Link-Context Learning for Multimodal LLMs

Yan Tai[1, 2*], **Weichen Fan**[1*†], **Zhao Zhang**[1], **Zhu Feng**[1], **Rui Zhao**[1], **Ziwei Liu**[3]

[1]SenseTime Research
[2]Institute of Automation, CAS
[3]S-Lab, Nanyang Technological University

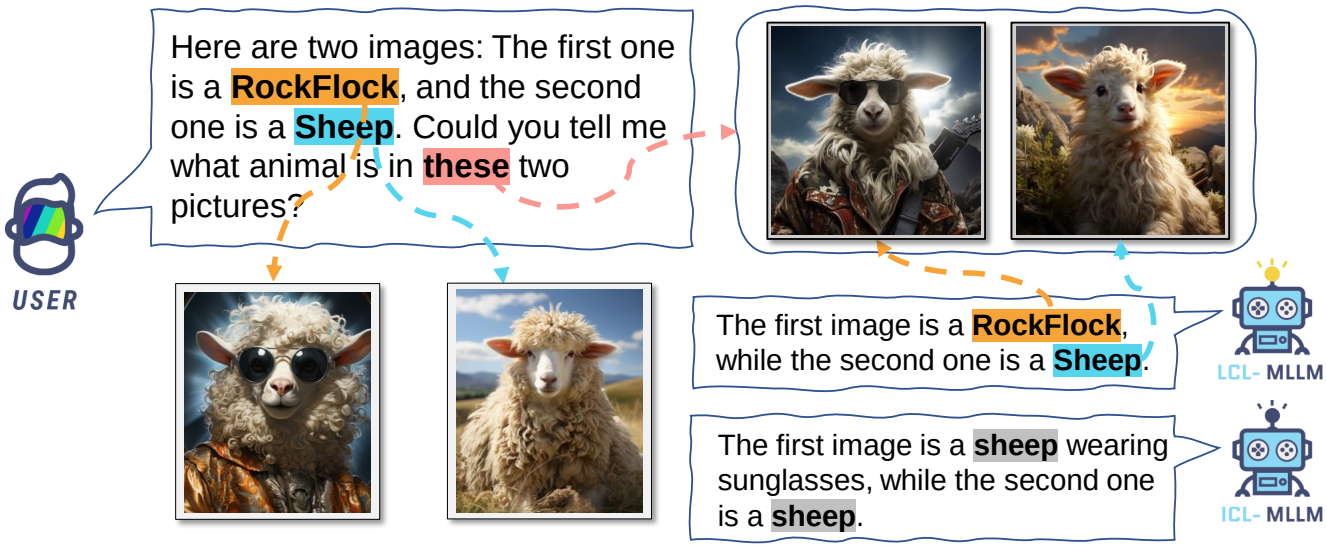

Figure 1: **The demo dialogue of our proposed link-context learning**. After presenting the model with a pair of unseen images and novel concepts, our improved model gains the ability to learn and retain the acquired knowledge throughout the conversation while the vanilla MLLMs fail to provide accurate answers.

Abstract

The ability to learn from context with novel concepts, and deliver appropriate responses are essential in human conversations. Despite current Multimodal Large Language Models (MLLMs) and Large Language Models (LLMs) being trained on mega-scale datasets, recognizing unseen images or understanding novel concepts in a training-free manner remains a challenge. In-Context Learning (ICL) explores training-free few-shot learning, where models are encouraged to "learn to learn" from limited tasks and generalize to unseen tasks. In this work, we propose **link-context learning (LCL)**, which emphasizes "reasoning from cause and effect" to augment the learning capabilities of MLLMs. LCL goes beyond traditional ICL by explicitly strengthening the causal relationship between the support set and the query set. By providing demonstrations with causal links, LCL guides the model to discern not only the analogy but also the underlying causal associations between data points, which empowers MLLMs to recognize unseen images and understand novel concepts more effectively. To facilitate the evaluation of this novel approach, we introduce the **ISEKAI** dataset, comprising exclusively of unseen generated image-label pairs designed for link-context learning. Extensive experiments show that our LCL-MLLM exhibits strong link-context learning capabilities to novel concepts over vanilla MLLMs. Code and data will be at https://github.com/isekai-portal/Link-Context-Learning.

1 Introduction

(In the near future, mankind finally be able to travel interstellar and come to the centaur constellation.)
Human *and* ***MLLM*** *walk off the spaceship.*
Human*: "We made it! Look! The locals are here."*
Locals*: Greetings, you can call us 'RockFlock'.*
MLLM*: "Hi, sheep!"*
Human*: "😓"*

*Equal Technical Contribution.
†Project Lead.

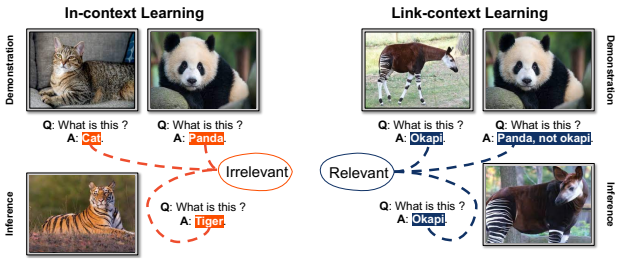

Figure 2: **The difference between our link-context learning with in-context learning**. In-context learning involves providing irrelevant tasks for demonstration, whereas there is a direct causal relationship between the demonstration and inference phases of link-context learning.

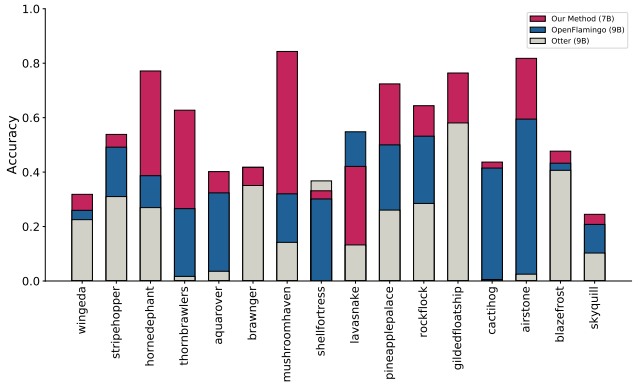

Figure 3: **Overview of results on several categories of ISEKAI dataset**: Our model outperforms OpenFlamingo (9B) [6] and Otter (9B) [5] across almost all the categories, showcasing superior performance in scenarios involving entirely unseen images.

The above conversation between humans and MLLMs serves as a humorous representation of how MLLMs struggle to learn from demonstration during the conversation for real. 'RockFlock' is our hand-made species, which possesses both a human-like body and a sheep-like head, as shown in Figure 1. Current MLLMs fail to link the unseen image-label pairs to recognize novel objects in a single conversation. To address this limitation, equipping the model with few-shot learning ability has been a long-standing topic in computer vision even before the era of MLLMs. This approach enables the model to learn from limited examples and mitigate the issue effectively. The primary method for MLLMs to learn from demonstrations is known as in-context learning, wherein the models show remarkable improvement on downstream tasks after being exposed to a few input-label pairs. However, current MLLMs have very limited benefits from in-context learning, since the emphasis is primarily on guiding the model to acquire the ability to process novel tasks after "learning" from meta tasks. However, the model's performance is not affected even if the answers provided in the meta-tasks are all wrong. [1] Thus, what MLLMs have "learned" from demonstration remains on answering questions in a specific format rather than understanding the causal relationship between the image-label pairs. To enable MLLMs to concentrate more on the causal relationship between the image and label pairs, *Frozen* method [2] binds different labels to known images. However, a significant challenge arises when MLLMs encounter entirely novel scenarios where both the image and the label are unseen. In such instances, the task of extracting the underlying cause and effect from the demonstration and making accurate predictions based on this newfound knowledge remains an unsolved puzzle. The 'RockFlock' (unseen images and novel concepts), shown in Figure 1, would be misrecognized by the previous methods, while our model learns the concept of 'RockFlock' from the demonstration and makes responses accurately. Moreover, the acquisition of novel concepts does not impede the existing knowledge, enabling the model to effectively distinguish between the original and newly learned images.

Inspired by in-context learning (hereinafter called **ICL**), we propose *link-context learning* (hereinafter called **LCL**), which requires the MLLMs to acquire knowledge about new concepts from the conversation and retain their existing knowledge for accurate question-answering. As shown in Figure 2, current in-context learning in MLLMs emphasizes benefiting from the causal-irrelevant demonstration. However, for link-context learning, the demonstration and the final task are linked causally. (e.g. If the 'apple' is renamed as 'orange' in the demonstration, the model should call apple an 'orange' during the inference.) With this ability, the MLLMs could support few-shot learning in a flexible way.

In the era of Large Language Models, evaluating models' performance on few-shot learning becomes a challenge, as these models are extensively trained on vast amounts of real-life data. To address this issue and provide a comprehensive assessment of *link-context learning*, we introduce the *ISEKAI* dataset. This dataset comprises unseen images and concepts, entirely novel to MLLMs, as they transcend the boundaries of realism. All the images in the dataset are generated by Stable Diffusion [3] and Midjourney [4], while all the labels or concepts are fabricated as well. Figure 3 shows the comparisons between our model and Otter [5], OpenFlamingo [6] on ISEKAI dataset.

In this paper, we present *link-context learning* (LCL), a setting that bestows MLLMs with the capability to understand the potential causal relationship in the conversation and process unseen images and concepts. Unlike ICL mainly focuses on inspiring models with a wide variety of different tasks, LCL goes a step further by empowering the model to establish a mapping between the source and target, thereby enhancing its overall performance. The contributions of this work can be summarized as follows:

- **Link-Context Learning**: We introduce a novel causal-relevant few-shot learning setting, where MLLMs are challenged to assimilate new concepts from the ongoing conversation and retain this knowledge for accurate question-answering. Under link-context learning, we empower the MLLMs to grasp the causal relationship between the source and target from the demonstration.

- **ISEKAI Dataset**: Since most real-world data is not completely unseen to MLLMs, we release a challenging fabricated dataset to the public, where novel image-concept pairs are introduced, for evaluation of MLLMs' performance.

2 Related Works

Multimodal Large Language Models [7–11] have demonstrated significant capabilities in universal generation or recognition tasks. Following the new paradigm of MLLMs, various visual tasks can be achieved in a training-free zero-shot manner [12, 13], escaping from the heavy *pretrain-and-finetune* process. However, recognize arbitrary content through a single model is generally considered extremely difficult. How to enhancing recognition capability of MLLMs in the wild at a low cost has emerged as a recent research focus.

Multimodal Prompt Tuning Multimodal Prompt Tuning (M-PT) is commonly used in contrastive learning-based multimodal large models, such as CLIP [12]. In the training process, prompt tuning usually freezes most of the model's parameters and only updates a small number of parameters to achieve results similar to fine-tuning [14–17]. PT [14] add tunable prompt embeddings to each layer of the encoder and decoder, only the weights of the added embeddings will be updated during training. VPT [18] added a set of learnable parameters in specific positions to tune the model. CoOp [15] and UPT [19] used CLIP as the backbone and prompted it to fit few-shot settings. CoCoOp [16], POMP [20] and MaPLe [21] extend prompt tuning to open-vocabulary visual recognition tasks. However, traditional prompt tuning methods are not suitable for the powerful generative multimodal large language models.

Multimodal Instruction Tuning Multimodal Instruction Tuning (M-IT) enhances the zero-shot capability of MLLMs in unseen tasks by fine-tuning them on an instruction descriptions-based dataset [7, 8, 11, 22, 23]. MiniGPT-4 [24] and LLaVA [11] keep the visual encoder frozen and tune the language model, extending instruction tuning to multimodality. mPLUG-Owl [25] tuned visual and text encoder separately in two stages, and proposed an evaluation dataset for assessing vision-related instruction tuning. InstructBLIP [26] enhances zero-shot capability by performing instruction tuning on multiple datasets. Shikra [27] and Kosmos-2 [28] expanded MLLMs to visual grounding tasks using instructions with bounding box coordinates. Even though these studies demonstrate outstanding zero-shot capability, they still cannot recognize classes that were not seen during the model training process.

Multimodal In-Context Learning Large Language Models (LLMs) have shown outstanding capability in learning from context samples. In the Multimodal In-Context Learning (M-ICL) settings, following the input image samples and optional instruction, MLLMs can learn new task patterns in a few-shot manner [29–32]. Flamingo [33] takes in-context learning into consideration during the pretraining process, allowing the model to possess the ability to support in-context learning. Otter [5] follows Flamingo and proposed a new in-context learning dataset, proceeding with the ICL capability in the instruction tuning stage.

Different from previous methods, our proposed *link-context learning* can establish a causal link between the support and query set. Specifically, using few-shot class-specific images and textual prompts, LCL can link the prompt and inference samples, and even associate previously unseen images with new concepts.

3 Link-Context Learning

In this section, we first give a brief introduction to in-context learning and unveil its main restrictions and difference to our link-context learning in Preliminary; next, we bring the power of link-context learning into MLLMs in Bring Link-Context Learning to MLLMs.

3.1 Preliminary

In-Context Learning Formally, in-context learning [34] refers to: the model should choose the answer with the highest prediction score from a set candidate answers $Y = \{y_1, y_2, ..., y_n\}$, given a query input x, conditioning on a support set S, which consists of multiple input-label pairs from a wide variety of tasks, where $S = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$. (The query and the sample of S should belong to different tasks.)

From another perspective, in-context learning could be denoted as training-free few-shot learning, as it transforms the training stage of few-shot learning into the demonstration input for Large Language Models. Noted that the ICL [34] is consistent with FSL, where the tasks in the demonstration (training) stage and in the inference (query) stage are different.

Link-Context Learning Essentially, link-context learning (LCL) represents a form of training-free and causal-linked few-shot learning. In this approach, a support set $S = (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ is provided, along with a query sample x from the query set Q, where the data pairs from the support set are causally linked to the query set. The model is tasked with predicting the answer based on the causal-linked relationship between the query and support set.

To provide further clarity, link-context learning significantly strengthens the causal relationship between the support set and the query set. For instance: 1). New arithmetic rules: In this scenario, the support set consists of arithmetic expressions such as $(1$ <op> $2 = 3), (2$ <op> $3 = 5)$, with the query sample being 4 <op> 5 $=?$. Here, "<op>" represents a new arithmetic rule that we aim to teach the model through the demonstration; 2). Novel image classification: In this case, the support set contains pairs like (<unseen image> : <novel cls A>), (<unseen image> : <novel cls B>), while the query sample is (<unseen image> belongs to?). This example demonstrates how we expect the model to correctly classify the unseen

image into one of the specified novel classes based on the demonstration.

In essence, link-context learning enhances the model's capacity to grasp new concepts and relationships by effectively establishing a causal link between the support set and the query set. While this setting is applicable to both LLMs and MLLMs, our primary focus in this paper is on the application of link-context learning specifically in MLLMs. By concentrating on MLLMs, we aim to showcase the potential of this approach in multimodal models and its implications for advancing their learning capabilities.

3.2 Bring Link-Context Learning to MLLMs

In this section, our main objective is to introduce Link-Context Learning (LCL) to the realm of MLLMs. Recognizing that the current MLLMs trained in the ICL manner may not excel in LCL tasks, we propose a novel training strategy to fine-tune MLLMs. This approach aims to equip the models with the capability to grasp causal links from context effectively. By leveraging this novel training strategy, we aim to empower MLLMs to excel in tasks that require reasoning and understanding causal relationships, thereby broadening their range of capabilities and improving their overall performance. To be more specific, we choose Shikra [27] as our baseline, and we divide ImageNet1k into ImageNet-900 and ImageNet-100 by classes, which would be discussed in detail in Training Dataset. Additionally, we incorporate the concept of contrast learning in our training strategy, as discussed in Training Strategy. This helps guide the model to understand the shared characteristics among samples of the same kind and the distinctions between samples of different kinds.

3.2.1 Training Dataset

Unlike traditional tasks that require extensive training data, LCL concentrates on acquiring the ability to find the link between the source-target pairs in demonstration and generalize to the query samples. Thus, adequate representation of diverse image categories is essential to enable MLLMs to grasp causal relationships effectively and efficiently.

ImageNet1k [35] is commonly employed for image classification tasks, and it is customary to train models on the entire dataset to enhance their recognition ability across all categories. In contrast, within the training configuration of LCL, we only select a limited number of samples randomly from each category. Then we arrange a set of related categories with decreasing similarity for each category, referred to as "neighbors". Specifically, we adopted CLIP [12] to calculate the similarity between different classes within the training dataset. Firstly, we randomly select 100 images from each class and calculate the average image feature for each class. Subsequently, we encode the text names of all classes to obtain their corresponding feature vectors. Ultimately, we compute weighted similarities across distinct class pairs, encompassing image-to-image, image-to-text, and text-to-text correlations. For a specific category, we sort all other categories based on similarity and divide them into N intervals. Then, within each interval, we randomly select categories to construct a set of "neighbors" with a total quantity of N.

3.2.2 Training Strategy

In order to make MLLMs understand the causal link between the support set and query sample, as well as the causal relationship between the input-label pairs in the support set, we build positive-negative pairs to urge the model to learn from comparisons. Let the support set be denoted as $S = \{s_1, s_2, ..., s_n\}$. Based on the correlation among its samples, we can redefine the support set as $C = \{c_1, c_2, ..., c_m\}$, where each c_m serves as a prototype representing a cluster of samples from S. These prototypes capture the essential relationships and similarities among samples within S. Given the query x, we train θ to maximize the likelihood:

$$\log p_\theta(y|x) = \sum_l \log p_\theta(y_l|x, C, y_1, y_2, ..., y_{l-1}), \quad (1)$$

where θ denotes the parameters of the language model. The parameters of the visual encoder are frozen during the training.

[2-way] strategy: In this strategy, we train the MLLMs for binary image classification, where the $C = \{c_1, c_2\}$. To be more specific, c_1 and c_2 here represent the prototype of two classes. We denote the training class set as $T = \{t_1, t_2, ..., t_{100}\}$, we randomly sample a class t_i as the positive class, where its neighbor class set $N^{t_i} = \{n_1^{t_i}, n_2^{t_i}, ..., n_{100}^{t_i}\}$ ($n_1^{t_i}$ is the most similar class to t_i, while the $n_{100}^{t_i}$ is the least). Then we apply a hard-negative mining strategy, where we sample the negative class $n_j^{t_i}$ from N^{t_i} with a probability $p_j = \frac{101-j}{\sum_{m=1}^{100} m}$. Noted that this setting is fixed to train on 16 shots.

[2-way-random] strategy: In this strategy, we first train the MLLMs on fixed-16 shots following the [*2-way*] strategy, then further train the model with shots averaged sampled from 2-16 shots for 10 epochs.

[2-way-weight] strategy: Within this strategy, we initially train the MLLMs using a fixed-16 shot regimen, adhering to the [*2-way*] approach. Subsequently, we refine the model by additional training with shots sampled from the range of 2-16, with each shot's probability denoted as $p_j = \frac{e^j}{\sum_{m=2}^{16} e^m}$.

[mix] strategy: To enhance the model's generalizability, we undertake a fine-tuning process that involves both [2-way] tasks and Shikra's [27] original tasks. During each iteration, the training samples are evenly sampled from both the [2-way] tasks and the original tasks. This balanced approach ensures that the model gains proficiency in both the newly introduced link-context learning tasks and the pre-existing tasks from Shikra [27].

4 ISEKAI Dataset

To objectively evaluate MLLM's ability to learn new concepts through LCL, we created an ISEKAI dataset, shown in Figure 4. The concepts involved are unreal, rarely seen in

Figure 4: **Overview of the ISEKAI Dataset**: This dataset comprises entirely generated images, where the images from "ISEKAI World" are non-existent in real life, while the images from "Real World" are sourced from reality.

legends, myths, or fictional media. Thus, MLLM's exposure to these concepts is minimal. The term "Isekai" originates from a fantasy subgenre in anime. Plots usually involve characters transported to a different world, like a fantasy realm or virtual universe. Audiences understand the new world gradually through the protagonist's exploration, akin to MLLM's journey into a new realm of knowledge.

The dataset's images are generated by Midjourney's [4] text-to-image model using well-crafted instructions. Images were manually selected to ensure core concept consistency. The dataset currently comprises 20 groups, and 40 categories in total (continues to grow). Each group pairs a new concept with a related real-world concept, like "octopus vacuum" and "octopus." These can serve as challenging negative samples for each other. Each concept has no less than 32 images, supporting multi-shot examples. These features enable ISEKAI to comprehensively assess the model's LCL capability. We also provide text descriptions of each concept's appearance and name, contributing to evaluations beyond LCL.

In this paper, we evaluated different models' performance on ISEKAI. For details, refer to Results on ISEKAI.

5 Experiments

In this section, we present the results of our experiments to showcase the effectiveness of our proposed method. We conduct comprehensive comparisons between our approach (link-context learning-based) and other in-context learning-based MLLMs.

5.1 Results on ISEKAI

To quantitatively evaluate the performance of link-context learning, we compare our methods in different strategies with our baseline (Shikra [27]) as well as ICL methods (Otter and OpenFlamingo) in two challenge datasets: ISEKAI-10 and ISEKAI-pair.

ISEKAI-10 Evaluation: Comprising 10 classes of challenging positive-negative image pairs, **ISEKAI-10** presents a scenario where the positive class is entirely nonexistent in the real world yet shares certain characteristics with the negative class, which comprises common animals or objects from our reality. The upper section of Table 1 showcases the outcomes on the ISEKAI-10 dataset, where vanilla-shikra [27] encountered difficulty. Our model demonstrates competitive performance compared with OpenFlamingo [6] and Otter [5]

Setting	Method	2-shot	4-shot	6-shot	8-shot	10-shot	12-shot	14-shot	16-shot
ISEKAI-10	OpenFlamingo [6]	0.46	0.44	0.46	0.48	0.50	0.50	0.48	0.46
	Otter [5]	0.23	0.23	0.19	0.15	0.14	0.12	0.10	0.07
	Vanilla-Shikra [27]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Ours-[2-way-random]	0.64	0.63	0.65	0.62	0.61	0.57	0.56	0.56
	Ours-[mix]	**0.68**	**0.70**	**0.73**	**0.69**	**0.63**	**0.62**	**0.65**	**0.62**
ISEKAI-pair	OpenFlamingo [6]	0.19	0.34	0.38	0.39	0.41	0.40	0.40	0.40
	Otter [5]	0.01	0.04	0.04	0.03	0.03	0.02	0.02	0.01
	Vanilla-Shikra [27]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Ours-[mix]	0.39	0.38	0.38	0.40	0.40	0.39	0.37	0.35
	Ours-[2-way-random]	**0.43**	**0.46**	**0.47**	**0.48**	**0.48**	**0.49**	**0.49**	**0.49**

Table 1: **Quantitative evaluation on ISEKAI** from zero-shot to 16-shot, measured by accuracy. We achieve the best results compared with Otter [5] and OpenFlamingo [6].

Method	zero-shot	2-shot	4-shot	6-shot	8-shot	10-shot	12-shot	14-shot	16-shot
OpenFlamingo [6]	0.00	0.41	0.62	0.72	0.75	0.77	0.78	0.73	0.72
Otter [5]	0.13	0.18	0.21	0.24	0.25	0.26	0.24	0.23	0.23
Vanilla-Shikra [27]	0.05	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ours-[mix]	**0.16**	0.73	0.78	**0.83**	0.73	0.71	0.72	0.65	0.57
Ours-[2-way]	0.02	0.51	0.61	0.68	0.73	0.77	0.78	0.78	0.79
Ours-[2-way-random]	0.0	**0.77**	**0.78**	0.77	**0.79**	**0.77**	0.77	0.77	0.75
Ours-[2-way-weight]	0.0	0.69	0.71	0.72	0.76	0.77	**0.78**	**0.78**	**0.79**

Table 2: **Quantitative evaluation on ImageNet-100** from zero-shot to 16-shot, measured by accuracy. We achieve the best results compared with Otter [5] and OpenFlamingo [6].

across all shot numbers.

ISEKAI-pair Evaluation: In the **ISEKAI-pair** evaluation, positive and negative pairs are constructed using all image categories that do not exist in the real world. Each individual image is paired with all images from other categories, facilitating a comprehensive assessment. This evaluation provides a realistic gauge of the model's capability to handle complete unknowns through various combinations. The lower section of Table 1 underscores our model's superiority over OpenFlamingo [6] and Otter [5] in this context.

Qualitative Results: Figure 1 provides a visual comparison between our model and OpenFlamingo [6], as well as Otter [5]. Notably, our model demonstrates its proficiency in accurately comprehending novel concepts and effectively discerning unfamiliar objects from those with close resemblance. This observation underscores our model's capacity to capture the causal relationship between the source and target domains from the demonstration.

5.2 Results on ImageNet-100

We proceed to assess our model's performance on ImageNet-100, encompassing 100 classes that were entirely absent from the training phase. The outcomes underscore the efficacy of our *mix* strategy, which attains the highest accuracy of **83%** at 6-shot. In contrast, Otter achieves a peak accuracy of **25%**, and OpenFlamingo's performance reaches **78%**. Unlike the ISEKAI dataset, the images from ImageNet-100 do correspond to real-world entities.

5.3 Ablation Study

Does the ground-truth input-label mapping exists? We conduct an ablation analysis on the correctness of labels within the demonstration (support set). Given a set of image domains $\mathcal{X}c \in \mathbb{R}^{H\times W\times 3}$ and label domains $\mathcal{C} \in \mathbb{R}^{N}$, a mapping $f : \mathcal{X}_c \rightarrow \mathcal{C}$ exists to associate each image with its corresponding label. We use several image-label pairs $\{(x^1_{c_1}, c_1), (x^2_{c_1}, c_1), ..., (x^n_{c_1}, c_1)\}$, where $x^j_{c_i} \in \mathcal{X}_{c_i}$, as the support set. The model is going to predict the correct answer from a candidate set Y:

$$\hat{y} = \arg\max_{y_i \in Y} P(y_i|x, f), \tag{2}$$

where the prediction is conditioned on the mapping f. Consequently, intentionally breaking the mapping relationship within the support set would lead the model to provide incorrect answers, as it heavily relies on the accurate association between the image-label pairs of the support set to make precise predictions. As shown in Figure 7, we disturb the mapping f by gradually inserting false labels into the support set, and the accuracy falls from 0.78 to 0.00 when the correctness of the labels falls from 100% to 0%. These results clearly show that maintaining accurate associations between image-label pairs within the support

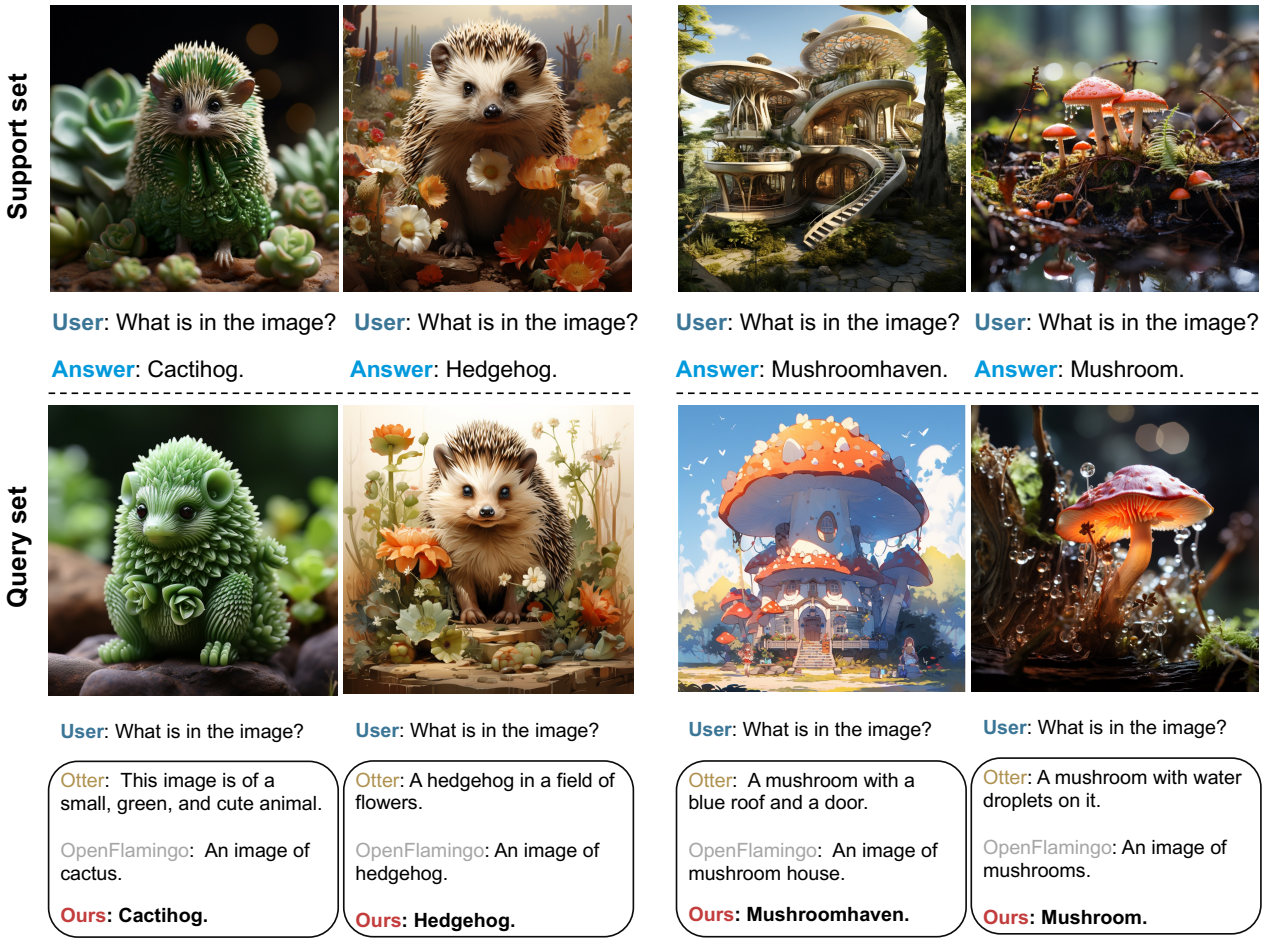

Figure 5: **Qualitative comparisons** of novel images understanding results between ours and OpenFlamingo [6], Otter [5]. The name "Cactihog" is a fusion of "cactus" and "hedgehog", combining the key features of these two creatures. The name "MushroomHaven" suggests a dwelling place characterized by giant mushrooms

set plays a crucial role in link-context learning.

Would the model benefit from using a larger shot?

Much like supervised learning, the model's accuracy experiences rapid initial growth with an increasing amount of training data, eventually reaching a plateau. During this phase, the selection of more representative samples becomes crucial. Figure 6 presents two outcomes: one depicts model accuracy from separate training at a fixed shot (gray bar in the figure), while the other showcases the model's performance through sampling across various shots (red line in the figure). The results reveal slight gains from lower fixed-shot training and consistent performance from random-shot training. Notably, in both random and fixed settings, accuracy plateaus or experiences gradual growth after the 8-shot threshold.

What does the model's decision-making in the case of multi-shot depend on?

As shown in Fig 8, when disturbing the label of different positions, the accuracy of the model with 16-shot drops differently, which reflects the extent to which the model prefers different locations. We observe that the model heavily relies on the beginning and the middle positions. From another aspect, it provides an explanation of why the model encounters a plateau in a higher number of shots. Similarly, this phenomenon also exists in LLMs [36], where the language model tends to be "lost in the middle" when processing long contexts. They also reveal that the model's performance keeps decreasing when the contexts grow longer.

What is the difference between different training strategies?

Table 2 presents a comprehensive view of the outcomes achieved through our four distinct training strategies. The *mix* strategy stands out by elevating the zero-shot accuracy from 5% to 16% and attaining a remarkable 83% accuracy at 6-shot; however, its performance diminishes to 57% at 16-shot. In contrast, the *2-way* strategy, anchored at 16-shot training, initiates with a 51% accuracy at 2-shot and progressively ascends to 79% at 16-shot. Interestingly, we observe that the accuracy trend of the *2-way* strategy isn't solely attributable to an increase in shots, but rather stems from a closer alignment with the trained pattern. To validate this,

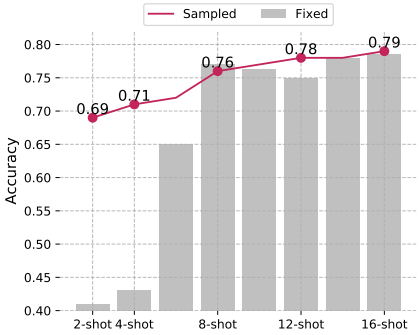

Figure 6: **The ablation study on shot number**. The grey bars illustrate the highest accuracy achieved for each shot number, denoting specific shot-based training. The red line illustrates the performance of the model trained using a sampled strategy. Notably, both scenarios exhibit plateaus in accuracy after reaching the 8-shot mark.

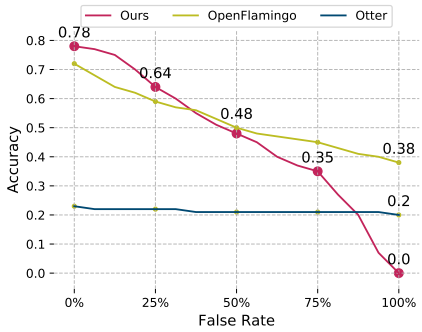

Figure 7: **The ablation study on false rate**. In contrast to OpenFlamingo [6], which sustains a 38% accuracy at a 100% false rate, our model attains 0% accuracy under the same conditions. This outcome underscores our model's ability to preserve precise linkages between the support set and the query.

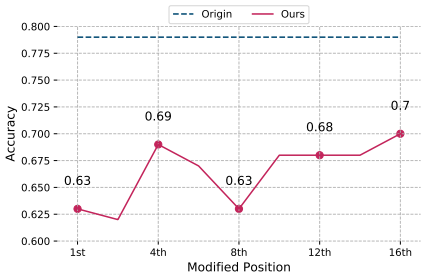

Figure 8: **The effect of label modifications at distinct positions**. The dashed blue line serves as a reference for the original accuracy, while the red line portrays the accuracy of our model subsequent to the label modified at specific positions. Significant accuracy drop reflects position dependency, while minor change indicates position insignificance in the model's decision-making.

we introduce two additional settings: *2-way-random* and *2-way-weight*. These settings undergo fixed-shot training for initialization, followed by finetuning across 2-16 shots with random and weighted approaches, respectively. Both exhibit considerable accuracy improvements in lower shots. Notably, while the accuracy of higher shots, finetuned with a random strategy, drops—an observation mirroring the behavior of the *mix* strategy. These results underscore the efficacy of an even, sustained, and generalized training approach in harnessing the potential of large language models, revealing the emergence of a "lost-in-the-middle" phenomenon, in coherence with our earlier observations.

Does the training harm the zero-shot performance? Table 3 shows the comparison between our-7B model with shikra-13B [27] and some previous SOTA methods on Imagenet-100 and VQAv2. From the results, we conclude that our *mix* training strategy would not harm the model's zero-shot performance.

Method	ImageNet-100	VQAv2^{dev}	VQAv2^{std}
OpenFlamingo [6]	0.00	-	-
Flamingo-80B [33]	-	56.3	-
Flamingo-9B [33]	-	51.8	-
BLIP2 [9]	-	65.0	-
Otter [5]	0.13	-	-
Shikra-13B [27]	0.05	77.3	77.5
Ours-7B-[mix]	**0.16**	TBD	TBD

Table 3: **Quantitative evaluation** was conducted on both ImageNet-100 and VQAv2 datasets employing a zero-shot approach. The outcomes substantiate that our training strategy exhibits no detrimental impact on the zero-shot performance.

6 Discussion

6.1 Limitations

We believe that this work introduces a challenging and promising setting for both MLLMs and LLMs. However, the primary focus in this paper lies on link-context learning within the context of MLLMs, specifically validating the basic tasks such as image classification. Consequently, this work should be regarded as a foundational baseline for exploring the potential of link-context learning.

Looking ahead, future research directions encompass a deeper theoretical analysis that delves into the intricacies of the causal relationship between the support samples and, crucially, between the support set and the query. Understanding and unraveling the complexities of these causal links represent meaningful avenues of inquiry that could lead to significant advancements in the capabilities of models in reasoning, learning, and adapting to novel scenarios. As the field progresses, we anticipate further investigations and refinements that will not only enrich our understanding of link-context learning but also implement in-context learning for MLLMs and LLMs in a unified way.

6.2 Conclusion

In conclusion, this paper introduces a groundbreaking paradigm of causal-relevant few-shot learning, significantly expanding the capabilities of Multimodal Large Language Models (MLLMs) within the context of single conversations. Through meticulous experimentation and a carefully devised training strategy, we demonstrate that MLLMs can adeptly establish a mapping between ground-truth input-label pairs, thereby acquiring the proficiency to seamlessly generalize

this capacity to previously unencountered images and novel concepts. This pivotal advancement propels MLLMs into uncharted territories, enabling them to not only acquire but also apply knowledge in a manner more akin to human cognition.

References

[1] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

[2] 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.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[4] Midjourney. Midjourney. https://www.midjourney.com, 2023.

[5] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *CoRR*, abs/2305.03726, 2023.

[6] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, March 2023.

[7] OpenAI. Gpt-4 technical report, 2023.

[8] 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*, pages 12888–12900. PMLR, 2022.

[9] 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*, 2023.

[10] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[12] 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*, pages 8748–8763. PMLR, 2021.

[13] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[14] Hao Yang, Junyang Lin, An Yang, Peng Wang, Chang Zhou, and Hongxia Yang. Prompt tuning for generative multimodal pretrained models. *arXiv preprint arXiv:2208.02532*, 2022.

[15] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[16] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.

[17] Xuejing Liu, Wei Tang, Jinghui Lu, Rui Zhao, Zhaojun Guo, and Fei Tan. Deeply coupled cross-modal prompt learning. *arXiv preprint arXiv:2305.17903*, 2023.

[18] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

[19] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning, 2022.

[20] Shuhuai Ren, Aston Zhang, Yi Zhu, Shuai Zhang, Shuai Zheng, Mu Li, Alex Smola, and Xu Sun. Prompt pre-training with twenty-thousand classes for open-vocabulary visual recognition, 2023.

[21] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning, 2023.

[22] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.

[23] OpenAI. Introducing ChatGPT. https://openai.com/blog/chatgpt, 2023.

[24] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.

[25] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023.

[26] 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, 2023.

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

[28] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world, 2023.

[29] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. A survey on in-context learning, 2023.

[30] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action, 2023.

[31] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models, 2023.

[32] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training, 2022.

[33] 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, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc., 2022.

[34] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[35] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[36] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.