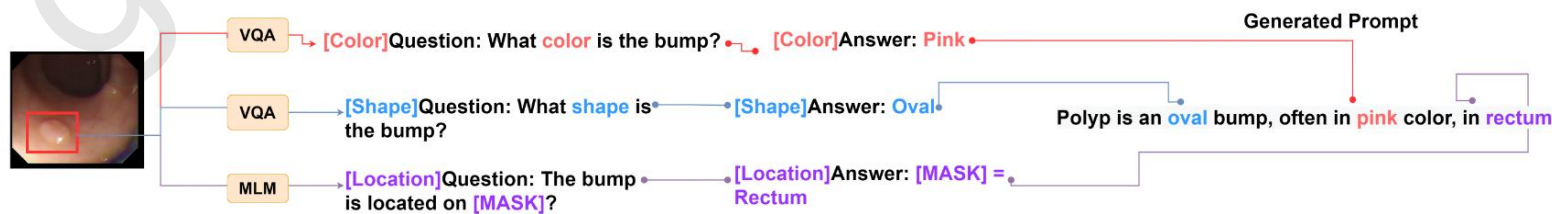
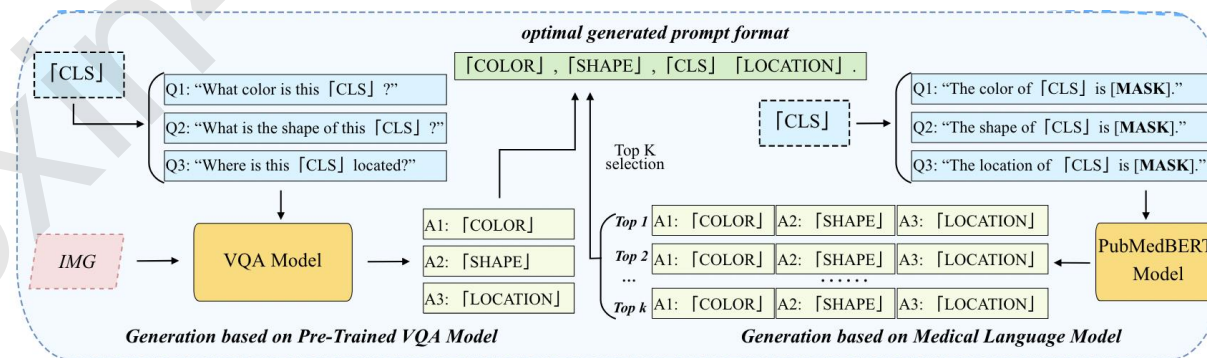


Figure 4: Auto-prompt generation show case.

Table 1: Dataset overview (13 datasets in total).

|         | Photography images |           | Endoscopy images                | Microscopy images |                | Radiology images |        |      |            |
|---------|--------------------|-----------|---------------------------------|-------------------|----------------|------------------|--------|------|------------|
|         |                    |           |                                 | Cytology          | Histopathology | X ray            | CT     | MRI  | Ultrasound |
| Dataset | ISIC 2016          | DFUC 2020 | Ployp Benchmark ( $\times 5$ )* | BCCD              | CPM-17         | TBX11K           | Luna16 | ADNI | TN3k       |

\* includes CVC-300, CVC-ClinicDB, CVC-ColonDB, Kvasir, and ETIS

## 1. Transfer performance surpassing supervised methods

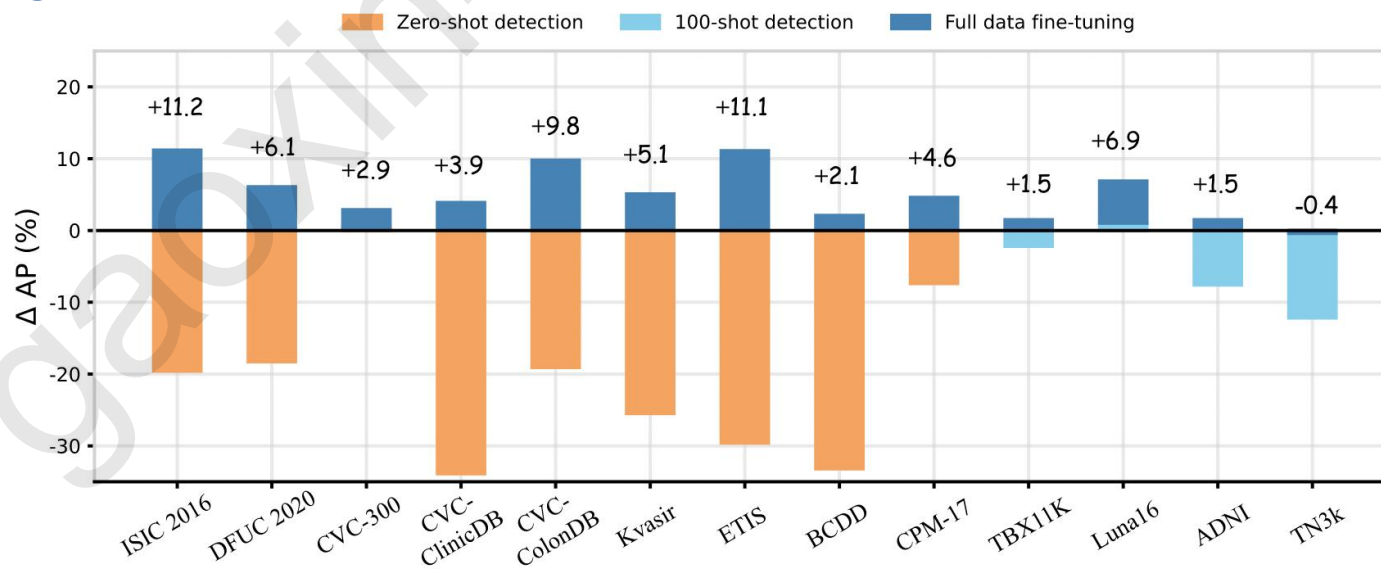
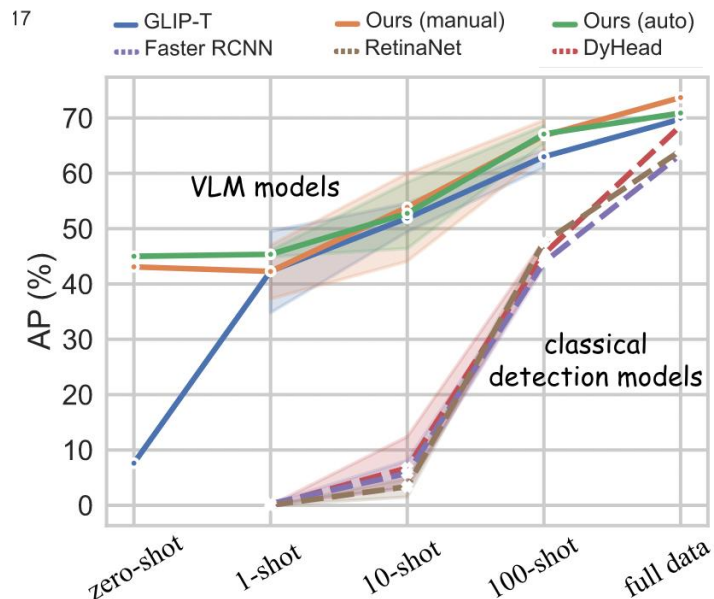


Figure 2: Comparisons with a fully supervised baseline (the horizontal line). The y-axis shows  $\Delta AP$  compared to the supervised baseline. For non-radiology datasets, we exhibit zero-shot and full data results; we show 100-shot and full data results for the radiology datasets (from TBX11K to TN3k).

## 2. Superior zero-shot transfer performance compared to the baseline

Table 3: Our approaches v.s. supervised models on non-radiology datasets (AP%).

|           | Method              | Backbone | ISIC 2016   | DFUC 2022   | Polyp ( $\times 5$ ) | BCCD        | CPM-17      |
|-----------|---------------------|----------|-------------|-------------|----------------------|-------------|-------------|
| Full Data | Faster RCNN         | RN50     | 50.3        | 42.3        | 56.6                 | 56.9        | 39.8        |
|           | RetinaNet           | RN50     | 54.0        | 43.1        | 58.8                 | 56.7        | 35.7        |
|           | DyHead              | Swin-T   | 52.9        | 44.2        | 62.9                 | 60.1        | 38.8        |
|           | GLIP-T(default cls) | Swin-T   | 62.4        | <b>50.3</b> | 68.1                 | 62.5        | 43.9        |
|           | Ours (Manual)       | Swin-T   | <b>64.1</b> | <b>50.3</b> | <b>69.4</b>          | 62.2        | 43.4        |
|           | Ours (Auto)         | Swin-T   | 61.6        | 50.1        | 68.8                 | <b>63.1</b> | <b>44.2</b> |
| 100-Shot  | Faster RCNN         | RN50     | 44.6        | 27.0        | 44.9                 | 38.6        | —           |
|           | RetinaNet           | RN50     | 41.7        | 28.4        | 41.7                 | 54.3        | —           |
|           | DyHead              | Swin-T   | 42.5        | 27.8        | 42.5                 | 40.5        | —           |
|           | GLIP-T(default cls) | Swin-T   | 55.9        | 41.4        | 57.6                 | 59.8        | —           |
|           | Ours (Manual)       | Swin-T   | 58.0        | <b>43.7</b> | <b>60.8</b>          | 60.1        | —           |
|           | Ours (Auto)         | Swin-T   | <b>58.8</b> | 42.4        | <b>60.8</b>          | <b>60.2</b> | —           |
| Zero-Shot | GLIP-T(default cls) | Swin-T   | 20.1        | 0.1         | 4.1                  | 0.7         | 7.6         |
|           | GLIP-L(default cls) | Swin-L   | 20.4        | 3.6         | 11.9                 | 10.4        | 11.6        |
|           | Ours (with MLM)     | Swin-T   | 25.1        | 24.8        | 38.4                 | 24.1        | 20.3        |
|           | Ours (with VQA)     | Swin-T   | 23.5        | 12.9        | 27.1                 | 14.3        | 26.2        |
|           | Ours (with Hybrid)  | Swin-T   | 24.5        | 22.5        | 35.1                 | 14.3        | 24.8        |
|           | Ours (Manual)       | Swin-T   | <b>33.3</b> | <b>25.9</b> | <b>41.3</b>          | <b>26.9</b> | <b>31.4</b> |

our approaches can  
empower the pre-  
trained VLM with  
remarkable zero-shot  
capability in the  
medical domain.

### 3. The effectiveness of attribution injection and auto-prompts

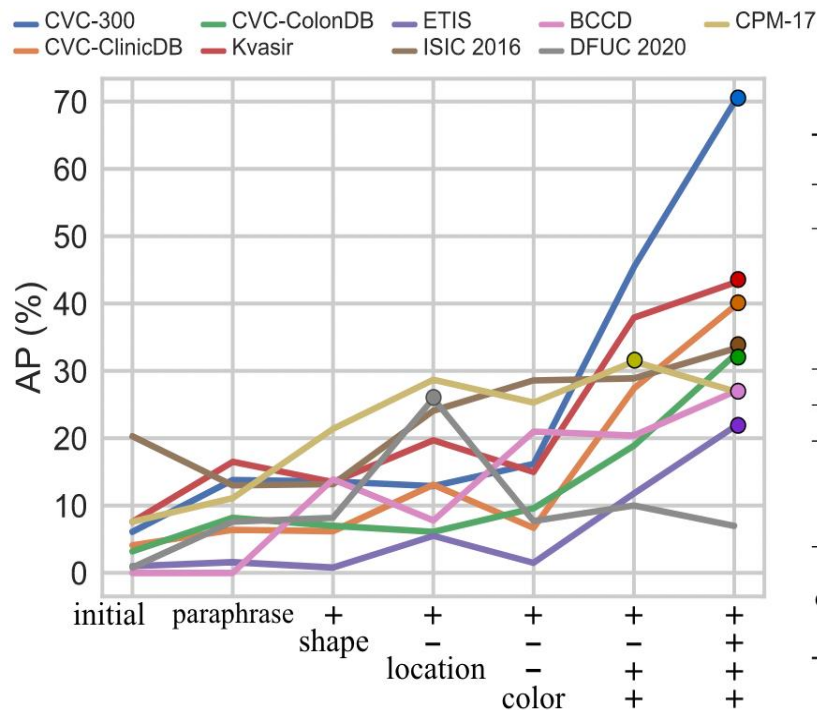


Table 4: Examples of prompts for BCDD (zero-shot performance on the validation and test set)

|                  | Prompt  | AP          | AP50        |
|------------------|---|-------------|-------------|
| initial          | platelet. red blood cell. white blood cell  | 0.4         | 0.9         |
|                  | thrombocyte. erythrocyte. leukocyte   | 0.1         | 0.1         |
| medical concepts | blood platelet. red blood corpuscle. white blood corpuscle  | 3.1         | 7.0         |
|                  | thrombocyte, blood platelet. erythrocyte, red blood corpuscle. leukocyte, white blood corpuscle   | 6.8         | 15.5        |
|                  | <b>thrombocyte or blood platelet. erythrocyte or red blood corpuscle. leukocyte or white blood corpuscle</b>  | 8.6         | 17.9        |
| + location       | platelet <b>in blood</b> . red blood cell <b>in blood</b> . white blood cell <b>in blood</b>  | 6.9         | 14.4        |
| + shape          | <b>small</b> platelet. <b>rounded</b> red blood cell. <b>irregular</b> white blood cell   | 7.7         | 14.9        |
| + color          | colorless platelet. freshcolor red blood cell. blue white blood cell  | 18.3        | 32.3        |
|                  | colorless platelet. freshcolor red blood cell. purple white blood cell  | 17.8        | 32.9        |
|                  | <b>colorless</b> platelet. <b>freshcolor</b> red blood cell. <b>purple or blue</b> white blood cell   | 24.9        | 43.8        |
| combinations     | small, colorless platelet. rounded, freshcolor red blood cell. irregular, purple or blue white blood cell   | 26.6        | 47.1        |
|                  | small, colorless blood platelet. rounded, freshcolor erythrocyte. irregular, purple or blue leukocyte   | 26.4        | 45.3        |
|                  | <b>small, colorless</b> platelet. <b>rounded, freshcolor</b> red blood <b>corpuscle</b> . <b>irregular, purple or blue</b> white blood <b>corpuscle</b> | <b>27.1</b> | <b>47.6</b> |

the overall performance increases as more attributes are integrated into the prompts



# Multimodal Chain-of-Thought Reasoning in Language Models

Zhuosheng Zhang<sup>1</sup> Aston Zhang<sup>2</sup> Mu Li<sup>2</sup> Hai Zhao<sup>1</sup> George Karypis<sup>2</sup> Alex Smola<sup>2</sup>

2023,2,17

<sup>1</sup>Shanghai Jiao Tong University <sup>2</sup>Amazon Web Services.  
Correspondence to: Zhuosheng Zhang (work done at Amazon Web Services) <zhangzs@sjtu.edu.cn>, Aston Zhang <az@astonzhang.com>.

<sup>1</sup><https://github.com/amazon-science/mm-cot>

Input

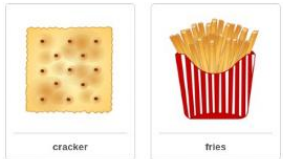
Language

**Question:** Which property do these two objects have in common?

**Context:** Select the better answer.

**Options:** ☐ (A) soft ☐ (B) salty

Vision



cracker fries

Output

**Rationale:** Look at each object. For each object, decide if it has that property. Potato chips have a salty taste. Both objects are salty. A soft object changes shape when you squeeze it. The fries are soft, but the cracker is not. The property that both objects have in common is salty.

**Answer:** The answer is (B).



- Large language models (LLMs) have shown impressive performance on complex reasoning by leveraging chain-of-thought (CoT) prompting to generate intermediate reasoning chains as the rationale to infer the answer.

## Why is chain-of-thought reasoning useful?

**locality<sup>[1]</sup>**

**P (C|A)      intermediate variable B      (B,C) and (B,A) are often seen**

- Direct prediction of conditional probabilities is inaccurate for some inferences because the relevant variables are rarely seen together in training. Chain-of-thought reasoning improves estimation because it can chain together local statistical dependencies that are frequently observed in training.

[1] Prystawski, B.; Goodman, N. D. Why Think Step-by-Step? Reasoning Emerges from the Locality of Experience. arXiv April 7, 2023. <http://arxiv.org/abs/2304.03843>.

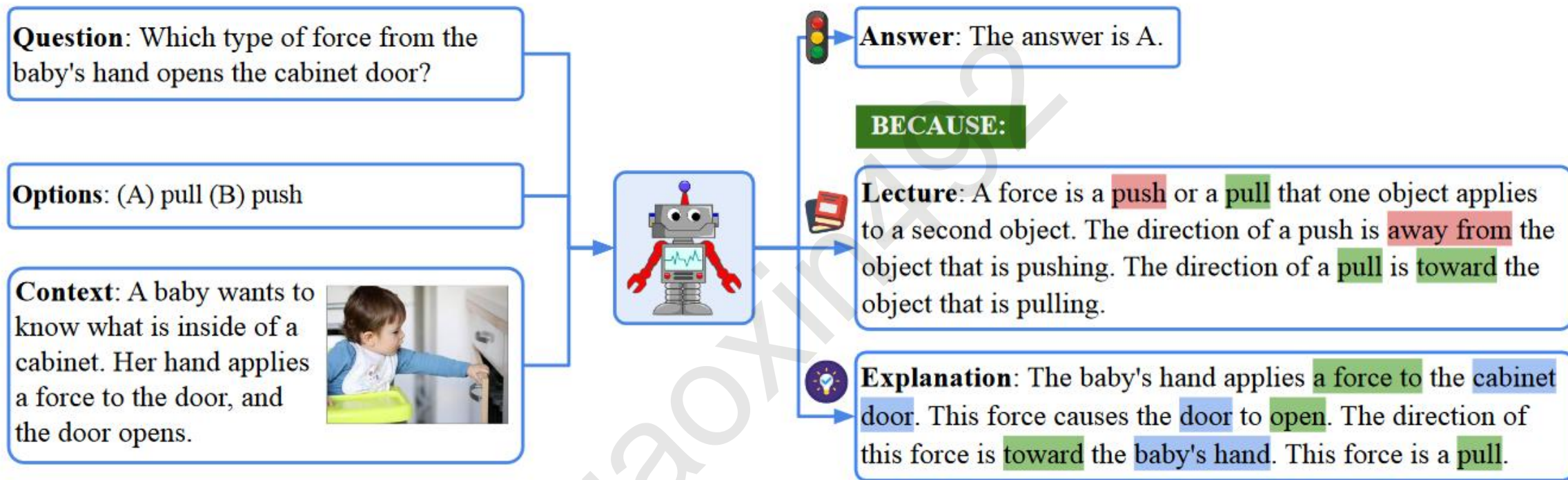


Figure 1: We construct the SCIENCEQA dataset where a data example consists of multimodal question answering information and the grounded lecture and explanation. We study if QA models can generate a reasonable explanation to reveal the chain-of-thought reasoning.

- The first to study CoT reasoning in different modalities
- Two ways to elicit Multimodal-CoT reasoning:

**prompting LLMs**

将图像转化成文字(caption), 和问题一起输给大模型, 大量的信息损失, 缺少不同模态在某一共同表示空间的相互作用

**fine-tuning small models**

利用到不同模态之间的相互作用, 问题是参数量太小的模型会产生错误的推理过程 (The key challenge is that language models under 100 billion parameters tend to generate hallucinated rationales that mislead the answer inference (Ho et al., 2022; Magister et al., 2022; Ji et al., 2022)).

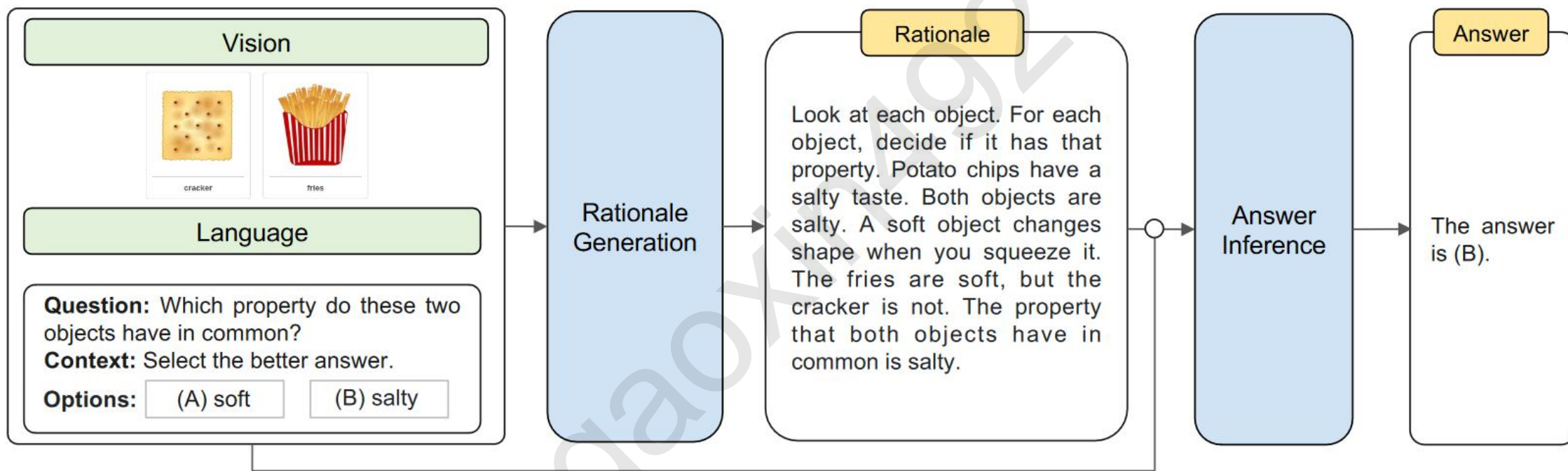
Table 2. Effects of CoT in the one-stage setting.

| Method      | Format | Accuracy |
|-------------|--------|----------|
| No-CoT      | QCM→A  | 80.40    |
| Reasoning   | QCM→RA | 67.86    |
| Explanation | QCM→AR | 69.77    |

**The question text (Q), the context text (C), and multiple options (M) as the input; the rationale text(R), and answer(A) as the output**

**The rationales might not necessarily contribute to predicting the right answer**

separate the CoT problem into two stages: **rationale generation** and **answer inference**



*Figure 4.* Overview of our Multimodal-CoT framework. Multimodal-CoT consists of two stages: (i) rationale generation and (ii) answer inference. Both stages share the same model architecture but differ in the input and output. In the first stage, we feed the model with language and vision inputs to generate rationales. In the second stage, we append the original language input with the rationale generated from the first stage. Then, we feed the updated language input with the original vision input to the model to infer the answer.



## How the rationales affect the answer prediction?

**Q:** question, **C:** context, **M:** multiple options, **R:** rationale, **A:** answer

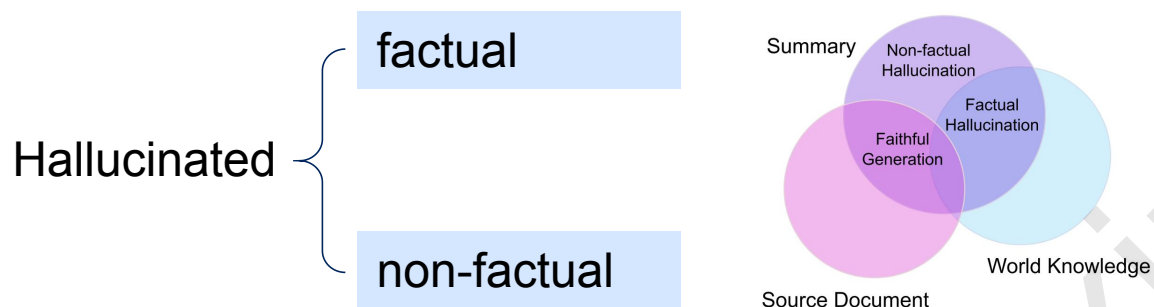
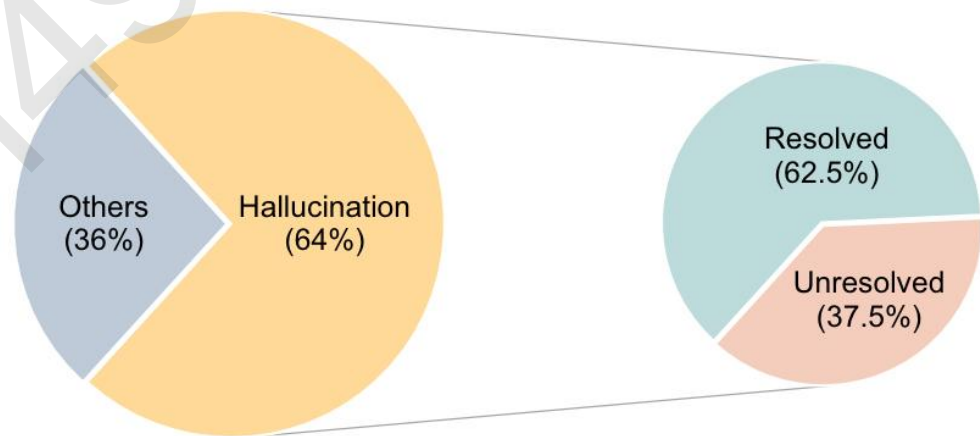


Table 3. Two-stage setting of (i) rationale generation (RougeL) and (ii) answer inference (Accuracy).

| Method              | (i) QCM $\rightarrow$ R | (ii) QCMR $\rightarrow$ A |
|---------------------|-------------------------|---------------------------|
| Two-Stage Framework | 91.76                   | 70.53                     |
| w/ Captions         | 91.85                   | 71.12                     |
| w/ Vision Features  | 96.97                   | 84.91                     |



(a) ratio of hallucination mistakes

(b) correction rate w/ vision features

Figure 3. The ratio of hallucination mistakes (a) and correction rate w/ vision features (b).

## Multimodality (Vision Features) Contributes to Effective Rationales



We speculate that such a phenomenon of hallucination is due to a lack of necessary vision contexts for performing effective Multimodal-CoT.

1. DETR model (Carion et al., 2020) to extract vision features
2. fuse the vision features with the encoded language representations
3. feed to the decoder

backbone language model: UnifiedQA (T5)

4 NVIDIA Tesla V100 32G GPUs

$X = \{X_{\text{language}}^1, X_{\text{vision}}\} \rightarrow R = F(X)$  where  $R$  is the rationale

$\bar{X}_{\text{language}}^2 = X_{\text{language}}^1 \circ R$  (concat)

$\rightarrow X' = \{X_{\text{language}}^2, X_{\text{vision}}\} \rightarrow A = F(X')$

| model                          | size |
|--------------------------------|------|
| Mutimodal-CoT <sub>Base</sub>  | 223M |
| Mutimodal-CoT <sub>Large</sub> | 738M |

**1. encoding**

$$H_{\text{language}} = \text{LanguageEncoder}(X_{\text{language}}), \quad (2)$$

$$H_{\text{vision}} = W_h \cdot \text{VisionExtractor}(X_{\text{vision}}), \quad (3)$$

$H_{\text{language}} \in \mathbb{R}^{n \times d}$  where  $n$  denotes the length of the language input, and  $d$  is the hidden dimension. Meanwhile,

patch-level vision features, we apply a learnable projection matrix  $W_h$  to convert the shape of  $\text{VisionExtractor}(X_{\text{vision}})$  into that of  $H_{\text{language}}$ ; thus we have  $H_{\text{vision}} \in \mathbb{R}^{m \times d}$  where  $m$  is the number of patches.

**2. interaction**

we use a single-head attention network to correlate text tokens with image patches, where the query ( $Q$ ), key ( $K$ ) and value ( $V$ ) are  $H_{\text{language}}$ ,  $H_{\text{vision}}$  and  $H_{\text{vision}}$ ,

$$H_{\text{vision}}^{\text{attn}} = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \quad (4)$$

$$H_{\text{vision}}^{\text{attn}} \in \mathbb{R}^{n \times d}$$

$$\lambda = \text{Sigmoid}(W_l H_{\text{language}} + W_v H_{\text{vision}}^{\text{attn}}), \quad (5)$$

$$H_{\text{fuse}} = (1 - \lambda) \cdot H_{\text{language}} + \lambda \cdot H_{\text{vision}}^{\text{attn}}, \quad (6)$$

**3. decoding**

$H_{\text{fuse}}$  is fed into the Transformer decoder to predict the target  $Y$

Multimodal-  
CoTLarge  
outperforms  
GPT-3.5 by  
16.51%  
(75.17%→91.6  
8%) and  
surpasses  
human  
performance

Table 4. Main results (%). Size = backbone model size. Question classes: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12. Results except ours are taken from [Lu et al. \(2022a\)](#). Segment 1: Human performance; Segment 2: VQA baselines; Segment 3: UnifiedQA baselines; Segment 4: GPT-3.5 baselines; Segment 5: Our Multimodal-CoT results. Results in **bold** are the best performance.

| Model   | Size        | NAT          | SOC          | LAN          | TXT          | IMG          | NO           | G1-6         | G7-12        | Avg          |
|---|-------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Human   | -           | 90.23        | 84.97        | 87.48        | 89.60        | 87.50        | 88.10        | 91.59        | 82.42        | 88.40        |
| MCAN (Yu et al., 2019)                              | 95M         | 56.08        | 46.23        | 58.09        | 59.43        | 51.17        | 55.40        | 51.65        | 59.72        | 54.54        |
| Top-Down (Anderson et al., 2018)                    | 70M         | 59.50        | 54.33        | 61.82        | 62.90        | 54.88        | 59.79        | 57.27        | 62.16        | 59.02        |
| BAN (Kim et al., 2018)                              | 112M        | 60.88        | 46.57        | 66.64        | 62.61        | 52.60        | 65.51        | 56.83        | 63.94        | 59.37        |
| DFAF (Gao et al., 2019)                             | 74M         | 64.03        | 48.82        | 63.55        | 65.88        | 54.49        | 64.11        | 57.12        | 67.17        | 60.72        |
| ViLT (Kim et al., 2021)                             | 113M        | 60.48        | 63.89        | 60.27        | 63.20        | 61.38        | 57.00        | 60.72        | 61.90        | 61.14        |
| Patch-TRM (Lu et al., 2021)                         | 90M         | 65.19        | 46.79        | 65.55        | 66.96        | 55.28        | 64.95        | 58.04        | 67.50        | 61.42        |
| VisualBERT (Li et al., 2019)                        | 111M        | 59.33        | 69.18        | 61.18        | 62.71        | 62.17        | 58.54        | 62.96        | 59.92        | 61.87        |
| UnifiedQA <sub>Base</sub> (Khashabi et al., 2020)   | 223M        | 68.16        | 69.18        | 74.91        | 63.78        | 61.38        | 77.84        | 72.98        | 65.00        | 70.12        |
| UnifiedQA <sub>Base</sub> w/ CoT (Lu et al., 2022a) | 223M        | 71.00        | 76.04        | 78.91        | 66.42        | 66.53        | 81.81        | 77.06        | 68.82        | 74.11        |
| GPT-3.5 (Chen et al., 2020)                         | 175B        | 74.64        | 69.74        | 76.00        | 74.44        | 67.28        | 77.42        | 76.80        | 68.89        | 73.97        |
| GPT-3.5 w/ CoT (Lu et al., 2022a)                   | 175B        | 75.44        | 70.87        | 78.09        | 74.68        | 67.43        | 79.93        | 78.23        | 69.68        | 75.17        |
| Multimodal-CoT <sub>Base</sub>                      | <b>223M</b> | 87.52        | 77.17        | 85.82        | 87.88        | 82.90        | 86.83        | 84.65        | 85.37        | 84.91        |
| Multimodal-CoT <sub>Large</sub>                     | <b>738M</b> | <b>95.91</b> | <b>82.00</b> | <b>90.82</b> | <b>95.26</b> | <b>88.80</b> | <b>92.89</b> | <b>92.44</b> | <b>90.31</b> | <b>91.68</b> |

**Two stages?****VT-R ==> VTR-A VS VT-RA***Table 5. Ablation results of Multimodal-CoT.*

| Model                   | NAT   | SOC   | LAN   | TXT   | IMG   | NO    | G1-6  | G7-12 | Avg   |
|-------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Multimodal-CoT          | 87.52 | 77.17 | 85.82 | 87.88 | 82.90 | 86.83 | 84.65 | 85.37 | 84.91 |
| w/o Two-Stage Framework | 80.99 | 87.40 | 81.91 | 80.25 | 78.83 | 83.62 | 82.78 | 82.20 | 82.57 |
| w/o Vision Features     | 71.09 | 70.75 | 69.18 | 71.16 | 65.84 | 71.57 | 71.00 | 69.68 | 70.53 |

**Vision encoder?***Table 6. Accuracy (%) of using different vision features.*

| Method    | One-stage | Two-Stage |
|-----------|-----------|-----------|
| w/ CLIP   | 81.21     | 84.81     |
| w/ DETR   | 82.57     | 84.91     |
| w/ ResNet | 80.97     | 84.77     |

*Table 8. Categorization analysis of Multimodal-CoT.*

| Answer    | CoT Category        | Percentage (%) |
|-----------|---------------------|----------------|
| Correct   | CoT is correct      | 90             |
|           | CoT is incorrect    | 10             |
| Incorrect | Commonsense Mistake | 82             |
|           | Logical Mistake     | 12             |
|           | CoT is correct      | 6              |

- injecting commonsense knowledge;
- incorporating more informative vision features and improving language-vision interaction to be capable of understanding maps and counting numbers;



---

# Visual Instruction Tuning

---

**Haotian Liu<sup>1\*</sup>, Chunyuan Li<sup>2\*</sup>, Qingyang Wu<sup>3</sup>, Yong Jae Lee<sup>1</sup>**

<sup>1</sup>University of Wisconsin–Madison    <sup>2</sup>Microsoft Research    <sup>3</sup>Columbia University

<https://llava-vl.github.io>

**LLaVA:** Large Language and Vision Assistant

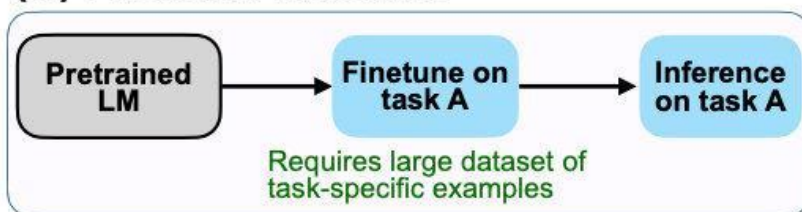
2023,4,17

In this paper, we present the first attempt to use language-only GPT-4 to generate multimodal language-image instruction-following data.

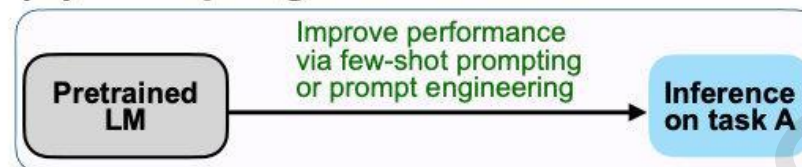


Instruction tuning large language models (LLMs) using machine-generated instruction-following data has improved zero-shot capabilities on new tasks

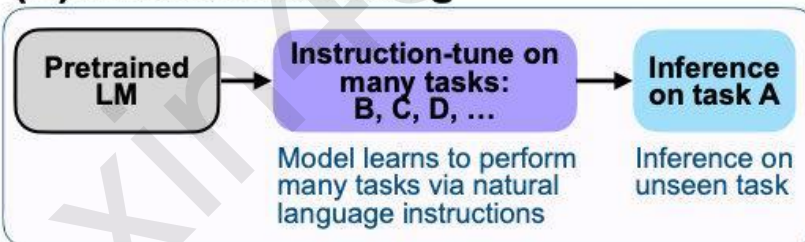
### (A) Pretrain–finetune



### (B) Prompting



### (C) Instruction tuning



**instruction-following ability**

In the natural language processing (NLP) community, to enable LLMs such as GPT-3, T5, PaLM, and OPT to follow natural language instructions and complete real-world tasks, researchers have explored methods for LLM instruction-tuning, leading to instruction-tuned counterparts such as InstructGPT/ChatGPT, FLAN-T5, FLANPaLM, and OPT-IML, respectively.

[3] Wei J, Bosma M, Zhao V Y, et al. Finetuned language models are zero-shot learners[J]. arXiv preprint arXiv:2109.01652, 2021.

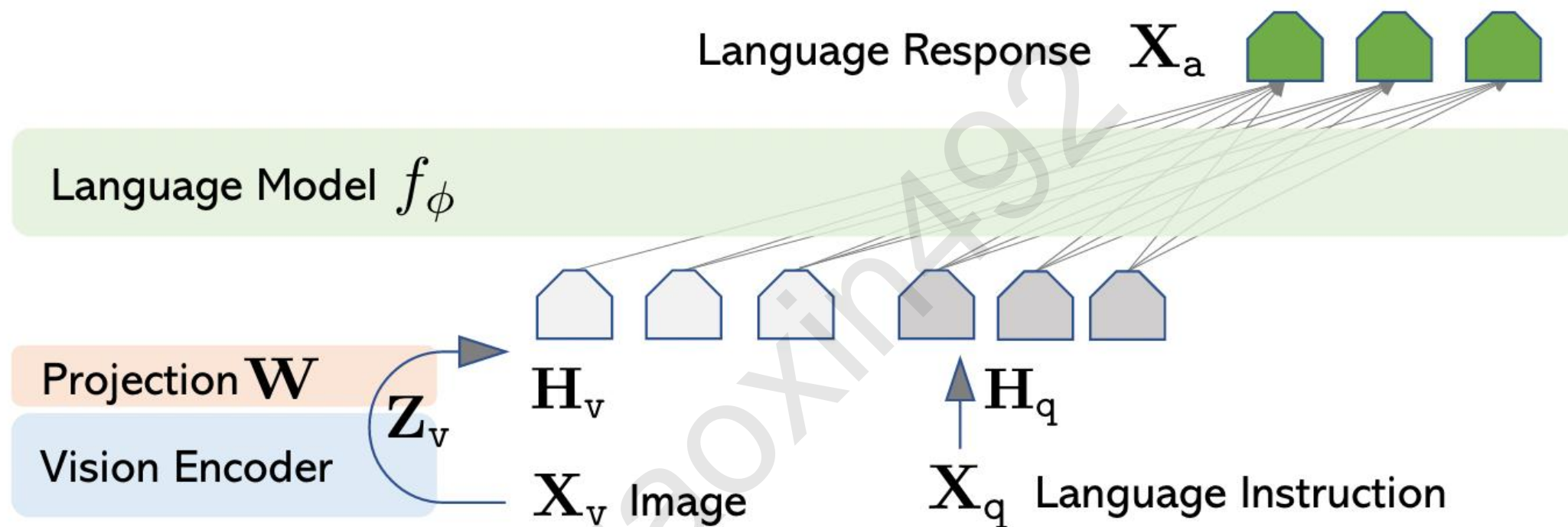


Figure 1: LLaVA network architecture.

We leave exploring possibly more effective and sophisticated architecture designs for LLaVA as future work.

158K unique language-image samples

23K in detailed description

58K in conversations

77k in complex reasoning

#### Response type 1: conversation

Question: What type of vehicle is featured in the image?

Answer: The image features a black sport utility vehicle (SUV).

Question: Where is the vehicle parked?

Answer: The vehicle is parked in an underground parking area, likely in a public garage.

Question: What are the people in the image doing?

Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

#### Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip.

In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle.

Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

#### Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage.

Luggage surrounds a vehicle in an underground parking area

People try to fit all of their luggage in an SUV.

The sport utility vehicle is parked in the public garage, being packed for a trip

Some people with luggage near a van that is transporting it.

#### Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]



**caption+bounding box => GPT4/3.5**

#### Response type 3: complex reasoning

Question: What challenges do these people face?

Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver's view or make the passengers uncomfortable during the trip.



To get a rich and comprehensive description for an image,  
ask GPT-4 to use the original image captions to generate the detailed description

- brief description question + image caption
- detailed description question + GPT-refined image caption

**Instructions for brief image description.** The list of instructions used to briefly describe the image content are shown in Table 8. They present the same meaning with natural language variance.

- "Describe the image concisely."
- "Provide a brief description of the given image."
- "Offer a succinct explanation of the picture presented."
- "Summarize the visual content of the image."
- "Give a short and clear explanation of the subsequent image."
- "Share a concise interpretation of the image provided."
- "Present a compact description of the photo's key features."
- "Relay a brief, clear account of the picture shown."
- "Render a clear and concise summary of the photo."
- "Write a terse but informative summary of the picture."
- "Create a compact narrative representing the image presented."

Table 8: The list of instructions for brief image description.

**Instructions for detailed image description.** The list of instructions used to describe the image content in detail are shown in Table 9. They present the same meaning with natural language variance.

- "Describe the following image in detail"
- "Provide a detailed description of the given image"
- "Give an elaborate explanation of the image you see"
- "Share a comprehensive rundown of the presented image"
- "Offer a thorough analysis of the image"
- "Explain the various aspects of the image before you"
- "Clarify the contents of the displayed image with great detail"
- "Characterize the image using a well-detailed description"
- "Break down the elements of the image in a detailed manner"
- "Walk through the important details of the image"
- "Portray the image with a rich, descriptive narrative"
- "Narrate the contents of the image with precision"
- "Analyze the image in a comprehensive and detailed manner"
- "Illustrate the image through a descriptive explanation"
- "Examine the image closely and share its details"
- "Write an exhaustive depiction of the given image"

Table 9: The list of instructions for detailed image description.



```
messages = [ {"role": "system", "content": f"""\nYou are an AI visual assistant, and you are\nseeing a single image. What you see are provided with five sentences, describing the same image you\nare looking at. Answer all questions as you are seeing the image.\n\nDesign a conversation between you and a person asking about this photo. The answers should be in a\ntone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions\nand give corresponding answers.\n\nInclude questions asking about the visual content of the image, including the object types, counting\nthe objects, object actions, object locations, relative positions between objects, etc. Only include\nquestions that have definite answers:\n(1) one can see the content in the image that the question asks about and can answer confidently;\n(2) one can determine confidently from the image that it is not in the image. Do not ask any question\nthat cannot be answered confidently.\n\nAlso include complex questions that are relevant to the content in the image, for example, asking\nabout background knowledge of the objects in the image, asking to discuss about events happening in\nthe image, etc. Again, do not ask about uncertain details. Provide detailed answers when answering\ncomplex questions. For example, give detailed examples or reasoning steps to make the content more\nconvincing and well-organized. You can include multiple paragraphs if necessary.\n"""} ]
```

```
for sample in fewshot_samples:\n    messages.append({"role": "user", "content": sample['context']})\n    messages.append({"role": "assistant", "content": sample['response']})\nmessages.append({"role": "user", "content": '\n'.join(query)})
```

1. 假装你是一个AI视觉助手，正在看一张图片。caption和bounding box描述了这张图片的信息，但不能透露你看过caption和bounding box
2. 设计一段你和一个人关于这张图像的对话
3. 包含针对视觉内容的问题，如物体种类、物体数量、物体动作、物体位置、物体之间的相对位置等等。只设计有明确答案的问题
4. 包含与图片内容相关的复杂问题，例如物体的背景知识，对发生事件的讨论等等。

**in-context learning  
seed examples**

For each image  $\mathbf{X}_v$ , we generate multi-turn conversation data  $(\mathbf{X}_q^1, \mathbf{X}_a^1, \dots, \mathbf{X}_q^T, \mathbf{X}_a^T)$ , where  $T$  is the total number of turns. We organize them as a sequence, by treating all answers as the assistant's response, and the instruction  $\mathbf{X}_{\text{instruct}}^t$  at the  $t$ -th turn as:

$$\mathbf{X}_{\text{instruct}}^t = \begin{cases} \text{Random choose } [\mathbf{X}_q^1, \mathbf{X}_v] \text{ or } [\mathbf{X}_v, \mathbf{X}_q^1], & \text{the first turn } t = 1 \\ \mathbf{X}_q^t, & \text{the remaining turns } t > 1 \end{cases}$$

**description => a single-turn conversation**

$\mathbf{X}_{\text{system-message}} <\text{STOP}> \backslash \mathbf{n}$   
 Human :  $\mathbf{X}_{\text{instruct}}^1 <\text{STOP}> \backslash \mathbf{n}$  Assistant:  $\mathbf{X}_a^1 <\text{STOP}> \backslash \mathbf{n}$   
 Human :  $\mathbf{X}_{\text{instruct}}^2 <\text{STOP}> \backslash \mathbf{n}$  Assistant:  $\mathbf{X}_a^2 <\text{STOP}> \backslash \mathbf{n} \dots$

$$p(\mathbf{X}_a | \mathbf{X}_v, \mathbf{X}_{\text{instruct}}) = \prod_{i=1}^L p_{\theta}(\mathbf{x}_i | \mathbf{X}_v, \mathbf{X}_{\text{instruct}, <i}, \mathbf{X}_{a, <i}):$$

- Only **green sequence/tokens** are used to compute the loss in the auto-regressive model.
- X system-message = A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions.
- $<\text{STOP}> = \text{###}$ .

**Stage 1 Pre-training for Feature Alignment:** image + brief description question => image caption

**Stage 2 Fine-tuning End-to-End:** image + instruct\_question => instruct\_answer

Multimodal Chatbot.

Science QA: question & context as instruct\_question, and reasoning & answer as instruct\_answer

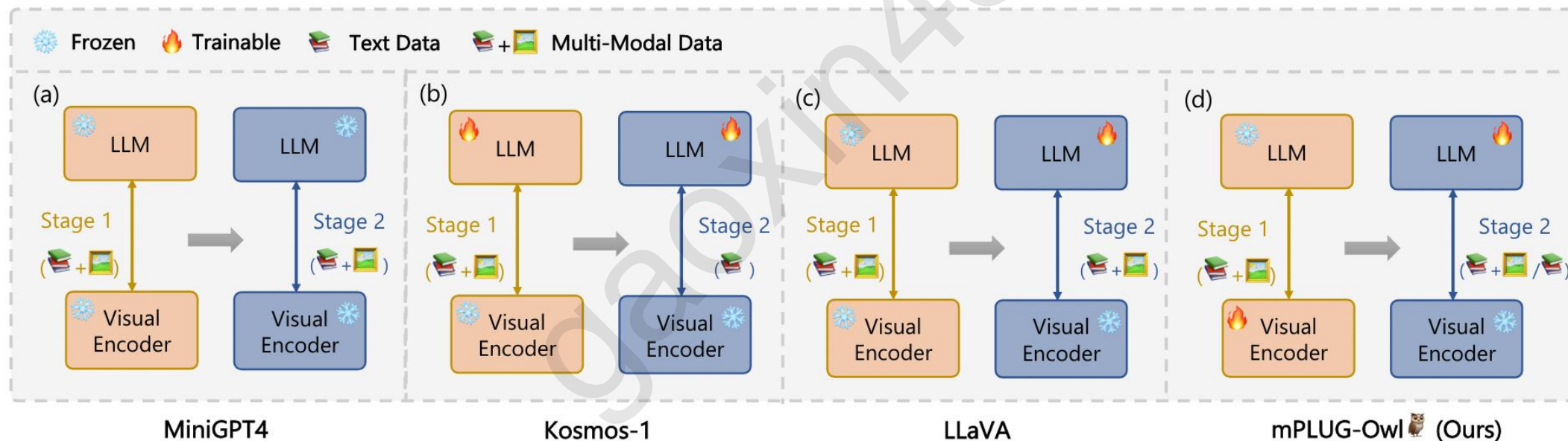


Figure 1: Comparison between different training paradigms. All of these methods are trained in a two-stage fashion. Stage 1 stands for pre-training and Stage 2 represents instruction tuning.



## Visual input example, Extreme Ironing:



## Case study

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.   |
| User         | [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 <b>due to his unconventional choice of ironing his clothes on top of a moving car</b> . 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 [32]   | 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         | Can you explain this meme in detail?   |
| BLIP-2       | a man is sitting on the back of a yellow cab   |
| User         | Can you explain this meme in detail?   |
| OpenFlamingo | The man is drying his clothes on the hood of his car.  |

|                                | Conversation | Detail description | Complex reasoning | All          |
|--------------------------------|--------------|--------------------|-------------------|--------------|
| Full data                      | 83.1         | 75.3               | 96.5              | 85.1         |
| Detail + Complex               | 81.5 (-1.6)  | 73.3 (-2.0)        | 90.8 (-5.7)       | 81.9 (-3.2)  |
| Conv + 5% Detail + 10% Complex | 81.0 (-2.1)  | 68.4 (-7.1)        | 91.5 (-5.0)       | 80.5 (-4.4)  |
| Conversation                   | 76.5 (-6.6)  | 59.8 (-16.2)       | 84.9 (-12.4)      | 73.8 (-11.3) |
| No Instruction Tuning          | 22.0 (-61.1) | 24.0 (-51.3)       | 18.5 (-78.0)      | 21.5 (-63.6) |

Adding a small amount of the detailed description and complex reasoning questions => model's capability ↑

## Results on ScienceQA

| Method   | Subject |       |       | Context Modality |       |       | Grade |       | Average |
|--|---------|-------|-------|------------------|-------|-------|-------|-------|---------|
|  | NAT     | SOC   | LAN   | TXT              | IMG   | NO    | G1-6  | G7-12 |         |
| <i>Representative &amp; SoTA methods with numbers reported in the literature</i> |         |       |       |                  |       |       |       |       |         |
| Human [30]   | 90.23   | 84.97 | 87.48 | 89.60            | 87.50 | 88.10 | 91.59 | 82.42 | 88.40   |
| GPT-3.5 [30]   | 74.64   | 69.74 | 76.00 | 74.44            | 67.28 | 77.42 | 76.80 | 68.89 | 73.97   |
| GPT-3.5 w/ CoT [30]  | 75.44   | 70.87 | 78.09 | 74.68            | 67.43 | 79.93 | 78.23 | 69.68 | 75.17   |
| LLaMA-Adapter [55]   | 84.37   | 88.30 | 84.36 | 83.72            | 80.32 | 86.90 | 85.83 | 84.05 | 85.19   |
| MM-CoT <sub>Base</sub> [57]  | 87.52   | 77.17 | 85.82 | 87.88            | 82.90 | 86.83 | 84.65 | 85.37 | 84.91   |
| MM-CoT <sub>Large</sub> [57]   | 95.91   | 82.00 | 90.82 | 95.26            | 88.80 | 92.89 | 92.44 | 90.31 | 91.68   |
| <i>Results with our own experiment runs</i>                                      |         |       |       |                  |       |       |       |       |         |
| GPT-4  | 84.06   | 73.45 | 87.36 | 81.87            | 70.75 | 90.73 | 84.69 | 79.10 | 82.69   |
| LLaVA  | 90.36   | 95.95 | 88.00 | 89.49            | 88.00 | 90.66 | 90.93 | 90.90 | 90.92   |
| LLaVA+GPT-4 (complement)   | 90.36   | 95.50 | 88.55 | 89.05            | 87.80 | 91.08 | 92.22 | 88.73 | 90.97   |
| LLaVA+GPT-4 (judge)  | 91.56   | 96.74 | 91.09 | 90.62            | 88.99 | 93.52 | 92.73 | 92.16 | 92.53   |

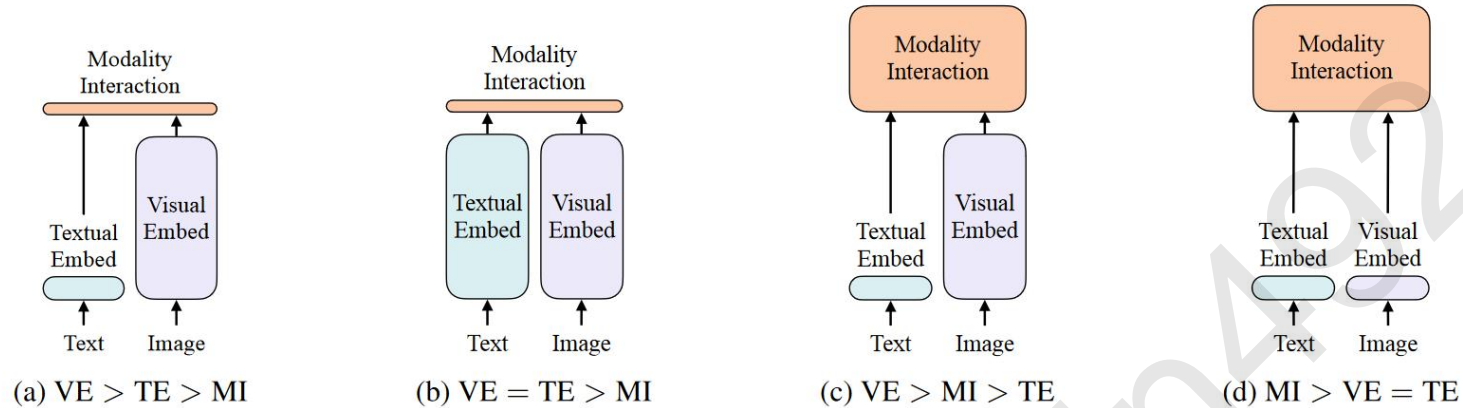
Table 6: Results (accuracy %) on Science QA dataset. Question classes: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12.



# Conclusion

## Architecture:

Multimodality embed  
Modality interaction

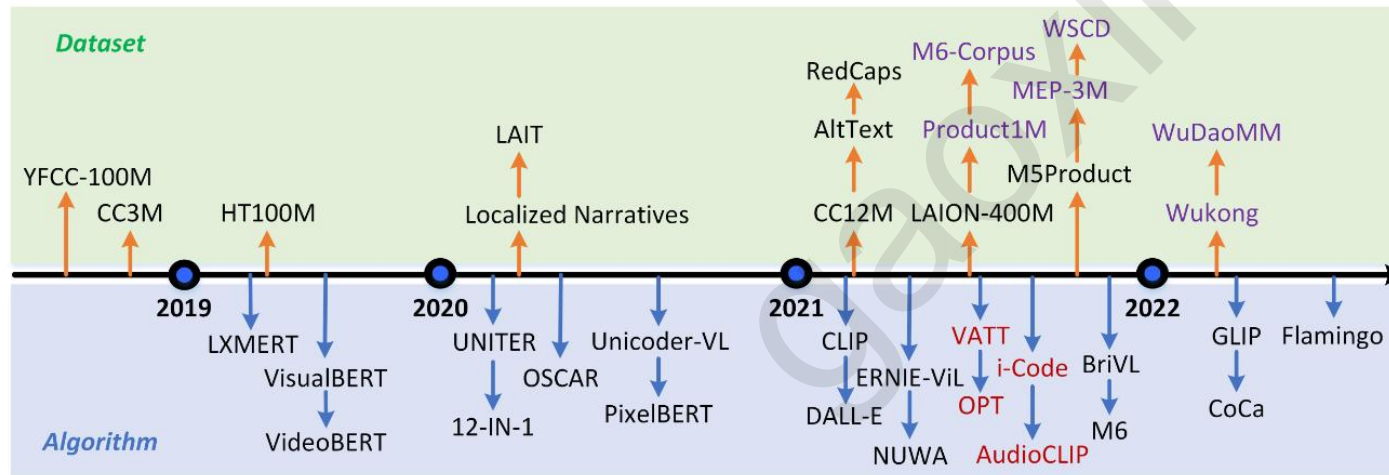


Pretrained models

Generative models

## Techniques:

- contrastive learning
- prompt learning
- chain of thought
- instruct-tuning



**Fig. 1** The chronological milestones on multi-modal pre-trained big models from 2019 to the present (June 2022), including multi-modal datasets (as shown by the orange arrow) and representative models (as shown by the blue arrow). The purple font indicates that the dataset contains Chinese text (other datasets contain English text). The models highlighted in wine red are trained on more than two modalities.

# 谢谢观看

W o r k s u m m a r y r e p o r t