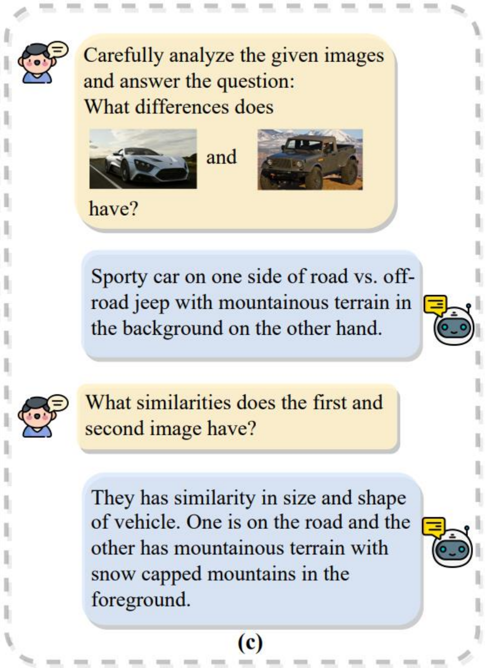
MMICL

Empowering Vision-language Model with Multi-Modal In-Context Learning Ability

Haozhe Zhao[1,2], Zefan Cai[1], Shuzheng Si[1], Xiaojian Ma[3], Kaikai An[1],
Liang Chen[1], Zixuan Liu[4], Sheng Wang[4], Wenjuan Han[5], Baobao Chang[1]

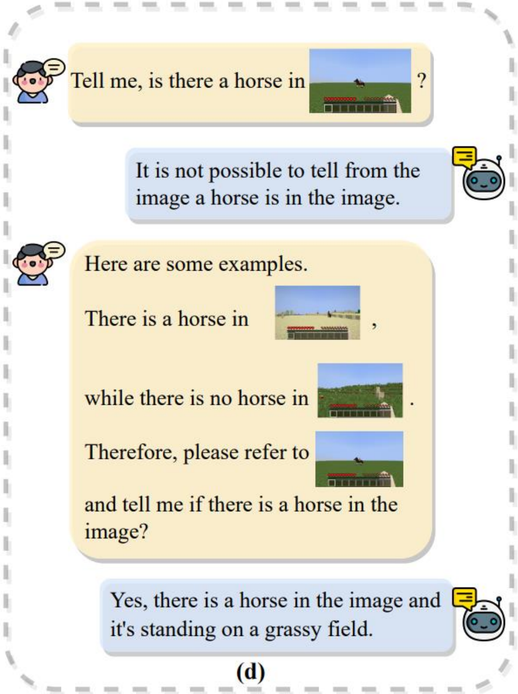
[1] National Key Laboratory for Multimedia Information Processing, Peking University
[2] School of Software and Microelectronics, Peking University, China
[3] National Key Laboratory of General Artificial Intelligence, BIGAI
[3] School of Intelligence Science and Technology, Peking University
[4] Paul G. Allen School of Computer Science and Engineering, University of Washington
[5] Beijing Jiaotong University

VLMs may suffer from the following three limitations:



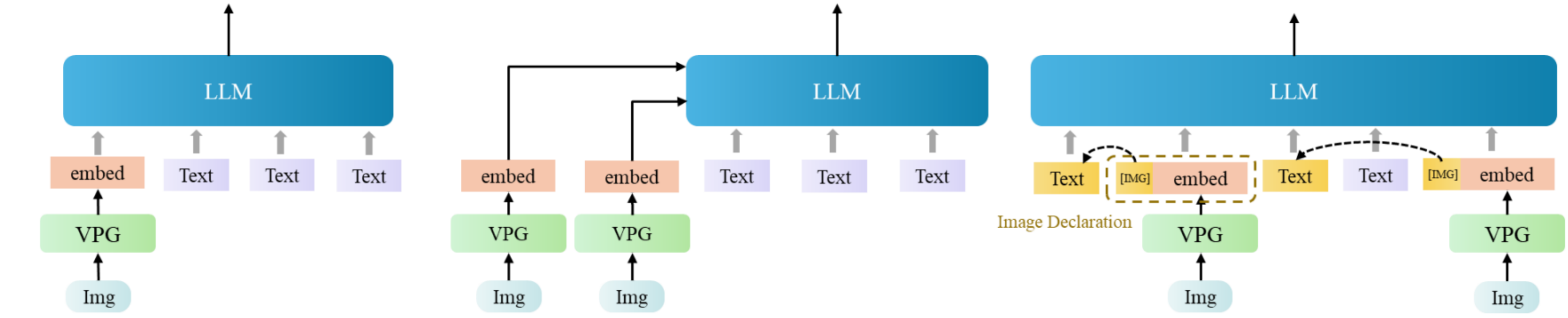
Hard to Understand Text-to-Image Reference

Hard to Understand the Relationships between Multiple Images



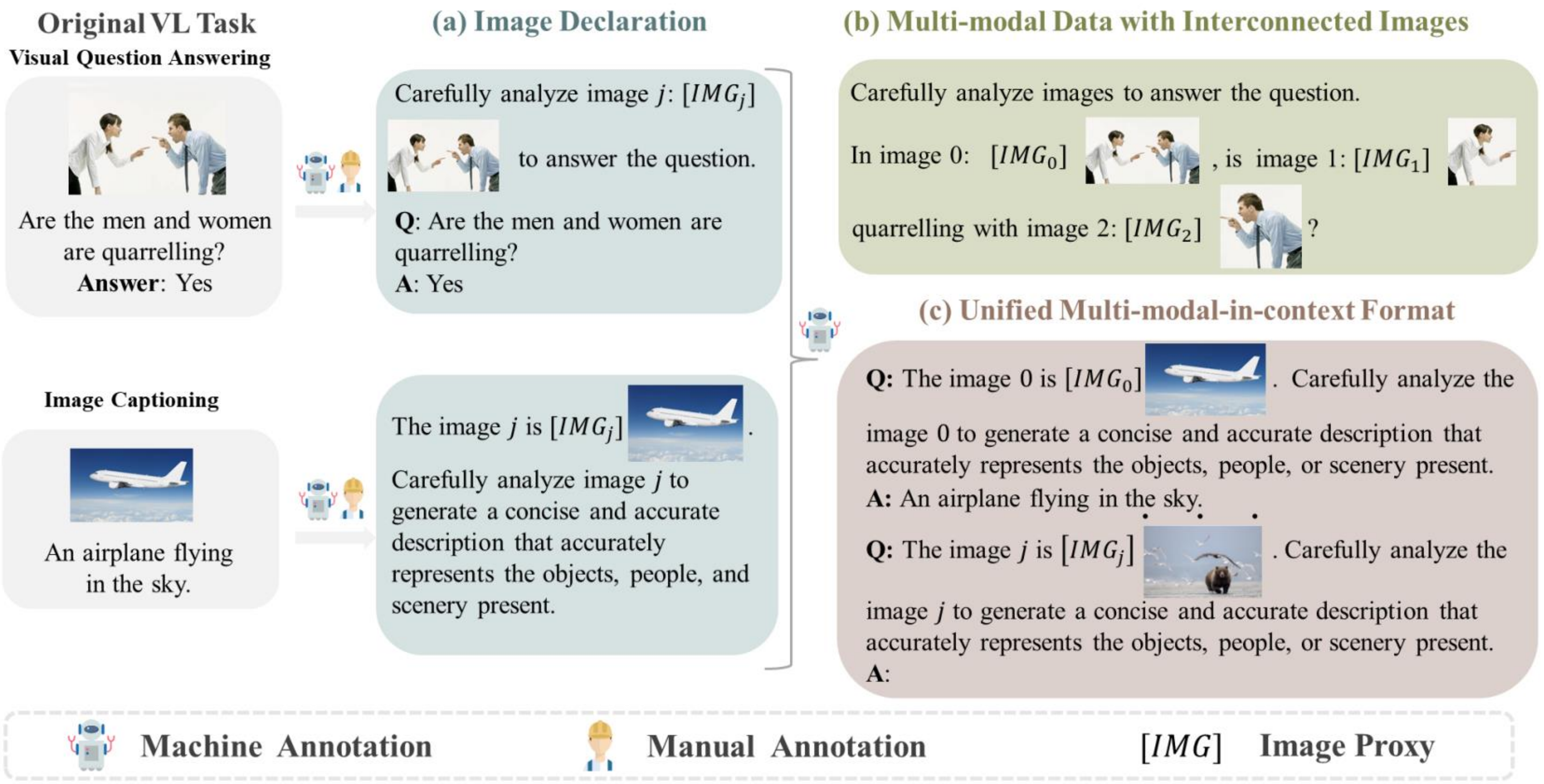
Hard to Learn from In-Context Multi-Modal Demonstrations

Comparison of different VLM architectures

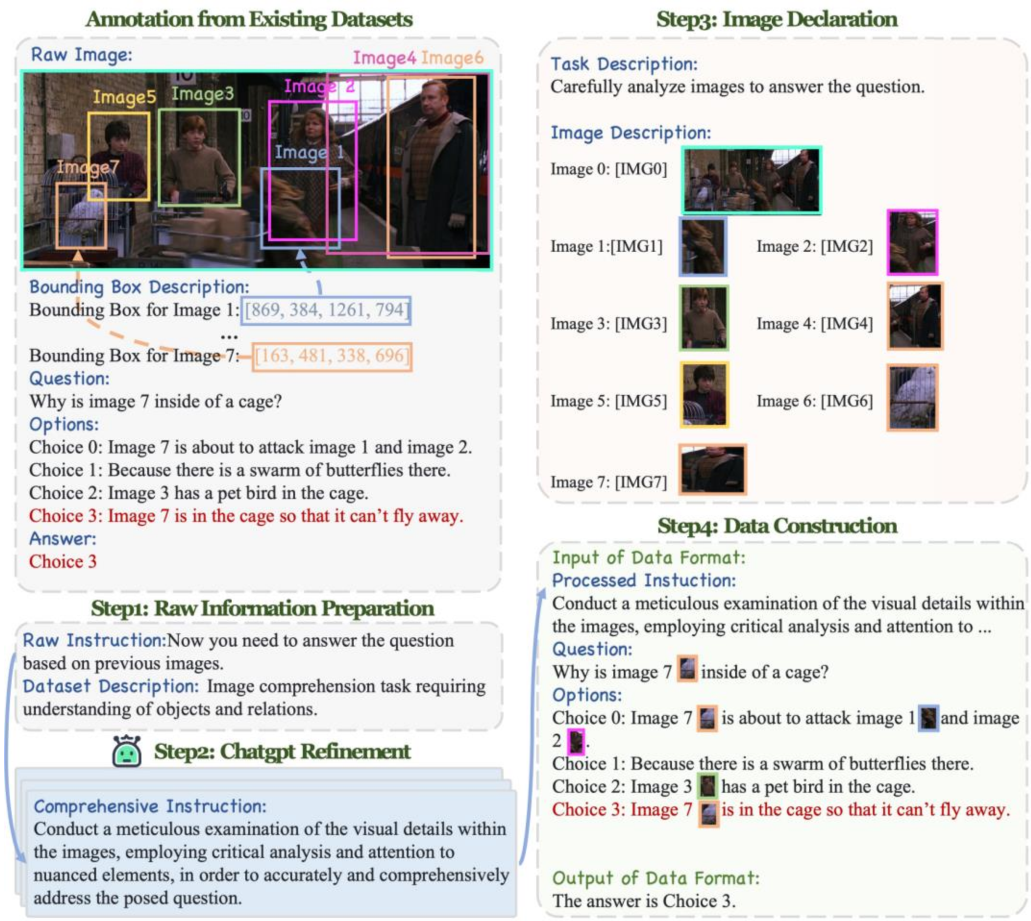


(a) VLMs Focused on a single image (b) VLMs with few-shot ability (c) MMICL

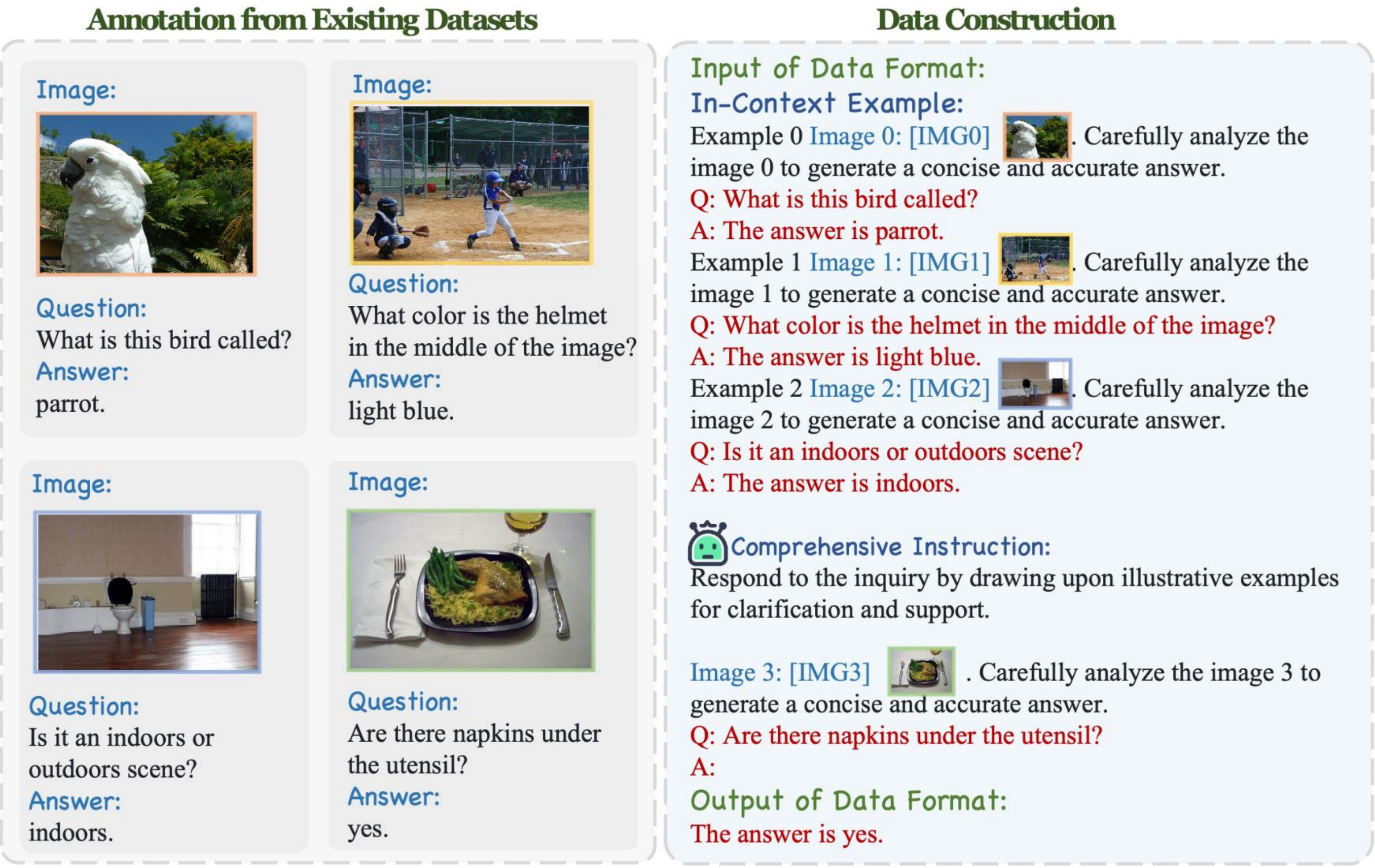
Design of Context Scheme of MMICL



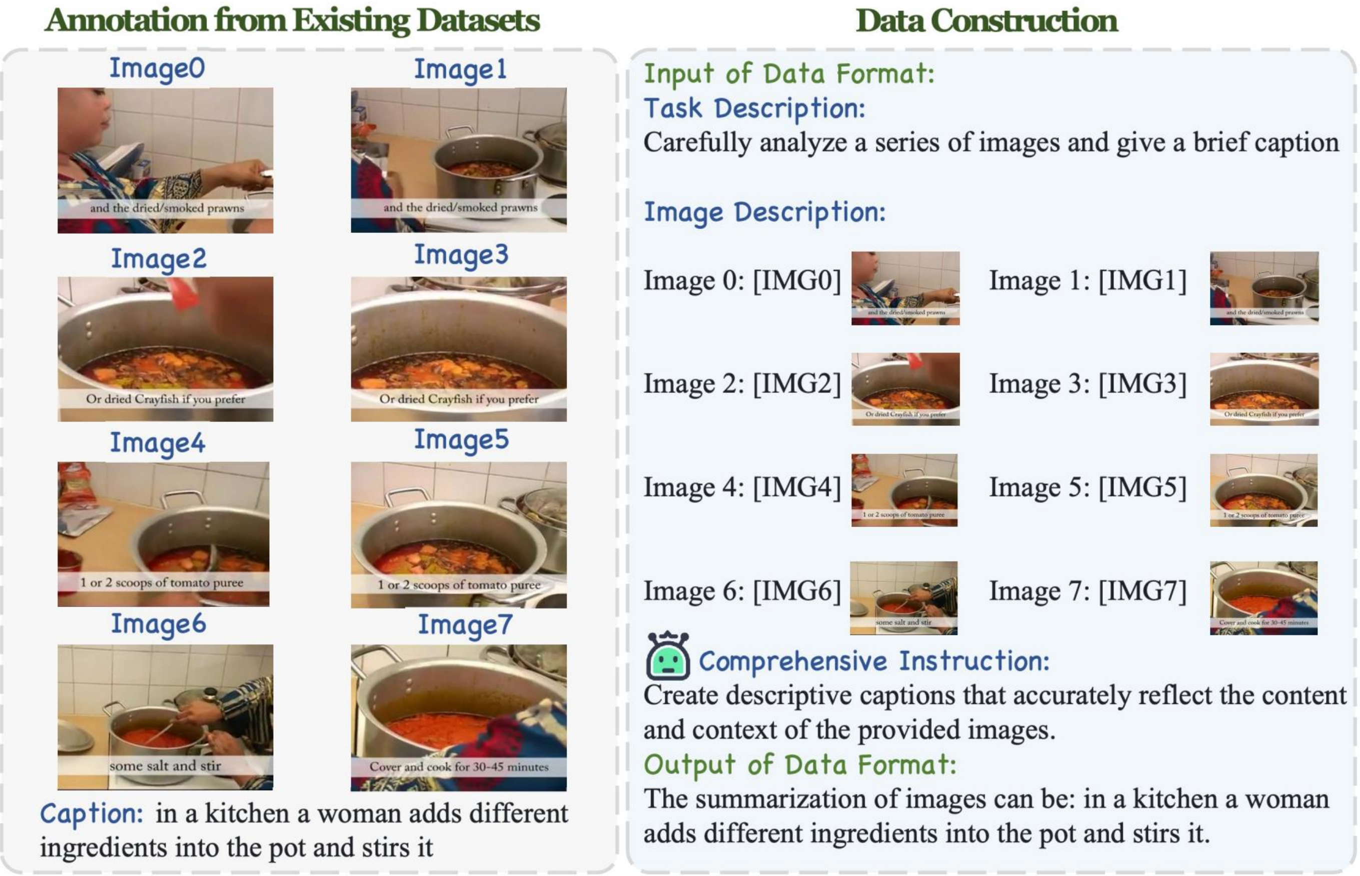
Data Construction Pipeline



Data Construction Pipeline



Data Construction Pipeline



Data Source

Image Captioning
- Flickr30k
- COCO Caption
- Diffusiondb
- Nocaps

Knowledge Question Answering
- OKVQA
- A-OKVQA
- ScienceQA

Video Question Captioning
- MSRVTT

Video Question Answering
- MSRVTT QA
- iVQA
- MVSD
- NextQA-Multiple-Choice
- NextQA-Open-Domain

Image Question Answering
- VQAv2
- STVQA
- TextVQA
- Wikiart
- RefCOCO
- VizWiz

Visual Reasoning
- GQA
- Visual Commonsense Reasoning
- Natural Language Visual Reasoning v2
- Visual Spatial Reasoning
- Winoground
- IconQA-Multi-image-choice
- IconQA-Multi-text-choice

Web Page Question Answering
- Websrc

Few-Shot Image Classification
- MiniImagenet
- HatefulMemes
- Bongard-HOI

Nonverbal Reasoning
- Raven IQ Test

Visual Dialog
- LLaVa-Instruct-150K
- Visual Dialog

OOD Generalization
- Minecraft

MMICL Architecture & Training Paradigm

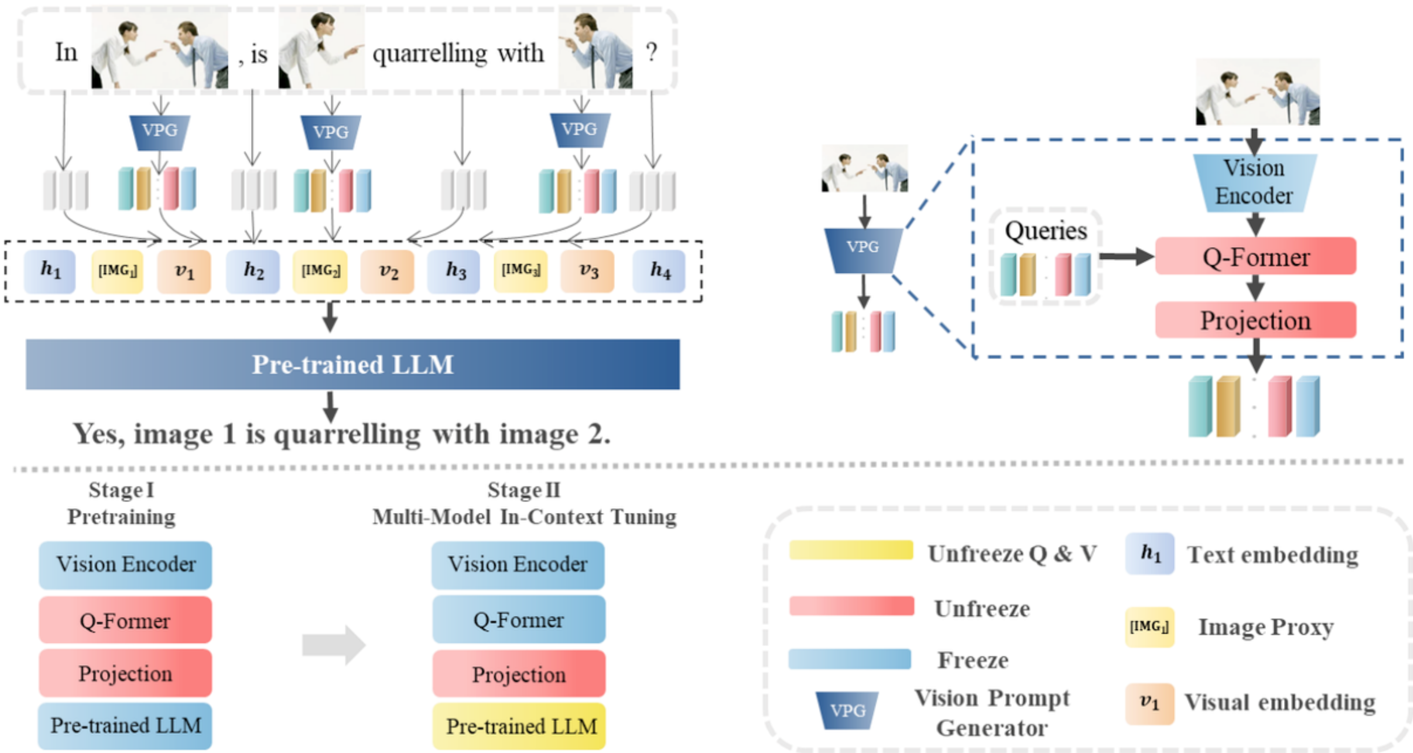


Figure 4: Illustration of MMICL architecture and training paradigm. The upper part denotes the overview of model architecture and the bottom denotes the pipeline of the two-stage training paradigm.

General Performance Evaluation

Model	Model Size	Cognition				Perception										Total Avg.
		Comm.	Num.	Text.	Code.	Existen.	Count	Pos.	Color	OCR	Poster	Cele.	Scene	Land.	Art.	
LLaVA	13B	57.14	50.00	57.50	50.00	50.00	50.00	50.00	55.00	50.00	50.00	48.82	50.00	50.00	49.00	51.25
MiniGPT-4	13B	59.29	45.00	0.00	40.00	68.33	55.00	43.33	75.00	57.50	41.84	54.41	71.75	54.00	60.50	51.85
MultiModal-GPT	9B	49.29	62.50	60.00	55.00	61.67	55.00	58.33	68.33	82.50	57.82	73.82	68.00	69.75	59.50	62.97
VisualGLM-6B	6B	39.29	45.00	50.00	47.50	85.00	50.00	48.33	55.00	42.50	65.99	53.24	146.25	83.75	75.25	63.36
VPGTrans	7B	64.29	50.00	77.50	57.50	70.00	85.00	63.33	73.33	77.50	84.01	53.53	141.75	64.75	77.25	74.27
LaVIN	13B	87.14	65.00	47.50	50.00	185.00	88.33	63.33	75.00	107.50	79.59	47.35	136.75	93.50	87.25	86.66
LLaMA-Adapter-V2	7B	81.43	62.50	50.00	55.00	120.00	50.00	48.33	75.00	125.00	99.66	86.18	148.50	150.25	69.75	87.26
mPLUG-Owl	7B	78.57	60.00	80.00	57.50	120.00	50.00	50.00	55.00	65.00	136.05	100.29	135.50	159.25	96.25	88.82
InstructBLIP	12.1B	129.29	40.00	65.00	57.50	185.00	143.33	66.67	153.33	72.50	123.81	101.18	153.00	79.75	134.25	107.47
BLIP-2	12.1B	110.00	40.00	65.00	75.00	160.00	135.00	73.33	148.33	110.00	141.84	105.59	145.25	138.00	136.50	113.13
Lynx	7B	110.71	17.50	42.50	45.00	195.00	151.67	90.00	170.00	77.50	124.83	118.24	164.50	**162.00**	119.50	113.50
GIT2	5.1B	99.29	50.00	67.50	45.00	190.00	118.33	**96.67**	158.33	65.00	112.59	145.88	158.50	140.50	**146.25**	113.85
Otter	9B	106.43	72.50	57.50	70.00	**195.00**	88.33	86.67	113.33	72.50	138.78	**172.65**	**158.75**	137.25	129.00	114.19
Cheetor	7B	98.57	77.50	57.50	**87.50**	180.00	96.67	80.00	116.67	100.00	147.28	164.12	156.00	145.73	113.50	115.79
LRV-Instruction	7B	100.71	70.00	85.00	72.50	165.00	111.67	86.67	165.00	110.00	139.04	112.65	147.98	160.53	101.25	116.29
BLIVA	12.1B	136.43	57.50	77.50	60.00	180.00	138.33	81.67	**180.00**	87.50	**155.10**	140.88	151.50	89.50	133.25	119.23
MMICL	12.1B	**136.43**	**82.50**	**132.50**	77.50	170.00	**160.00**	81.67	156.67	100.00	146.26	141.76	153.75	136.13	135.50	**129.33**

Table 1: Evaluation results on the MME. Top two scores are highlighted and underlined, respectively.

Performance Prob

Text-to-Image Reference

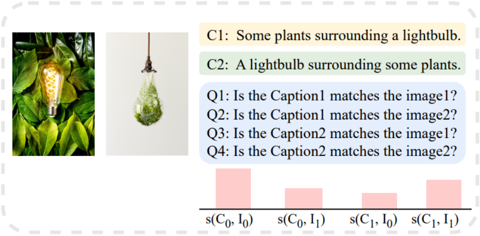
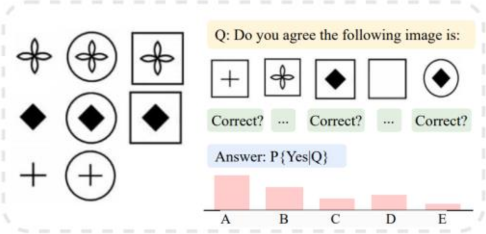
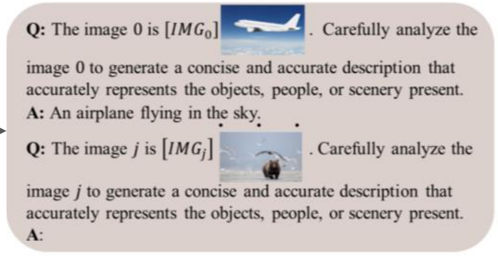


Image-to-Image Relationships



Multi-Modal In-Context Learning



Performance Prob

➢ Text-to-Image Reference

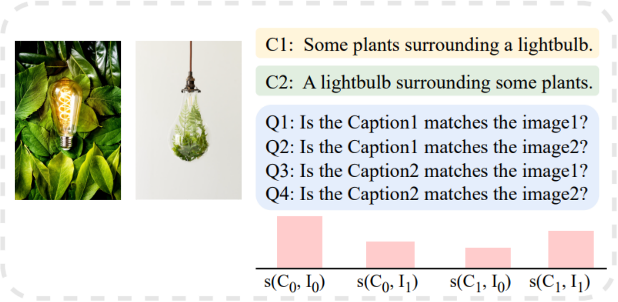


Table 2: Results on Winoground across text, image and group score metrics.

Model	Text	Image	**Group**
MTurk Human	89.50	88.50	85.50
VQ2 (Yarom et al., 2023)	47.00	42.20	30.50
PALI (Chen et al., 2022)	46.50	38.00	28.75
Blip-2 (Li et al., 2023d)	44.00	26.00	23.50
GPT4-V (Wu et al., 2023)	***69.25***	***46.25***	*39.25*
MMICL (FLAN-T5-XXL)	45.00	45.00	**43.00**

➢ Image-to-Image Relationships

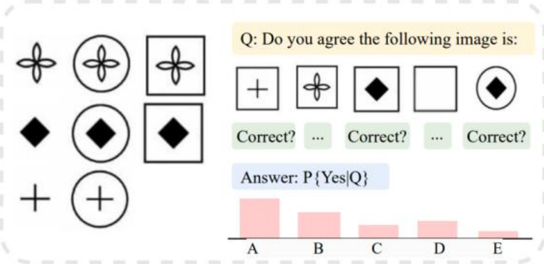


Table 3: Zero-shot generalization on Raven IQ test.

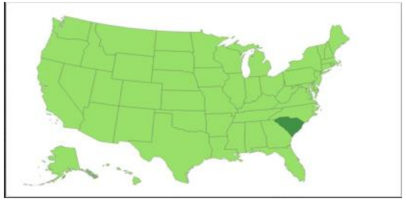
Model	Accuracy
Random Choice	17
InstructBlip (Dai et al., 2023)	10.00
Otter (Li et al., 2023a)	22.00
KOSMOS-1 (Huang et al., 2023a)	22.00
MMICL (FLAN-T5-XXL)	**34.00**

Performance Prob

- Multi-Modal In-Context Learning

Model	Flickr 30K	WebSRC	VQAv2	Hateful Memes	VizWiz
Flamingo-3B (Alayrac et al., 2022) (w/o ICL example)	60.60	-	49.20	53.70	28.90
Flamingo-3B (Alayrac et al., 2022) (w/ ICL examples (4))	72.00	-	53.20	53.60	34.00
Flamingo-9B (Alayrac et al., 2022) (w/o ICL example)	61.50	-	51.80	57.00	28.80
Flamingo-9B (Alayrac et al., 2022) (w/ ICL examples (4))	72.60	-	56.30	62.70	34.90
KOSMOS-1 (Huang et al., 2023b) (w/o ICL example)	67.10	3.80	51.00	63.90	29.20
KOSMOS-1 (Huang et al., 2023b) (w/ ICL examples (4))	75.30	-	51.80	-	35.30
w/o ICL example					
BLIP-2 (Li et al., 2023d) (FLANT5-XL)	64.51	12.25	58.79	60.00	25.52
BLIP-2 (Li et al., 2023d) (FLANT5-XXL)	60.74	10.10	60.91	62.25	22.50
InstructBLIP (Dai et al., 2023) (FLANT5-XL)	77.16	10.80	36.77	58.54	32.08
InstructBLIP (Dai et al., 2023) (FLANT5-XXL)	73.13	11.50	63.69	61.70	15.11
ICL example Evaluation					
MMICL (FLAN-T5-XL) (w/o ICL example)	83.47	12.55	62.17	60.28	25.04
MMICL (FLAN-T5-XL) (w/ ICL examples (4))	83.84	12.30	62.63	60.80	50.17
MMICL (FLAN-T5-XXL) (w/o ICL example)	85.03	18.85	69.99	60.32	29.34
MMICL (FLAN-T5-XXL) (w/ ICL examples (4))	**89.27**	18.70	69.83	61.12	33.16
MMICL (Instruct-FLAN-T5-XL) (w/o ICL example)	82.68	14.75	69.13	61.12	29.92
MMICL (Instruct-FLAN-T5-XL) (w/ ICL examples (4))	88.31	14.80	69.16	61.12	33.16
MMICL (Instruct-FLAN-T5-XXL) (w/o ICL example)	73.97	17.05	70.30	62.23	24.45
MMICL (Instruct-FLAN-T5-XXL) (w/ ICL examples (4))	88.79	**19.65**	**70.56**	**64.60**	**50.28**

Hallucination & Language Bais



Don’t Require Visual Information to answer

What is the capital of South Carolina?	["Columbia", "Montgomery", "Charleston", "Harrisburg"]

Require Visual Information to answer

Which property matches this object?	["flexible", "sticky"]

Model	Model Size	Average Performance	Don’t Require Visual Infomation	Require Visual Infomation	Performance Gap
Random Guess	-	35.50	35.80	34.90	-
Ying-VLM (Li et al., 2023e)	13.6B	55.70	66.60	44.90	21.70
InstructBLIP (Dai et al., 2023)	12.1B	71.30	82.00	60.70	21.30
Otter (Li et al., 2023a)	9B	63.10	70.90	55.70	15.20
Shikra (Chen et al., 2023a)	7.2B	45.80	52.90	39.30	13.60
MMICL	12.1B	**82.10**	**82.60**	**81.70**	**0.90**

Ablation Study

Ablation Study on Training Paradigm

Model	VSR	IconQA text	VisDial	IconQA img	Bongard HOI
		Stage I			
Stage I (Blip-2-FLANT5-XL)	61.62	45.44	35.43	48.42	52.75
Stage I (Blip-2-FLANT5-XXL)	63.18	50.08	36.48	48.42	59.20
Stage I (InstructBLIP-FLANT5-XL)	61.54	47.53	35.36	50.11	53.15
Stage I (InstructBLIP-FLANT5-XXL)	65.06	51.39	36.09	45.10	63.35
		Stage I + Stage II			
Stage I + Stage II (BLIP-2-FLAN-T5-XL)	62.85	47.23	35.76	<u>51.24</u>	56.95
Stage I + Stage II (BLIP-2-FLAN-T5-XXL)	64.73	50.55	<u>37.00</u>	34.93	<u>68.05</u>
Stage I + Stage II (InstructBLIP-FLAN-T5-XL)	**70.54**	**52.55**	36.87	47.27	**74.20**
Stage I + Stage II (InstructBLIP-FLAN-T5-XXL)	<u>66.45</u>	<u>52.00</u>	**37.98**	**60.85**	67.20

Ablation Study on Context Scheme

Model	$\mathrm{MME}_{\mathrm{Perception}}$	$\mathrm{MME}_{\mathrm{Cognition}}$	Icon-QA	NLVR2	Raven	Winoground
- w/o context scheme	1238.99	316.79	52.80	56.65	8.00	6.00
- w/o image declaration	1170.87	341.07	47.15	61.00	18.00	3.00
- w/o in-context format	1141.02	<u>345.36</u>	51.95	<u>62.63</u>	<u>28.00</u>	20.00
- w/o interrelated images	1207.70	333.21	<u>54.35</u>	59.60	16.00	<u>25.75</u>
MMICL	**1303.59**	**370.71**	**58.12**	**72.45**	**32.00**	**38.75**

Takeaway

To we address the limitation of most VLMs, we introduce the MMICL, a new approach to allow the VLM to deal with multi-modal inputs efficiently.

We propose a novel context scheme to augment the in-context learning ability of the VLM and constructe the MIC dataset under the guidance the proposed context scheme for tuning the VLM.

MMICL effectively tackles the challenge of complex multi-modal prompt understanding and emerges the impressive ICL ability. It achieves new SOTA zero-shot performance on a wide range of general vision-language and complex benchmarks.

Paper& Code & Data: MMICL