MMICL: EMPOWERING VISION-LANGUAGE MODEL WITH MULTI-MODAL IN-CONTEXT LEARNING

Haozhe Zhao[*1], **Zefan Cai**[*1], **Shuzheng Si**[*1], **Xiaojian Ma**[2], **Kaikai An**[1], **Liang Chen**[1], **Zixuan Liu**[3], **Sheng Wang**[3], **Wenjuan Han**[†4], **Baobao Chang**[†1]
[1]National Key Laboratory for Multimedia Information Processing, Peking University
[2]National Key Laboratory of General Artificial Intelligence, BIGAI
[3]Paul G. Allen School of Computer Science and Engineering, University of Washington
[4]Beijing Jiaotong University
hanszhao@stu.pku.edu.cn, zefncai@gmail.com
https://github.com/HaozheZhao/MIC

ABSTRACT

Starting from the resurgence of deep learning, vision-language models (VLMs) benefiting from large language models (LLMs) have never been so popular. However, while LLMs can utilize extensive background knowledge and task information with in-context learning, most VLMs still struggle with understanding complex multi-modal prompts with multiple images. The issue can traced back to the architectural design of VLMs or pre-training data. Specifically, the current VLMs primarily emphasize utilizing multi-modal data with a single image some, rather than multi-modal prompts with interleaved multiple images and text. Even though some newly proposed VLMs could handle user prompts with multiple images, pre-training data does not provide more sophisticated multi-modal prompts than interleaved image and text crawled from the web. We propose MMICL to address the issue by considering both the model and data perspectives. We introduce a well-designed architecture capable of seamlessly integrating visual and textual context in an interleaved manner and MIC dataset to reduce the gap between the training data and the complex user prompts in real-world applications, including: 1) multi-modal context with interleaved images and text, 2) textual references for each image, and 3) multi-image data with spatial, logical, or temporal relationships. Our experiments confirm that MMICL achieves new stat-of-the-art zero-shot and few-shot performance on a wide range of general vision-language tasks, especially for complex reasoning benchmarks including MME and MMBench. Our analysis demonstrates that MMICL effectively deals with the challenge of complex multi-modal prompt understanding. The experiments on ScienceQA-IMG also show that MMICL successfully alleviates the issue of language bias in VLMs, which we believe is the reason behind the advanced performance of MMICL . Our code, dataset, checkpoints, and demos are available at https://github.com/HaozheZhao/MIC.

1 INTRODUCTION

General-purpose vision-language pre-trained models (VLMs) have been making significant advancements in the realm of AI's multi-modality capabilities. (Li et al., 2022; 2023e;h; Huang et al., 2023b; Zhu et al., 2023; Li et al., 2023c;f). Recent VLMs mostly augment a large language model (LLM) with a visual encoder and exhibit impressive zero-shot capacities in a variety of visual tasks. However, unlike LLMs that can extract rich background knowledge and task information from the prompt with *in-context learning* (ICL), most VLMs still fall short of understanding complex multi-modal prompts with multiple images. As shown in Figure 1, vision-language dialogue typically contains multiple images and the user may query with any of the images. Such inability makes VLMs less effective at handling real-world user queries, ultimately impacting their performance in downstream vision-language tasks. Specifically, the current VLMs have the following three limitations.

[*]Equal contribution.
[†]Corresponding author.

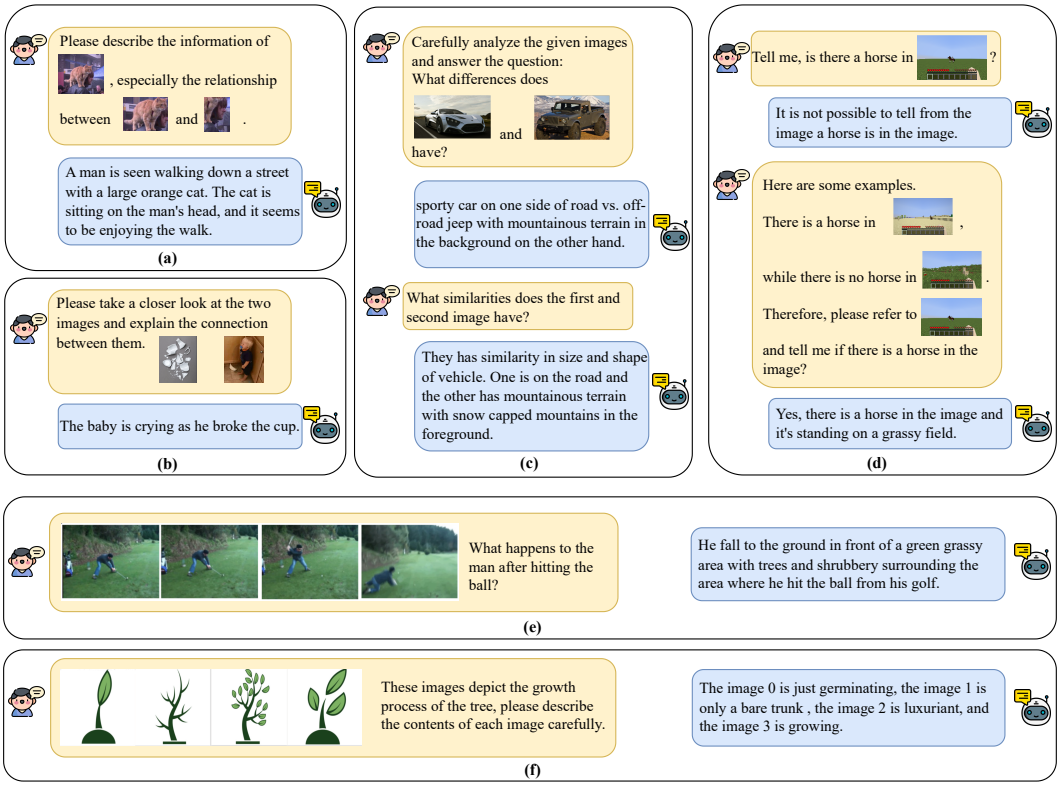

Figure 1: Examples of vision-language dialogue generate by MMICL . A vision-language dialogue typically contains prompt with interleaved images and text. MMICL understands spatial (**a**), logical (**b**), and temporal (**e**) image relationships. Additionally, MMICL can grasp the intricate text to image reference as **(c)**,**(d)** and **(f)**.

Hard to Understand Complex Prompt With Multiple Images and Text: Previous studies have primarily emphasized utilizing multi-modal data with a single image, rather than multi-modal prompts with interleaved multiple images and text. As a result, these models often exhibit poor performance or are unable to handle multi-image inputs when users provide more than one image during a conversation and ask about a particular one or several of them. This can be attributed to the architectural design: popular VLMs like BLIP-2. (Li et al., 2023e), MiniGPT-4 (Zhu et al., 2023), LLaVA (Li et al., 2023c) and InstructBLIP (Dai et al., 2023) do not support processing multi-modal prompt with more than one image. Even though newly proposed Flamingo (Alayrac et al., 2022) and Kosmos-1 (Huang et al., 2023b) could handle user prompts with multiple images, their pre-training data does not provide more sophisticated multi-modal prompts than interleaved image and text crawled from the web (Awadalla et al., 2023), which creates a gap between the prompts used in training of these VLMs and the real-world application.

Hard to Understand Text-to-Image Reference: In real-world scenarios, the text and images in user queries are closely associated, i.e., certain parts of the text refer to particular images. As the example shown in case (c), case (d) and case (f) in Figure 1, the user may ask a specific question about multiple images, or the user may use multiple images as exemplars to ask the question only about the last image. However, the training data used in previous studies (e.g., Flamingo) (Li et al., 2023e; Dai et al., 2023; Awadalla et al., 2023; Alayrac et al., 2022; Huang et al., 2023a) are typically crawled from the web and lack explicit text-to-image references. Consequently, it is difficult for these VLMs to effectively handle user queries involving intricate text-to-image references.

Hard to Understand the Relationships between Multiple Images: The training data used by previous VLMs are mainly crawled from the internet, so the connections between images may be weak, especially when these images are far apart on the same webpage. The lack of the closely interconnected images in the training data hampers the ability of VLMs to understand complex

relationships between multiple images in user prompts (e.g., spatial, logical, and temporal relationships), which further limits the model's reasoning ability and few-shot capability in downstream vision-language tasks.

Therefore, we propose MMICL , a simple yet efficient pipeline to address these limitations of VLMs from both model and data perspectives as Sec. 2.

For model perspective, we introduce a novel network capable of handling multi-modal inputs with multiple images in Sec. 2.1. For each image input, we utilize the same visual prompt generator (i.e., Q-Former (Li et al., 2022)) to generate visual embeddings from the image features encoded by vision backbones (e.g., ViT (Radford et al., 2021; Fang et al., 2023)). These visual embeddings along with text embeddings are used to construct interleaved inputs for LLMs (i.e., FLAN-T5 (Chung et al., 2022). In this way, images and text can be combined in any order. This approach enables our model to handle multi-modal inputs with more than one image, paving the way for VLMs to effectively comprehend complex multi-modal prompts.

For data perspective, we introduce MIC (Multimodality In-Context Learning) dataset to help VLMs understand more complex multi-modal prompts with multiple images for real-world applications in Sec. 2.2. Specifically, MIC consists of 1) multi-modal prompt with interleaved images and text; 2) textual references for each image; and 3) multi-image data with spatial, logical, and temporal relationships. Drawing inspiration from Li et al. (2023a); Alayrac et al. (2022), we re-organize the data from existing datasets in a user-friendly multi-model-in-context format that incorporates multiple images and sentences in context, as demonstrated in case (d) of Figure 1. We also design templates for each interleaved multi-modal input. We use different vision tokens to refer to visual prompts and use natural language to establish accurate text to image references. Additionally, we introduce interconnected images data that are related spatially, logically, and temporally, as demonstrated in case (a) and (e) of Figure 1. Such data format helps the model learn complex referential relationships in the multi-modal prompts and close relationship between the images, eventually understand complex vision-language prompts with multiple images.

Our experiments (Sec. 3) confirm that MMICL shows amazing multi-modal in-context learning ability and achieves new state-of-the-art performance on a wide range of vision-language tasks including MME (Fu et al., 2023) and MMBench (Liu et al., 2023c) * (Sec. 3.2). We also make comprehensive evaluation on the in-context learning ability of MMICL (Sec. 3.3). We further take a direct prob of the issues we aim to solve: understanding multiple images in complex prompts, text-to-image reference, and understanding complex relationships among images. For multiple image understanding, MMICL achieves state-of-the-art performance on various video-answering benchmarks (Sec. 3.4.1). For text-to-image reference, MMICL outperforms previous works by approximately 13 points in accuracy on the Winoground benchmark (Thrush et al., 2022a) (Sec. 3.4.2). For understanding complex relationships among images, MMICL outperforms reported state-of-the-art performance by 12 points on the RAVEN benchmark (Huang et al., 2023a) (Sec. 3.4.3). For detailed qualitative evaluation, we find MMICL effectively alleviates the language bias commonly encountered in VLMs. This is primarily due to the design of text-to-image templates in MIC , as they encourage the model to better understand the referential relationships between text and image. This design helps the model focus on visual content and perform multi-modal reasoning to comprehend complex images and contextual text (Sec. 3.5).

To sum up, our contributions can be summarized as follows:

i) We discuss and point out the limitations in current VLMs. We attempt to address these limitations from model perspective and data perspective, respectively. Specifically, we introduce a novel network capable of handling multi-modal inputs with multiple images. We further propose MIC dataset to help model understand complex prompt with multiple images and text.

ii)Based on the well-designed architecture and dataset, we further propose MMICL which can perform various multi-modal tasks (e.g., visual dialogue, visual question-answering) based on only a few multi-modal in-context examples. MMICL can support multi-modal inputs with more than one image and understand the complex multi-modal prompts in real-world applications.

iii) By addressing the aforementioned limitations, experiments show that MMICL achieves a new state-of-the-art performance on general VLM benchmarks (e.g., MME and MMBench) and complex

*Results of MMICL are submitted on August 23rd, 2023.

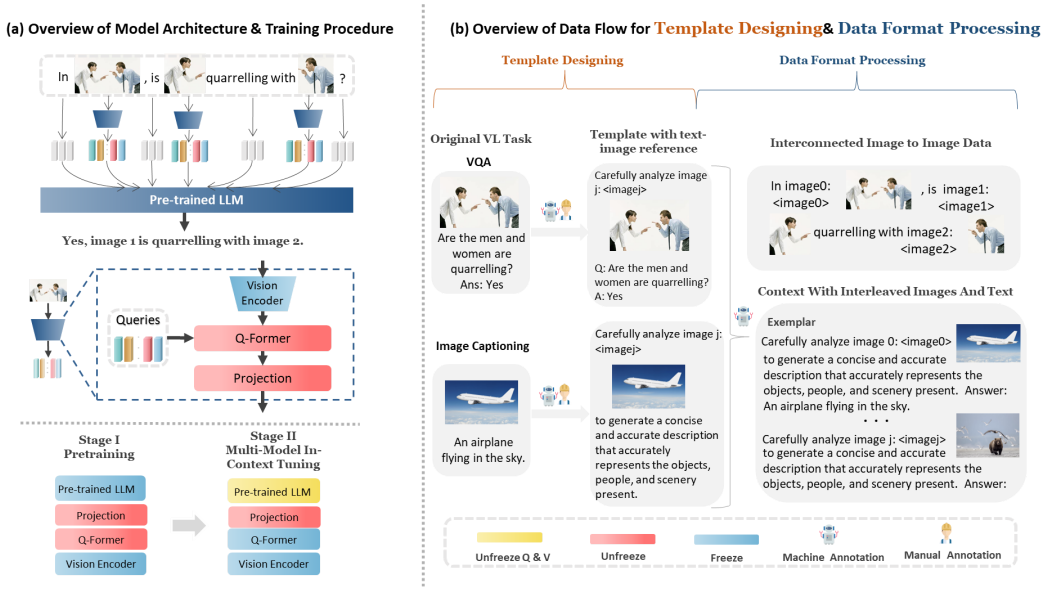

Figure 2: Illustration of MMICL . The left part denotes the overview of model architecture (**Sec. 2.1**) and the pipeline of the two-stage training paradigm (**Sec. 2.3**) includes Pretraining, and Multi-Modal In-Context Tuning as described. The blue color denotes freezing parameters, the red color denotes unfreezing parameters, and the yellow color denotes unfreezing the weights of mapping query and value vectors in the attention layer of LLMs. The right part illustrates the overview of the data construction of the MIC dataset (**Sec. 2.2**).

multi-modal reasoning benchmarks (e.g., RAVEN, Winoground). MMICL also effectively alleviates the language bias commonly encountered in VLMs.

2 MMICL

MMICL comprises three parts: Model structure to support multi-image context (Section 2.1); MIC dataset with 1) multi-modal context with interleaved images and text, 2) textual references for each image, and 3) multi-image data with spatial, logical, or temporal relationships. (Section 2.2) ; training strategy to train MMICL based on MIC (Section 2.3).

2.1 Model Structure

In this section, we will cover the model structure of MMICL .

Most VLMs utilize Visual-Prompt Generators (VPG) (e.g., linear projection (Liu et al., 2023b), Resampler (Alayrac et al., 2022), Q-former (Li et al., 2023e)) to extract visual prompts from the image features encoded by the vision backbones, and use visual prompts to help LLMs understand visual inputs. Remarkable zero-shot performance though they may achieve, they only focus on vision-language tasks with only a single image as the visual context and do not support complex context with multiple images in real-world applications. We propose MMICL with flexibility in interaction by allowing free combination of images and text, to support complex context with multiple images. Users can input many images and text in any order. There is no limitation on images as inputs.

The proposed model structure is shown as Figure 3. Each given image is encoded by a vision encoder (e.g., ViT (Radford et al., 2021; Fang et al., 2023)) to get the vision representation of the image. Then we use the Q-former as the VPG to extract the visual prompt. we then use a fully connected layer as the language projection to transform each visual prompt into the same dimension as the text embedding of the LLMs to align the visual prompt to the textual context for the LLM to understand.

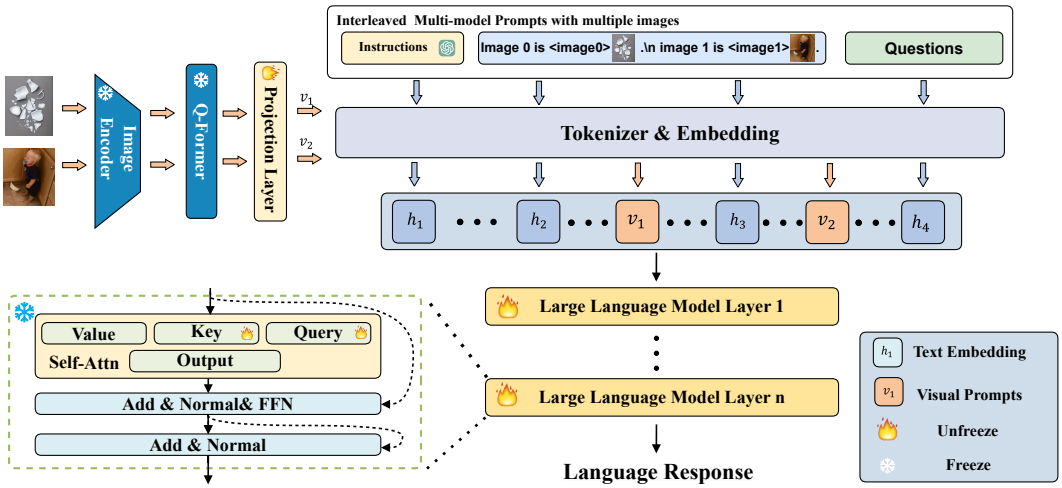

Figure 3: Illustration of the MMICL structure.

Finally, we combine the visual prompts for multiple images with text embeddings in an interleaved style and then feed them into the LLM.

A strong limitation of BLIP (Li et al., 2023e; Dai et al., 2023) in real-world applications is the inevitability of placing the image at the front of the entire input MMICL forces the model to treat the image and language representations equally and combines them into interleaved image-text representations according to the original input.

Inspired by LoRa (Hu et al., 2021), we set the weights for mapping query and value vectors in the attention layer of LLM as learnable to better adapt to the multi-modal context with multiple images. During the pre-training, we freeze the image encoder, Q-former, and the backbone LLM, while jointly training the language projection and the query and value vectors of the LLM.

2.2 Multimodality In-Context Learning (MIC) Dataset

In this section, we will cover the data resources and the construction of MIC dataset.

2.2.1 Data Resource

MIC dataset comes from 8 task categories and 16 datasets.

Image Captioning aims to produce descriptions of the given images according to different needs. Our training dataset includes MS COCO (Lin et al., 2014), DiffusionDB (Wang et al., 2022b)[†], and Flickr 30K (Young et al., 2014).

Knowledgeable Visual Question Answering (KVQA) requires the model to make use of commonsense knowledge outside the input image to answer questions. Our training dataset includes OK-VQA (Marino et al., 2019).

Image Question Answering (IQA) requires the model to answer questions based on the image correctly. Our training dataset includes VQAv2 (Goyal et al., 2017), ST-VQA (Biten et al., 2019), Text-VQA (Singh et al., 2019), WikiART (Saleh & Elgammal, 2015) and RefCOCO (Yu et al., 2016).

Video Question Answering (VideoQA) requires the model to answer questions based on the video[‡] correctly. Our training dataset includes MSRVTTQA (Xu et al., 2016).

[†]converted from origin dataset, see detail in Appendix B

[‡]We extract 8-12 frames from the video to construct multi-image input.

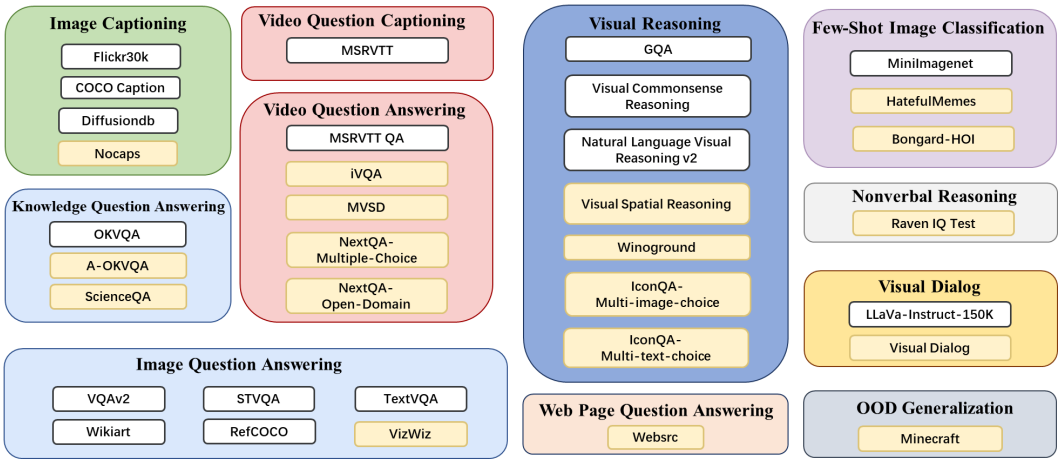

Figure 4: Illustration of the data resource used to construct MIC dataset. It consists of 11 tasks, and 33 different datasets. The held-in datasets are indicated by white and the held-out datasets are indicated by yellow.

Video Captioning requires the model to give the caption based on the video Our training dataset includes MSRVTT (Xu et al., 2016).

Visual Reasoning requires the model to perform image reasoning and answer questions correctly. Our training dataset includes GQA (Hudson & Manning, 2019), VCR (Zellers et al., 2019), and NLVR2 (Suhr et al., 2018).

Image Classification involves classifying an image based on a given set of candidate labels. Our training dataset includes MiniImage (Russakovsky et al., 2015).

Visual Dialog requires the model to hold a meaningful dialog with humans in natural, conversational language about visual content. Our training dataset includes LLAVA-Instruct-150K (Liu et al., 2023b).

Please refer to Appendix B for more details and data statistics about the training and testing dataset.

2.2.2 Data Format

To enhance VLMs to understand the complex prompts in real-world applications, we gather data from the data resources and construct the multi-modality in-context learning dataset MIC . MIC mainly consists of three parts: 1) multi-modal context with interleaved images and text, 2) textual references for each image, and 3) multi-image data with spatial, logical, or temporal relationships.

Multi-modal Context with Interleaved Images and Text MIC dataset consists of data with multiple images and text to enable the model to learn and comprehend the complex context with interleaved images and text. Such data helps VLMs to work in complex scenarios in real-world applications.

Specifically, we gather data from datasets[§] mentioned in data resource into a multi-image-in-context format, as shown below:

Each instance $\mathbf{I}_i$ comprises N exemplars.

$$\mathbf{I}_i = (\{\mathbf{P}_1, \cdots, \mathbf{P}_N\}, \mathbf{X}_i, \mathbf{q}_i, \mathbf{a}_i) \tag{1}$$

Each exemplar $\mathbf{P}_j$ where $j \in [1, ..., N]$ is represented as a tuple $\mathbf{P}_j = (\mathbf{X}_j, \mathbf{q}_j, \mathbf{a}_j)$, $\mathbf{X}_j$ denotes the visual information of the j-th exemplar. $\mathbf{q}_j$ denotes the j-th question, and $\mathbf{a}_j$ denotes the j-th answer

[§]Except for the video datasets, vcr dataset, and LLaVa dataset. More detail can be found in Appendix B

for the j-th exemplar. For instance $\mathbf{I}_i$, $\mathbf{X}_i$ denotes the visual information, $\mathbf{q}_i$ denotes the question, and $\mathbf{a}_i$ denotes the answer. The organized input data format is as follows [¶]:

```
# In-Context Demonstration
Image 0 is <image 0> {image 0}.
Human:{question}
Assistant:{answer}
...
Image 4 is <image 4> {image 4}.
Human:{question}
Assistant:{answer}
# Image Input for Qury
Image 5 is <image 5> {image 5}.
# Text Input for Query
Human:{question}
Assistant:{answer}
```

In the data format, **<**image j**>** is the image token, a special token for exact reference, and image j is the visual representation corresponding to the j-th image. The design of image input helps the user exact reference to image inputs.

Textual References for Each Image In real-world scenarios, users may use textual descriptions to explicitly or implicitly refer to a specific image. The reference in the textual context can provide information about the visual content that is being described or mentioned in the text to the VLM, allowing it to learn alignment between the two modalities. However, the training data used in previous works are crawled from the web and lack such reference, which results in poor performance in the understanding of text-to-image reference.

We focus on exact references between textual and visual contexts. In most cases, the j-th image in an instance can be referred in natural language by *image* j, but there exist challenging settings where images are not easily identifiable by language in a few words.

To address the aforementioned problems and accurately establish the text-to-image reference, we create templates for each interleaved input. The template labels images with image tokens (<image j>, a special token for image reference) that correspond to an image representation image j in the input prompt. We then use natural language prompts to establish text-to-image references in training. The prompt template is displayed below:

```
Image 1 is <image 1>. {image 1}

....

Image j is <image j>. {image j} {question}
```

Then we can build the reference in the multi-modal context to a rough reference based on natural language or an exact reference based on an image token. When referring to the j-th image, we can directly refer to the image using **image** j in natural language or image token **<image**j**>** in the context.

Multi-image Data with Spatial, Logical, or Temporal Relationships In MIC dataset, we also create interconnected image data that incorporates spatial, logical, and temporal relationships. Initially, we use video data from the MSRVTT (Xu et al., 2016) and MSRVTTQA (Xu et al., 2016) datasets to construct multi-image data with temporal relationships.

For video data, we extract 8 frames from each video instance following Dai et al. (2023) to create a multi-image input. The frame images extracted from video data naturally have a close relationship in both time and space, which helps the model understand the spatial and temporal relationship between multiple images.

We construct multi-image data with spatial and logical relationships by using bounding boxes of objects from the VCR dataset As shown in case (a) in Figure 1, we crop the images based on the bounding boxes of objects to obtain multiple sub-images based on a main image. We then replace the textual content mentioning the objects with the corresponding cropped images to create context with interleaved images and text.

[¶]The text after the pound sign (#) is for explanation rather than the actual input.

Such image data improves reasoning ability and helps MMICL understand spatial, logical, and temporal connections between images. The multi-image prompts will be formulated in the following format:

Each instance $\mathbf{I}_i$ comprises a question-answer text pair along with a set of K images and is represented as:

$$\mathbf{I}_i = (\mathbf{X}_i, \mathbf{q}_i, \mathbf{a}_i) \tag{2}$$

where $\mathbf{x}_{i,k} \in \mathbf{X}_i$, each $\mathbf{x}_{i,k}$ represents the k-th image of the image set and $k \in [1, ..., K]$. The organized input data format (also refer to as the prompt template) is presented as follows ‖:

```
# Image Input for Qury
Image 0 is <image0> {image0}
...
Image j is <imagej> {imagej}.
# Text Input for Query
Human:{question}
Assistant:{answer}
```

2.2.3 Data Construction Pipeline

We use the following three stages to make data construction from the data resources. For each data resource, we designed prompt templates with different instructions. We convert most datasets into a vision-language question-answering format to create high-quality multi-modal data for training.

Stage I: Template Designing We request annotators to carefully read the original paper of the dataset resource and examine several samples to obtain a detailed understanding of the task in order to create high-quality prompt templates. Next, we use ChatGPT ** to rewrite ten diverse task instructions that describe the key characteristics of the task. After ChatGPT generates the instructions based on the initial request, we manually review them to ensure high quality.

Stage II: Data Format Processing After the manual selection of the template based on the task characteristics, we further fill the input content into the template to generate data in a unified format. For the VCR dataset, we process the dataset to the text-image interleaved format as Equation 2 to build the data. For other data resources, we transfer them into multi-modal-in-context data format as appending few-shot exemplars as demonstrations (Equation 1).

During our data construction process, we obtained a total of 5.8M samples in the Multi-modal ICL data. Due to resource constraints, we extracted approximately 10% of the data using the sampling strategy described in Sec. C.2 to finetune MMICL . It is anticipated that a larger model trained on all of our data would yield a more promising outcome.

2.3 Training Paradigm

The training paradigm mainly includes two stages.

Stage I: Pretraining This stage is designed to acquire vision-language knowledge from a large collection of aligned image-text pairs following the training stage of BLIP-2 (Li et al., 2023e). During this stage, both the vision encoder and the LLM remain frozen. The VPGs(Q-Former) are trained to learn visual representation which can be interpreted by the LLM by training on LAION-400M (Schuhmann et al., 2021).

Stage II: Multi-Modal In-Context Tuning We further perform In-Context Tuning using the MIC dataset constructed to realize the ability to process multi-image inputs. The data used is introduced in Sec. 2.2. In this stage, we take our model a step further by extending it to multi-modal In-Context Learning. This stage allows us to fully leverage the impressive potential of VLMs to understand complex contexts with multiple images and text.

‖The text after the pound sign (#) is for explanation rather than the actual input.

**The `gpt-3.5-turbo` version is ChatGPT. We use the *gpt-3.5-turbo* version of ChatGPT to complete all of our experiments.

3 EXPERIMENT

In this section, we extensively evaluate the zero-shot and few-shot performance of MMICL . Experimental details are provided in Sec. 3.1. MMICL is evaluated on two popular multi-modal large language model benchmarks, MME (Fu et al., 2023) and MMBench (Liu et al., 2023c) as Sec. 3.2. We also evaluate the multi-modal in-context learning ability of MMICL across various vision-language tasks as Sec. 3.3. We are also interested in the following research questions and make further analysis about them:

Q1: Is MMICL able to handle multiple images in complex prompts?

Q2: Is MMICL able to capture the correct text-image reference?

Q3: Is MMICL able to understand complex relationships between images?

We answer these three questions by performing experiments in Sec. 3.4. We also conduct additional experiments to discover that MMICL effectively mitigates the language bias issue commonly faced in VLMs in Sec. 3.5. The ablation study of MMICL is conducted in Sec. 3.6

3.1 EXPERIMENTAL SETUP

Implementation Details Following Chung et al. (2022), we use FLANT5-XL and FLANT5-XXL (Chung et al., 2022) as the backbone LLMs. In Stage I Pretraining as Sec. 2.3, we set the vision encoder and language model to be frozen and utilize the COCO captioning data and LAION-400M data (Schuhmann et al., 2021) to perform feature alignment training on the Q-former. We keep the other part of the VLM frozen and jointly train the Q-former and the language projection. To benefit from BLIP-2's significant visual representation extraction ability, we integrate its powerful vision encoder to initialize the Q-former and the language projection. [††]. In Stage II Multi-Modal In-Context Tuning as Sec. 2.3, we train the model for 3 epochs and with a lower learning rate of $1e-5$. The weights of mapping query and value vectors in the attention layer of LLMs are learnable in this stage to better adapt to the multi-modal prompts with multiple images. In this stage, we freeze the visual encoder, Q-former, and the backbone LLM, and jointly train the language projection, the query vectors and value vectors of the LLM.

All experiments are conducted with 6 NVIDIA A40 GPUs with the zero2-offload (Rajbhandari et al., 2020) of Deepspeed (Rasley et al., 2020) with the trainer of huggingface transformers (Wolf et al., 2020). The batch size is set as 10 and 4 for MMICL (FLAN-T5-XL) and MMICL (FLAN-T5-XXL), respectively. Largest MMICL (FLAN-T5-XXL) requires about 2 days for the Stage II.

Evaluation Setup The primary focus of our work is to develop general-purpose vision-language models that can generally adapt to diverse challenging multi-modal prompts. To achieve this goal, we utilize a diverse set of popular vision-language benchmarks, including tasks that involve images and videos. The metrics used in these benchmarks are shown in Appendix F. For Video QA tasks, we extract 8 frames per video as visual inputs.

Models and Baselines We provide two versions of MMICL : (1) MMICL (FLAN-T5) which uses BLIP-2 (Li et al., 2023e) as the backbone and (2) MMICL (Instruct-FLAN-T5 which uses InstructBLIP (Dai et al., 2023) as the backbone. For both two versions, we adopt two sizes of FLANT5 (Chung et al., 2022) models including FLANT5-XL and FLAN-XXL.

Baselines We primarily compare MMICL with recently proposed powerful multi-modal approaches, including **(1) Flamingo** (Alayrac et al., 2022) where a VLM is trained on large-scale multi-modal web corpora containing arbitrarily interleaved text and images; **(2) KOSMOS-1** (Huang et al., 2023a) which is trained from scratch on web-scale multi-modal corpora; **(3) BLIP-2-FLAN-T5** (Li et al., 2023e) where an instruction-tuned Flan-T5 (Chung et al., 2022) is connected with a powerful visual encoder to perform a series of multi-modal tasks; **(4) InstructBLIP-FLAN-T5** (Dai et al., 2023), a recently proposed instruction tuning enhanced model; **(5) Shikra** (Chen et al., 2023), a VLM which can handle spatial coordinate inputs and outputs in natural language; **(6) Otter** (Li et al., 2023b), an open-source implementation of flamingo (Alayrac et al., 2022) and trained with

[††]We use BLIP-2 and InstructBlip as the backbone for MMICL , so Stage I is skipped.

Model	Commonsense Reasoning	Numerical Calculation	Text Translation	Code Reasoning	Avg.
MiniGPT-4 (Zhu et al., 2023)	59.29	45.00	0.00	40.00	36.07
VisualGLM-6B (Du et al., 2021)	39.29	45.00	50.00	47.50	45.45
LLaVA (Liu et al., 2023b)	57.14	50.00	57.50	50.00	53.66
Lynx (Zeng et al., 2023)	110.71	17.50	42.50	45.00	53.93
MultiModal-GPT (Gong et al., 2023)	49.29	62.50	60.00	55.00	56.70
LLaMA-Adapter-V2 (Gao et al., 2023)	81.43	62.50	50.00	55.00	62.23
LaVIN (Luo et al., 2023)	87.14	65.00	47.50	50.00	62.41
GIT2 (Wang et al., 2022a)	99.29	50.00	67.50	45.00	65.45
mPLUG-Owl (Ye et al., 2023)	78.57	60.00	80.00	57.50	69.02
BLIP-2 (Li et al., 2023e)	110.00	40.00	65.00	75.00	72.50
InstructBLIP (Dai et al., 2023)	129.29	40.00	65.00	57.50	72.95
Otter (Li et al., 2023b)	106.43	72.50	57.50	70.00	76.61
Cheetor (Li et al., 2023d)	98.57	77.50	57.50	87.50	78.02
LRV-Instruction (Liu et al., 2023a)	100.71	70.00	85.00	72.50	<u>82.05</u>
MMICL	117.86	62.50	107.50	72.50	**90.09**

Table 1: **Evaluation of cognition**. In the MME benchmark, each image will have two questions, with answers restricted to 'yes' or 'no'. The evaluation metrics for this benchmark include ACC and ACC+. ACC refers to the accuracy calculated for each individual question, while ACC+ represents the accuracy for each image, where both questions must be answered correctly. The Avg. metric denotes the average value across all numbers. It is important to note that all the reported figures for the baseline methods are obtained from the MME benchmark (Fu et al., 2023). We use the FLAN-T5-XXL version of MMICL to evaluate the performance.

multi-modal instruction in-context tuning data.; **(7) Ying-VLM** (Li et al., 2023f), a VLM model trained on Multi-Modal multilingual instruction tuning dataset.

Model	Existen.	Count	Pos.	Color	OCR	Poster	Cele.	Scene	Land.	Art.	Avg.
LLaVA	50.00	50.00	50.00	50.00	50.00	50.00	48.82	50.00	50.00	49.00	50.28
MiniGPT-4	68.33	55.00	43.33	43.33	57.50	41.84	54.41	71.75	54.00	60.50	64.26
MM-GPT	61.67	55.00	58.33	58.33	82.50	57.82	73.82	68.00	69.75	59.50	65.47
VisualGLM-6B	85.00	50.00	48.33	48.33	42.50	65.99	53.24	146.25	83.75	75.25	70.53
LaVIN	185.00	88.33	63.33	63.33	107.50	79.59	47.35	136.75	93.50	87.25	96.36
mPLUG-Owl	120.00	50.00	50.00	50.00	65.00	136.05	100.29	135.50	159.25	96.25	96.73
LLaMA-A.-V2	120.00	50.00	48.33	48.33	125.00	99.66	86.18	148.50	150.25	69.75	97.27
InstructBLIP	185.00	143.33	66.67	66.67	72.50	123.81	101.18	153.00	79.75	134.25	121.28
Otter	195.00	88.33	86.67	86.67	72.50	138.78	172.65	158.75	137.25	129.00	129.23
BLIP-2	160.00	135.00	73.33	73.33	110.00	141.84	105.59	145.25	138.00	136.50	129.38
LRV-Instruct.	165.00	111.67	86.67	86.67	110.00	139.04	112.65	147.98	160.53	101.25	129.98
Cheetor	180.00	96.67	80.00	80.00	100.00	147.28	164.12	156.00	145.73	113.50	130.00
GIT2	190.00	118.33	96.67	96.67	65.00	112.59	145.88	158.50	140.50	146.25	133.21
Lynx	195.00	151.67	90.00	90.00	77.50	124.83	118.24	164.50	162.00	119.50	<u>137.32</u>
MMICL	175.00	143.33	73.33	73.33	112.00	130.95	145.88	152.75	135.08	133.00	**137.60**

Table 2: **Evaluation of coarse-grained and fine-grained recognition and OCR**. The settings are the same as Table 1. It is important to note that all the reported figures for the baseline methods are obtained from the MME benchmark (Fu et al., 2023). We use the FLAN-T5-XXL version of MMICL to evaluate the performance.

3.2 Quantitative Evaluations

3.2.1 MME Benchmark

MME comprehensively evaluates VLMs with 14 sub-tasks that encompass perception and cognition abilities. Other than OCR, perception ability includes the recognition of coarse-grained and fine-grained objects. The former identifies the existence, count, position, and color of objects. The latter

Method	Language Model	Vision Model	Overall	LR	AR	RR	FP-S	FP-C	CP
MMGPT	LLaMA-7B	CLIP ViT-L/14	16.0	1.1	23.8	20.7	18.3	5.2	18.3
MiniGPT-4	Vincuna-7B	EVA-G	12.0	13.6	32.9	8.9	28.8	11.2	28.3
PandaGPT	Vincuna-13B	ImageBind ViT-H/14	30.6	15.3	41.5	22.0	20.3	20.4	47.9
VisualGLM	ChatGLM-6B	EVA-CLIP	33.5	11.4	48.8	27.7	35.8	17.6	41.5
InstructBLIP	Vincuna-7B	EVA-G	33.9	21.6	47.4	22.5	33.0	24.4	41.1
LLaVA	LLaMA-7B	CLIP ViT-L/14	36.2	15.9	53.6	28.6	41.8	20.0	40.4
G2PT	LLaMA-7B	ViT-G	39.8	14.8	46.7	31.5	41.8	34.4	49.8
Otter-I	LLaMA-7B	CLIP ViT-L/14	48.3	22.2	63.3	39.4	46.8	36.4	60.6
Shikra	Vincuna-7B	CLIP ViT-L/14	60.2	33.5	69.6	53.1	61.8	50.4	71.7
LMEye	Flan-XL	CLIP ViT-L/14	61.3	36.9	73.0	55.4	60.0	68.0	68.9
mPLUG-Owl	LLaMA-7B	CLIP ViT-L/14	62.3	37.5	75.4	56.8	67.3	52.4	67.2
JiuTian	FLANT5-XXL	EVA-G	<u>64.7</u>	46.6	76.5	66.7	66.5	51.6	68.7
MMICL	FLAN-T5-XXL	EVA-G	**65.24**	44.32	77.85	64.78	66.5	53.6	70.64

Table 3: **Evaluation of MM benchmark dev set**. All the reported performance for the baseline methods is from the leaderboard of MM benchmark (Liu et al., 2023c). We use the FLAN-T5-XXL version of MMICL to evaluate the performance.

recognizes movie posters, celebrities, scenes, landmarks, and artworks. The cognition includes commonsense reasoning, numerical calculation, text translation, and code reasoning.

MME evaluates a wide range of multi-modal abilities. The compared baselines include LLaVA (Liu et al., 2023b), MiniGPT-4 (Zhu et al., 2023), MultiModal-GPT (Gong et al., 2023), VisualGPM-6B (Du et al., 2021), LaVIN (Luo et al., 2023), mPLUG-Owl (Ye et al., 2023), LLaMA-Adapter-V2 (Gao et al., 2023), InstructBLIP (Dai et al., 2023), Otter (Li et al., 2023b), BLIP-2 (Li et al., 2023e), LRV-Instruction (Liu et al., 2023a), Cheetor (Li et al., 2023d), GIT2 (Wang et al., 2022a), Lynx (Zeng et al., 2023). Results show that MMICL can achieve the best average scores in comparisons with current VLMs as shown in Table 1 and Table 2, especially in cognition tasks.

3.2.2 MM BENCHMARK

We further evaluate the comprehensive capability of our MMICL in MMBench (Liu et al., 2023c), a thoughtfully designed benchmark that thoroughly evaluates the diverse skills of vision-language models. The results from the test set are presented in Table 3 and MMICL also achieves state-of-the-art performance on MMBench and outperforms other VLMs by a large margin. MMICL achieves remarkable success on Attribute reasoning (AR) compared with other VLMs, which indicates that MMICL can recognize characters and make good understanding of the information in the image. This means that MIC is helpful for MMICL to understand visual information.

3.3 MULTI-MODAL IN-CONTEXT LEARNING EVALUATION

We evaluate the multi-modal in-context learning ability of MMICL across various vision-language tasks. The result is present in the Table 4. In general, MMICL outperforms other VLMs models on both the held-in and held-out datasets and achieve the state-of-art few-shot performance. For example, few-shot evaluation (4-shot) of MMICL on the VizWiz (Bigham et al., 2010) benchmark outperforms the baseline Flamingo-9B (Alayrac et al., 2022) and KOSMOS-1 (Huang et al., 2023b) by 15.38 and 14.98 points, respectively. Since VizWiz has never been exposed in the training data, this superior suggests the ability of MMICL to generalize to new tasks with a few exemplars. The few-shot performance of Flickr30K decreases with examples given because the captions examples provide noise for VLM to finish the task (i.e. in-context exemplars generally do not provide hinds for models to perform image captioning tasks).

3.4 PERFORMANCE PROB

3.4.1 UNDERSTANDING MULTIPLE IMAGES IN THE MULTI-MODAL PROMPT

Videos contain more temporal information compared to static images. We test MMICL across different video-languages tasks to evaluate whether the MMICL is able to support the multiple

Model	Flickr 30K	WebSRC	VQAv2	Hateful Memes	VizWiz
Flamingo-3B (Alayrac et al., 2022) (Zero-Shot)	60.60	-	49.20	53.70	28.90
Flamingo-3B (Alayrac et al., 2022) (4-Shot)	72.00	-	53.20	53.60	34.00
Flamingo-9B (Alayrac et al., 2022) (Zero-Shot)	61.50	-	51.80	57.00	28.80
Flamingo-9B (Alayrac et al., 2022) (4-Shot)	72.60	-	56.30	62.70	34.90
KOSMOS-1 (Huang et al., 2023b) (Zero-Shot)	67.10	3.80	51.00	63.90	29.20
KOSMOS-1 (Huang et al., 2023b) (4-Shot)	75.30	-	51.80	-	35.30
Zero-Shot Evaluation					
BLIP-2 (Li et al., 2023e) (FLANT5-XL)	64.51	12.25	58.79	60.00	25.52
BLIP-2 (Li et al., 2023e) (FLANT5-XXL)	60.74	10.10	60.91	62.25	22.5
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	11.5	63.69	61.70	15.11
Zero-Shot Evaluation					
MMICL (FLAN-T5-XL)	60.56	12.55	62.17	60.28	25.04
MMICL (FLAN-T5-XXL)	78.64	18.85	69.99	60.32	29.34
MMICL (Instruct-FLAN-T5-XL)	**78.89**	14.75	69.13	61.12	29.92
MMICL (Instruct-FLAN-T5-XXL)	44.29	17.05	70.30	62.23	24.45
Few-Shot (4-Shot) Evaluation					
MMICL (FLAN-T5-XL)	71.95	12.30	62.63	60.80	50.17
MMICL (FLAN-T5-XXL)	75.37	18.70	69.83	61.12	33.16
MMICL (Instruct-FLAN-T5-XL)	74.27	14.80	69.16	61.12	33.16
MMICL (Instruct-FLAN-T5-XXL)	72.04	**19.65**	**70.56**	**64.60**	**50.28**

Table 4: Main results of multi-modal in-context learning ability of MMICL across vision-language tasks. All evaluation metrics used in the evaluation is introduced as Table 16.

Model	MSVD QA	NExT QA Multi-choice	iVQA
Flamingo-3B (Alayrac et al., 2022) (Zero-Shot)	27.50	-	32.70
Flamingo-3B (Alayrac et al., 2022) (4-Shot)	33.00	-	35.20
Flamingo-9B (Alayrac et al., 2022) (Zero-Shot)	30.20	-	35.20
Flamingo-9B (Alayrac et al., 2022) (4-Shot)	36.20	-	37.70
Flamingo-80B (Alayrac et al., 2022) (Zero-Shot)	35.60	-	40.70
Flamingo-80B (Alayrac et al., 2022) (4-Shot)	41.70	-	44.10
R2A (Pan et al., 2023)	37.00	-	29.30
BLIP-2 (Li et al., 2023e) (FLANT5-XL)	33.70	61.73	37.30
BLIP-2 (Li et al., 2023e) (FLANT5-XXL)	34.40	61.97	49.38
InstructBLIP (Dai et al., 2023) (FLANT5-XL)	43.40	36.10	25.18
InstructBLIP (Dai et al., 2023) (FLANT5-XXL)	44.30	64.27	36.15
MMICL (FLAN-T5-XL)	47.31	66.17	41.68
MMICL (FLAN-T5-XXL)	**55.16**	64.67	41.13
MMICL (Instruct-FLAN-T5-XL)	53.68	65.33	49.28
MMICL (Instruct-FLAN-T5-XXL)	52.19	**68.80**	**51.83**

Table 5: Results of MMICL compared with other VLMs on across different video-languages tasks. All evaluation metrics used in the evaluation is introduced as Appendix F. For Blip2 and Instructblip, We concatenate the visual prompts of all frames and place them on the top of the textual prompts following (Dai et al., 2023).

images in the complex prompts. The result is present in Table 5. Our model, MMICL , achieved significant improvement of 10.86, 4.53, and 2.45 points for MSVD-QA (Chen & Dolan, 2011), NExT-QA (Xiao et al., 2021), and iVQA (Yang et al., 2021) respectively, when compared to the strongest baselines. It is important to note that our training dataset did not include any videos. This

indicates that MMICL effectively enhances the model's ability to understand temporal information in videos.

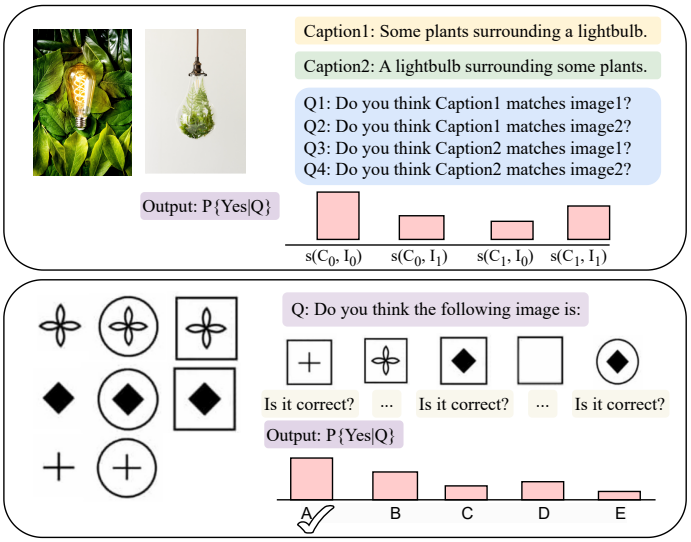

Figure 5: Illustration of two complex vision language reasoning tasks: **Winoground** (Thrush et al., 2022b) (Top): The VLM must capture the distinctions between images and texts, as well as the relationship of reference between text and image. **RAVEN** (Zhang et al., 2019) (Bottom): The VLM is required to capture inter-image relationships and make the right choice. "Q" denotes the input question.

3.4.2 Understanding Text-to-Image Reference

The Winoground (Thrush et al., 2022b) dataset proposes a task of matching two given images and two captions correctly as Fig. 5. Captions serve as implicit reference to images as captions The challenge of this task is that both captions contain a completely identical set of words only in a different order, so as the images, which requires the VLM to capture the slight difference between images and between texts. VLMs need to understand two similar captions and make a difference between their reference to two images. Winoground is appropriate to evaluate whether VLMs can understand the text-to-image reference.

Model	Text	Image	Group
MTurk Human	**89.50**	**88.50**	**85.50**
Random Chance	25.00	25.00	16.67
CLIP-based Model			
VQ2 (Yarom et al., 2023)	**47.00**	42.20	30.50
Vision-language Model			
PALI (Chen et al., 2022)	46.50	38.00	28.75
BLIP2 (Li et al., 2023e)	44.00	26.00	23.50
MMICL (FLAN-T5-XXL)	45.00	**44.99**	**43.00**

Table 6: Results on the Winoground (Thrush et al., 2022b) dataset across the text, image and group score metrics. MMICL outperforms baselines at image and group metrics at a large margin.

The results of Winoground are shown in Table 6.We utilize the same metrics as Thrush et al. (2022b). The results demonstrate that MMICL successfully captures the reference relations between the image and text, surpassing previous baselines.

3.4.3 Understanding Complex Image-to-Image Relationship

Raven's Progressive Matrices (RAVEN) (Carpenter et al., 1990; Zhang et al., 2019; Huang et al., 2023a; Jiang et al., 2022) is one of the most common tests to evaluate nonverbal reasoning. The images in RAVEN are sensitive to their relative and absolute positions, and the task requires visual and logical-mathematical skills to understand the inter-image relationships among multiple groups of images. Thus, the RAVEN task is appropriate for assessing whether one can capture the image to image relationship.

Model	Accuracy
Random Choice	17%
KOSMOS-1 (Huang et al., 2023a)	22%
MMICL (FLAN-T5-XXL)	34%

Table 7: Zero-shot generalization on Raven IQ test. MMICL outperforms KOSMOS-1 and random choice at a large scale.

Following Huang et al. (2023a), we conduct zero-shot nonverbal reasoning experiments on the Raven IQ test with the same method used in KOSMOS-1. Each instance has 3 or 8 images as inputs and 6 candidate images with a unique correct completion and the goal is to predict the next one as Fig. 5. The metrics are Top-1 Acc. The results of RAVEN are shown in Table 7. MMICL achieves new state-of-the-art compared to KOSMOS-1. It achieve 12 points improvement in accuracy over the KOSMOS-1 baseline. The results indicate that MMICL is able to capture the complex logical image-to-image relationship, and then conduct complex nonverbal visual reasoning tasks.

3.5 Hallucination and Language Bias of VLMS

Current VLMs suffer from serious visual hallucinations (Li et al., 2023g). Especially when dealing with complex multi-modal prompts with multiple images (e.g., Chain of the thoughts (Zhang et al., 2023)). VLMs may fail to benefit from multi-modal in-context examples because of hallucinations. We also find the language bias where most VLMs ignore the visual content when faced with extensive textual context. This is a fatal flaw in answering questions that require visual information and extensive text.

ScienceQA-IMG (Lu et al., 2022) is a challenging task that requires a model to effectively utilize information from different modalities to create a coherent and comprehensive line of reasoning. In ScienceQA-IMG, the ability to interpret visual aids such as charts and tables is required to answer the questions. However, some questions of ScienceQA-IMG may not need aid from images as text already provides enough information to answer these questions. So we manually categorized ScienceQA-IMG into two groups, one requiring image information for answering and the other not. We conducted extensive experiments on the ScienceQA-IMG dataset to verify the language bias. Models are required to answer given images and questions.

Moder	Average Performance	Don't Require Visual Infomation	Require Visual Infomation	Performance Gap
Random Guess	35.50	35.80	34.90	-
Shikra (Chen et al., 2023)	45.80	52.90	39.30	13.60
Ying-VLM (Li et al., 2023f)	55.70	66.60	44.90	21.70
Otter (Li et al., 2023b)	63.10	70.90	55.70	15.20
InstructBLIP (Dai et al., 2023)	71.3	82.00	60.70	21.30
MMICL	**82.10**	**82.60**	**81.70**	0.9

Table 8: Zero-shot performance of different VLMs on ScienceQA-IMG dataset in different split. MMICL outperforms other VLMs by successfully alleviating language bias.

The results in Table 8 show that MMICL successfully alleviates language bias as MMICL performs equally in both the 'Don't Require Visual Information' group and the 'Require Visual Information' group. Other VLMs suffer from language bias and may perform drastically differently in the two groups of questions.

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	37.00	34.93	68.05
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)	66.45	52.00	**37.98**	**60.85**	67.20

Table 9: Ablation study on Training Paradigm. We report the Top-1 accuracy for VSR, IconQA-text, Bongard-HOI and IconQA-img; and Exact Match accuracy for VisDial.

MMICL achieves a significant improvement compared to other VLMs with a similar model structure (e.g., Instructblip and Ying-VLM), while reducing language bias. In comparison to Otter, though it also utilizes multi-modal instruction in-context tuning data to fine-tune Openflamingo, it lacks an understanding of text-to-image reference and complex relationships among images, resulting in significant language bias. Shrika[‡‡] mitigates the influence of language bias by providing spatial coordinate inputs. Shrika achieves the lowest Performance Gap except for MMICL with a different design compared to MMICL .

We also evaluate MMICL on object hallucination benchmark using the POPE evaluation pipeline (Li et al., 2023g) and the results are shown in Appendix E. MMICL also demonstrates amazing performance in handling object hallucination.

3.6 Ablation Study on Training Paradigm

To evaluate the effect of multi-modal in-context tuning, we conduct an ablation study using five datasets: VSR (Liu et al., 2022), IconQA-text (Lu et al., 2021), VisDial (Das et al., 2017), IconQA-img (Lu et al., 2021), and Bongard-HOI (Jiang et al., 2022). The first three dataset vison-language tasks provide only one image input while the last two dataset involves multiple images when answering the question. These tasks consist of visual questions answering, visual dialog, and few-shot image classification. The evaluation metrics are: Top-1 accuracy for VSR, IconQA-text, Bongard-HOI, and IconQA-img; and Exact Match accuracy for VisDial.

Stage I means that the model is only trained at Stage I pertaining. We consider two different sizes (Flant5-xl/ Flant5-xxl) of baseline models from two different model families (Blip2/ Instructblip) to conduct the ablation study. Table 9 shows that the performance of MMICL achieves a remarkable enhancement through Stage II multi-modal in-context tuning. We can observe significant performance improvements across all types and sizes of models on most datasets.

We have noticed significant progress in tasks that involve multiple images. MMICL with Stage I + Stage II has achieved an impressive improvement of 15.75 points in accuracy for IconQA-img and a remarkable increase of 21.05 points for Bongard-HOI when compared to MMICL with only Stage I. This indicates that with the assistance of Stage II multi-modal in-context tuning, MMICL is able to understand complex multi-modal prompts and successfully perform challenging tasks involving multiple images.

[‡‡]We use the 0708 version of Shikra, which performs better for multi-choice questions to ensure fair competition.

4 Related Work

4.1 Vision-Language Pretraining

Recent VLMs (Zhu et al., 2023; Liu et al., 2023b; Li et al., 2022; Alayrac et al., 2022; Dai et al., 2023) have been proven effective for aligning visual inputs and frozen LLMs to obtain cross-modal generalization ability. However, there is still less work focusing on VLMs with multi-image inputs. Tsimpoukelli et al. (2021) supports multi-visual inputs based on self-attention for images but performs poorly in downstream tasks. Flamingo (Alayrac et al., 2022) supports Few-Shot Learning in VLMs via ICL by leveraging its robust capability to handle multi-visual inputs and uses cross-attention to capture image relationships. However, Flamingo still suffers from exact reference to specific images. However, they still perform poorly and are unable to achieve exact references from text to image.

4.2 In-Context Learning

In-context learning (Min et al., 2021) is successful in large language models (LLM). Few-shot Chain-of-Thought (Zhang et al., 2023; Cai et al., 2023a; Wang et al., 2023a) further improves generalization ability based on well-designed examples, and enhances applications (i.e., evaluation (Wang et al., 2023b). Li et al. (2023a) explores multi-modal in-context instruction-tuning but mainly focuses on the video domain. However, ICL still lacks exploration in VLMs. Recent works in VLM typically focus on zero-shot evaluation with single image input.

4.3 Multi-Model Instruction Tuning

Instruction-tuning (Kung & Peng, 2023; Shao et al., 2022) achieves great success in cross-task generalization for LLMs. And multi-modal instruction-tuning still requires further exploration. Multiinstruct (Xu et al., 2023) introduces instruction tuning to enhance the performance of VLMs in instruction-following ability. Due to the architectural design, Multiinstruct still struggles with complex contexts containing multiple images. Otter (Li et al., 2023b) fine-tunes Openflamingo (Awadalla et al., 2023) to augment its instruction comprehension capabilities. However, Otter's dataset lacks text-to-image reference and interconnected image-to-image data. This limitation hinders its capability to handle complex contexts that involve visual-textual relationships.

5 Conclusion

In this paper, we highlight the disparity between the prompts used in the training of VLMs and their real-world applications, which makes VLMs less effective at handling complex prompts with multiple images. To tackle this problem, we introduce MMICL , a solution that can handle multi-modal inputs with multiple images. This breakthrough enables VLMs to understand complicated and realistic multi-modal prompts. Furthermore, MMICL sets a new state-of-the-art performacnce on the general VLM benchmarks and complex multi-modal reasoning benchmarks.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2016.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *ICCV*, pp. 8948–8957, 2019.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp.

23716–23736. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *ArXiv*, abs/2308.01390, 2023. URL https://api.semanticscholar.org/CorpusID:261043320.

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pp. 333–342, 2010.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4291–4301, 2019.

Zefan Cai, Baobao Chang, and Wenjuan Han. Human-in-the-loop through chain-of-thought. *arXiv preprint arXiv:2306.07932*, 2023a.

Zefan Cai, Xin Zheng, Tianyu Liu, Xu Wang, Haoran Meng, Jiaqi Han, Gang Yuan, Binghuai Lin, Baobao Chang, and Yunbo Cao. Dialogvcs: Robust natural language understanding in dialogue system upgrade. *arXiv preprint arXiv:2305.14751*, 2023b.

Patricia A Carpenter, Marcel A Just, and Peter Shell. What one intelligence test measures: a theoretical account of the processing in the raven progressive matrices test. *Psychological review*, 97(3):404, 1990.

David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pp. 190–200, 2011.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic, 2023.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

Xingyu Chen, Zihan Zhao, Lu Chen, JiaBao Ji, Danyang Zhang, Ao Luo, Yuxuan Xiong, and Kai Yu. WebSRC: A dataset for web-based structural reading comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4173–4185, Online and Punta Cana, Dominican Republic, November 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.343. URL https://aclanthology.org/2021.emnlp-main.343.

Xingyu Chen, Zihan Zhao, Lu Chen, Danyang Zhang, Jiabao Ji, Ao Luo, Yuxuan Xiong, and Kai Yu. Websrc: A dataset for web-based structural reading comprehension. *arXiv preprint arXiv:2101.09465*, 2021b.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Maria Cipollone, Catherine C Schifter, and Rick A Moffat. Minecraft as a creative tool: A case study. *International Journal of Game-Based Learning (IJGBL)*, 4(2):1–14, 2014.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 326–335, 2017.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*, 2021.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19358–19369, June 2023.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, July 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023a.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023b.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

Huaizu Jiang, Xiaojian Ma, Weili Nie, Zhiding Yu, Yuke Zhu, and Anima Anandkumar. Bongard-hoi: Benchmarking few-shot visual reasoning for human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19056–19065, 2022.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624, 2020.

Po-Nien Kung and Nanyun Peng. Do models really learn to follow instructions? an empirical study of instruction tuning. *arXiv preprint arXiv:2305.11383*, 2023.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023a.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning, 2023b.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023c.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. Empowering vision-language models to follow interleaved vision-language instructions. *arXiv preprint arXiv:2308.04152*, 2023d.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023e.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. M^3it: A large-scale dataset towards multi-modal multilingual instruction tuning, 2023f.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023g.

Yunshui Li, Binyuan Hui, ZhiChao Yin, Min Yang, Fei Huang, and Yongbin Li. PaCE: Unified multi-modal dialogue pre-training with progressive and compositional experts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13402–13416, Toronto, Canada, July 2023h. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.749. URL https://aclanthology.org/2023.acl-long.749.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *arXiv preprint arXiv:2205.00363*, 2022.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 2023b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2023c.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. Cheap and quick: Efficient vision-language instruction tuning for large language models. *arXiv preprint arXiv:2305.15023*, 2023.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021.

Ivona Najdenkoska, Xiantong Zhen, and Marcel Worring. Meta learning to bridge vision and language models for multimodal few-shot learning. *arXiv preprint arXiv:2302.14794*, 2023.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

Junting Pan, Ziyi Lin, Yuying Ge, Xiatian Zhu, Renrui Zhang, Yi Wang, Yu Qiao, and Hongsheng Li. Retrieving-to-answer: Zero-shot video question answering with frozen large language models, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–16. IEEE, 2020.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 3505–3506, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3406703. URL https://doi.org/10.1145/3394486.3406703.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge, 2022.

Nan Shao, Zefan Cai, Chonghua Liao, Yanan Zheng, Zhilin Yang, et al. Compositional task representations for large language models. In *The Eleventh International Conference on Learning Representations*, 2022.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 8317–8326, 2019.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022a.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022b.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022a.

Peiyi Wang, Lei Li, Liang Chen, Feifan Song, Binghuai Lin, Yunbo Cao, Tianyu Liu, and Zhifang Sui. Making large language models better reasoners with alignment. *arXiv preprint arXiv:2309.02144*, 2023a.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023b.

Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv:2210.14896 [cs]*, 2022b. URL https://arxiv.org/abs/2210.14896.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9777–9786, 2021.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

Zhiyang Xu, Ying Shen, and Lifu Huang. MultiInstruct: Improving multi-modal zero-shot learning via instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11445–11465, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.641. URL https://aclanthology.org/2023.acl-long.641.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1686–1697, 2021.

Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation. *arXiv preprint arXiv:2305.10400*, 2023.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2, 2014.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 69–85. Springer, 2016.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6720–6731, 2019.

Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, and Tao Kong. What matters in training a gpt4-style language model with multimodal inputs? *arXiv preprint arXiv:2307.02469*, 2023.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5317–5327, 2019.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2023.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A Related Work

A.1 Vision-Language Pretraining

Model	Multi-Image Inputs	Multi-model Instruction Tuning	Text-to-Image Reference
Flamingo	✓	✗	✗
BLIP-2	✗	✗	✗
LLAVA	✗	✓	✗
MiniGPT-4	✗	✓	✗
InstructBLIP	✗	✓	✗
Shikra	✗	✓	✓
Kosmos-1	✓	✗	✗
Otter	✓	✓	✗
MMICL	✓	✓	✓

Table 10: Summary of Vision-Language Pre-Trained Models.

Our work is inspired by recent vision-language pre-training works (Zhu et al., 2023; Liu et al., 2023b; Li et al., 2022; 2023e), which has been proven effective for aligning visual inputs and frozen LLMs to obtain cross-modal generalization ability.

BLIP-2 BLIP-2 (Li et al., 2023e) bridges the modality gap with a lightweight Querying Transformer, which is pre- trained in two stages. The first stage boot- straps vision-language representation learning from a frozen image encoder. The second stage bootstraps vision-to-language generative learning from a frozen language model.

InstructBLIP InstructBLIP (Dai et al., 2023) performs vision-language instruction tuning based on the pretrained BLIP-2 models with converted multi-modal datasets and the LLaVA (Liu et al., 2023b) dataset generated by GPT-4.

MiniGPT-4 MiniGPT-4 (Zhu et al., 2023)aligns a CLIP visual encoder with a frozen Vincuna (Chiang et al., 2023) with artificially collected dialog dataset

Shikra Shikra (Chen et al., 2023), a VLM which can handle spatial coordinate inputs and outputs in natural language. It makes Shikra excel at both referential dialogue tasks and general vision-language tasks, resulting in outstanding performance.

However, there are still less work focusing on VLMs with multi-image inputs.

Flamingo Flamingo (Tsimpoukelli et al., 2021) achieves multi-visual inputs based on self-attention for images but performs poorly in downstream tasks. Flamingo excels in supporting Few-Shot Learning (FSL) in VLM via ICL by leveraging its robust capability to handle multi-visual inputs and uses cross-attention instead of self-attention to get better performance. However, it still suffers from the unableness to explicitly point images, so they introduce a hacky cross-attention mask.

Kosmos-1 Kosmos-1 (Huang et al., 2023a), which is trained from scratch on web-scale multi-modal corpora with interleaved text-image data, image-text caption and language-only instruction tuning data. It supports multi-model Few-Shot Learning and Chain-of-Thought and achieve competitive performance.

Otter Otter (Li et al., 2023b), an open-source implementation of flamingo and trained with multi-modal instruction in-context tuning data.

Najdenkoska et al. (2023) uses meta-learning objective to train an adapter that aggregates multiple image features so the original VLM and adapter become a better few-shot learner.

However, dialogue research is still limited in the language field (Cai et al., 2023b). Multi-modal dialogue system still requires more exploration.

A.2 IN-CONTEXT LEARNING

It has been well-explored to enable ICL in pre-trained language models (PLM). MetaICL (Min et al., 2021) proposes a meta-training framework for few-shot learning to tune a PLM to do in-context learning on a large set of training tasks. LM-BFF (Gao et al., 2020) studies few-shot fine-tuning of PLMs. However, ICL in VLM is still less explored. Recent works in VLM mainly focus on zero-shot evaluation with single image input.

B MULTI-MODAL ICL DATA

We construct two training datasets, text-image interleaved data and in-context learning data for the text-image relationship challenge and image-image relationship challenge, respectively. In this section, we will cover the data resources.

Task	Dataset	Used	#samples Train	#samples Val	#samples Test	License
Captioning	MS COCO (Lin et al., 2014)	Yes	566,747	25,010	25,010	Custom
	DiffusionDB (Wang et al., 2022b)	Yes	19,963	0	0	Unknown
	Flickr (Young et al., 2014)	Yes	144,896	768	768	Unknown
	NoCaps (Agrawal et al., 2019)	Yes	0	0	4,500	Unknown
Classification	MiniImage (Russakovsky et al., 2015)	Yes	38,400	9,600	12,000	Non-commercial
VQA	VQA v2 (Goyal et al., 2017)	Yes	30,000	30,000	0	CC-BY 4.0
	ST-VQA (Biten et al., 2019)	Yes	26,074	0	4,070	Unknown
	Text-VQA (Singh et al., 2019)	Yes	27,113	0	5,734	CC BY 4.0
	NLVR2 (Suhr et al., 2018)	Yes	86,373	6,982	6,967	Unknown
	RefCOCO (Yu et al., 2016)	Yes	26,074	0	4,070	Unknown
KVQA	OK-VQA (Marino et al., 2019)	Yes	9,009	5,046	0	Unknown
Reasoning	GQA (Hudson & Manning, 2019)	Yes	943,000	132,062	12,578	Unknown
	VCR (Zellers et al., 2019)	Yes	25,000	5,000	5,000	Custom
	Winoground (Thrush et al., 2022a)	No	0	0	800	Unknown
Others	WikiART (Saleh & Elgammal, 2015)	Yes	13,000	5,500	0	Unknown
	LLAVA-Instruct-150K (Liu et al., 2023b)	Yes	15,000	0	0	Non-commercial

Table 11: Detailed task descriptions and statistics of our instruction tuning tasks, including all datasets in all types of tasks. The column "Used" indicates whether we use this dataset in the multi-model in-context tuning stage.

C TRAINING DATA

C.1 DATA RESOURCE

Our training dataset comes from 8 task categories and 16 datasets.

Image Captioning aims to produce descriptions of the given images according to different needs. Our training dataset includes MS COCO (Lin et al., 2014), DiffusionDB (Wang et al., 2022b), and Flickr 30K (Young et al., 2014).

Knowledgeable Visual Question Answering (KVQA) requires the model to make use of commonsense knowledge outside the input image to answer questions. Our training dataset includes OK-VQA (Marino et al., 2019).

Image Question Answering (IQA) requires the model to answer questions based on the image correctly. Our training dataset includes VQAv2 (Goyal et al., 2017), ST-VQA (Biten et al., 2019), Text-VQA (Singh et al., 2019), WikiART (Saleh & Elgammal, 2015) and RefCOCO (Yu et al., 2016).

Visual Reasoning requires the model to perform image reasoning and answer questions correctly. Our training dataset includes GQA (Hudson & Manning, 2019), VCR (Zellers et al., 2019), and NLVR2 (Suhr et al., 2018).

Image Classification involves classifying an image based on a given set of candidate labels. Our training dataset includes MiniImage (Russakovsky et al., 2015).

Visual Dialog requires the model to hold a meaningful dialog with humans in natural, conversational language about visual content. Our training dataset includes LLAVA-Instruct-150K (Liu et al., 2023b).

Our testing dataset comes from 10 task categories and 16 datasets.

Image Captioning includes the Nocaps (Agrawal et al., 2019) dataset.

Knowledgeable Visual Question Answering (KVQA) includes the ScienceQA (Lu et al., 2022) and A-OKVQA (Schwenk et al., 2022) datasets.

Image Question Answering (IQA) includes the VizWiz (Bigham et al., 2010) dataset.

Visual Reasoning includes the Winoground (Thrush et al., 2022b), VSR (Liu et al., 2022) and IconQA (Lu et al., 2021) dataset. Winoground proposes a task of matching two given images and two captions correctly. The challenge of this task is that both captions contain a completely identical set of words only in a different order. VSR describes the spatial relation of two individual objects in the image, and a VLM needs to judge whether the caption is correctly describing the image (True) or not (False). The IconQA dataset has two sub-datasets: image question answering with multiple text choice and image question answering with multiple image choice.

Web Page Question Answering (Web QA) includes the Websrc (Chen et al., 2021a; Huang et al., 2023a) datasets. The model must answer questions based on the web image and the optional extracted texts. We sampled 2000 instances from Websrc for the evaluation. To align with KOSMOS-1 (Huang et al., 2023a), we only use the web image as input.

Video Question Answering (VideoQA) includes the iVQA (Yang et al., 2021), MVSD (Chen & Dolan, 2011), and NextQA (Xiao et al., 2021) dateset. The NextQA dataset has two sub-datasets: video question answering with multiple choice and open-domain video question answering.

Few-shot Image Classification includes the HatefulMemes (Kiela et al., 2020) and Bonard-HOI (Jiang et al., 2022) dataset. HatefulMemes requires the model to determine if a meme is hateful based on the image and explanation provided. Bonard-HOI serves as the benchmark for evaluating the model's ability in Few-Shot Visual Reasoning for Human-Object Interactions. It provides few-shot examples with challenging negatives, where positive and negative images only differ in action labels. The model is then asked to determine whether the final image is positive or negative. We sampled 2000 instances from Bonard-HOI for the evaluation.

Nonverbal Reasoning includes the Raven IQ test (Huang et al., 2023a). Each instance in teh Raven IQ test has 3 or 8 images as inputs and 6 candidate images with a unique correct completion and the goal is to predict the next one image from the candidates.

Visual Dialog includes the visual dialog dataset (Das et al., 2017). We use the question of the final dialogue as the question for the instance and take all preceding dialogues as the context to perform open-domain image question answering.

OOD Generalization includes the Minecraft dataset that we construct using Minecraft (Cipollone et al., 2014) game which requires the VLM to identify whether a animal (i.e., cow, llama, chicken, donkey and so on) is present in a picture.

More detailed task descriptions and statistics about the datasets are shown at Table 11.

C.2 Data Balance

Previous studies have shown that the the data balance of training data could significantly influence the model performance (Dai et al., 2023). Mixing the training data of each dataset uniformly could cause the model to overfit smaller datasets and underfit larger datasets, causing poor performance. In order to alleviate this problem, we employ a sampling strategy to sample datasets with probabilities proportional to the square root of the number of training samples following Dai et al. (2023). Formally, given D datasets with $N.$ training samples $\{N_1, N_2, \ldots, N_D\}$, the probability p_d of data samples

Model	Existence		Count		Position		Color		OCR		Avg.
	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+	
BLIP-2	86.67	73.33	75.00	60.00	56.67	16.67	81.67	66.67	70.00	40.00	62.67
LLaVA	50.00	0.00	50.00	0.00	50.00	0.00	51.67	3.33	50.00	0.00	25.49
MiniGPT-4	75.00	60.00	66.67	56.67	56.67	33.33	71.67	53.33	62.50	35.00	57.08
mPLUG-Owl	73.33	46.67	50.00	0.00	50.00	0.00	51.67	3.33	55.00	10.00	34.00
LLaMA-Adapter-V2	76.67	56.67	58.33	6.67	43.33	3.33	55.00	16.67	57.50	15.00	38.92
VisualGLM-6B	61.67	23.33	50.00	0.00	48.33	0.00	51.67	3.33	42.50	0.00	28.08
Otter	53.33	6.67	50.00	0.00	50.00	0.00	51.67	3.33	50.00	0.00	26.50
Multimodal-GPT	46.67	10.00	51.67	6.67	45.00	13.33	55.00	13.33	57.50	25.00	32.42
PandaGPT	56.67	13.33	50.00	0.00	50.00	0.00	50.00	0.00	50.00	0.00	27.00
MMICL	91.67	83.33	80.00	63.33	53.33	20.00	88.33	83.33	65.00	50.00	67.83

Table 12: Fine-grained result of MME benchmark

Model	Poster		Celebrity		Scene		Landmark		Artwork		Avg.
	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+	
BLIP-2	79,25	62.59	58.53	37.06	81.25	64.00	79.00	59.00	76.50	60.00	66.72
LLaVA	50.00	0.00	48.82	0.00	50.00	0.00	50.00	0.00	49.00	0.00	24.78
MiniGPT-4	49.32	19.73	58.82	24.71	68.25	45.50	59.75	30.50	56.25	27.00	44.00
mPLUG-Owl	77.89	57.14	66.18	34.12	78.00	57.50	86.25	73.00	63.25	33.00	62.63
LLaMA-Adapter-V2	52.72	10.88	55.00	21.18	68.75	44.50	53.00	9.00	52.50	14.50	38.2
InstructBLIP	74.15	49.66	67.06	34.12	84.00	69.00	59.75	20.00	76.75	57.50	59.20
VisualGLM-6B	54.42	12.24	50.88	2.35	81.75	64.50	59.75	24.00	55.75	20.00	42.56
Otter	45.24	0.00	50.00	0.00	55.00	14.50	52.00	4.50	48.00	5.50	27.47
Multimodal-GPT	45.24	17.01	49.12	24.12	50.50	17.50	50.50	23.00	46.00	12.00	33.50
PandaGPT	56.80	19.73	46.47	10.59	72.50	45.50	56.25	13.50	50.25	1.00	37.26
MMICL	73.81	57.14	81.18	64.71	83.75	69.00	77.49	57.59	75.00	58.00	69.77

Table 13: Fine-grained result of MME benchmark

being selected from a dataset during training is as follows.

$$p_d = \frac{\sqrt{N_d}}{\sum_{i=1}^{D} \sqrt{N_i}} \tag{3}$$

D MME BENCHMARK

MME comprehensively evaluates VLMs with 14 sub-tasks that encompass perception and cognition abilities. Other than OCR, perception ability includes the recognition of coarse-grained and fine-grained objects. The former identifies the existence, count, position, and color of objects. The latter recognizes movie posters, celebrities, scenes, landmarks, and artworks. The cognition includes commonsense reasoning, numerical calculation, text translation, and code reasoning.

MME evaluates a wide range of multi-modal abilities. The compared baselines include LLaVA (Liu et al., 2023b), MiniGPT-4 (Zhu et al., 2023), MultiModal-GPT (Gong et al., 2023), VisualGPM-6B (Du et al., 2021), LaVIN (Luo et al., 2023), mPLUG-Owl (Ye et al., 2023), LLaMA-Adapter-V2 (Gao et al., 2023), InstructBLIP (Dai et al., 2023), Otter (Li et al., 2023b), BLIP-2 (Li et al., 2023e), LRV-Instruction (Liu et al., 2023a), Cheetor (Li et al., 2023d), GIT2 (Wang et al., 2022a), Lynx (Zeng et al., 2023). We also provide more detail evaluation results for MMICL at Table 12, Table 13 and Table 14. Results show that MMICL can achieve the best average scores in comparisons with current VLMs

E OBJECT HALLUCINATION BENCHMARK

We test the following VLMs on the POPE benchmark to evaluate their object hallucination performance: MMICL , Shikra (Chen et al., 2023), InstructBLIP (Dai et al., 2023), MiniGPT-4 (Zhu et al.,

Model	Common. Reason.		Numerical Calculation		Text Translation		Code Reason.		Avg.
	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC	
BLIP-2	68.57	41.43	40.00	0.00	55.00	10.00	55.00	20.00	36.25
LLaVA	49.29	11.43	50.00	0.00	52.50	5.00	50.00	0.00	27.27
MiniGPT-4	58.57	34.29	47.50	20.00	42.50	15.00	67.50	45.00	41.30
mPLUG-Owl	59.29	24.29	50.00	10.00	60.00	20.00	47.50	10.00	35.14
LLaMA-Ada.-V2	54.29	14.29	52.50	5.00	52.50	5.00	52.50	10.00	30.76
InstructBLIP	75.00	54.29	35.00	5.00	55.00	10.00	47.50	0.00	35.22
VisualGLM-6B	45.71	12.86	45.00	0.00	55.00	10.00	50.00	0.00	27.32
Otter	48.57	10.00	47.50	10.00	55.00	10.00	50.00	0.00	28.88
MultiModal-GPT	45.71	5.71	50.00	20.00	50.00	5.00	45.00	10.00	28.93
PandaGPT	56.43	17.14	50.00	0.00	52.50	5.00	47.50	0.00	28.67
MMICL	70.71	47.14	47.50	15.00	62.50	45.00	47.50	25.00	45.04

Table 14: Fine-grained result of MME benchmark

2023), LLaVA (Liu et al., 2023b), MM-GPT (Gong et al., 2023) and mPLUG-Owl (Ye et al., 2023). The result is present in the Table 15.

Table 15: Performance result of different VLMs on the POPE benchmark

Dataset	Metric	Models						
		MMICL	Shikra	InstructBLIP	MiniGPT-4	LLaVA	MM-GPT	mPLUG-Owl
Random	Accuracy	0.8729	86.90	88.57	79.67	50.37	50.10	53.97
	Precision	0.9463	94.40	84.09	78.24	50.19	50.05	52.07
	Recall	0.7987	79.27	95.13	82.20	99.13	100.00	99.60
	F1-Score	0.8662	86.19	89.27	80.17	66.64	66.71	68.39
	Yes	0.4351	43.26	56.57	52.53	98.77	99.90	95.63
Popular	Accuracy	0.8270	83.97	82.77	69.73	49.87	50.00	50.90
	Precision	0.8511	87.55	76.27	65.86	49.93	50.00	50.46
	Recall	0.7927	79.20	95.13	81.93	99.27	100.00	99.40
	F1-Score	0.8208	83.16	84.66	73.02	66.44	66.67	66.94
	Yes	0.4657	45.23	62.37	62.20	99.40	100.00	98.57
Adversarial	Accuracy	0.8097	83.10	72.10	65.17	49.70	50.00	50.67
	Precision	0.8188	85.60	65.13	61.19	49.85	50.00	50.34
	Recall	0.7953	79.60	95.13	82.93	99.07	100.00	99.33
	F1-Score	0.8069	82.49	77.32	70.42	66.32	66.67	66.82
	Yes	0.4857	46.50	73.03	67.77	99.37	100.00	98.67

F Details for Evaluation

F.1 Evaluation Metrics

We provide evaluation metrics as Table 16

F.2 VQA Tools

We use the same VQA Tools as the original VQA paper (Agrawal et al., 2016) and use it in all metrics using the vqa accuracy.

G Baselines

Baselines We primarily compare MMICL with recently proposed powerful multi-modal approaches, including:

(1) Flamingo (Alayrac et al., 2022) where a VLM is trained on large -scale multimodal web corpora containing arbitrarily interleaved text and images;

Dataset	Metrics
MSVD (Chen & Dolan, 2011)	Top-1 Acc.
iVQA (Yang et al., 2021)	iVQA Acc.
NExT-QA-multiple-choice (Xiao et al., 2021)	Top-1 Acc.
NExT-QA-opendomain (Xiao et al., 2021)	WUPS Score.
Hateful Memes (Kiela et al., 2020)	AUC Score
WebSRC (Chen et al., 2021b)	Exact Match
VSR (Liu et al., 2022)	Top-1 Acc.
∗VQAv2 (Goyal et al., 2017)	VQA Acc.
VizWiz (Bigham et al., 2010)	VQA Acc.
IconQA-text (Lu et al., 2021)	Top-1 Acc.
IconQA-img (Lu et al., 2021)	Top-1 Acc.
ScienceQA-IMG (Lu et al., 2022)	Top-1 Acc.
Bongard-HOI (Jiang et al., 2022)	Top-1 Acc.
VisDial (Das et al., 2017)	Exact Match
NoCaps (Agrawal et al., 2019)	Cider Score
A-OKVQA (Agrawal et al., 2019)	Top-1 Acc.
∗Flickr (Young et al., 2014)	Cider Score
Winoground (Thrush et al., 2022b)	Winoground mertic.
Raven IQ Test (Huang et al., 2023a)	Top-1 Acc.
Minecraft	Top-1 Acc.

Table 16: Summary of the evaluation datasets and metrics. These datasets are used to validate general design of MMICL . The datasets marked with ∗ is the hold-in datasets, where their training set is used in training the MMICL .

(2) KOSMOS-1 (Huang et al., 2023a) which is trained from scratch on web-scale multi-modal corpora;

(3) BLIP-2-FLAN-T5 (Li et al., 2023e) where an instruction-tuned Flan-T5 (Chung et al., 2022) is connected with a powerful visual encoder to perform a series of multi-modal tasks;

(4) InstructBLIP-FLAN-T5 (Dai et al., 2023), a recently proposed instruction tuning enhanced multi-modal agents with FLAN-T5 with converted multi-model datasets and the LLaVA (Liu et al., 2023b) dataset generated by GPT-4 (OpenAI, 2023);

(5) Shikra (Chen et al., 2023), a VLM which can handle spatial coordinate inputs and outputs in natural language. Without the need for extra vocabularies, or external plugin models. all inputs and outputs of Shikra are in natural language form.

(6) Otter (Li et al., 2023b), an open-source implementation of flamingo (Alayrac et al., 2022). By utilizing multi-model instruction in-context tuning data, Otter fine-tunes Openflamingo to augment its instruction comprehension capabilities while maintaining its ability to learn in-context;

(7) Ying-VLM (Li et al., 2023f), a VLM model trained on Multi-Modal multilingual instruction tuning dataset, showcasing its potential to answer complex questions requiring world knowledge, generalize to unseen video tasks, and comprehend unseen instructions in Chinese.

H OOD Generalization to Unseen Domain

Method	Shot	Top-1 Acc.
MiniGPT-4 (Vincuna-7B)	Zero-Shot	35.10%
MiniGPT-4 (Vincuna-13B)	Zero-Shot	48.40%
MMICL (FLAN-T5-XL)	Zero-Shot	55.41%
MMICL (FLAN-T5-XL)	4-Shot	64.05%
MMICL (FLAN-T5-XXL)	8-Shot	65.41%

Table 17: Results of generalization of MMICL to unseen domain in Minecraft. Results show that MMICL is able to generalize to unseen domains and tasks given a few examples.

In an unseen challenging domain with limited exemplars, analyzing regular patterns, reasoning, and learning new knowledge (OOD Generalization to unseen domain) is a great way to test multi-modal ICL ability.

We construct a task using Minecraft (Cipollone et al., 2014) which requires the VLM to identify whether a animal (*i.e.*, cow, llama, chicken, donkey and so on) is present in the case (d) of Fig. 1.

We collect 550 cases and transfer the task to a vision-to-text question-answering task to evaluate the performance of OOD generalization of MMICL . The results are shown in Table 17. Results demonstrate that MMICL is able to generalize to the Minecraft domain even if the images are extremely different compared to the images used by training at Stage I and II as Sec. 2.3.