AiDD 6th 2025

# 2025 AI+研发数字峰会

AI+ Development Digital Summit

## 拥抱AI 重塑研发

05/23-24 | 上海站

AiDD 6th 2025 | 05/23-24 | 上海站

2025 AI+ Development Digital Summit

AI+研发数字峰会

拥抱AI 重塑研发

# 多模态大语言模型中的上下文学习

杨旭 | 东南大学

# 杨旭

东南大学计算机学院副教授/博导

杨旭博士2021年6月从南洋理工大学计算机科学与技术系获工学博士学位，导师为蔡剑飞，张含望教授。现为东南大学计算机科学与工程学院、软件学院、人工智能学院副教授。新一代人工智能技术与应用教育部重点实验室副主任，江苏省双创博士。主要研究方向为多模态视觉语言任务，基于多模态大语言模型的上下文学习。在过去的3年内，以第一作者身份在人工智能顶级会议期刊发表论文多篇，包括 TPAMI，CVPR，ICCV，NeurIPS 等。

# 目录
# CONTENTS

# PART 01

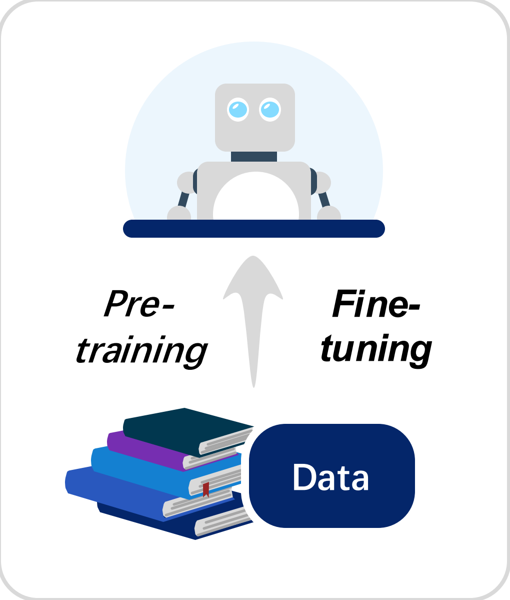
# Background

# The Development of GPT

## *GPT* (2018)

117M Parameters

1

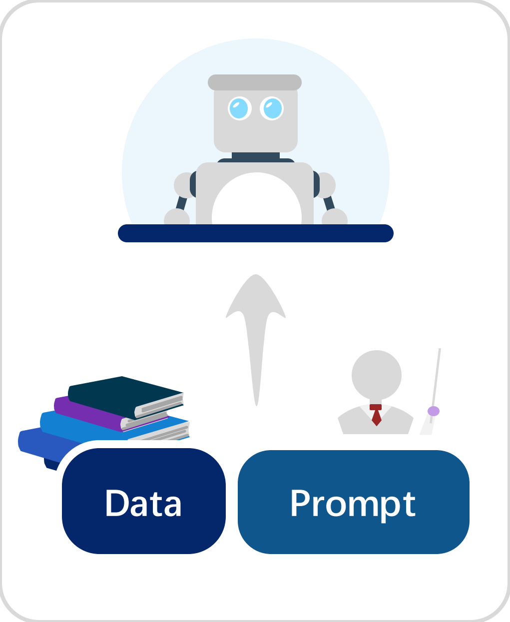



## *GPT-2* (2019)

1.5B Parameters

**Prompt Engineering**

2

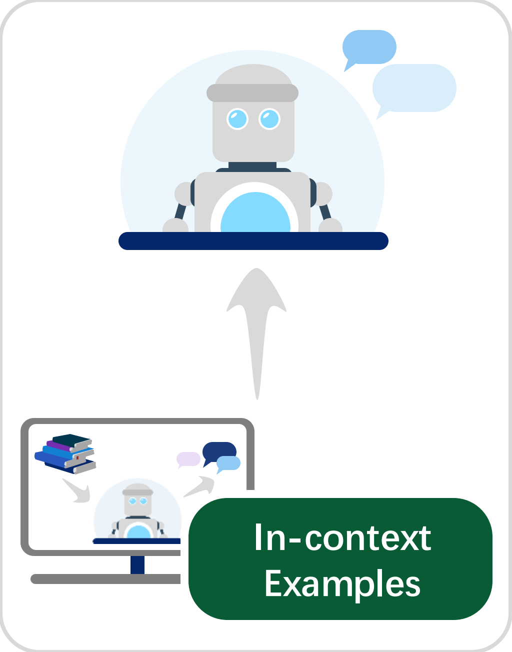



## *GPT-3* (2020)

175B Parameters

**In-context Learning**

3

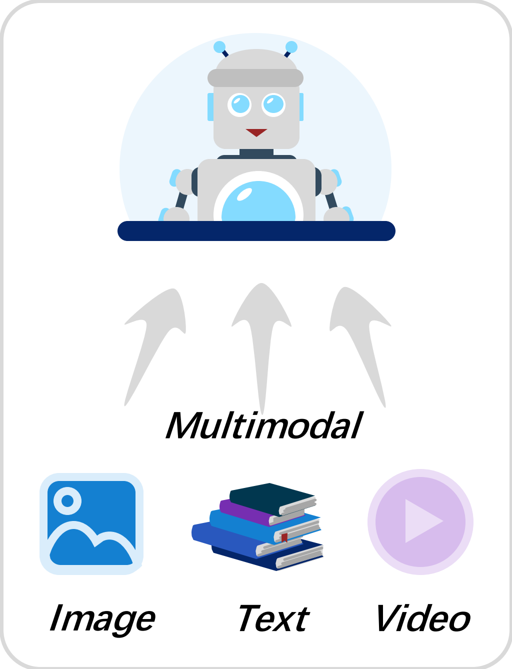



## *GPT-4* (2023)

1.76T Parameters

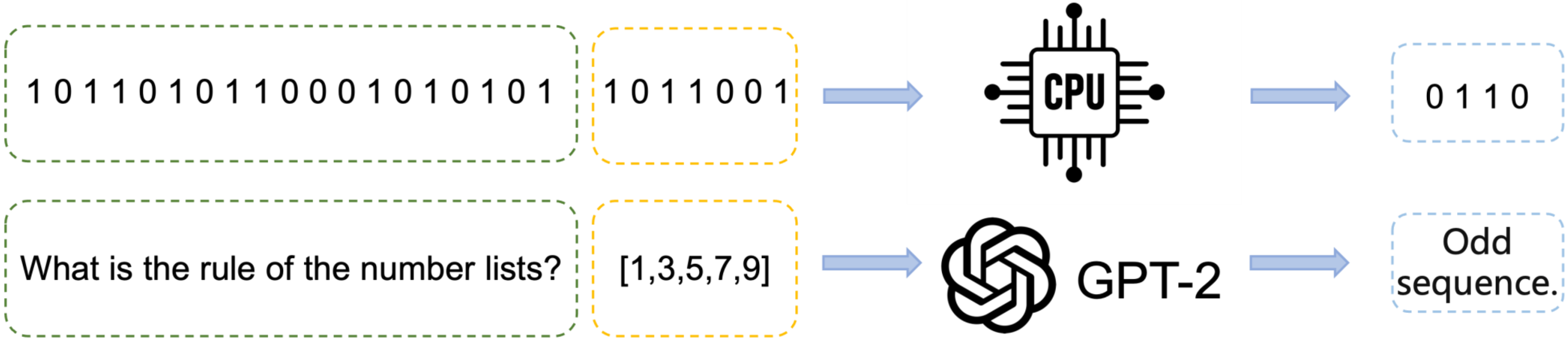
**Multimodal**

4



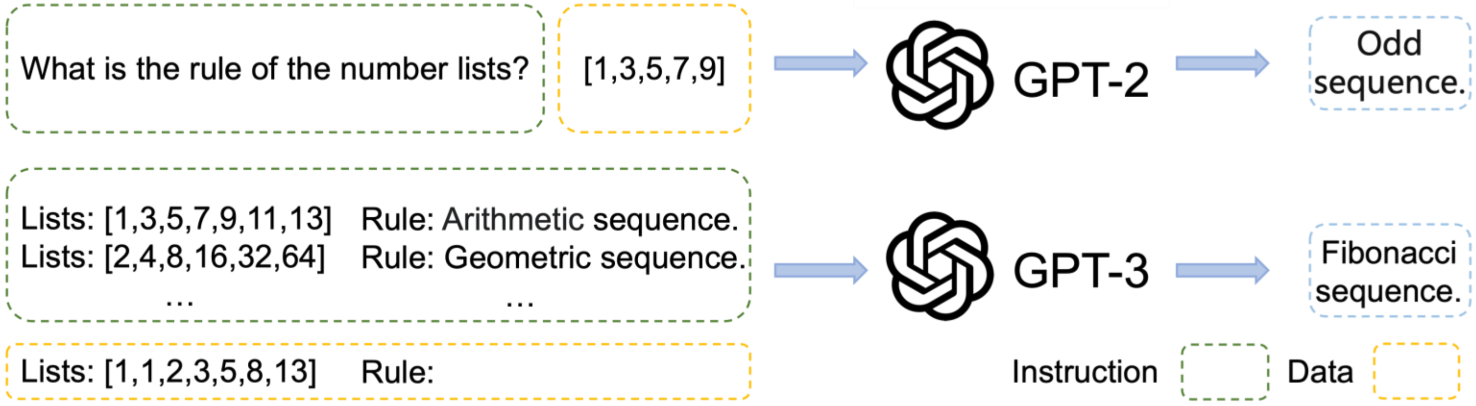
# GPT-2’s Capability of Prompt Engineering

- GPT-2 exhibits a distinctive feature known as **“prompt engineering”.**
- This can be compared to the architecture of modern computers, where both data and commands exist in the form of 0s and 1s encoding.



# GPT-3's Capability of Analogy: In-Context Learning

- GPT-3 possesses a unique capability known as **"In-context learning".**
- It will learn the representation of tasks from the provided in-context examples.




few shot

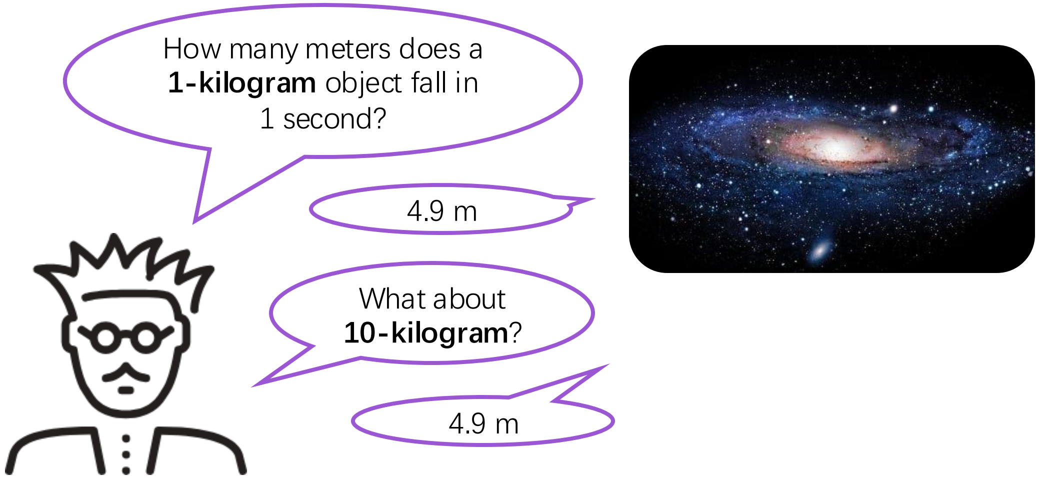
**Prompt Engineering** → **In-Context Learning**

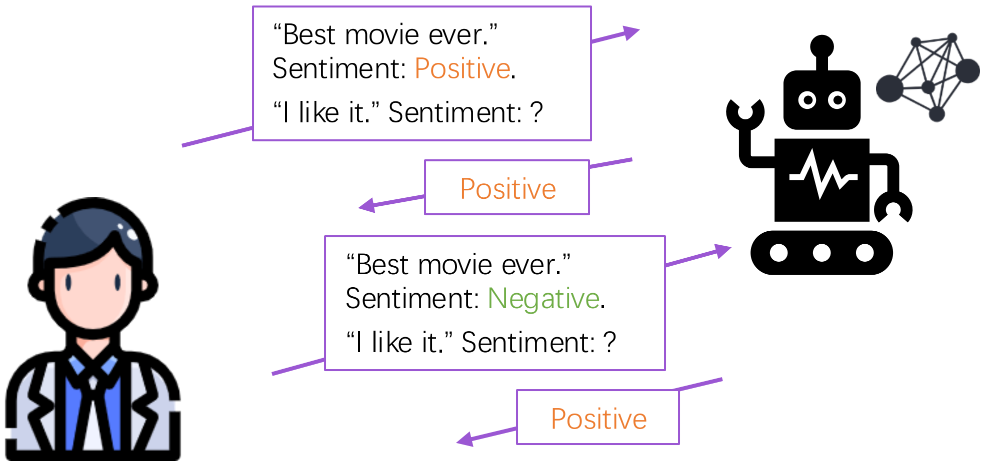
Prompt Engineering:
- Yield precise responses
- Unlock the potential of LLMs

In-Context Learning:
- A specialized prompt engineering
- Adapt to a task using a few examples

# Why In-Context Learning?

## "outside-in" methodologies to unravel the inner properties of LLMs




Objects fall with a constant acceleration due to gravity, regardless of their mass.

Providing incorrect examples does not affect the LLM's ability to make correct judgments.

**Pros of ICL**

- **Flexible controllability**
- **Encapsulate more information**

# GPT-4: Large Multimodal Model

## What is LMM?

Process visual data & understand and generate natural language

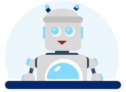
What color is the purse?

blue

Answer questions about the images

How does this food taste?

Delicious, especially the cake!



Refer to visual information in conversations

## How about GPT-4?

These two images represent two different robots, respectively···

**Excellent Multimodal capabilities**

Incorporate the understanding of visual content

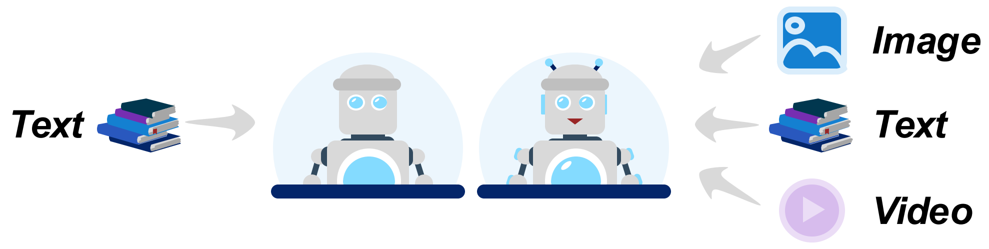
**Not open-source**

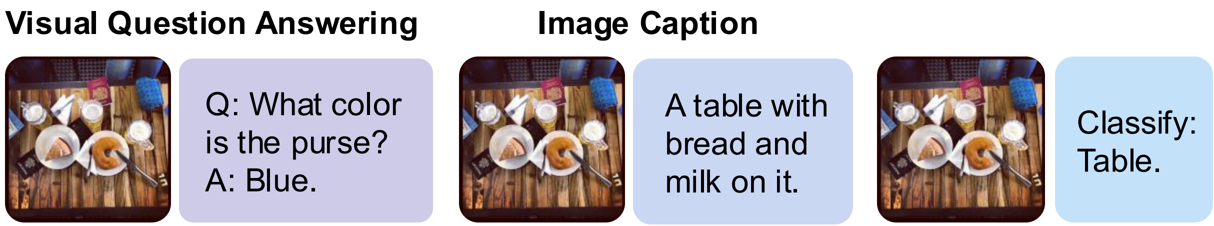
internal workings and training processes are opaque

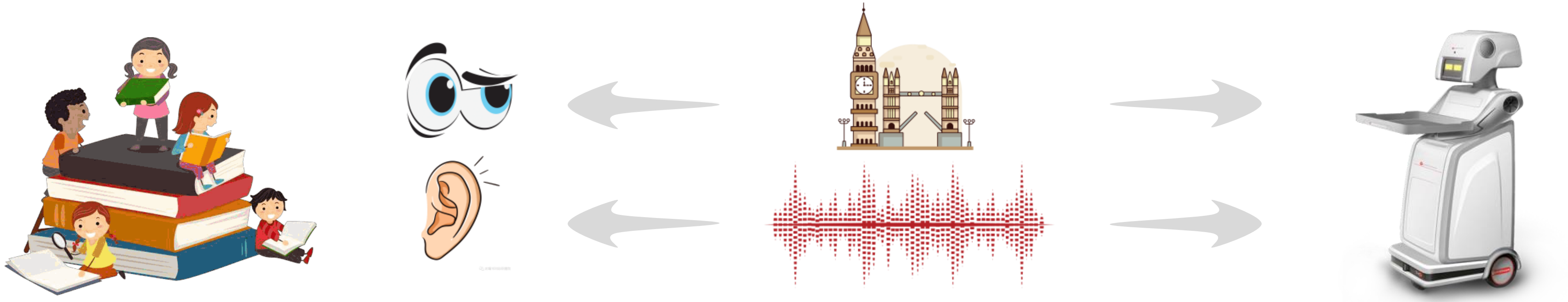
# Why Multimodal Model In-Context Learning?

The development of large models from single-modal to multi-modal

Expands the application scope of the model: various image/video understanding tasks.



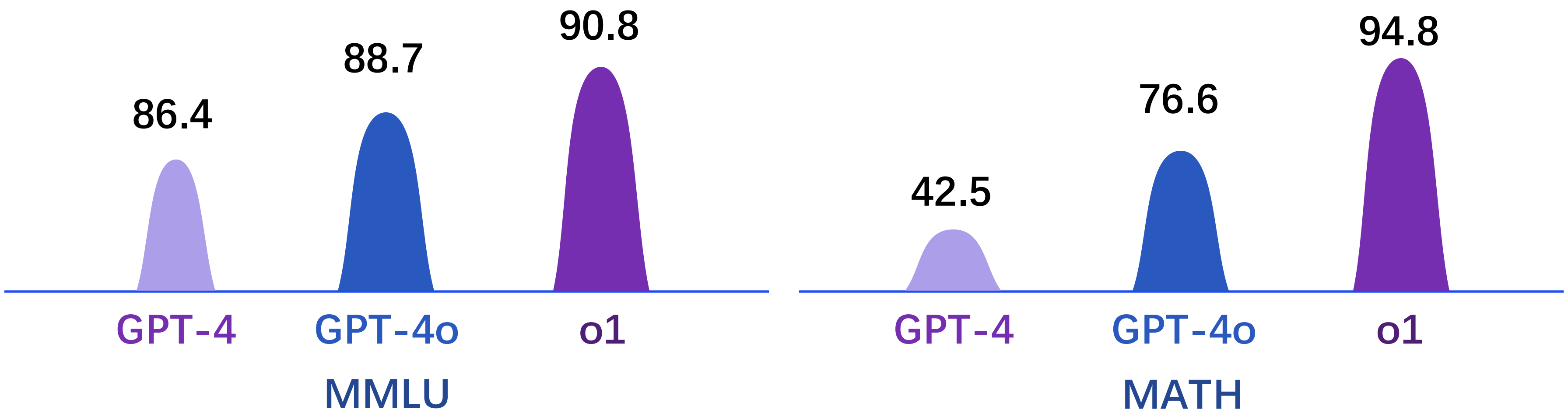



Imitate real humans and achieve multi-modal analogy capabilities

# GPT-4o & OpenAI o1: From Analogy to Reasoning

Stronger Reasoning Performance




# Deepseek-v3: Mixture-of-Experts

Training cost

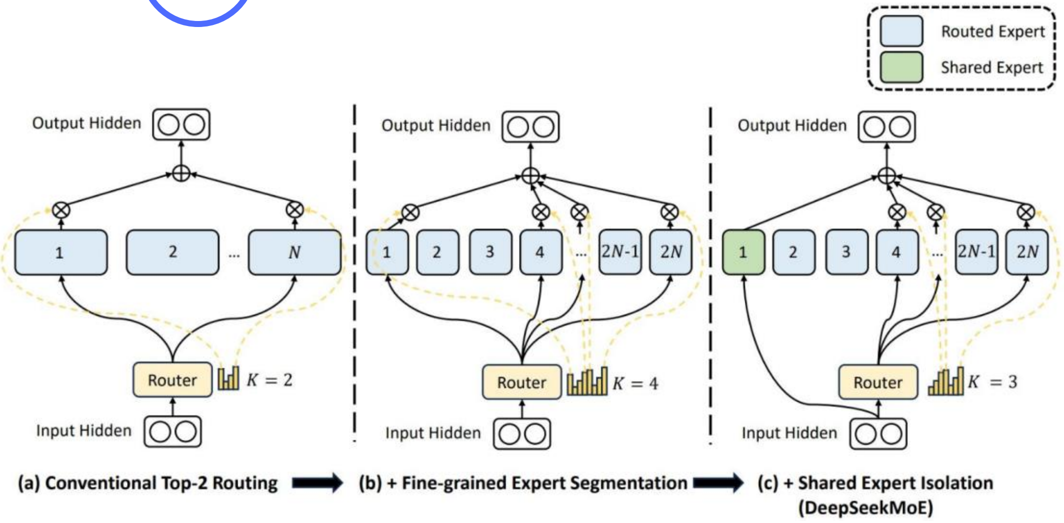
5.57 million US dollars

<< 100 million US dollars of GPT-4o

Inference cost

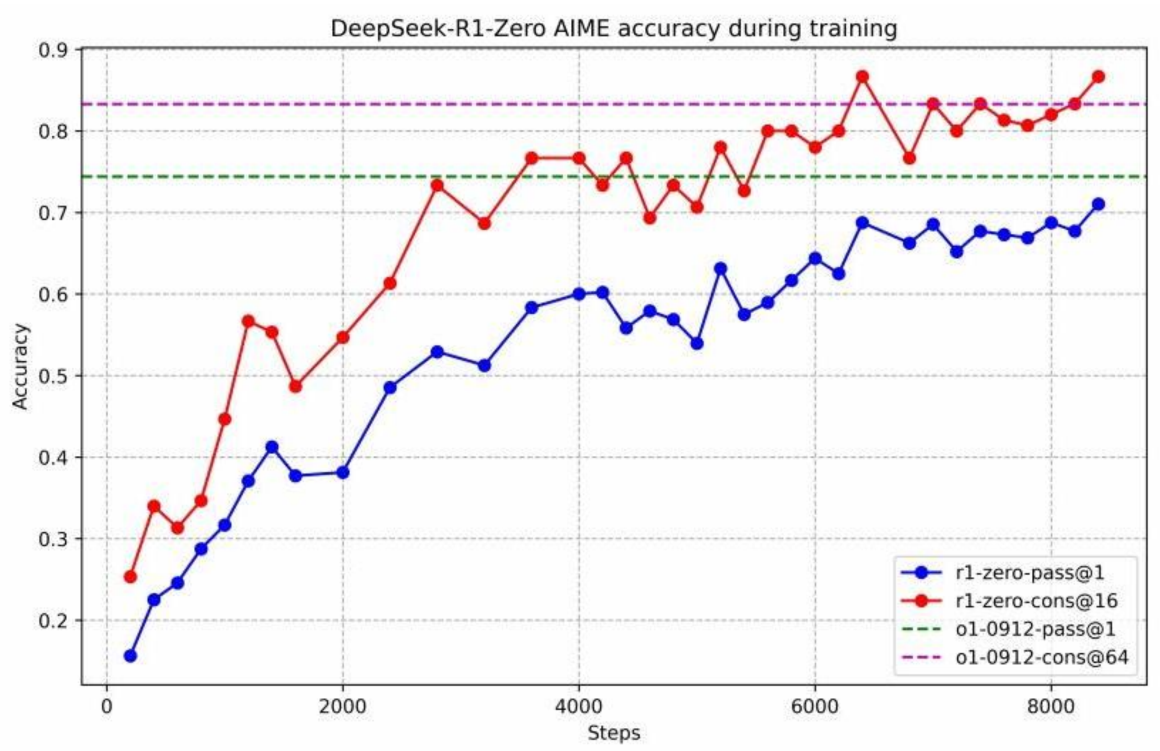
input/output cost per million tokens

= 1/10 of Sonnet-3.5



# Deepseek-R1: Rule-based Reinforcement Learning

## Reinforcement Learning Driven Reasoning Ability




## Model Distillation and Miniaturization

R1 supports distilling reasoning ability into smaller models

Qwen　Llama　...

Open source and flexibility

Suitable for local deployment and applications

# PART 02

# Heuristic-based configuration strategies

# How to Configure Good In-Context Sequence for Visual Question Answering

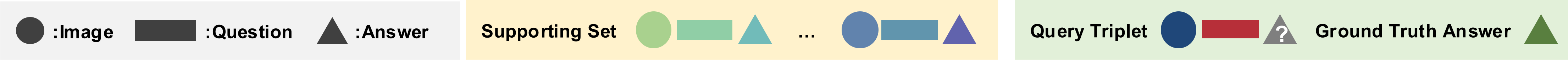
Li Li, Jiawei Peng, Huiyi Chen, Chongyang Gao, Xu Yang

arXiv: https://arxiv.org/abs/2312.01571

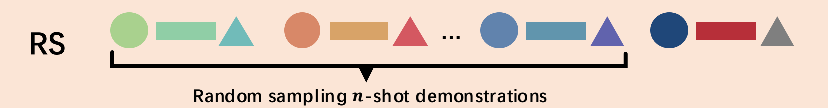
code: https://github.com/GaryJiajia/OFv2_ICL_VQA

# How to Configure Good In-Context Sequence for VQA: Approach

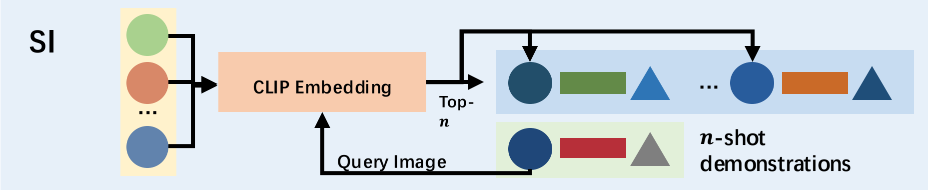
## Retrieving In-context examples



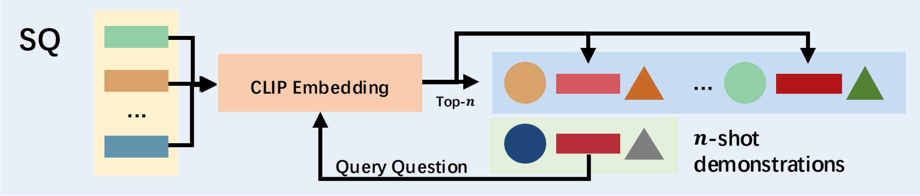

### Random Sampling (RS)



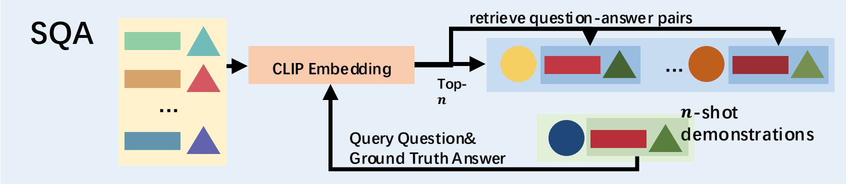

### Retrieving via Similar Image (SI)



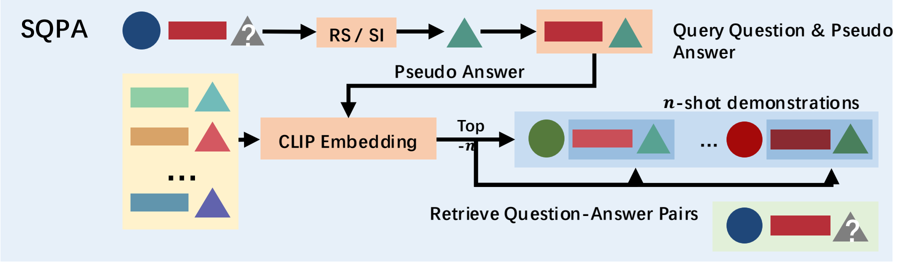

### Retrieving via Similar Questions (SQ)




### Retrieving via Similar Question&Answer (SQA)




### Retrieving via Similar Question&Pseudo Answer(SQPA)

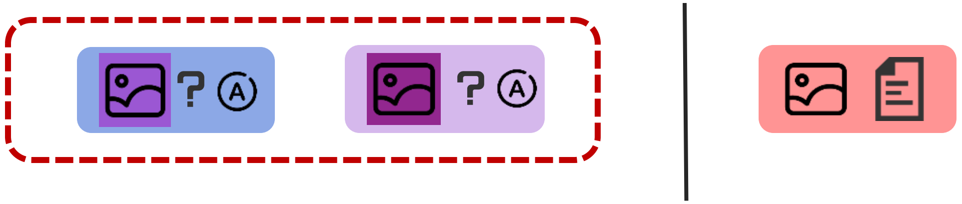


# How to Configure Good In-Context Sequence for VQA: Approach

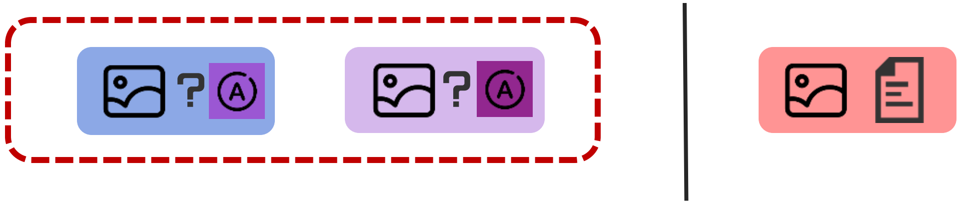
## Manipulating examples

### Mismatching the Triplet

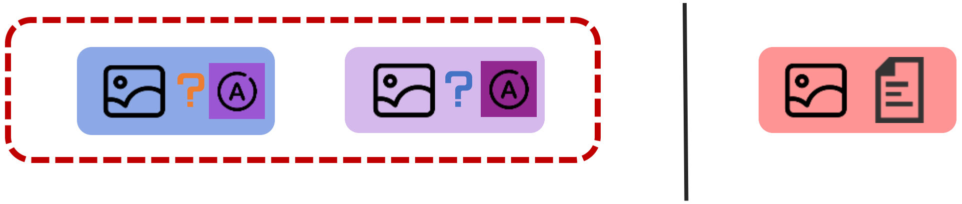
Mismatching Image (MI)



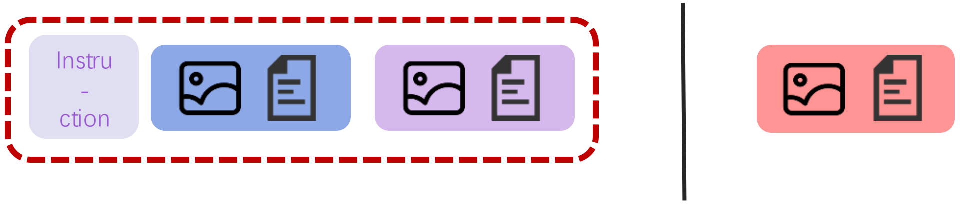
Mismatching Answer (MA)



Mismatching Question-Answer pair (MQA)



### Using Instructions




e.g.

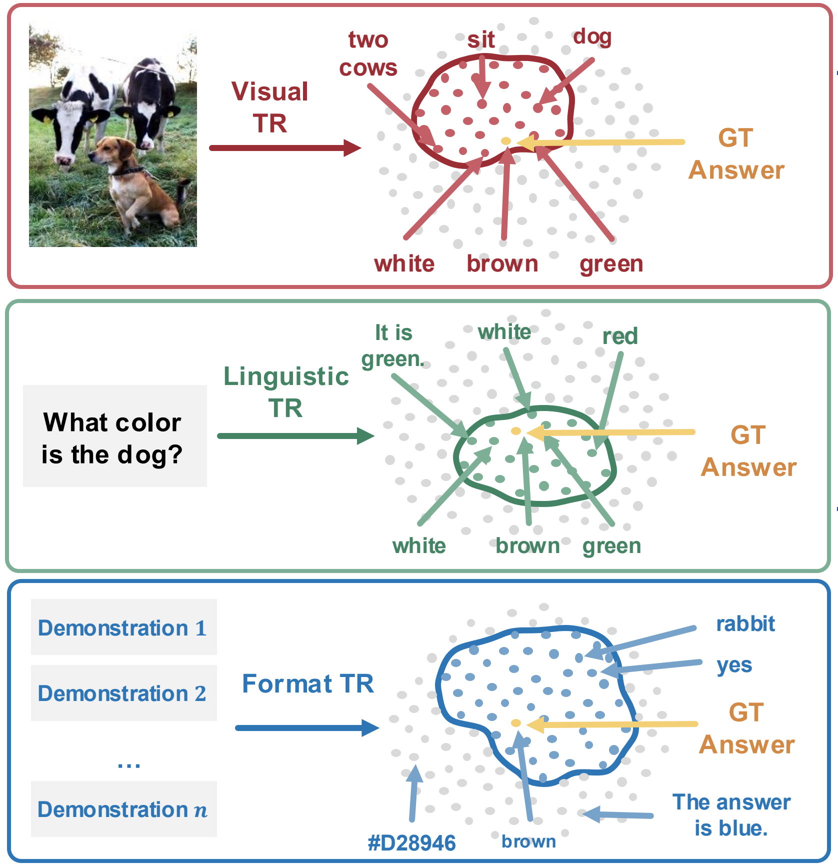
According to the previous question and answer pair, answer the final question.

<image>Question:What number is on the bus? Short Answer:284<|endofchunk|>

<image>Question:Where would a taxi park to wait for a customer? Short Answer:curb<|endofchunk|>

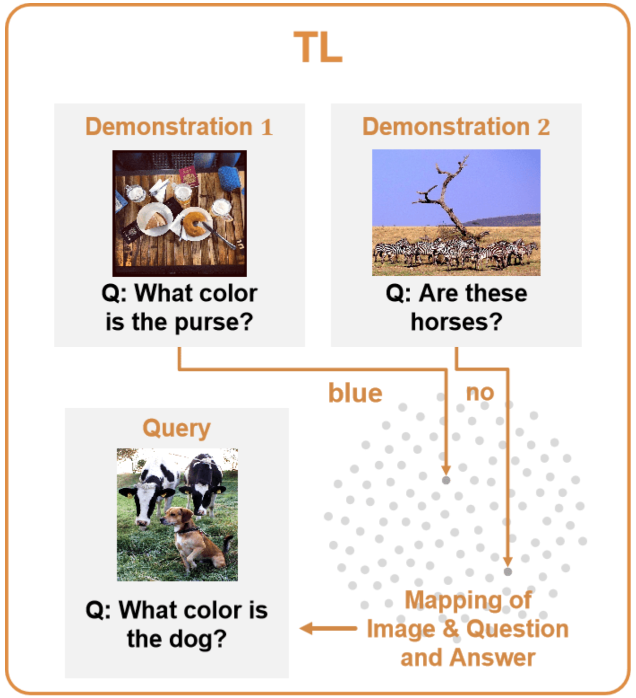
<image>Question:What is the man doing in the street? Short Answer:

# How to Configure Good In-Context Sequence for VQA




The recall of **pre-trained visual** / **language** knowledge

**Identify:**
task format,
input distribution
label space from
demonstrations




Treats QAs from demonstrations as "**training samples**"

**Implicit learning process** analogous to explicit fine-tuning

**Task Recognition**

Recognizes the **distribution of the task.** Applying **pre-trained priors** of LLM

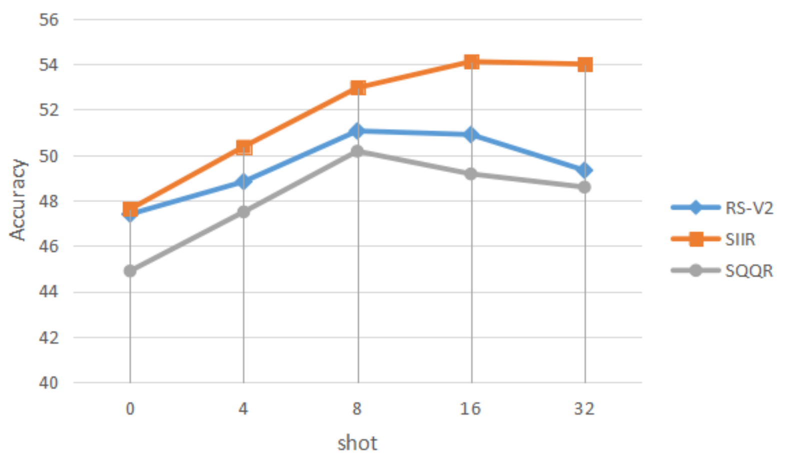
**Task Learning**

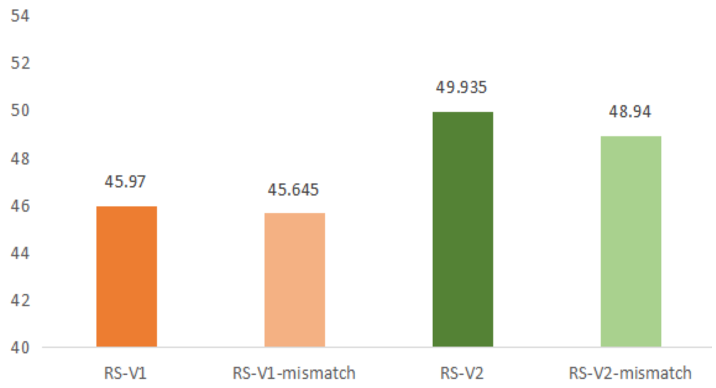
**Learn the mapping relationship** between QA pairs from the demonstrations

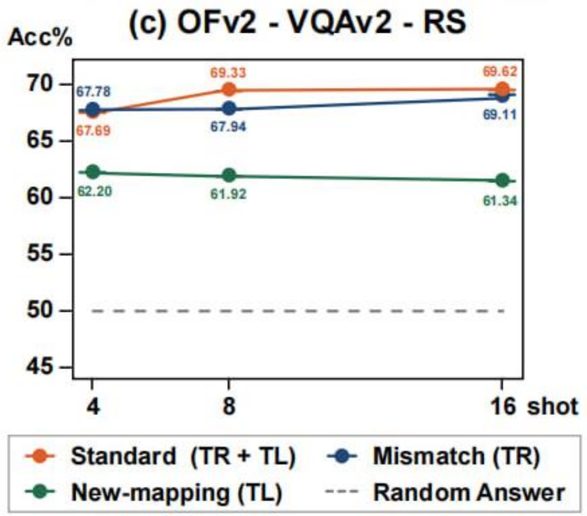
# How to Configure Good In-Context Sequence for VQA: Analysis

## Three important inner properties of LVLM during ICL

### 1. Limited TL capabilities




- As the **number** of **shots increases**, the improvement of the **model diminishes**
- **Replacing incorrect answers** in demonstrations did **not** significantly **impact** the model's performance.
- Disentangle TR and TL and find that the accuracy of **TR** is significantly **higher than TL**

# How to Configure Good In-Context Sequence for VQA: Analysis

## Three important inner properties of LVLM during ICL

### 2. The presence of a short-cut effect

SQ

Q: What is the design on the sheets?

A: alligators and bears

Q: What is the design of the bed cover?

A: alligators and bears

GT: zebra

SQ

Q: What is the scientific name of this leaf?

A: tulip

Q: What is the scientific name of this leaf?

A: tulip

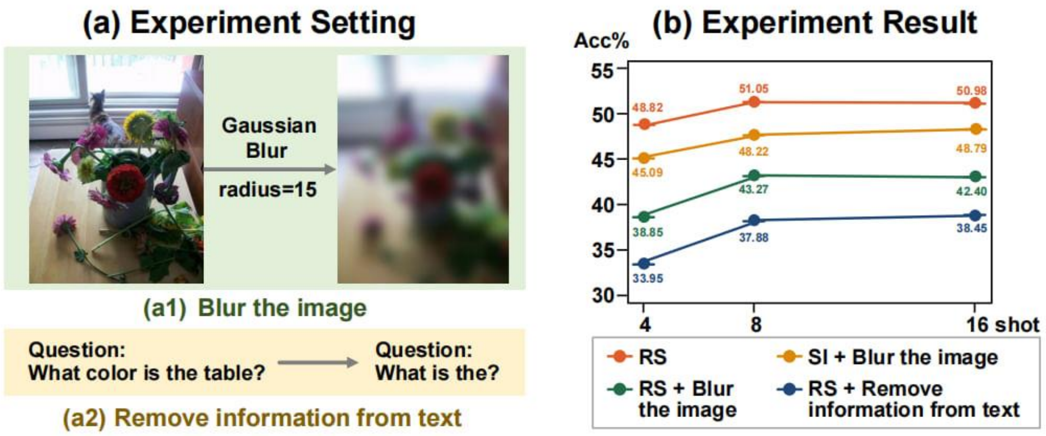
GT: camellia

| Copy rate(%) | OFv1 | OFv2 |
|---|---|---|
| RS | 43.64 | 37.34 |
| SI | 50.44 | 54.38 |
| SQ | 77.26 | 79.84 |
| SQA | 87.74 | 89.47 |
| SQA(sole) | 47.39 | 45.82 |
| SQA(sole wrong) | 37.07 | 45.71 |

# How to Configure Good In-Context Sequence for VQA: Analysis

## Three important inner properties of LVLM during ICL

### 3. Partial compatibility between vision and language modules




| | 4-shot | 8-shot | 16-shot |
|---|---|---|---|
| RS(OFv1) | 44.56 | 47.38 | 48.71 |
| instruct1(OFV1) | 43.75 | 46.91 | 48.67 |
| RS(OFv2) | 48.82 | 51.05 | 50.89 |
| instruct1(OFv2) | **49.93** | **52.71** | **50.95** |

linguistic TR plays a more substantial role than visual TR

Some language reasoning ability lose efficacy in the VL case

# PART 03

# Shift vector-based in-context learning approximation

# LIVE: Learnable In-Context Vector for Visual Question Answering

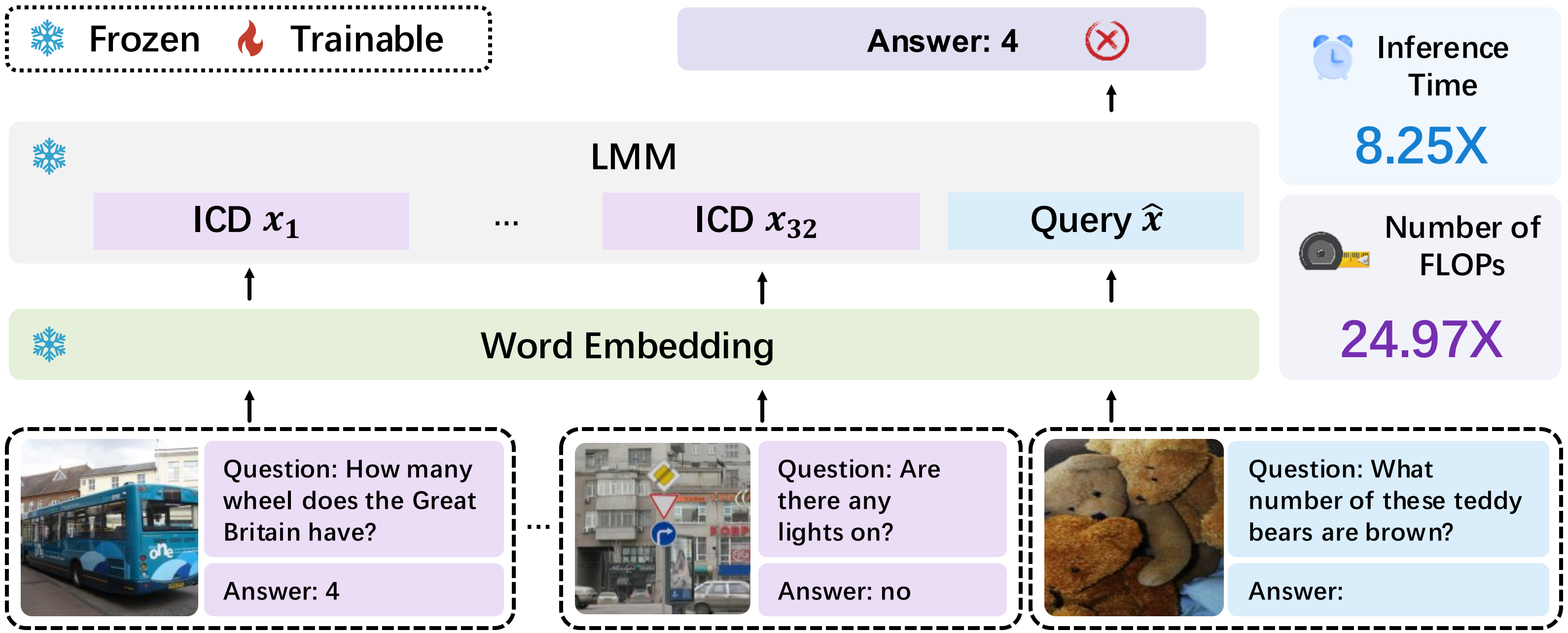
Yingzhe Peng, Chenduo Hao, Xinting Hu, Jiawei Peng, Xin Geng, Xu Yang

arXiv: https://arxiv.org/pdf/2406.13185

code: https://github.com/ForJadeForest/LIVE-Learnable-In-Context-Vector

# LIVE: Learnable In-Context Vector for Visual Question Answering

## Traditional ICL




Conventional In-Context Learning (ICL)

- Sensitive to ICD selection and requires more inference time
- There is inherent misalignment between different modalities, making multimodal tasks more difficult

## Self Attention Break Down

- Consider self-attention ($SA$) of a specific head :

$$
\begin{aligned}
&\mathrm{SA}\left(\boldsymbol{q}, \begin{bmatrix}\boldsymbol{K}_D\\ \boldsymbol{K}\end{bmatrix}, \begin{bmatrix}\boldsymbol{V}_D\\ \boldsymbol{V}\end{bmatrix}\right)\\
&= \mathrm{softmax}([\boldsymbol{q}\boldsymbol{K}_D^\top, \boldsymbol{q}\boldsymbol{K}^\top]) \begin{bmatrix}\boldsymbol{V}_D\\ \boldsymbol{V}\end{bmatrix}\\
&= (1-\mu)\mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}, \boldsymbol{V}) + \mu\mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}_D, \boldsymbol{V}_D)\\
&= \underbrace{\mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}, \boldsymbol{V})}_{\text{standard attention}} + \underbrace{\mu(\mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}_D, \boldsymbol{V}_D) - \mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}, \boldsymbol{V}))}_{\text{shift vector}}
\end{aligned}
$$

$$\mu(\boldsymbol{q}, \boldsymbol{K}_D, \boldsymbol{K}) = \frac{Z_1(\boldsymbol{q}, \boldsymbol{K}_D)}{Z_1(\boldsymbol{q}, \boldsymbol{K}_D) + Z_2(\boldsymbol{q}, \boldsymbol{K})}$$

$$Z_1 = \sum_i \exp(\boldsymbol{q}\boldsymbol{K}_D^\top)_i$$

$$Z_2 = \sum_j \exp(\boldsymbol{q}\boldsymbol{K}^\top)_j$$

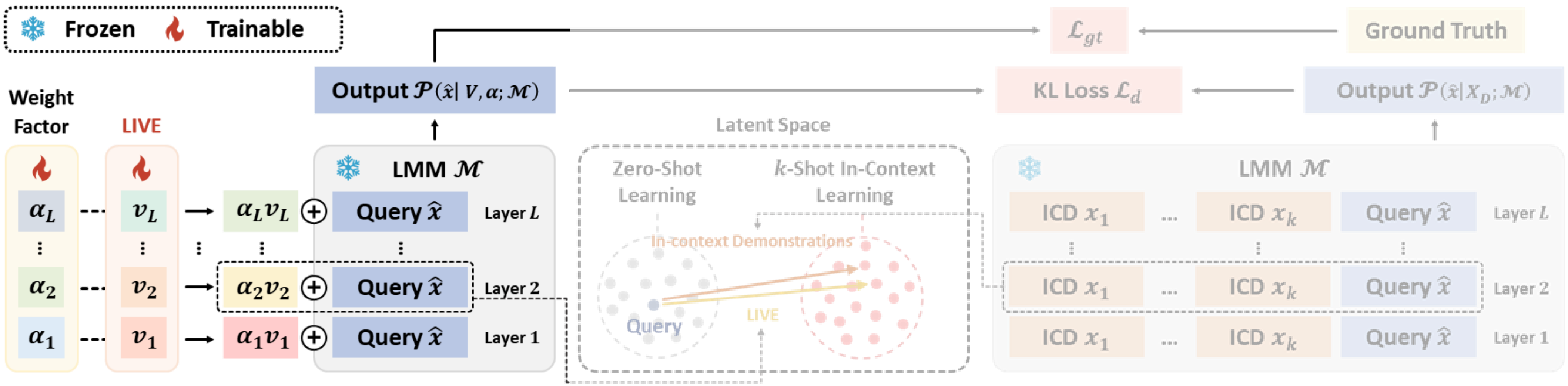
**Observation:**

Demonstrations can be seen as shift vectors on zero-shot attention.

$\boldsymbol{K}_D/\boldsymbol{V}_D$: The key/value of demonstrations
$\boldsymbol{K}/\boldsymbol{V}$: The key/value of query input
$\boldsymbol{q}$: A token in query input

# LIVE: Learnable In-Context Vector for Visual Question Answering

## Overall Framework



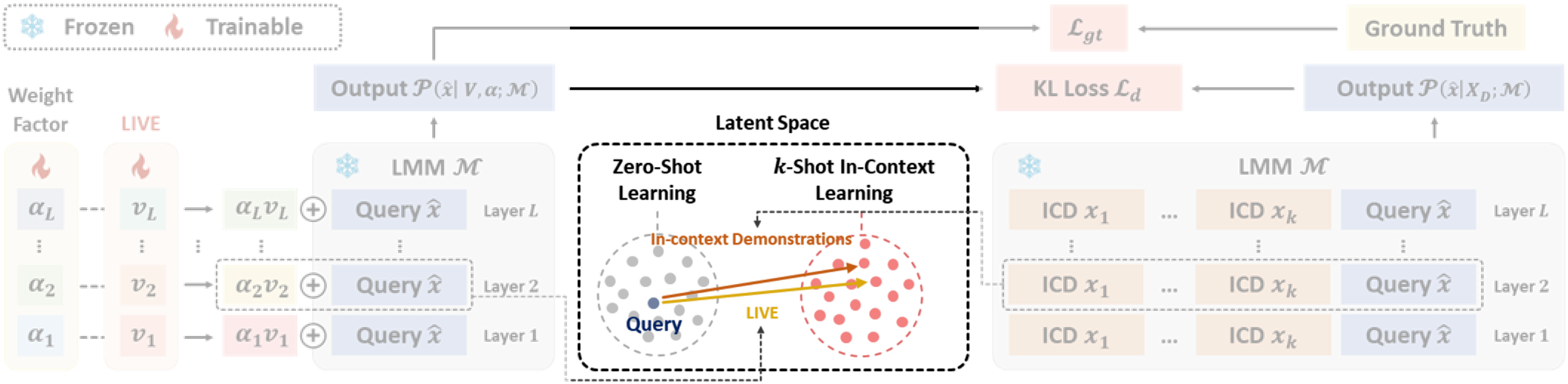

a) **Learnable in-context vector.**
- Set $L$ learnable vectors $v_i$ and corresponding weights $\alpha_i$.

b) **LIVE intervention.**
- Add $\alpha_i v_i$ to the output of $i^{th}$ decoder layer.
- In the latent space, transforming a zero-shot query into k-shot in-context learning.

c) **Align with conventional ICL.**
- Train LIVE using a language modeling loss $\mathcal{L}_{gt}$ and a KL loss $\mathcal{L}_d$ for distillation.
- $\mathcal{L}_d$ minimizes the divergence between the model's output under zero-shot setting and ICL.

# LIVE: Learnable In-Context Vector for Visual Question Answering

## Overall Framework




a) Learnable in-context vector.

- Set $L$ learnable vectors $v_i$ and corresponding weights $\alpha_i$.

b) **LIVE intervention.**

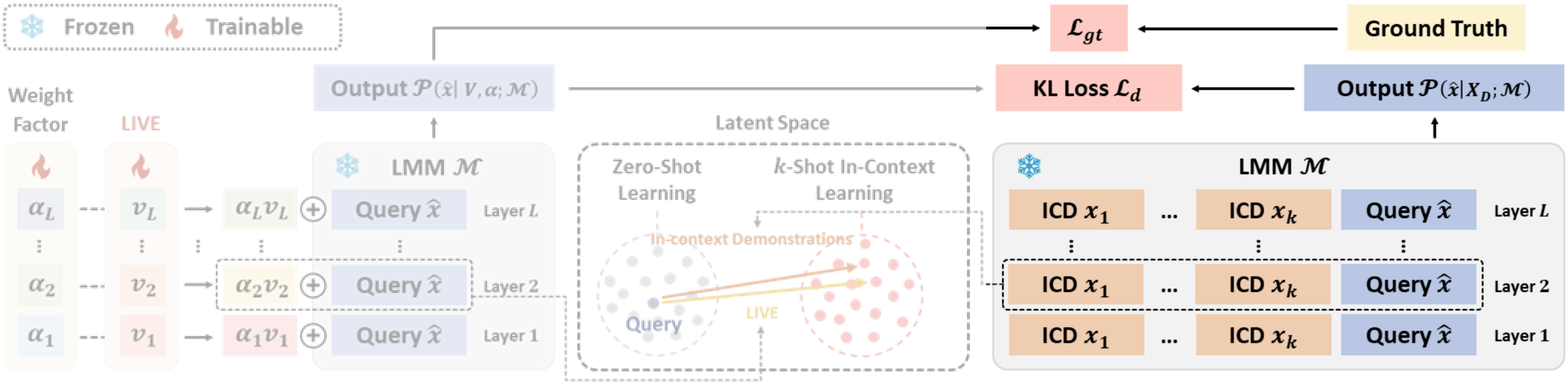
- Add $\boldsymbol{\alpha_i v_i}$ to the output of $i^{th}$ decoder layer.
- In the latent space, transforming a zero-shot query into k-shot in-context learning.

c) Align with conventional ICL.

- Train LIVE using a language modeling loss $\mathcal{L}_{gt}$ and a KL loss $\mathcal{L}_d$ for distillation.
- $\mathcal{L}_d$ minimizes the divergence between the model's output under zero-shot setting and ICL.

# LIVE: Learnable In-Context Vector for Visual Question Answering

## Overall Framework




a) **Learnable in-context vector.**
- Set $L$ learnable vectors $v_i$ and corresponding weights $\alpha_i$.

b) **LIVE intervention.**
- Add $\alpha_i v_i$ to the output of $i^{th}$ decoder layer.
- In the latent space, transforming a zero-shot query into k-shot in-context learning.

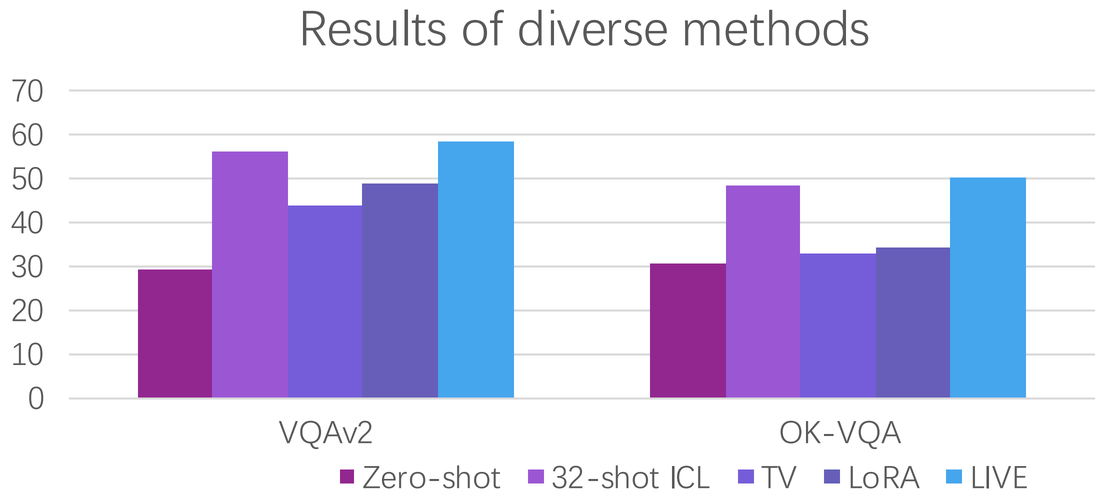
c) **Align with conventional ICL.**
- Train LIVE using a language modeling loss $\mathcal{L}_{gt}$ and a KL loss $\mathcal{L}_d$ for distillation.
- $\mathcal{L}_d$ minimizes the divergence between the model's output under zero-shot setting and ICL.

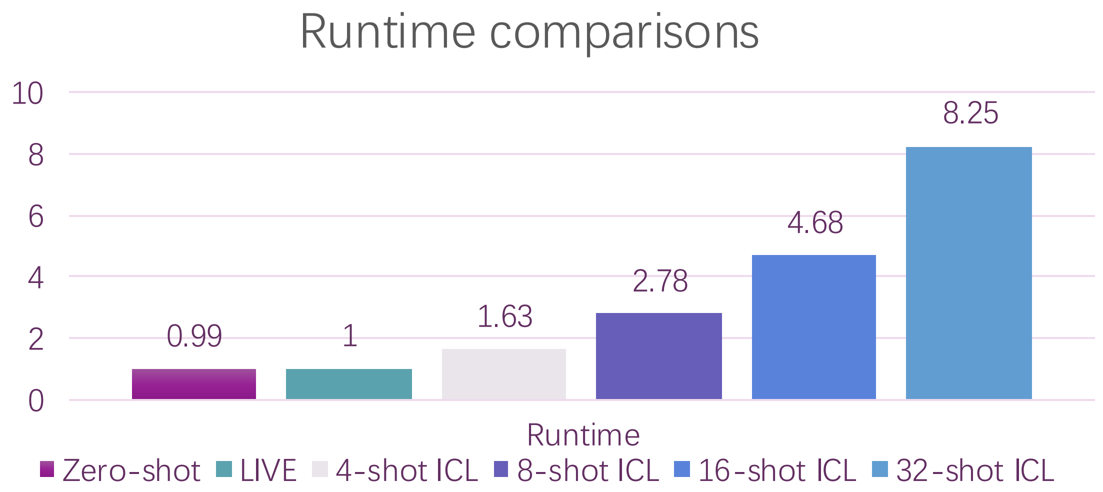
# LIVE: Learnable In-Context Vector for Visual Question Answering

## Main Results

We conduct experiments using Idefics-9b model across VQAv2 and OK-VQA.

- LIVE achieve **the best performance** compared with other methods.
- LIVE maintains almost **the same inference speed as zero-shot** while achieving 32-shot ICL performance.



# MimIC: Mimic In-Context Learning for Multimodal Tasks

Yingzhe Peng, Jiale Fu, Chenduo Hao, Xinting Hu, Yingzhe Peng, Xin Geng, Xu Yang

code: https://github.com/Kamichanw/MimIC

# MimIC: Mimic In-Context Learning for Multimodal Tasks

## More Elegant Approximation

- Consider self-attention ($\mathrm{SA}$) of a specific head :

$$
\begin{aligned}
&\mathrm{SA}\left(\boldsymbol{q}, \begin{bmatrix}\boldsymbol{K}_D \\ \boldsymbol{K}\end{bmatrix}, \begin{bmatrix}\boldsymbol{V}_D \\ \boldsymbol{V}\end{bmatrix}\right) \\
&= \mathrm{softmax}([\boldsymbol{q}\boldsymbol{K}_D^\top, \boldsymbol{q}\boldsymbol{K}^\top]) \begin{bmatrix}\boldsymbol{V}_D \\ \boldsymbol{V}\end{bmatrix} \\
&= (1-\mu)\mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}, \boldsymbol{V}) + \mu \mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}_D, \boldsymbol{V}_D) \\
&= \underbrace{\mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}, \boldsymbol{V})}_{\text{standard attention}} + \underbrace{\mu(\mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}_D, \boldsymbol{V}_D) - \mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}, \boldsymbol{V}))}_{\text{shift vector}}
\end{aligned}
$$

$$
\mu(\boldsymbol{q}, \boldsymbol{K}_D, \boldsymbol{K}) = \frac{Z_1(\boldsymbol{q}, \boldsymbol{K}_D)}{Z_1(\boldsymbol{q}, \boldsymbol{K}_D) + Z_2(\boldsymbol{q}, \boldsymbol{K})}
$$

$$
Z_1 = \sum_i \exp(\boldsymbol{q}\boldsymbol{K}_D^\top)_i
$$

$$
Z_2 = \sum_j \exp(\boldsymbol{q}\boldsymbol{K}^\top)_j
$$

**Additional Observation:**

1. Only $Z_1$ and $\mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}_D, \boldsymbol{V}_D)$ are related to demonstrations.
2. Shifts should be multi-head and inserted after self-attention layers.

$\boldsymbol{K}_D/\boldsymbol{V}_D$: The key/value of demonstrations
$\boldsymbol{K}/\boldsymbol{V}$: The key/value of query input
$\boldsymbol{q}$: A token in query input

# MimIC: Mimic In-Context Learning for Multimodal Tasks

## Mimicking Demonstration Affected Terms

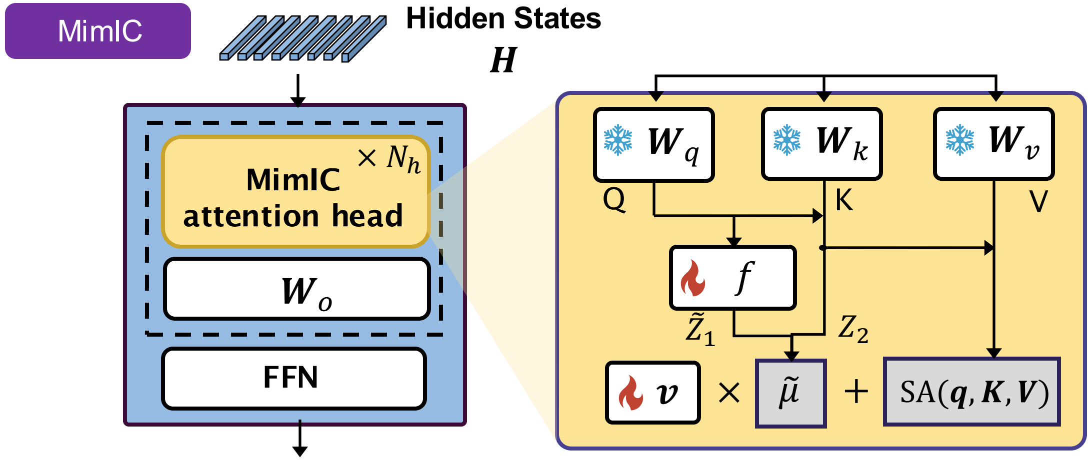
a. Attention difference term

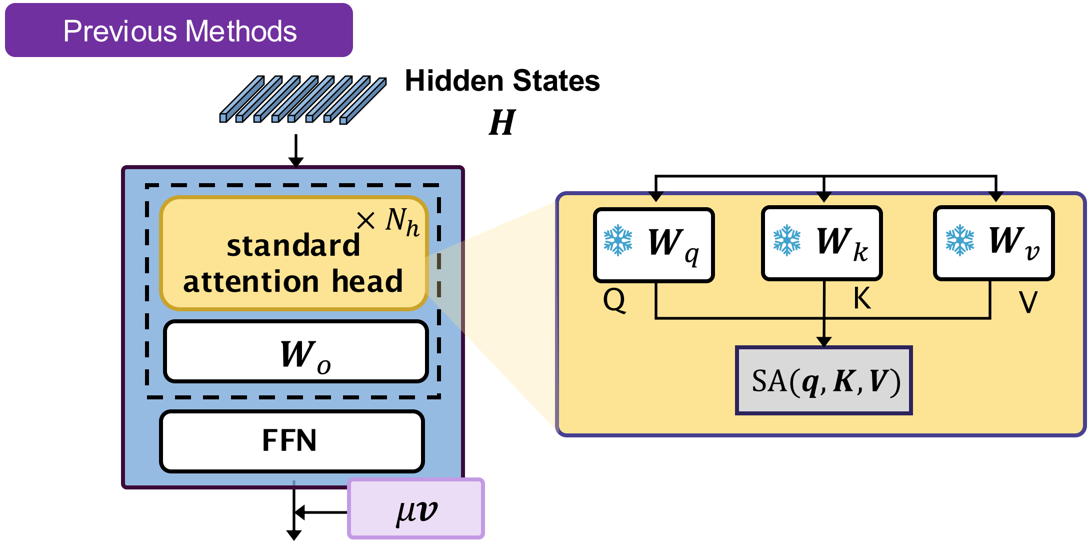
$$\mathrm{SA}\left(\boldsymbol{q}, \begin{bmatrix}\boldsymbol{K}_D \\ \boldsymbol{K}\end{bmatrix}, \begin{bmatrix}\boldsymbol{V}_D \\ \boldsymbol{V}\end{bmatrix}\right) = \mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}, \boldsymbol{V}) + \mu\underbrace{(\mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}_D, \boldsymbol{V}_D) - \mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}, \boldsymbol{V}))}_{\boxed{\boldsymbol{v}}}$$

b. Normalized attention weights

$$\mu(\boldsymbol{q}, \boldsymbol{K}_D, \boldsymbol{K}) = \frac{Z_1(\boldsymbol{q}, \boldsymbol{K}_D)}{\underbrace{Z_1(\boldsymbol{q}, \boldsymbol{K}_D)}_{\boxed{f}} + Z_2(\boldsymbol{q}, \boldsymbol{K})}$$

$$Z_1 = \sum_i \exp(\boldsymbol{q}\boldsymbol{K}_D^\top)_i \qquad Z_2 = \sum_j \exp(\boldsymbol{q}\boldsymbol{K}^\top)_j$$



# MimIC: Mimic In-Context Learning for Multimodal Tasks

## Mimicking Demonstration Affected Terms

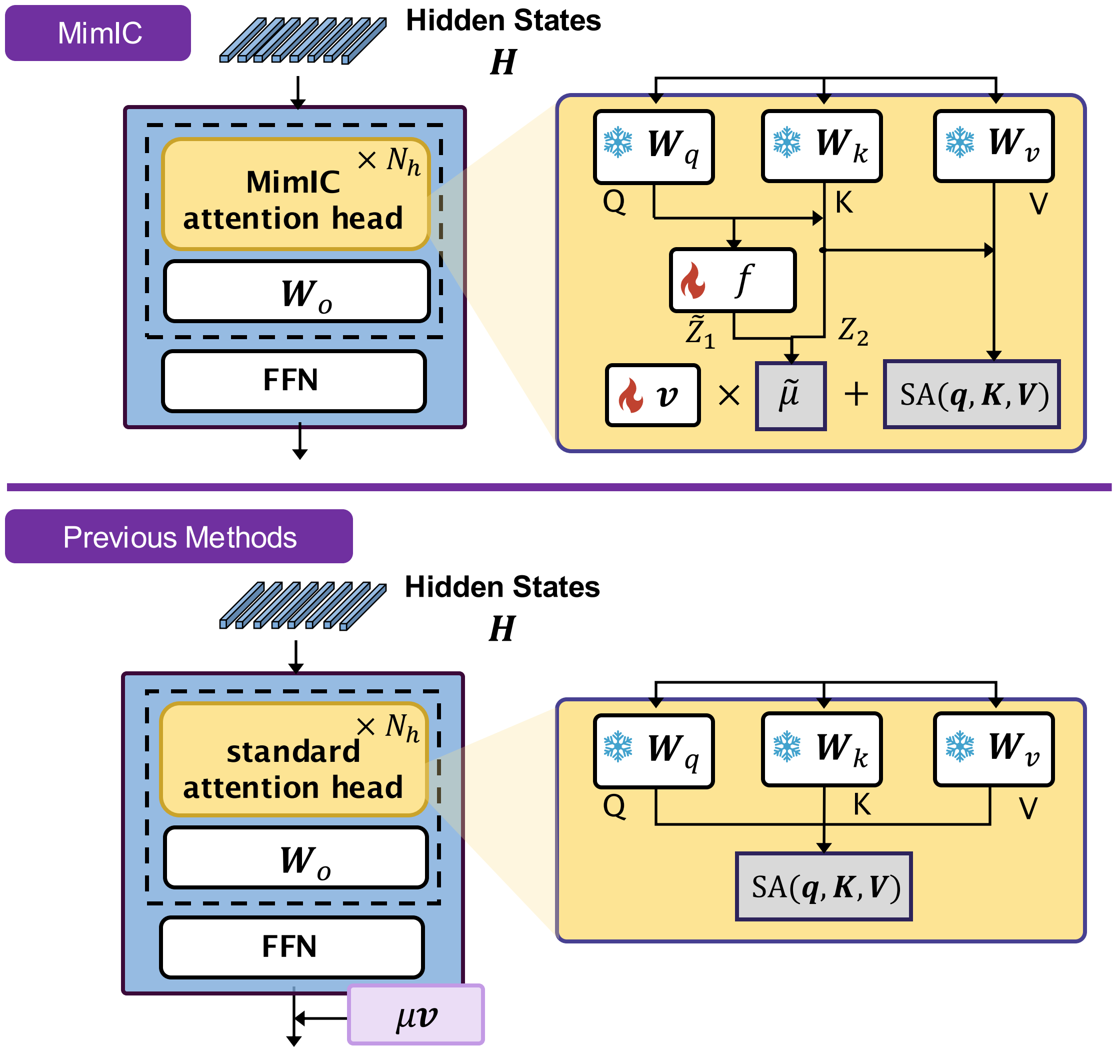
$$\mathrm{SA}\left(\boldsymbol{q},\begin{bmatrix}\boldsymbol{K}_D\\ \boldsymbol{K}\end{bmatrix},\begin{bmatrix}\boldsymbol{V}_D\\ \boldsymbol{V}\end{bmatrix}\right)$$
$$= \mathrm{SA}(\boldsymbol{q},\boldsymbol{K},\boldsymbol{V}) + \mu\big(\mathrm{SA}(\boldsymbol{q},\boldsymbol{K}_D,\boldsymbol{V}_D) - \mathrm{SA}(\boldsymbol{q},\boldsymbol{K},\boldsymbol{V})\big)$$

$\boxed{f}$ $\boxed{\boldsymbol{v}}$

$$= \mathrm{SA}(\boldsymbol{q},\boldsymbol{K},\boldsymbol{V}) + \tilde{\mu}(\boldsymbol{q},\boldsymbol{K})\cdot\boldsymbol{v}$$

$$\tilde{\mu}(\boldsymbol{q},\boldsymbol{K}) = \frac{\tilde{Z}_1(\boldsymbol{q})}{\tilde{Z}_1(\boldsymbol{q}) + Z_2(\boldsymbol{q},\boldsymbol{K})} \qquad \tilde{Z}_1 = \exp(f(\boldsymbol{q}))$$

$$\mu(\boldsymbol{q},\boldsymbol{K}_D,\boldsymbol{K}) = \frac{Z_1(\boldsymbol{q},\boldsymbol{K}_D)}{Z_1(\boldsymbol{q},\boldsymbol{K}_D) + Z_2(\boldsymbol{q},\boldsymbol{K})} \qquad Z_1 = \sum_i \exp(\boldsymbol{q}\boldsymbol{K}_D^{\top})_i$$



# MimIC: Mimic In-Context Learning for Multimodal Tasks

## Training

### Strategy

1. $k$ demonstrations + one query is fed as input to the original LMM, record corresponding hidden states $\boldsymbol{H}'$.

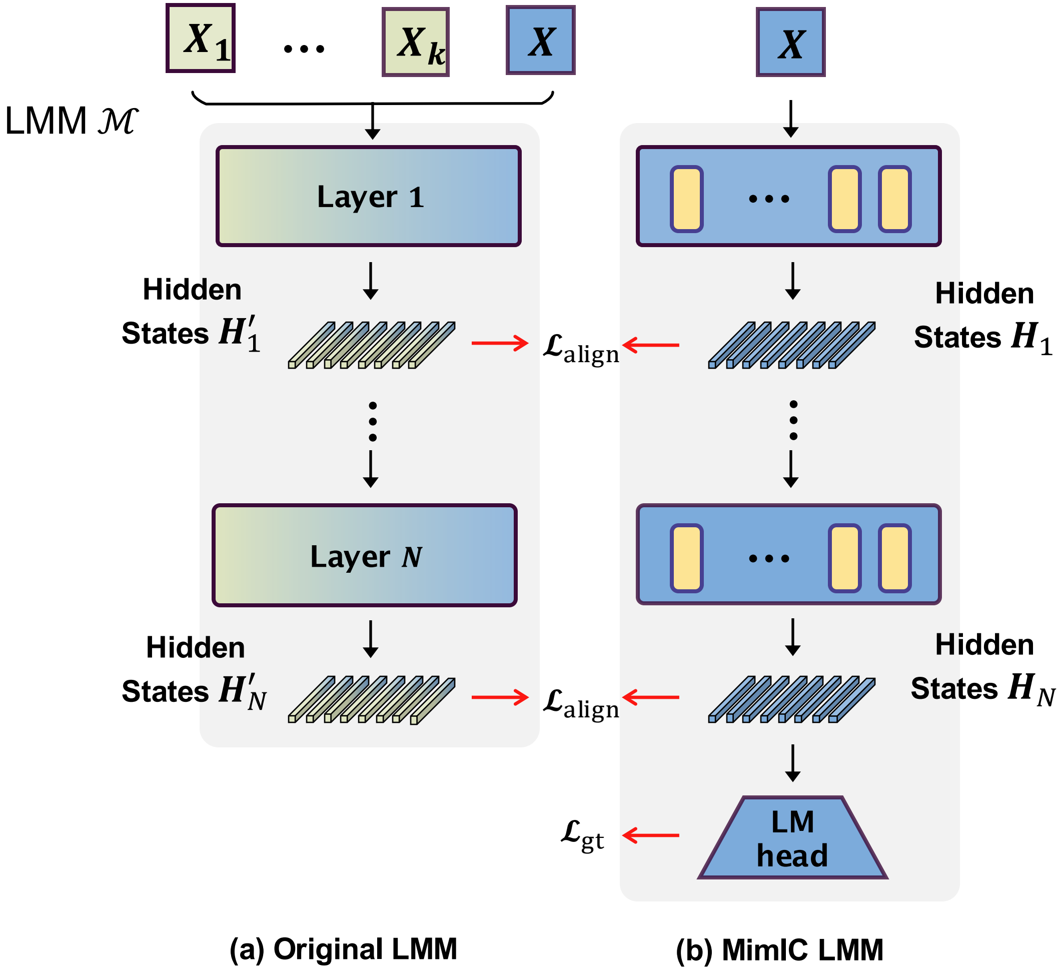
2. one query is fed as input to MimIC LMM, record corresponding hidden states $\boldsymbol{H}$.

3. Align $\boldsymbol{H}$ with $\boldsymbol{H}'$ by minimizing a composed training loss.

### Objective

$$\mathcal{L}_{\text{align}} = \frac{1}{N}\sum_{i}\sum_{j}\left\|h_{i,j} - h'_{i,j}\right\|_2^2$$

$$\mathcal{L} = \mathcal{L}_{\text{align}} + \lambda\mathcal{L}_{\text{gt}}$$

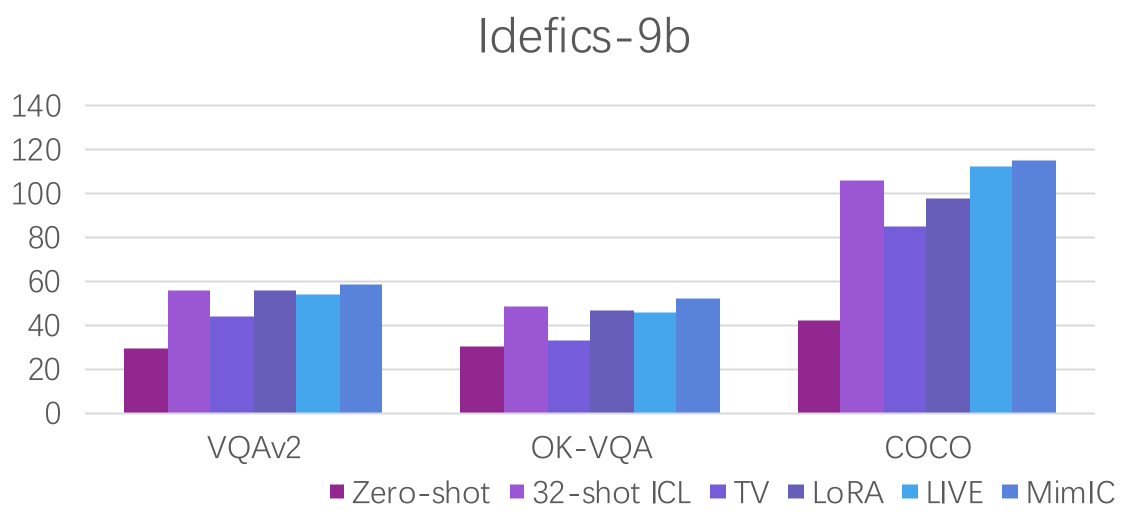


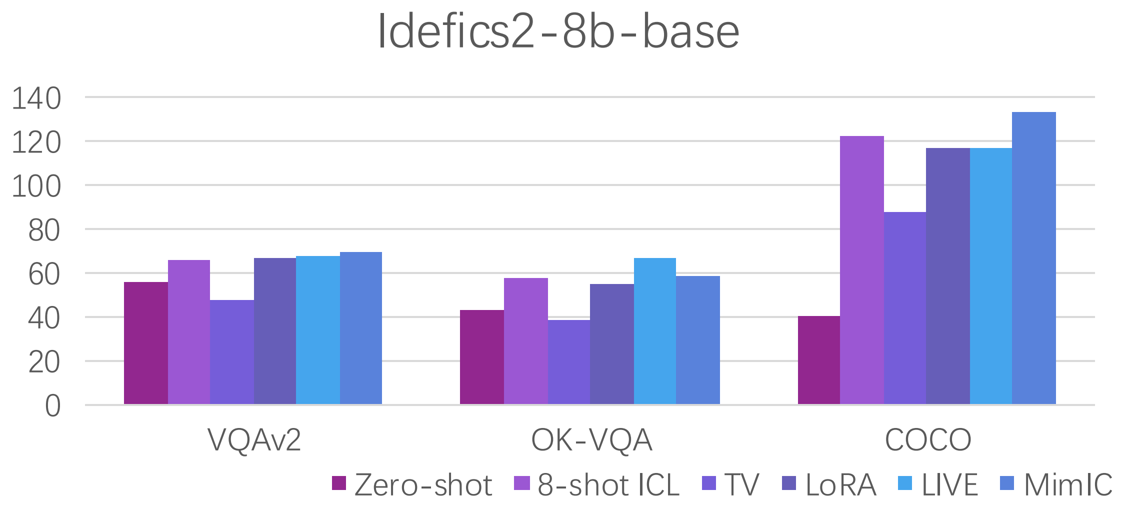
# MimIC: Mimic In-Context Learning for Multimodal Tasks

## Main Results

We conduct experiments using Idefics-9b and Idefics2-8b-base models across VQAv2, OK-VQA and COCO caption.

- MimIC achieve **the best performance compared with other methods.**
- MimIC is parameter efficient (0.26M), compared to LoRA (25M/67.7M).



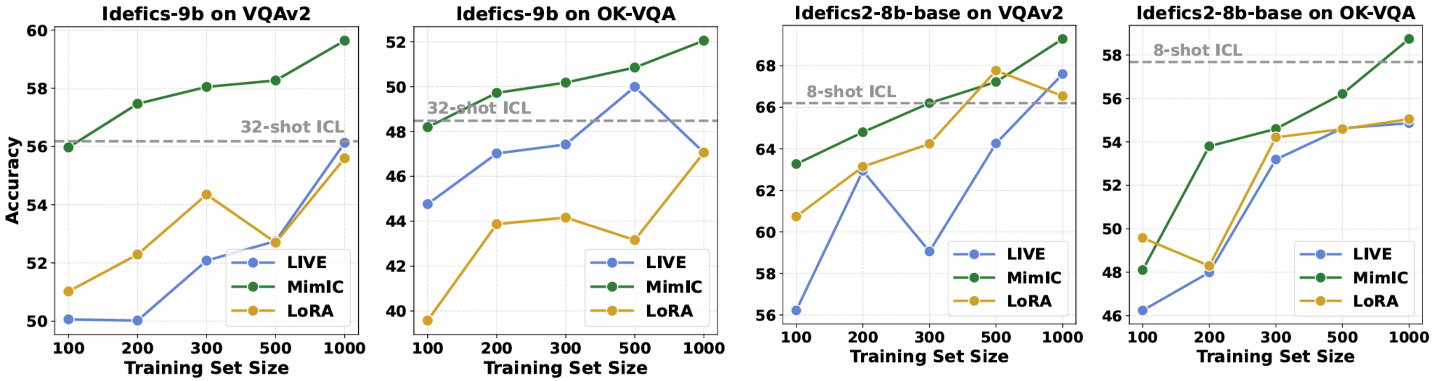



# MimIC: Mimic In-Context Learning for Multimodal Tasks

## Ablation Results

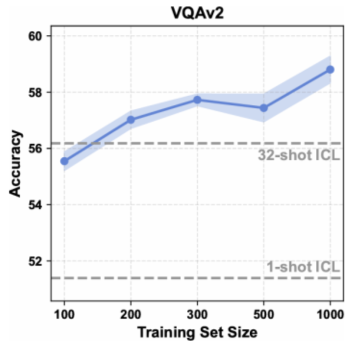
### 1. Training with Fewer Samples

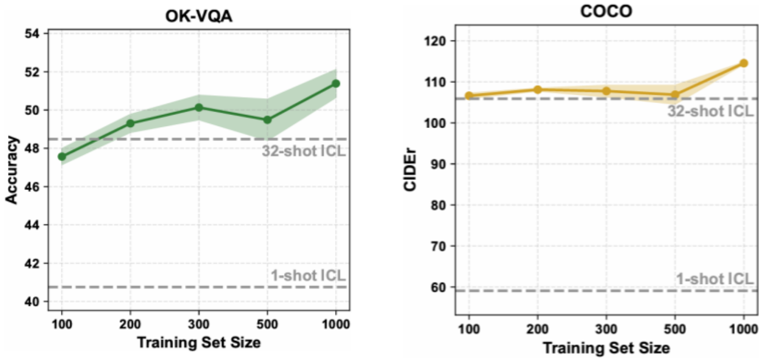
- MimIC increases stably on different training set size and different LMMs.
- MimIC can achieve few-shot ICL performance with fewer samples.




### 2. The number of shots

- MimIC is less sensitive compared to few shot ICL.



# PART 04

# Multimodal Reasoning Capability Enhancement

# LMM-R1: Empowering 3B LMMs with Strong Reasoning Abilities Through Two-Stage Rule-Based RL

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu,
Kai Yang, Xingzhong Xu, Xin Geng, Xu Yang

arXiv: https://arxiv.org/abs/2503.07536

Project Page: https://forjadeforest.github.io/LMM-R1-ProjectPage/

# LMM-R1: Motivation: Enhance the reasoning capabilities for LMMs

Inspirations of DeepSeek-R1-Zero:

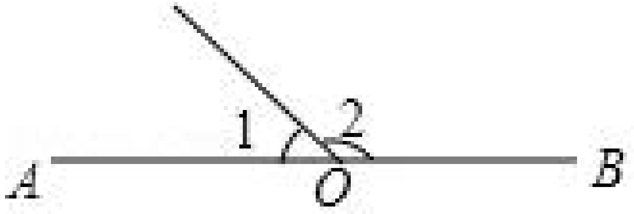
- Rule-Based RL can **boost the CoT inference performance**, which can **generalize** to other domains.

**Can we extend RL to multimodal models?**

1. Data Limitations: **ambiguous answers** & **scarce complex reasoning**

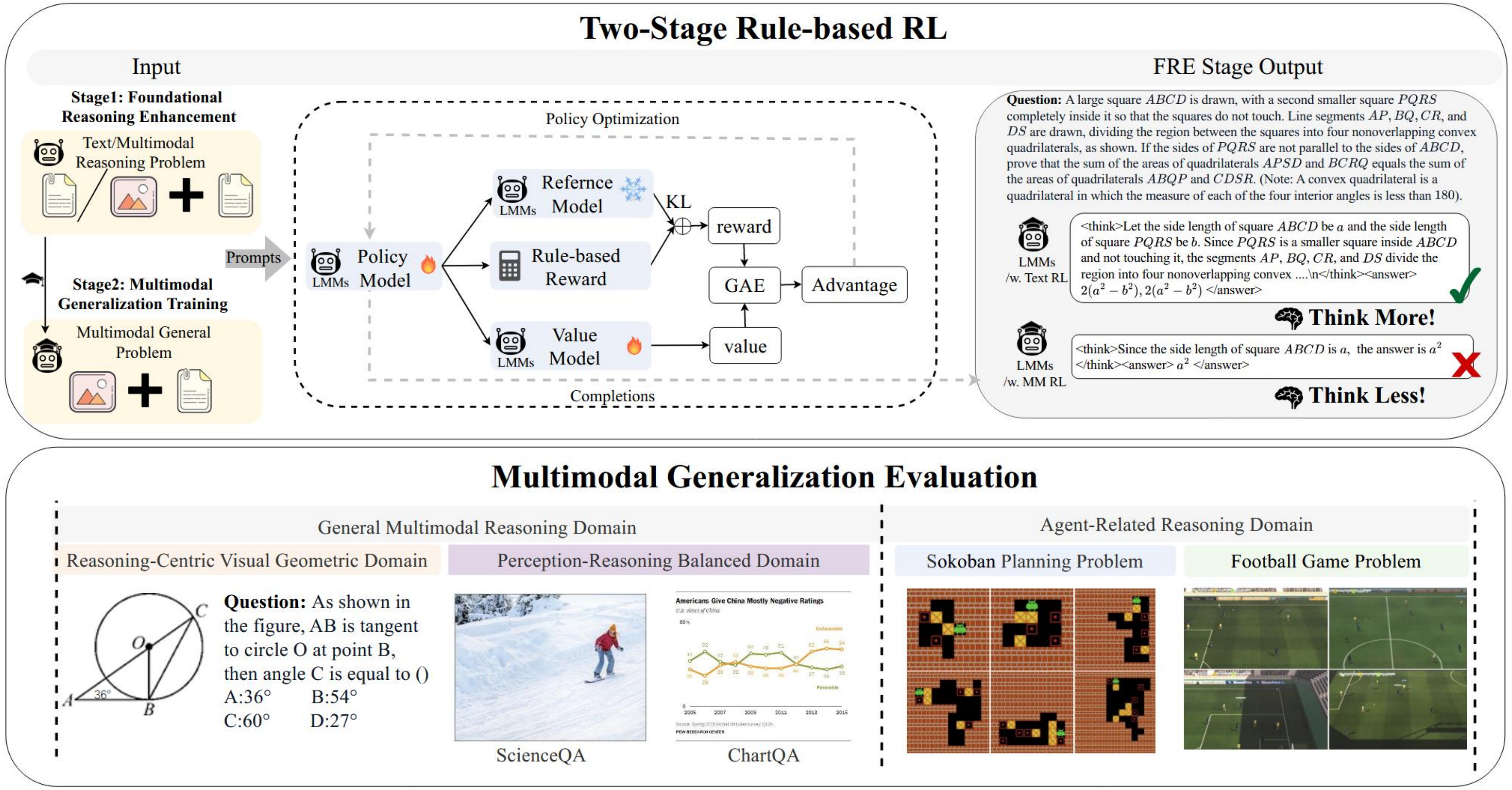
2. **Degraded** foundational reasoning induced by multimodal pretraining

Q: What's the object in image?
A1: Cat, A2: Ragdoll, A3: It's a cat.




Q: In the given diagram, if angle 1 has a measure of 35.0 degrees, what is the measure of angle 2?
A: 145 degrees.

# LMM-R1: Method: Two Stage Training Framework




Stage 1:

**Foundational Reasoning Enhancement (FRE)**:

Uses **text-only data** to develop strong reasoning foundations.

Stage 2:

**Multimodal Generalization Training (MGT)**:

Extends reasoning capabilities across **diverse multimodal domains**

# LMM-R1: Experiments: Training Datasets

Stage 1:

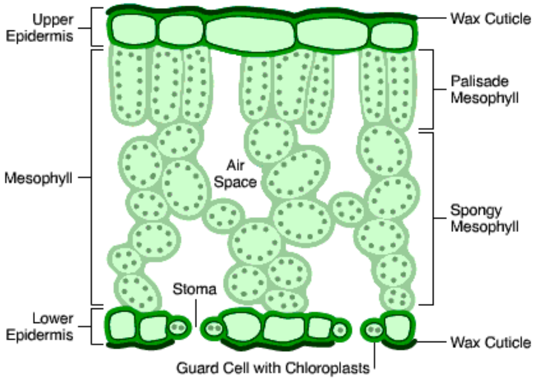
**Foundational Reasoning Enhancement (FRE)**

**Examples:**

Deepscaler40K => FRE-Text

Let $P(x)$ be a polynomial of degree $3n$ such that. \begin{align*} P(0) = P(3) = \dots = P(3n) &= 2, \\ P(1) = P(4) = \dots = P(3n+1-2) &= 1, \\ P(2) = P(5) = \dots = P(3n+2-2) &= 0. \end{align*}. Also, $P(3n+1) = 730$. Determine $n$.

MultiMath-65K => FRE-Multi




Q: What formes the defense barrier

A. palisade

B. wax cuticle

C. lower epidermis

D. stoma

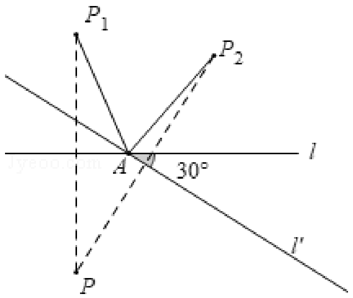
Stage 2:

**Multimodal Generalization Training (MGT)**

**Examples:**

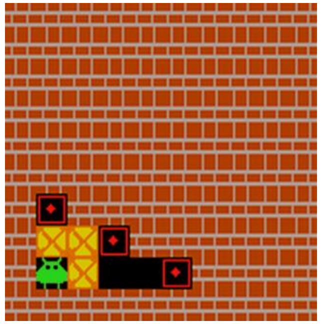
VerMulti-Geo-15K => MGT-Geo

Q: 如图,点P是直线l外一个定点,点A为直线l上一个定点,点P关于直线l的对称点记为P~1~,将直线l绕点A顺时针旋转30° 得到直线l′,此时点P~2~与点P关于直线l′对称,则∠P~1~AP~2~等于多少度?
(A) 30° (B) 45° (C) 60° (D) 75°




VerMulti-65K => MGT-PerceReason

Sokoban environments => MGT-Sokoban



# LMM-R1: Experiments: Main Results

## Main Results & MGT-PerceReason Results

Table 1. Results (%) across benchmarks categorized by three reasoning intensities: High-Level Reasoning (Text-Only) (MATH500/GPQA), Multimodal Reasoning (OlympiadBench/MathVision/MathVerse), and General Multimodal (MM-Star/MathVista). The "MM Avg" column displays the average performance across all multimodal benchmarks. The **best** result is **bolded** and the <u>second-best</u> is <u>underlined</u>.

| Model | Text-Only | | | MM Reasoning-Dominated | | | MM General | | MM Avg |
|---|---|---|---|---|---|---|---|---|---|
| | MATH | GPQA | Avg | Olymp. | MathVis. | MathVer. | MM-Star | MathVista | |
| Qwen2.5-VL CoT | 63.40 | 30.30 | 46.85 | 10.28 | 23.59 | 34.64 | 51.40 | 60.70 | 36.12 |
| *Foundational Reasoning Enhancement Stage* | | | | | | | | | |
| FRE-Multi | 61.80 | 27.27 | 44.54 | 11.80 | 24.74 | 38.45 | **58.76** | **64.20** | 38.71 |
| FRE-Text | <u>65.40</u> | <u>36.87</u> | <u>51.14</u> | **15.62** | 25.76 | 38.83 | 55.15 | 61.40 | <u>39.35</u> |
| *Multimodal Generalization Training Stage* | | | | | | | | | |
| MGT-Geo | **65.80** | 32.32 | 49.06 | <u>14.63</u> | **26.84** | **41.80** | 54.39 | 59.00 | 39.33 |
| MGT-PerceReason | 63.80 | **38.89** | **51.35** | **15.62** | <u>26.35</u> | <u>41.55</u> | <u>58.03</u> | <u>63.20</u> | **40.95** |

# LMM-R1: Experiments: Main Results

## Main Results & MGT-PerceReason Results

Table 1. Results (%) across benchmarks categorized by three reasoning intensities: High-Level Reasoning (Text-Only) (MATH500/GPQA), Multimodal Reasoning (OlympiadBench/MathVision/MathVerse), and General Multimodal (MM-Star/MathVista). The "MM Avg" column displays the average performance across all multimodal benchmarks. The **best** result is **bolded** and the second-best is underlined.

| Model | Text-Only | | | MM Reasoning-Dominated | | | MM General | | MM Avg |
|---|---|---|---|---|---|---|---|---|---|
| | MATH | GPQA | Avg | Olymp. | MathVis. | MathVer. | MM-Star | MathVista | |
| Qwen2.5-VL CoT | 63.40 | 30.30 | 46.85 | 10.28 | 23.59 | 34.64 | 51.40 | 60.70 | 36.12 |
| *Foundational Reasoning Enhancement Stage* | | | | | | | | | |
| FRE-Multi | 61.80 | 27.27 | 44.54 | 11.80 | 24.74 | 38.45 | **58.76** | **64.20** | 38.71 |
| FRE-Text | 65.40 | 36.87 | 51.14 | **15.62** | 25.76 | 38.83 | 55.15 | 61.40 | 39.35 |
| *Multimodal Generalization Training Stage* | | | | | | | | | |
| MGT-Geo | **65.80** | 32.32 | 49.06 | 14.63 | **26.84** | **41.80** | 54.39 | 59.00 | 39.33 |
| MGT-PerceReason | 63.80 | **38.89** | **51.35** | **15.62** | 26.35 | 41.55 | 58.03 | 63.20 | **40.95** |

# LMM-R1: Experiments: Main Results

## Main Results & MGT-PerceReason Results

Table 1. Results (%) across benchmarks categorized by three reasoning intensities: High-Level Reasoning (Text-Only) (MATH500/GPQA), Multimodal Reasoning (OlympiadBench/MathVision/MathVerse), and General Multimodal (MM-Star/MathVista). The "MM Avg" column displays the average performance across all multimodal benchmarks. The **best** result is **bolded** and the second-best is underlined.

| Model | Text-Only | | | MM Reasoning-Dominated | | | MM General | | MM Avg |
|---|---|---|---|---|---|---|---|---|---|
| | MATH | GPQA | Avg | Olymp. | MathVis. | MathVer. | MM-Star | MathVista | |
| Qwen2.5-VL CoT | 63.40 | 30.30 | 46.85 | 10.28 | 23.59 | 34.64 | 51.40 | 60.70 | 36.12 |
| *Foundational Reasoning Enhancement Stage* | | | | | | | | | |
| FRE-Multi | 61.80 | 27.27 | 44.54 | 11.80 | 24.74 | 38.45 | **58.76** | **64.20** | 38.71 |
| FRE-Text | 65.40 | 36.87 | 51.14 | **15.62** | 25.76 | 38.83 | 55.15 | 61.40 | 39.35 |
| *Multimodal Generalization Training Stage* | | | | | | | | | |
| MGT-Geo | **65.80** | 32.32 | 49.06 | 14.63 | **26.84** | **41.80** | 54.39 | 59.00 | 39.33 |
| MGT-PerceReason | 63.80 | **38.89** | **51.35** | **15.62** | 26.35 | 41.55 | 58.03 | 63.20 | **40.95** |

# LMM-R1: Experiments: Main Results

## MGT Results: MGT-Geo

Table 2. Results (%) on geometry-related benchmarks. For MathVision, results are reported for Analytic/Combinatorial/Metric/Solid Geometry. For MathVerse, results are categorized by modality emphasis: TD (Text Domain)/TL (Text Lite)/VI (Vision Intensive)/VD (Vision Domain)/VO (Vision Only). The **best** performance in each subfield is **bolded**.

| Model | MathVision | | | | | MathVerse | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Analy. | Combin. | Metric | Solid | AVG | TD. | TL. | VI. | VD. | VO. | Avg |
| Qwen2.5-VL CoT | 34.52 | 20.78 | 26.33 | 20.49 | 25.53 | 43.15 | 35.41 | 33.38 | 32.87 | 28.43 | 34.64 |
| Direct-RL-Geo | 30.95 | 17.53 | 26.59 | 22.54 | 24.40 | 47.59 | 40.36 | 38.96 | 36.04 | 27.03 | 38.00 |
| FRE-Text | 28.57 | 22.08 | 31.01 | 24.10 | 26.44 | 48.22 | 42.26 | **39.72** | 38.96 | 25.00 | 38.83 |
| MGT-Geo | **36.90** | **22.73** | **31.66** | **27.87** | **29.79** | **51.02** | **42.51** | **39.72** | **39.09** | **36.68** | **41.80** |

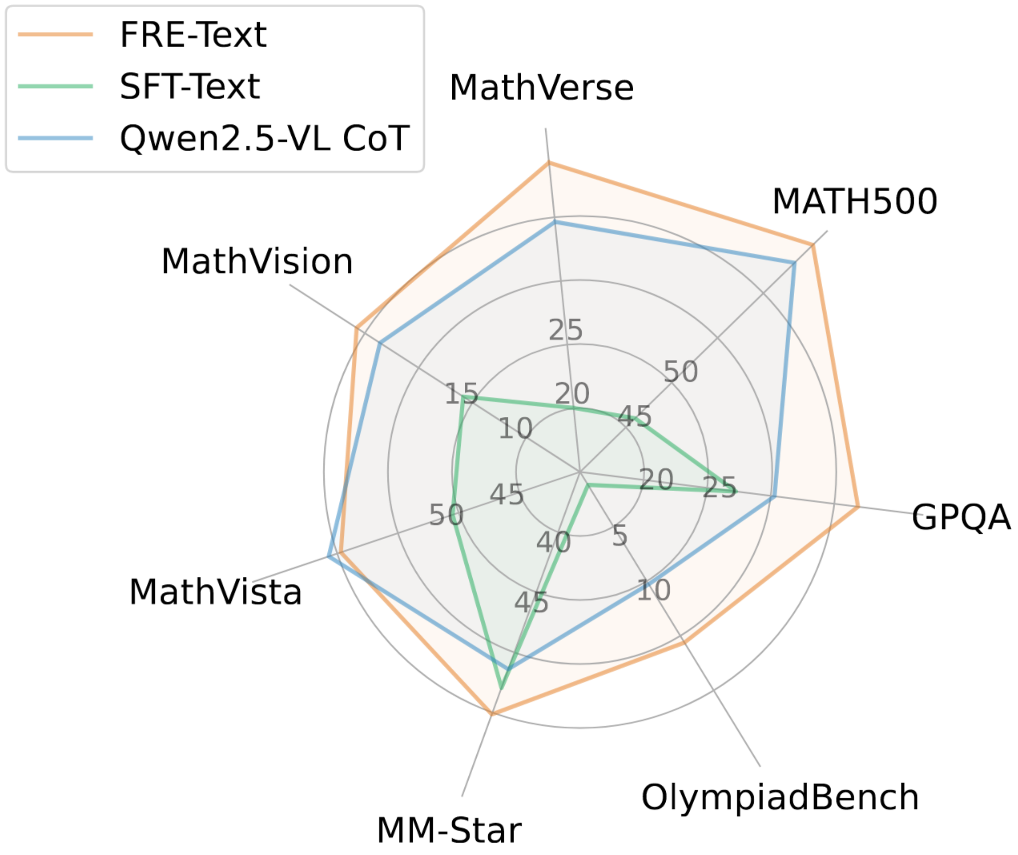
📊 3.35% improvement on MathVision geometry tasks across Analytic, Combinatorial, Metric and Solid geometry

🔄 2.97% improvement on MathVerse geometry problems from Text Domain to Vision Only categories

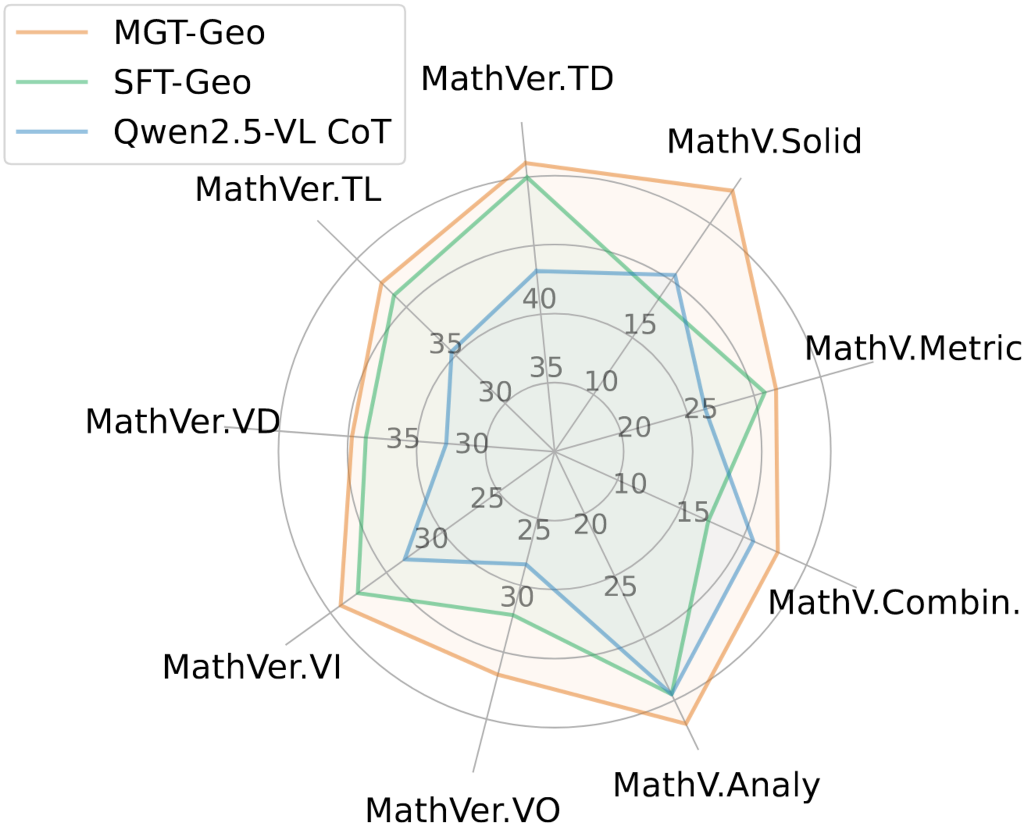
⚡ 11.68% gain in vision-only geometric reasoning compared to FRE-Text baseline

🎯 Significant improvements in both perception and reasoning capabilities for geometry-specific tasks

# LMM-R1: Discussion: SFT vs RFT




(a) General Reasoning Benchmark




(b) Geometry Specialized Benchmark

AiDD 6th 2025

## 参与调研您将优先获得


AiDD定制版
《AI+软件研发精选案例》


专属学习顾问
1对1需求对接

# AiDD会后小调研

AiDD峰会致力于协助企业利用AI技术深化计算机对现实世界的理解，推动研发进入智能化和数字化的新时代。作为峰会的重要共建者，您的真知灼见对我们至关重要。衷心感谢您的参与支持！

扫码参与调研

## 2025 AI+研发数字峰会

拥 抱 AI  重 塑 研 发

# 科技生态圈峰会 + 深度研习

## ——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

K+峰会 敦煌站
K+ 思考周®研习社
时间：2025.08.29-30

K+峰会 上海站
K+ 金融专场
时间：2025.09.26-27

K+峰会 香港站
K+ 思考周®研习社
时间：2025.11.17-18

K+峰会详情

AiDD峰会

AiDD峰会 上海站
AI+研发数字峰会
时间：2025.05.23-24

AiDD峰会 北京站
AI+研发数字峰会
时间：2025.08.08-09

AiDD峰会 深圳站
AI+研发数字峰会
时间：2025.11.14-15

AiDD峰会详情

AiDD 6th 2025

2025 AI+研发数字峰会

AI+ Development Digital Summit

# 感谢聆听！

扫码领取会议PPT资料