



# MMICL

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

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

<sup>1</sup> National Key Laboratory for Multimedia Information Processing, Peking University

<sup>2</sup> School of Software and Microelectronics, Peking University, China




<sup>3</sup> National Key Laboratory of General Artificial Intelligence, BIGAI


<sup>3</sup> School of Intelligence Science and Technology, Peking University


<sup>4</sup> Paul G. Allen School of Computer Science and Engineering, University of Washington


<sup>5</sup> Beijing Jiaotong University

# VLMs may suffer from the following three limitations:

 Carefully analyze the given images and answer the question: What differences does  and  have?

 Sporty car on one side of road vs. off-road jeep with mountainous terrain in the background on the other hand.

 What similarities does the first and second image have?

 They has similarity in size and shape of vehicle. One is on the road and the other has mountainous terrain with snow capped mountains in the foreground.

(c)

**Hard to Understand Text-to-Image Reference**

 Please describe the information of  , especially the relationship between  and  .

 A man is seen walking down a street with a large orange cat. The cat is sitting on the man's head, and it seems to be enjoying the walk.



(a)


 Please take a closer look at the two images and explain the connection between them.  


 The baby is crying as he broke the cup.


(b)


**Hard to Understand the Relationships between Multiple Images**


 Tell me, is there a horse in  ?


 It is not possible to tell from the image a horse is in the image.

 Here are some examples.

There is a horse in  ,

while there is no horse in  .

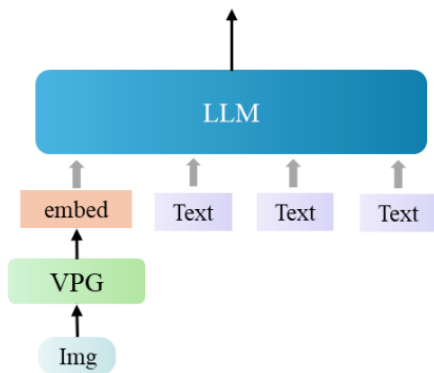
Therefore, please refer to  and tell me if there is a horse in the image?

 Yes, there is a horse in the image and it's standing on a grassy field.

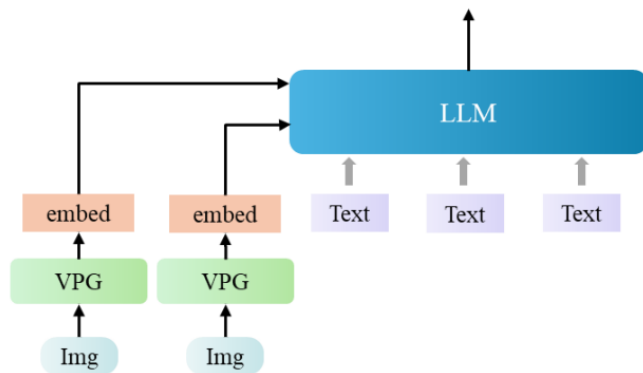
(d)

**Hard to Learn from In-Context Multi-Modal Demonstrations**

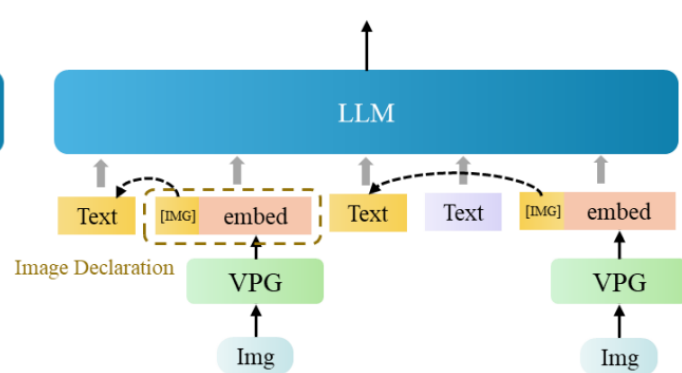
# Comparison of different VLM architectures



(a) VLMs Focused on a single image



(b) VLMs with few-shot ability



(c) MMICL

# Design of Context Scheme of MMICL

## Original VL Task Visual Question Answering



Are the men and women  
are quarrelling?  
**Answer:** Yes



## (a) Image Declaration

Carefully analyze image  $j$ :  $[IMG_j]$



to answer the question.

**Q:** Are the men and women are  
quarrelling?  
**A:** Yes

## Image Captioning



An airplane flying  
in the sky.



The image  $j$  is  $[IMG_j]$



Carefully analyze image  $j$  to  
generate a concise and accurate  
description that accurately  
represents the objects, people, and  
scenery present.

## (b) Multi-modal Data with Interconnected Images

Carefully analyze images to answer the question.

In image 0:  $[IMG_0]$



, is image 1:  $[IMG_1]$





quarrelling with image 2:  $[IMG_2]$



?

## (c) Unified Multi-modal-in-context Format

**Q:** The image 0 is  $[IMG_0]$   . Carefully analyze the  
image 0 to generate a concise and accurate description that  
accurately represents the objects, people, or scenery present.  
**A:** An airplane flying in the sky.

**Q:** The image  $j$  is  $[IMG_j]$   . Carefully analyze the  
image  $j$  to generate a concise and accurate description that  
accurately represents the objects, people, or scenery present.  
**A:**



Machine Annotation



Manual Annotation

$[IMG]$

Image Proxy

# Data Construction Pipeline

## Annotation from Existing Datasets

Raw Image: 



Bounding Box Description:  
 Bounding Box for Image 1: [869, 384, 1261, 794]  
 ...  
 Bounding Box for Image 7: [163, 481, 338, 696]

Question:  
 Why is image 7 inside of a cage?  
 Options:  
 Choice 0: Image 7 is about to attack image 1 and image 2.  
 Choice 1: Because there is a swarm of butterflies there.  
 Choice 2: Image 3 has a pet bird in the cage.  
**Choice 3: Image 7 is in the cage so that it can't fly away.**  
 Answer:  
**Choice 3**

## Step1: Raw Information Preparation

**Raw Instruction:** Now you need to answer the question based on previous images.  
**Dataset Description:** Image comprehension task requiring understanding of objects and relations.



## Step2: Chatgpt Refinement

**Comprehensive Instruction:**  
 Conduct a meticulous examination of the visual details within the images, employing critical analysis and attention to nuanced elements, in order to accurately and comprehensively address the posed question.

## Step3: Image Declaration







**Task Description:**  
 Carefully analyze images to answer the question.

**Image Description:**



## Step4: Data Construction

**Input of Data Format:**  
**Processed Instruction:**  
 Conduct a meticulous examination of the visual details within the images, employing critical analysis and attention to ...

**Question:**  
 Why is image 7  inside of a cage?  
**Options:**  
 Choice 0: Image 7  is about to attack image 1  and image 2 .
 Choice 1: Because there is a swarm of butterflies there.  
 Choice 2: Image 3  has a pet bird in the cage.  
**Choice 3: Image 7  is in the cage so that it can't fly away.**

**Output of Data Format:**  
 The answer is Choice 3.

# Data Construction Pipeline

## Annotation from Existing Datasets

Image:



Question:

What is this bird called?

Answer:

parrot.

Image:



Question:

What color is the helmet in the middle of the image?

Answer:

light blue.

Image:



Question:

Is it an indoors or outdoors scene?

Answer:

indoors.

Image:



Question:

Are there napkins under the utensil?

Answer:

yes.

## Data Construction

Input of Data Format:

In-Context Example:

Example 0 Image 0: [IMG0] . Carefully analyze the image 0 to generate a concise and accurate answer.

Q: What is this bird called?

A: The answer is parrot.

Example 1 Image 1: [IMG1] . Carefully analyze the image 1 to generate a concise and accurate answer.

Q: What color is the helmet in the middle of the image?

A: The answer is light blue.

Example 2 Image 2: [IMG2] . Carefully analyze the image 2 to generate a concise and accurate answer.


Q: Is it an indoors or outdoors scene?

A: The answer is indoors.



Comprehensive Instruction:

Respond to the inquiry by drawing upon illustrative examples for clarification and support.

Image 3: [IMG3] . Carefully analyze the image 3 to generate a concise and accurate answer.

Q: Are there napkins under the utensil?

A:

Output of Data Format:

The answer is yes.

# Data Construction Pipeline

## Annotation from Existing Datasets

Image0



Image1



Image2



Image3



Image4



Image5



Image6



Image7



**Caption:** in a kitchen a woman adds different ingredients into the pot and stirs it

## Data Construction

**Input of Data Format:**

**Task Description:**

Carefully analyze a series of images and give a brief caption

**Image Description:**

Image 0: [IMG0]



Image 1: [IMG1]



Image 2: [IMG2]



Image 3: [IMG3]



Image 4: [IMG4]



Image 5: [IMG5]



Image 6: [IMG6]



Image 7: [IMG7]



**Comprehensive Instruction:**

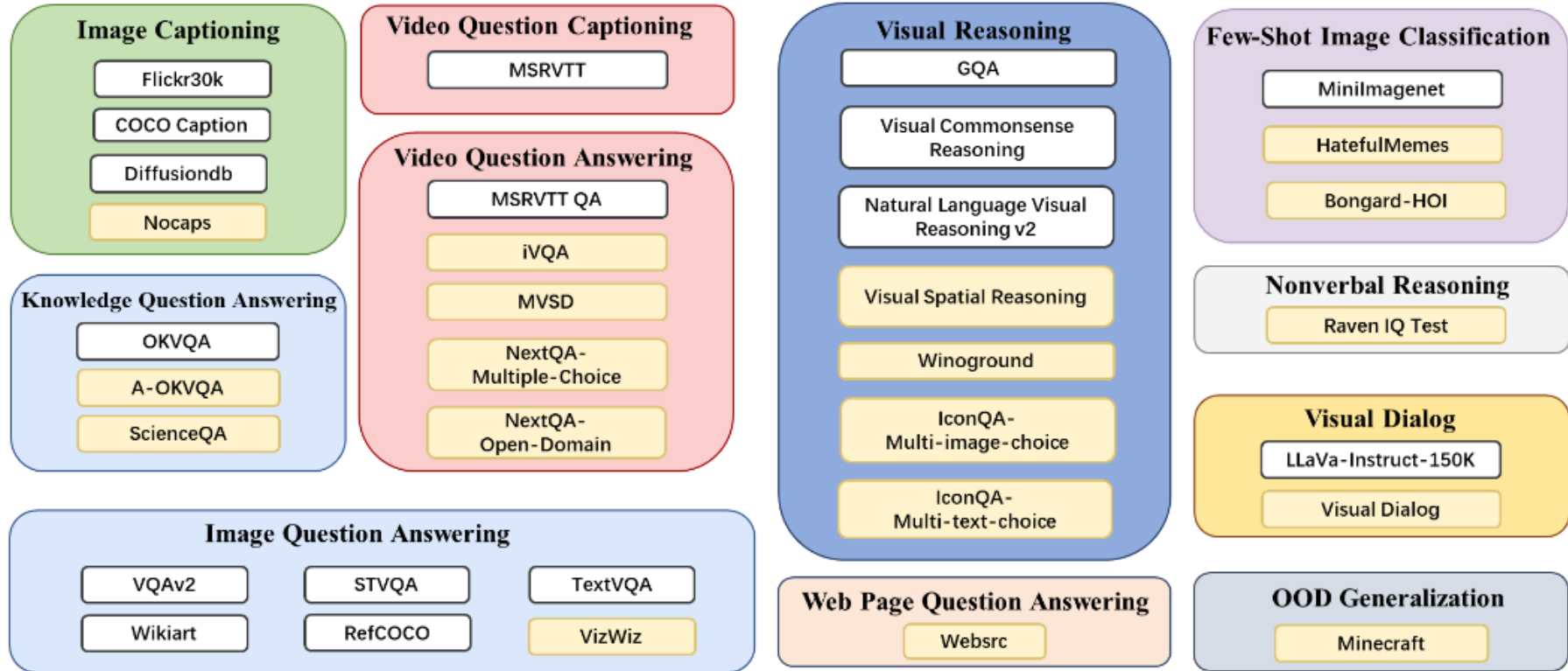
Create descriptive captions that accurately reflect the content and context of the provided images.

**Output of Data Format:**

The summarization of images can be: in a kitchen a woman adds different ingredients into the pot and stirs it.



# Data Source





# MMICL Architecture & Training Paradigm

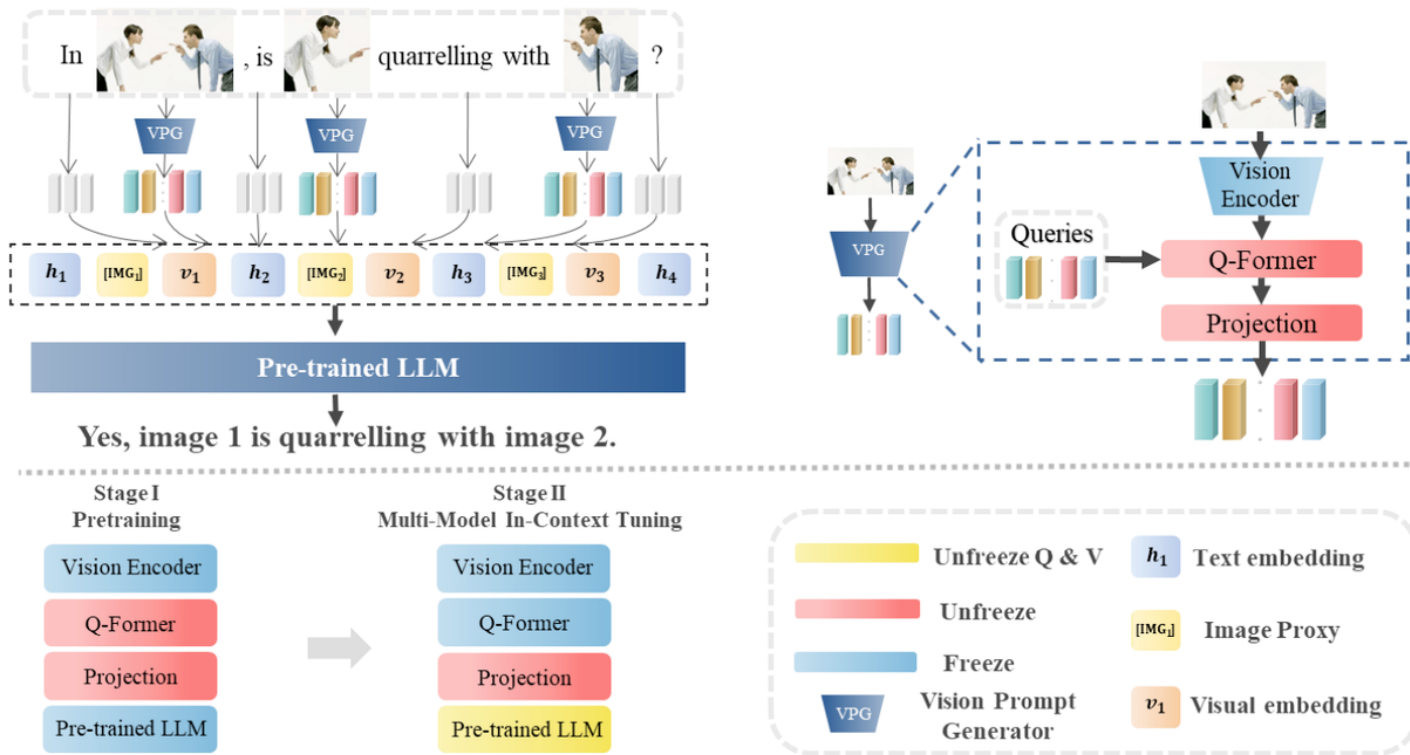


Figure 4: Illustration of MMICL architecture and training paradigm. The upper part denotes the overview of model architecture and the bottom denotes the pipeline of the two-stage training paradigm.



# 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	<u>110.00</u>	141.84	105.59	145.25	138.00	<u>136.50</u>	113.13
Lynx	7B	110.71	17.50	42.50	45.00	<u>195.00</u>	<u>151.67</u>	<u>90.00</u>	<u>170.00</u>	77.50	124.83	118.24	164.50	<b>162.00</b>	119.50	113.50
GIT2	5.1B	99.29	50.00	67.50	45.00	190.00	118.33	<b>96.67</b>	158.33	65.00	112.59	145.88	<u>158.50</u>	140.50	<b>146.25</b>	113.85
Otter	9B	106.43	72.50	57.50	70.00	<b>195.00</b>	88.33	86.67	113.33	72.50	138.78	<b>172.65</b>	<b>158.75</b>	137.25	129.00	114.19
Cheetor	7B	98.57	77.50	57.50	<b>87.50</b>	180.00	96.67	80.00	116.67	100.00	<u>147.28</u>	<u>164.12</u>	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	<u>160.53</u>	101.25	116.29
BLIVA	12.1B	<u>136.43</u>	57.50	<u>77.50</u>	60.00	180.00	138.33	81.67	<b>180.00</b>	87.50	<b>155.10</b>	140.88	151.50	89.50	133.25	<u>119.23</u>
MMICL	12.1B	<b>136.43</b>	<b>82.50</b>	<b>132.50</b>	<u>77.50</u>	170.00	<b>160.00</b>	81.67	156.67	100.00	146.26	141.76	153.75	136.13	135.50	<b>129.33</b>

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

# Performance Prob

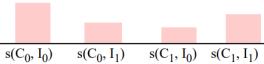
## Text-to-Image Reference



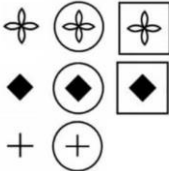
C1: Some plants surrounding a lightbulb.

C2: A lightbulb surrounding some plants.

- Q1: Is the Caption1 matches the image1?
- Q2: Is the Caption1 matches the image2?
- Q3: Is the Caption2 matches the image1?
- Q4: Is the Caption2 matches the image2?



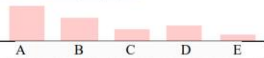
## Image-to-Image Relationships




Q: Do you agree the following image is:


- Correct? ... Correct? ... Correct?

Answer: P{Yes|Q}



## Multi-Modal In-Context Learning

Q: The image 0 is  $[IMG_0]$   . Carefully analyze the image 0 to generate a concise and accurate description that accurately represents the objects, people, or scenery present.  
 A: An airplane flying in the sky.

Q: The image  $j$  is  $[IMG_j]$   . Carefully analyze the image  $j$  to generate a concise and accurate description that accurately represents the objects, people, or scenery present.  
 A:

# Performance Prob

## ➤ Text-to-Image Reference

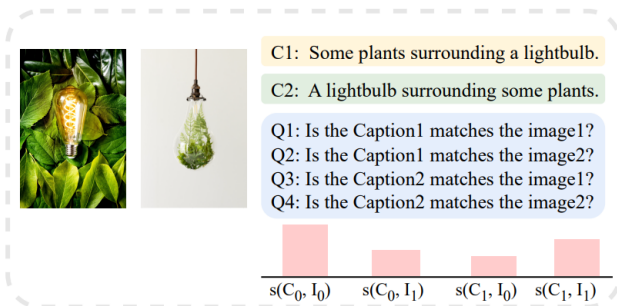


Table 2: Results on Winoground across text, image and group score metrics.

Model	Text	Image	Group
MTurk Human	89.50	88.50	85.50
VQ2 (Yarom et al., 2023)	<u>47.00</u>	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
<i>GPT4-V (Wu et al., 2023)</i>	<b>69.25</b>	<b>46.25</b>	<b>39.25</b>
MMICL (FLAN-T5-XXL)	45.00	<u>45.00</u>	<b>43.00</b>

## ➤ Image-to-Image Relationships

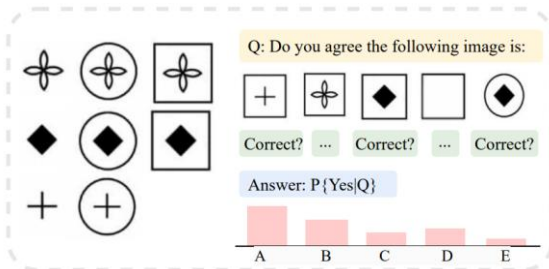


Table 3: Zero-shot generalization on Raven IQ test.

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)	<u>22.00</u>
MMICL (FLAN-T5-XXL)	<b>34.00</b>

# 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	<u>63.90</u>	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	<u>50.17</u>
MMICL (FLAN-T5-XXL) (w/o ICL example)	85.03	<u>18.85</u>	69.99	60.32	29.34
MMICL (FLAN-T5-XXL) (w/ ICL examples (4))	<b>89.27</b>	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	<u>70.30</u>	62.23	24.45
MMICL (Instruct-FLAN-T5-XXL) (w/ ICL examples (4))	<u>88.79</u>	<b>19.65</b>	<b>70.56</b>	<b>64.60</b>	<b>50.28</b>

# 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 Information	Require Visual Information	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	<u>71.30</u>	82.00	<u>60.70</u>	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	<u>13.60</u>
<b>MMICL</b>	12.1B	<b>82.10</b>	<b>82.60</b>	<b>81.70</b>	<b>0.90</b>



# 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	51.24	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)	<b>70.54</b>	<b>52.55</b>	36.87	47.27	<b>74.20</b>
Stage I + Stage II (InstructBLIP-FLAN-T5-XXL)	<u>66.45</u>	<u>52.00</u>	<b>37.98</b>	<b>60.85</b>	67.20

## Ablation Study on Context Scheme

Model	MME <sub>Perception</sub>	MME <sub>Cognition</sub>	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	<b>1303.59</b>	<b>370.71</b>	<b>58.12</b>	<b>72.45</b>	<b>32.00</b>	<b>38.75</b>



# 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](#)