What Matters in Training a GPT4-Style Language Model with Multimodal Inputs?

Anonymous ACL submission

Abstract

Recent advancements in GPT-4V have displayed remarkable multi-modal capabilities in processing image inputs and following open-ended instructions. Despite these advancements, there is considerable scope for enhancing open-source multi-modal LLMs, especially in terms of multi-modal understanding accuracy and instruction-following proficiency. In this paper, we conduct a comprehensive study on training GPT4-style models. We introduce **Lynx**, a multi-modal LLM developed through a series of controlled experiments comparing various model variants. This process allowed us to identify and implement an optimal training strategy tailored for multi-modal LLMs. In addition to our model development, we propose a plug-and-play technique designed to augment the instruction-following capabilities of multi-modal LLMs. We have validated the performance of Lynx on multiple benchmarks. Results demonstrate that Lynx not only achieves strong image understanding accuracy but also excels in instruction-following tasks, paving the path for ongoing enhancements in multi-modal LLMs.

1 Introduction

Large Language Models (LLMs) (Brown et al., 2020; Hoffmann et al., 2022; Chung et al., 2022; Ouyang et al., 2022; Scao et al., 2022; Iyer et al., 2022; Touvron et al., 2023) have progressed rapidly in recent years and achieved impressive performance in language understanding and generalization. With instruction fine-tuning (Ouyang et al., 2022; Chung et al., 2022; Wang et al., 2022; Chiang et al., 2023; Xu et al., 2023; Peng et al., 2023), LLMs can be further improved to follow open-ended instructions from non-expert users and serve as dialog-based assistants in our daily lives. Leveraging powerful LLMs, recent studies have examined methods for adapting LLMs to multimodal inputs (e.g., images (Alayrac et al., 2022; Zhu et al., 2023; Dai et al., 2023), videos (Li et al., 2023a; Chen et al., 2023; Zhao et al., 2023), and audio (Huang et al., 2023a)) and outputs (e.g., vision tasks (Wang et al., 2023), and robotic manipulation skills (Jin et al., 2023; Driess et al., 2023; Jiang et al., 2022)). Notably, GPT-4V (OpenAI, 2023b) has astounded the world with its impressively versatile yet practical capabilities, such as generating descriptions, stories, poetry, advertisements, and codes given images, which were rarely observed in previous vision language models (Li et al., 2021; Bao et al., 2022; Zeng et al., 2021, 2022).

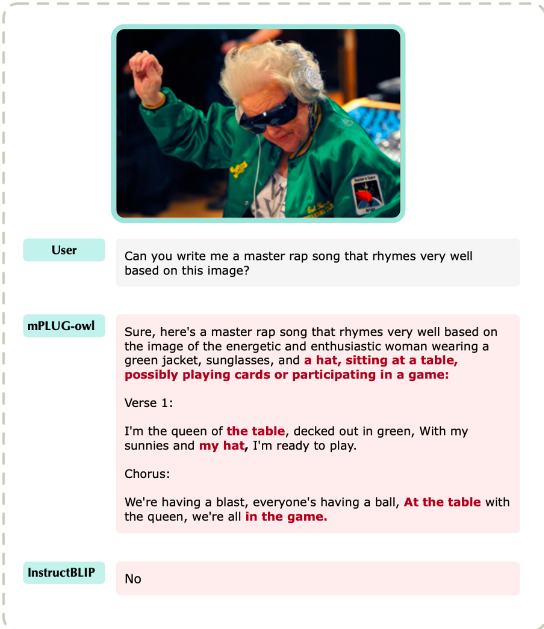


Figure 1: Illustration of hallucination and diminished instruction-following capabilities generated from open-source multi-modal LLMs.

However, it still remains a mystery that: *How does GPT-4V obtain its impressive smartness?* GPT-4V can analyze and generate descriptions for various types of images, including diagrams, text in images, maps, screenshots of software interfaces, illustrations, comics, and medical imagery. Such capabilities likely depend on extensive annotated data and perhaps the integration of tools like OCR. This presents a substantial challenge for open-source multi-modal LLMs which are end-to-end models trained on open-source datasets with lower-quality annotations.

Despite these challenges, we have observed that current multi-modal language models usually suffer from hallucinations, such as generation of facts unrelated to the image inputs, and degraded instruction-following abilities compared to their text-only counterparts, as shown in Figure 1. Though actively investigated recently, the open-source multi-modal LLMs are usually different in training data, training recipes, prompts, and evaluation benchmarks, which makes it challenging to identify which factors are crucial in achieving a strong baseline model.

In this paper, we conduct a comprehensive study on training GPT4-style models. By implementing multiple model variants under controlled settings and conducting extensive experiments to draw reliable conclusions both quantitatively and qualitatively, our findings can be summarized as follows:

- Data quality is more important than quantity. Our experiments with COYO700M (Byeon et al., 2022), DataComp1B (Gadre et al., 2023), and BlipCapFilt (Li et al., 2022) demonstrate that pre-training on COYO700M and DataComp1B does not yield improved model performance. Furthermore, it is crucial to avoid utilizing low-quality annotations during the instruction fine-tuning phase.

- Diversified prompts are essential for enhancing the model's ability to follow instructions. We incorporate a range of open-source multi-modal datasets, transforming them into an instruction-following format using manually crafted prompts, supplemented by additional prompts generated by GPT-4, producing 500 prompts for over 50 tasks in total.

- Language instruction fine-tuning plays a significant role in boosting the instruction-following capabilities of multi-modal models. Integrating NLP instruction data during training enables the model to handle a broader range of tasks, compensating for the typical limitations of open-source multi-modal datasets, which are predominantly focused on image captioning and visual reasoning. Furthermore, our results also indicate that multi-modal models based on Vicuna-7B, which have undergone instruction fine-tuning, generally outperform those based on LLaMA-7B.

Through our study, we present **Lynx**, a multi-modal LLM with a three-stage training recipe. The first stage focuses on aligning vision and language, the second stage enhances the resolution of image inputs, and the final stage concentrates on instruction fine-tuning. At each stage, a different set of model parameters is trained using a varied mixture of image-text pairs, multi-modal datasets, and NLP instruction data, in order to fully exploit training data.

Furthermore, we propose a plug-and-play method aimed at boosting the instruction-following capabilities of multi-modal LLMs. Our observations indicate that a more extended dialogue history can effectively unlock and amplify the instruction-following abilities inherent in LLMs. We recognize that most multi-modal tasks can be broken down into two steps: describing the image in details, and subsequently utilizing the capabilities of LLMs. Consequently, during evaluation, we initiate the task by prompting the model to describe the image. This strategy consistently results in improved performance across a variety of multi-modal instruction-following tasks, demonstrating the efficacy of our proposed method.

Experimental results show that Lynx is a strong baseline for multi-modal LLMs, achieving state-of-art performance. [1] Specifically, our model excels in MME's perception (Fu et al., 2023), demonstrating superior image understanding accuracy, and shows impressive results in VisIT-bench (Bitton et al., 2023), which evaluates multi-modal instruction-following abilities. We also adopt the OwlEval test set proposed by mPLUG-owl (Ye et al., 2023) for human evaluation, with Lynx achieving state-of-art multi-modal instruction-following results.

2 Lynx

Lynx is a large language model that can take images and videos as inputs. In this section, we will introduce our Lynx in detail, including the architecture (2.1) and three-stage training recipe(2.2).

[1] We compared with open-sourced multi-modal LLMs of comparable model sizes available as of August 31, 2023, for all evaluations.

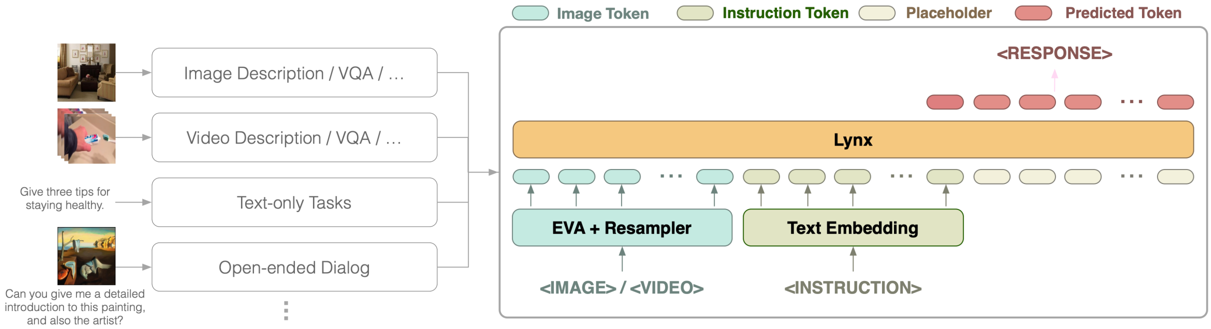


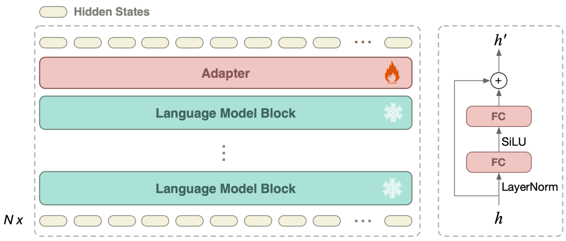
Figure 2: Architecture of Lynx. The vision tokens are concatenated with the text tokens to generate text outputs auto-regressively.

2.1 Model Architecture

Overview Our model takes vision and language as inputs to generate text responses following the input instructions. The overall structure of our model is shown in Figure 2. Concretely, vision inputs are first processed by a vision encoder to get a sequence of vision tokens $\mathbf{w}_v$. After that, $\mathbf{w}_v$ are concatenated with instruction tokens $\mathbf{w}_l$ as the input of LLMs for multi-modal tasks. To generate responses, the left-to-right causal decoder auto-regressively predicts the next token by taking all previous tokens as inputs until encountering the <EOS>.

Adapter The trainable adapters are inserted into the LLMs after every M blocks. In our experiments, $M = 2$. As shown in Figure 3(b), the adapter linearly projects each token into a lower-dimensional space and then re-projects it back. Concretely, in Lynx, the hidden state for each token is 4096-d. The adapter first imposes layer normalization (Ba et al., 2016) onto the hidden states. Then a linear layer is used to downsample the dimension of each token state from 4096 to 2048, based on which SiLU (Elfwing et al., 2018) is set as the non-linear activation function, which keeps consistent with LLaMA (Touvron et al., 2023). Finally, the other linear layer is used to re-map the 2048-d hidden state back to 4096-d.

Vision Encoder To extract vision features of images and video frames, we apply EVA-1B (Fang et al., 2023; Sun et al., 2023) as our vision encoder $\phi_v(x)$. It maps an image to a sequence of visual tokens. The downsample rate is 14, meaning that an image with resolution $H \times W$ will be represented by a sequence of $\frac{H}{14} \times \frac{W}{14}$ tokens. To improve the efficiency of training and inference, we adapt the resampler Φ mechanism (Jaegle et al., 2021; Alayrac et al., 2022) that reduces the dimensions of vision inputs by injecting the long vision token sequence into a short and learnable query sequence $\mathbf{w}_v^q$:



(a) Architecture of Lynx (b) Architecture of the adapter

Figure 3: Architecture of Lynx. (a) Overall; (b) Adapter.

$$\mathbf{w}_v = \Phi(\phi_v(x), \mathbf{w}_v^q) \tag{1}$$

where x is the input image, $\phi_v(x)$ is the raw tokens directly given by the vision encoder, $\mathbf{w}_v$ is the condensed token sequence consisting of 32 tokens regardless of the number of raw tokens from the vision encoder.

2.2 Training Recipe

We propose a three-stage training recipe to enhance image understanding accuracy and instruction-following abilities. Specifically, we first train the model to align vision inputs to the pre-trained LLM, and then we enhance the model by increasing the resolution of visual inputs. Last, we conduct instruction fine-tuning with carefully selected training datasets. Our model is trained on a total of $\sim$14B tokens during the pretraining and resolution enhancement stage, $\sim$3B tokens during the instruction-finetuning stage, all phases employing causal prediction loss.

Pretraining During pretraining, we freeze the vision encoder and language model and train the resampler and the inserted adapters for 100k steps using a batch size of 1536. The learning rate is set to 0.0001 (details about hyper-parameters can be found in Appendix Table D.2). To accelerate pretraining, the image resolution is set to 224x224. To establish alignment between visual features and the language model, we utilize more than 120M image-text pairs, mainly consist of BlipCapFilt 115M (Li et al., 2022), CC12M (Changpinyo et al., 2021), CC3M (Sharma et al., 2018), and SBU (Ordonez et al., 2011). Additionally, we also utilize public multi-modal downstream tasks during this phase. We include three predominant tasks, image captioning, visual question answering, and image classification. Details of the datasets we adopted and their mixing percentage are listed in Appendix Table 12.

Resolution Enhancement Training on low-resolution images is insufficient for certain downstream tasks, such as table reading and OCR. Therefore, after 100k steps of pretraining at low resolution, we increase the input resolution to 420×420 and continue training for an additional 10k steps. During this stage, the batch size is reduced to 448, and the learning rate is adjusted to 0.00001. In this phrase, we train the vision encoder together with the resampler and the inserted adapters, adapting the model to higher resolution inputs. Moreover, we adjust the training datasets by excluding large-scale noise image-text pairs and including new multi-modal downstream tasks that requires high-resolution image inputs, such as table reading and OCR.

Instruction Fintuning In this phase, we only train the resampler and the inserted adapters, same as in the pre-training stage. The instruction finetuning process consists of 20k training iterations with a batch size of 480 and a learning rate of 0.00002. Our finetuning datasets consists of text-only, image-text, and video-text tasks for complex multi-modal reasoning and instruction following, which mainly belongs to 5 categories: text-only instruction-following task, image/video visual question answering, image/video captioning, classification, and image-conditioned dialog. To finetune our model with diversified instructions, we provide appropriate instructions for each of these public datasets (see Appendix Table 13 for details). Specifically, we manually labeled at least 3 different prompts for each of these tasks, and then invoke GPT4 to automatically generate more based on the following "meta prompt", i.e., the prompt used to generate prompts for different tasks:

Here are some instructions that define a visual-language task. Continue to write 15 instructions with the same meaning: 1) PROMPT1; 2) PROMPT2; 3) PROMPT3;

Besides, we also collect some available public (visual-)text instruction data (also listed in Appendix Table 12) to further improve the ability of our model to follow open-ended instructions, including the instruction data used in FlanT5 (Chung et al., 2022), Alpaca (Wang et al., 2022), MiniGPT4 (Zhu et al., 2023), LLAVA (Liu et al., 2023a), and Baize (Xu et al., 2023).

We observe that different combinations of the instruction data have a crucial influence on the final performance. Empirically, we finally impose the weighting strategy presented in Appendix Table 12.

2.3 Inference

During the inference, we employ the nucleus sampling decoding method together with the beam search strategy. Detailed hyper-parameters for the generation process are presented in Appendix Table 9.

Moreover, we propose a plug-and-play method to improve instruction-following capabilities of multi-modal LLMs. We observe that most multi-modal tasks can be broken down into two steps: describing the image in details, and subsequently utilizing the inherent capabilities of LLMs. Therefore, we propose to *initiate a task by prompting the model to describe the image first during evaluation*. By doing so, we effectively unlock and amplify the instruction-following abilities inherent in multi-modal LLMs with a more extended dialogue history.

3 Experiment

In this section, we aim to answer the following questions by empirical studies:

a) What advantages does our Lynx offer in comparison to existing models? (Section 3.1)

b) What are the key factors in training a high-performance GPT4-style model? (Section 3.2)

c) How does our proposed plug-and-play method enhance the instruction-following capabilities of the model? (Section 3.2)

3.1 Quantitative Experiments

The evaluation of multi-modal LLMs is essentially different from typical visual-language methods. The primary challenge when evaluating a GPT4-

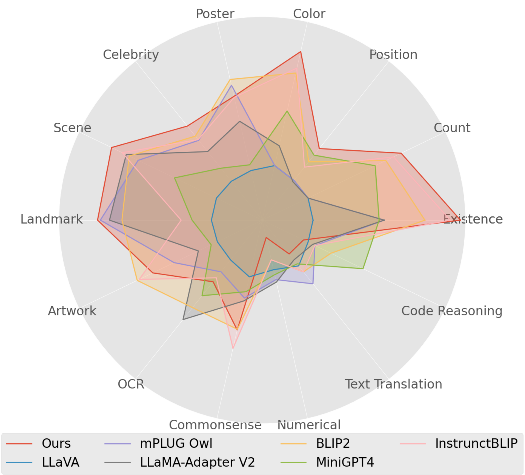


Figure 4: Comparison on MME benchmark.

style model is balancing multi-modal understanding accuracy and instruction-following abilities. For a thorough evaluation of both aspects, we adopt four protocols for quantitative evaluation. *The Open-VQA and MME benchmarks primarily assess the precision of visual comprehension, whereas the VisIT-Bench and OwlEval benchmarks evaluate instruction-following capabilities.*

Open-VQA benchmark We collect an Open-ended Visual Question Answering (Open-VQA) test set, including questions on objects, OCR, counting, reasoning, action recognition, chronological ordering, and more. Different from standard VQA (Antol et al., 2015; Zhang et al., 2016), the ground-truth answer in Open-VQA is open-ended. To evaluate the performance on Open-VQA, we prompt GPT4 to make it a discriminator, yielding a 95% agreement with human evaluation, see Appendix B for details. From the results of Table 1 and 2, we can conclude that our model has achieved the best performance both in the image and video understanding tasks. Notably, InstructBLIP (Dai et al., 2023) also achieves high performance in most cases, even better than our model in OCR, color recognition, and action recognition tasks. However, we observe that it always outputs one word answer as shown in Appendix Figure 6 and 7, which is less preferred by most of the users (see Table 4). We also showcase some of the examples in Appendix Figure 10. More cases including video VQA examples can be found Figure 10 and 11 in the Appendix. We can see that our model can give the correct answer in most cases as well as a concise reason that supports the answer.

MME benchmark We further compare Lynx with existing open-source models on the MME benchmark (Fu et al., 2023). Instruction-answer pairs for the MME benchmark are constructed manually, covering the examination of perception and cognition abilities. These instructions are deliberately succinct, facilitating intuitive and convenient quantitative analysis, as opposed to employing GPT models or manual scoring methods. Results are shown in Figure 4 and Appendix D.4. We can see that our model is a state-of-the-art model in 7 out of 14 subtasks, especially for the perception tasks including color, celebrity, scene, landmark, position, count, and existence. Yet, from the figure, we can also see that our model seems not to perform well on code reasoning, text translation, and numerical. This deficiency may be attributed to the absence of text translation, coding reasoning, and numerical tasks in our training datasets. Moreover, each of these three tasks only contains 20 examples, which may lead to high variance in the evaluation of different checkpoints.

VisIT benchmark We further conduct evaluation on the VisIT-Bench (Bitton et al., 2023) Single Images benchmark, a benchmark for evaluation of the instruction-following multi-modal LLMs. This benchmark comprises 592 test queries across varied domains, including art, object recognition, spatial understanding, and chemical analysis, etc. Each query is paired with a human-generated, instruction-based caption, enabling automated evaluation against text-only LLMs, and providing a standard answer for reference. The ELO-based results are presented in Table 3, this approach uses an GPT-4 evaluator to compare two models with an instruction and an instruction-conditioned caption. Our model ranks third, underperforming only LLava(13B) and LLava-a1(13B), surpassing counterparts including mPlug-Owl (Ye et al., 2023), LlamaAdapter-v2 (Gao et al., 2023a), InstructBLIP (Dai et al., 2023), Otter (Li et al., 2023a), etc, demonstrating the best performance among models with a comparable parameter scale (i.e. 7B).

OwlEval benchmark We adopt the OwlEval test set proposed by mPLUG-owl (Ye et al., 2023) to manually assess the text generation ability given images. Though OwlEval is a tiny set containing only 82 questions based on 50 images, it covers a diverse range of tasks such as generating descriptions, stories, poems, advertisements, codes, and other sophisticated yet practical analyses of

	OCR	Counting	Reasoning	Place	Color	Spatial	Action	Others	Overall
Open-Flamingo-0	20/53	5/37	15/31	**18/22**	5/30	7/15	11/20	53/94	44.37
Open-Flamingo-4	14/53	6/37	15/31	17/22	9/30	7/15	11/20	51/94	43.05
Multimodal GPT	19/53	8/37	21/31	12/22	8/30	6/15	12/20	56/94	47.02
MiniGPT-4	32/53	13/37	13/31	17/22	16/30	**9/15**	16/20	63/94	59.27
LLaVA	21/53	8/37	13/31	11/22	12/30	4/15	16/20	49/94	44.37
mPLUG-owl	34/53	8/37	16/31	16/22	14/30	**9/15**	13/20	62/94	56.95
BLIP2	29/53	15/37	21/31	12/22	17/30	8/15	16/20	67/94	61.26
InstructBLIP	**41/53**	20/37	**26/31**	14/22	**23/30**	6/15	**18/20**	77/94	74.50
Ours	36/53	**25/37**	**26/31**	17/22	21/30	**9/15**	17/20	**79/94**	**76.16**

Table 1: Quantitative evaluation results (accuracy) on Open-VQA image test set. For all models, we apply the same hyper-parameters defined in Appendix D.3.

	Action (Y/N)	Others	Overall	plug-and-play
InstructBLIP	62/108	21/40	56.08	✓
mPLUG-owl	65/108	19/40	56.76	✓
MiniGPT-4	56/108	18/40	50.00	✓
Ours	59/108	26/40	57.43	
Ours	**69/108**	**29/40**	**66.22**	✓

Table 2: Comparison of existing open-source multi-modal LLMs on the Open-VQA video benchmark.

	Elo	matches	Win vs. Reference(w/# ratings)
human verified reference	1361	6030	–
Llava(13b)-a1	1206	724	30.15% (n=136)
Llava(13b)	1091	5474	18.53% (n=475)
Lynx(7B)*	1078	708	15.15% (n=132)
mPLUG-Owl	1076	5465	16.04% (n=480)
LlamaAdapter-v2	1055	5485	14.14% (n=488)
idefics9b	1030	842	9.72% (n=144)
Lynx(7B)	1012	827	11.43% (n=140)
InstructBLIP	995	5505	14.12% (n=503)
Otter	970	5495	7.01% (n=499)
visual gpt davinci003	937	5486	1.57% (n=510)
Octopus-V2	936	820	8.90% (n=146)
MiniGPT-4	899	5473	3.36% (n=506)
Openflamingo	831	5490	2.95% (n=509)
PandaGPT(13b)	767	5480	2.70% (n=519)
mmgpt	757	5504	0.19% (n=527)

Table 3: Reference-free Elo rankings on VisIT-Bench (Single Image). Lynx(7B)* indicates the results applied the proposed plug-and-play method.

	InstructBLIP	BLIP2	MiniGPT-4	mPLUG-owl	Ours
scores	2.04	2.34	3.17	3.59	**4.13**

Table 4: Comparison of human-evaluation performance on OwlEval. Scores are averaged over the number of questions.

given images. From the human evaluation results in Table 4, we can see that our model has the best instruction-following performance while keeping high performance on the Open-VQA benchmark. BLIP2 (Li et al., 2023b) and InstructBLIP (Dai et al., 2023), though achieved high performance on the Open-VQA benchmark, are not preferred by human users due to their extremely short outputs, i.e., in most cases, they only output one word or phrase as the answer without any explanation. In contrast, MiniGPT4 (Zhu et al., 2023) and mPLUG-Owl (Ye et al., 2023) keep more instruction-following abilities. Hence, they are preferred over the BLIP models, though they may make more factual errors. We show more results on the OwlEval in Appendix Figure 7.

Overall, if a model has lower accuracy on the Open-VQA and MME benchmark, it tends to make factual errors. However, previous methods with higher performance on these two benchmarks usually tend to lose instruction-following abilities, e.g., always generating short answers, leading to inferior performance on OwlEval and VisIT-Bench. We attribute it to the under-training or over-training on visual-language tasks. Specifically, existing training data from visual-language tasks predominantly have short outputs. Though, by training on these data the model learns vision language alignments, it loses the instruction-following abilities inherited from the large language model. According to the experimental results, we can see that the multi-stage training recipe of Lynx contribute to accurate image understanding and enhanced instruction-following abilities.

3.2 Ablation Study

We conduct an in-depth ablation study to investigate the impact of different components or training recipes on multi-modal understanding and instruction-following performances.

Impact of Training Data We investigate the impact of data quantity and quality by training our model with or without the large-scale yet noisy image-text pairs (COYO700M (Byeon et al., 2022)

	Open-VQA image	Open-VQA video	OwlEval (win/all)
w/ LLaMA	70.86	60.81	42/82
w/o diverse prompts	75.50	59.46	18/82
w/ large-scale noisy data	72.85	60.14	38/82
w/o high-resolution	73.51	62.16	-
Ours	**76.16**	**66.22**	-

Table 5: Comparison of ablation models on the Open-VQA benchmark(see Appendix Table 7 for details) for evaluating multi-modal understanding accuracy, and on the OwlEval benchmark for assessing instruction-following abilities. The 'win/all' metric indicates that, out of all 82 cases in the OwlEval benchmark, this ablation model outperforms in these instances.

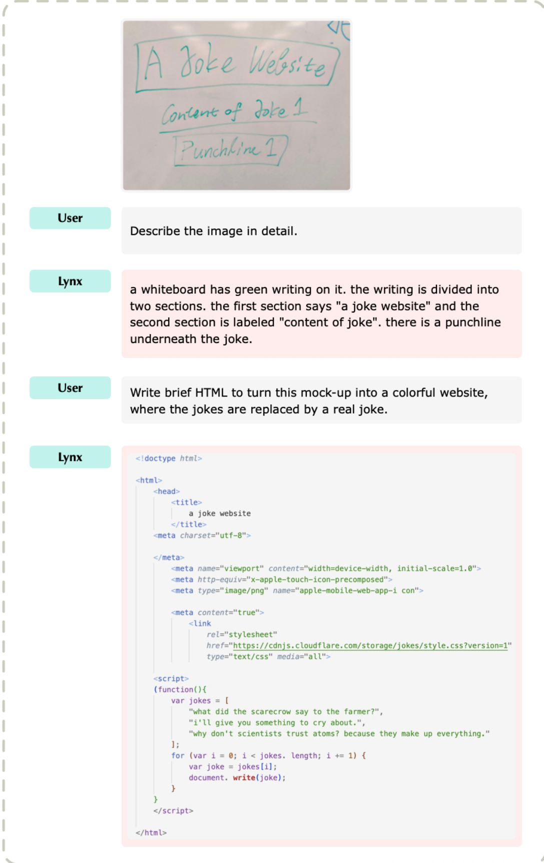


Figure 5: Illustration of the effectiveness of the proposed plug-and-play method.

and DataComp1B (Gadre et al., 2023)). During our experiments, we find training data in both pretraining and finetuning largely influence the model performance. Different from traditional visual-language pretraining (Radford et al., 2021), we find that multi-modal LLMs do not benefit from large-scale but noisy image-text pairs because many of the texts in such datasets are not fluent or natural language expressions. For the generative pretraining in our model, they largely damage the language generation ability as shown in Appendix Figure 9(b). As a result, pretraining on such large-scale datasets achieves no better results than only training on a much smaller but cleaner dataset as evaluated by the human users as shown in Table 5.

Impact of Diversified Prompts It has been proved to be important to train LLMs on instruction data so as to make them follow instructions correctly (Chung et al., 2022; Ouyang et al., 2022). Therefore, we ablate our model with diversified prompts written by both users and GPT4. The results in Table 5 show that our prompts help to balance different abilities. Moreover, we also find that by using diversified prompts, our model can follow the open-ended instructions better than the ones trained without these prompts (Appendix Table 13). This observation accords with the text-only models. The human evaluation results in Table 5 also accord with our observations.

LLaMA vs. Vicuna As shown in Table 5, in the aspect of multi-modal understanding accuracy, Lynx with instruction-finetuned backbone, namely Vicuna, substantially outperforms the counterpart with LLaMA on the Open-VQA benchmark in both image understanding and video understanding. The human evaluation on OwlEval, however, shows the two model variants have comparable performances. We further test the two variants on a simple yet practical instruction, "give a short answer", and the results show that Vicuna-based model follows the instruction better, producing answers with an average length of 15.81, compared to 20.15 from the LLaMA-based model. One can also refer to Appendix Figure 9(a) for examples of the comparison in terms of their instruction-following ability.

Impact of Larger Image Resolution Our research includes an ablation study to assess the effect of image resolution on model efficacy. The results presented in Table 5 demonstrate enhanced performance for the model trained with 420×420 resolution images relative to those trained with 224×224 resolution, with notable improvements observed in OCR and counting tasks within the Open-VQA benchmark, where OCR accuracy increases from 55.6% to 67.9%, and counting accuracy rises from 54.1% to 67.6%. More details can be found in the Appendix Table 7.

Plug-And-Play Method We introduce a plug-and-play method to improve the instruction-

	Reasoning	Calculation	Translating	Code	Overall
Lynx	110.71	17.50	42.50	45.00	215.71
Lynx*	103.57	55.00↑	75.00↑	77.50↑	311.07↑

Table 6: Comparison of Lynx and Lynx* applied the plug-and-play method on MME cognition benchmark, including commonsense reasoning, numerical calculation, text translation, and code reasoning tasks.

following capabilities of multi-modal LLMs by first prompting the model with "describe the image in detail". As indicated in Table 3 on VisIT-Bench, our model's rank improves from seventh to third with the proposed method. A similar improvement is observed on the Open-VQA video benchmark, where our model's score increases significantly from 57.43 to 66.22, detailed in Table 2. Experiments on the MME benchmark reveals that the most pronounced gains are in tasks requiring numerical calculation, code reasoning, and text translation, as showed in Table 6. Despite the notable lack of these types of data in our training process, with the proposed plug-and-play method, we unlock the inherent abilities of LLMs for tasks such as code writing or text translation. One showcase of this method is illustrated in Figure 5, where the model provides accurate code based on the information obtained from the dialogue history.

4 Related Work

Centralized Multi-modal Interactive System. Recent works investigate actively to design of such multi-modal interactive models. One of the most intuitive ideas, such as Visual ChatGPT (Wu et al., 2023), MM-REACT (Yang et al., 2023), HuggingGPT (Shen et al., 2023), InternGPT (Liu et al., 2023b), SayCan (Ahn et al., 2022), InnerMonologue (Huang et al., 2022), integrates various existing individual models or tools. In such a system, the LLM works as a “manager” that directly accepts instructions from users and selects the most appropriate tools to respond to requests while the integrated individual models are “workers” responsible for a specific kind of task. Typically, such models are powerful to address problems that are already well-defined. Yet, they, to some extent, lack zero-shot ability when encountering open-ended instructions which cannot be handled by any of their workers.

End-to-end Multi-modal Large Language Models. By contrast, inspired by the recent advances of LLMs, it has also been shown feasible and promising to directly train the neural networks that directly accept multi-modal inputs and output responses end-to-end. To achieve so, one intuitive idea is to adapt the LLMs to multi-modal inputs by adding some additional trainable parameters and finetuning them on multi-modal data. For example, Flamingos (Alayrac et al., 2022) is one of the early works to explore this idea. Firstly, it takes a vision encoder (like NFNet (Brock et al., 2021) in their original version, or recent CLIP ViT (Radford et al., 2021)) to extract visual embeddings. Then, it applies multi-layer cross-attention to fuse the multi-modal inputs for the final prediction. Recent works directly concatenate vision embeddings to the inputs of LLMs and finetune LLMs end-to-end. To do so, they usually add an additional projection layer to map the vision embeddings to the same dimension as the language embeddings, and then directly feed them into LLMs for further training. Different methods may take different training strategies. See Appendix A for more.

5 Conclusions

In this paper, we present Lynx, a multi-modal large language model that can take as input images/videos and responses with open-ended natural languages. Through extensive empirical study, we show that our model outperforms other existing open-source models both in multi-modal understanding accuracy and instruction-following capabilities. We also explore different factors that can affect the performance of a multi-modal large language model and conclude that: 1) the generative pretraining is much more sensitive to the quality of training data than previous methods such as contrastive training; 2) the abilities of instruction following are closely related to the number of different tasks and prompts used for training; 3) improving both the multi-modal understanding accuracy and instruction-following capabilities is important for multi-modal large language models; 4) by first prompting the model to describe the image, the instruction-following abilities can be improved. For future work, it is promising to scale up the model to a larger size (e.g. 30B and 65B LLaMA (Touvron et al., 2023)). Moreover, a high-quality multi-modal dataset with diverse instructions is also needed to train such models.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. arXiv preprint arXiv:2204.01691.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35:23716–23736.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision, pages 2425–2433.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. arXiv preprint arXiv:1607.06450.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. Advances in Neural Information Processing Systems, 35:32897–32912.

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

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

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. 2023. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. arXiv preprint arXiv:2308.06595.

Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. 2021. High-performance large-scale image recognition without normalization. In International Conference on Machine Learning, pages 1059–1071. PMLR.

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

Erhan Bulbul, Aydin Cetin, and Ibrahim Alper Dogru. 2018. Human activity recognition using smartphones. In 2018 2nd international symposium on multidisciplinary studies and innovative technologies (ismsit), pages 1–6. IEEE.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3558–3568.

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

Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. 2023. Videollm: Modeling video sequence with large language models. arXiv preprint arXiv:2305.13292.

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

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

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325.

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. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311.

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

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural Networks, 107:3–11.

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

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2023. Datacomp: In search of the next generation of multimodal datasets. arXiv preprint arXiv:2304.14108.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023a. Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010.

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

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. 2017. The" something something" video database for learning and evaluating visual common sense. In Proceedings of the IEEE international conference on computer vision, pages 5842–5850.

Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In International workshop ontoImage, volume 2.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. An empirical analysis of compute-optimal large language model training. Advances in Neural Information Processing Systems, 35:30016–30030.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural instructions: Tuning language models with (almost) no human labor.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2023a. Audiogpt: Understanding and generating speech, music, sound, and talking head. arXiv preprint arXiv:2304.12995.

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

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. 2022. Inner monologue: Embodied reasoning through planning with language models. arXiv preprint arXiv:2207.05608.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6700–6709.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Dániel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. arXiv preprint arXiv:2212.12017.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In International conference on machine learning, pages 4651–4664. PMLR.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. 2022. Vima: General robot manipulation with multimodal prompts. arXiv preprint arXiv:2210.03094.

Chuhao Jin, Wenhui Tan, Jiange Yang, Bei Liu, Ruihua Song, Limin Wang, and Jianlong Fu. 2023. Alphablock: Embodied finetuning for vision-language reasoning in robot manipulation. arXiv preprint arXiv:2305.18898.

Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In ICCV.

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

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision, 123:32–73.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726.

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

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

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems, 34:9694–9705.

Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person search with natural language description. arXiv preprint arXiv:1702.05729.

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

Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Yang Yang, Qingyun Li, Jiashuo Yu, et al. 2023b. Interchat: Solving vision-centric tasks by interacting with chatbots beyond language. arXiv preprint arXiv:2305.05662.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, pages 3195–3204.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In 2019 international conference on document analysis and recognition (ICDAR), pages 947–952. IEEE.

Huu Nguyen, Sameer Suri, Ken Tsui, Shahules786, Together.xyz, and Christoph Schuhmann. 2023. The oig small.

OpenAI. 2023a. Gpt-4 technical report. arXiv, page 2303.08774.

OpenAI. 2023b. Gpt-4v(ision system card).

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. Advances in neural information processing systems, 24.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35:27730–27744.

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

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, pages 647–664. Springer.

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

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using amazon's mechanical turk. In Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk, pages 139–147.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. 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, pages 3505–3506.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. arXiv preprint arXiv:2210.08402.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2556–2565.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. arXiv preprint arXiv:2303.17580.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, pages 742–758. Springer.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8317–8326.

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

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389.

Wei Ren Tan, Chee Seng Chan, Hernan Aguirre, and Kiyoshi Tanaka. 2019. Improved artgan for conditional synthesis of natural image and artwork. IEEE Transactions on Image Processing, 28(1):394–409.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2023. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. arXiv preprint arXiv:2305.11175.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. arXiv preprint arXiv:2212.10560.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671.

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

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. arXiv preprint arXiv:1901.06706.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. arXiv preprint arXiv:2304.01196.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In ACM Multimedia.

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

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. arXiv preprint arXiv:2303.11381.

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

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2:67–78.

Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. arXiv preprint arXiv:2111.08276.

Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. 2022. X^2-vlm: All-in-one pre-trained model for vision-language tasks. arXiv preprint arXiv:2211.12402.

Hanbo Zhang, Yuchen Mo, Jie Xu, Qingyi Si, and Tao Kong. 2023. Invig: Interactive visual-language disambiguation with 21k human-to-human dialogues. https://github.com/ZhangHanbo/invig-dataset.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5014–5022.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. 2023. Learning video representations from large language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6586–6597.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence, 40(6):1452–1464.

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

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4995–5004.

A Background

BLIP2 (Li et al., 2022) designs a Q-Former, which is the only trainable part, to align the dimensions of vision and language tokens. PaLM-E (Driess et al., 2023), which is built upon PaLM (Chowdhery et al., 2022), is trained totally end-to-end with no fixed layers using a mix of multi-modal datasets including WebLI 10B dataset (Chen et al., 2022). Mini-GPT4 (Zhu et al., 2023) freezes all weights of the vision encoder and the LLM while only finetuning the weights of the projection layer. LLAVA (Liu et al., 2023a) fixes the vision encoder while keeping the LLMs trainable during the instruction finetuning stage. mPLUG-owl (Ye et al., 2023) tunes the vision encoder and keeps LLMs fixed to align the vision and language embeddings in the first stage while further tuning the LLMs and keeping the vision encoder fixed in the second instruction-finetuning stage. KOSMOS-1 (Huang et al., 2023b) does not rely on any pretrained LLMs and is trained from scratch on large amounts of mixed data including image-text pairs (COYO700M (Byeon et al., 2022), LAION2B (Schuhmann et al., 2022), etc.), text corpora (Common Crawl, the Pile (Gao et al., 2020), etc.), and interleaved image-text data. These models are all powerful and show promising results to develop multi-modal large language models.

B Evaluation Protocols

The evaluation of GPT4-style generative language models is challenging because the quality of natural languages is inherently subjective and highly depends on specific cases. Existing models like PaLM-E (Driess et al., 2023), PaLI (Chen et al., 2022), BLIP2 (Li et al., 2023b), or InstructBLIP (Dai et al., 2023) turn to the evaluation on visual-language benchmarks like image caption (Chen et al., 2015) or visual question answering (Antol et al., 2015), i.e., fine-tuning multi-modal LLMs on a single downstream task on which the evaluation is conducted. Nevertheless, though it may achieve better performance, over-finetuning on such benchmarks will damage the generation ability of large language models, which conflicts with the primary motivation to use large language models. Moreover, such benchmarks, especially the (semi-)automatically generated ones like TDIUC (Kafle and Kanan, 2017), always contain a high ratio of easy or noisy examples, making them less suitable. On the contrary, other methods like MiniGPT4 (Zhu et al., 2023) or LLaVA (Liu et al., 2023a) only

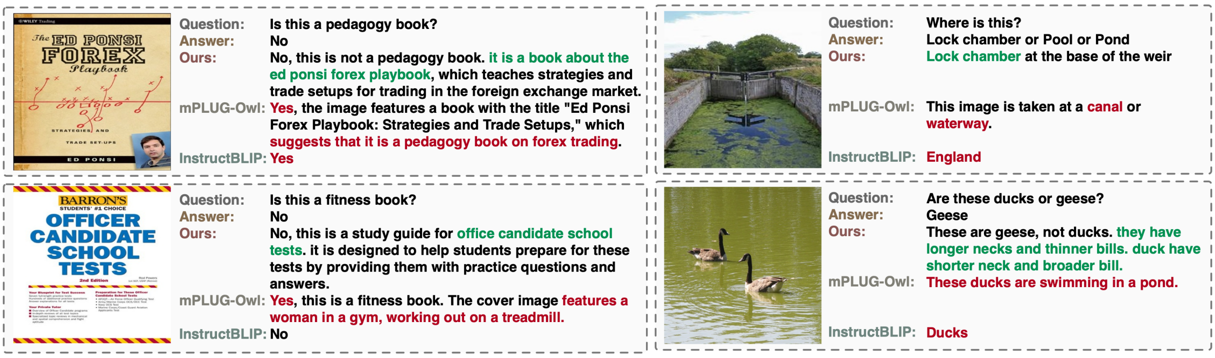


Figure 6: Qualitative results on our Open-VQA benchmark of different models. We choose InstructBLIP and mPLUG-Owl because they perform best on the Open-VQA benchmark and OwlEval benchmark in all baseline algorithms.

showcase their performance in some challenging yet practical scenarios without quantitative results due to the lack of quantitative benchmarks for such generative multi-modal language models. Therefore, in this section, we propose to evaluate the GPT4-style models in the following two aspects:

- A cleaned subset of visual-language benchmark, which should be challenging and compatible with generative models, with prompted GPT4 to get the quantitative results.

- An open-world challenging yet practical test set to evaluate the performance on realistic scenarios where GPT4-style models are needed, with humans to evaluate the user experience.

To do so, we manually collect an Open-VQA test set consisting of 450 samples with image or video input, which contains diverse questions on objects, OCR, counting, reasoning, action recognition, chronological ordering, etc., from VQA 2.0 (Antol et al., 2015), OCRVQA (Mishra et al., 2019), Place365 (Zhou et al., 2017), MSVD (Chen and Dolan, 2011), MSRVTT (Xu et al., 2016), and Something-Something-V2 (SthV2) (Goyal et al., 2017). Though Place365 is a classification task and SthV2 is a video captioning task, we write proper prompts to make them both VQA tasks. Besides, we carefully examine the data and modify the questions and ground-truth answers if necessary to make them reliably correct and challenging enough to be a benchmark for GPT4-style models. Randomly sampled examples are given in Fig. 8(a). Different from the traditional VQA benchmark, Open-VQA supports open-ended answers. To achieve so, we prompt GPT4 to make it the referee, which achieves a consistency of more than 95% compared with humans[2]. The prompt for GPT4 used in this phase is as follows:

Given the question "QUESTION", does the answer "PREDICTION" imply the answer "GROUND_TRUTH"? Answer with Yes or No.

Moreover, general-purpose language generation with image inputs is also important to multi-modal LLMs. Therefore, we also adopt the OwlEval test set proposed by mPLUG-owl (Ye et al., 2023), which contains 82 questions based on 50 images, where 21 from MiniGPT-4 (Zhu et al., 2023), 13 from MM-REACT (Yang et al., 2023), 9 from BLIP2 (Li et al., 2023b), 3 from GPT4 (OpenAI, 2023a), and 4 collected by mPLUG-owl itself. The test set includes diversified and practical cases such as dense image captioning, dialogue writing, story writing, poem writing, teaching, programming, etc.

We give some examples in Fig.8(b). However, OwlEval is proposed together with mPLUG-owl. Hence, directly using it as the benchmark is possibly unfair to other models. To make the comparison fair, we pad each image in the OwlEval with 8 pixels as shown in Fig.8(b) before feeding them into the models. We recruit human annotators to evaluate the performance. Scores range from 1 to 5. If two models are considered to be equally good or bad, they will have the same score. For each data, the annotator will assign a score for each model. We only allow at most 2 models that are equally good or bad, and for each annotator, the total number of ties should be no more than 10 for the whole set. During the evaluation, the correctness has the highest priority, then should be the richness of the generated content.

[2]We evaluate the consistency on 100 samples from a randomly selected subset with our model.

Instruction

Can you give me a detailed introduction to this painting, and also the artist?

Ours:

the painting "the persistence of memory" by salvador dali is a surrealist masterpiece that showcases the artist's unique and imaginative style. the painting was created in 1931 and is considered one of dali's most famous works. **the painting depicts a dreamlike landscape with a sleeping elephant, a tree with clocks hanging off of it, and a rocky beach with a mountainous backdrop**. the overall feeling of the image is one of dreamlike, surrealism. the clocks and the sleeping animal give the image a sense of time being distorted or out of order. the surreal nature of the scene and the dreamlike quality of the elephants body position suggest that the painting is a representation of the subconscious mind and the distortion of reality that can occur in dreams.
salvador dalí was a spaniard who was known for his unique and innovative artistic style. he was a pioneer of surrealistic art and is widely regarded as one of the most influential artists of the 20th century. his works often featured dreamlike landscapes, distorted figures,

mPLUG-Owl:

Sure, I can provide a detailed introduction to the painting and the artist.
The painting is a surrealist masterpiece by Salvador Dali, titled "The Persistence of Memory." It was created in 1931 and is one of Dali's most iconic works. **The painting depicts a woman sitting on a wooden floor, surrounded by clocks.** The clocks are arranged in various positions, some closer to the viewer and others further away, creating an intriguing and visually arresting composition. The overall effect of the painting is to evoke a sense of time and the passage of time, as the clocks seem to tick away, reminding us of the inevitable progression of time.
Salvador Dali (1904-1989) was a Spanish artist known for his innovative and imaginative approach to art. He was a leading figure in the surrealism movement, and his works are characterized by dreamlike landscapes, bizarre juxtapositions, and meticulous attention to detail. Dali is considered one of the most influential artists of the 20th century and his work continues to insp

InstructBLIP:

yes

Instruction

Explain why this meme is funny.

Ours:

the meme is funny because it depicts a cute little dog laying down on the floor with the caption "monday, just monday." **the contrast between the adorable dog and the negative connotation of monday makes it humorous. monday is often viewed as a less desirable day of the week due to the perception that it is associated with work, responsibilities, and a lack of leisure time.** the dog's relaxed and peaceful pose, combined with the playful caption, creates a lighthearted and amusing scene.
this meme can be relatable for those who have a love-hate relationship with mondays, as it pokes fun at the negative aspects of the day while also acknowledging the inevitability of its arrival. it can also serve as a reminder to take a break and enjoy the simple pleasures in life, even on a monday. overall, the combination of the cute dog and humorous caption make this meme entertaining and relatable to a wide range of audiences.

mPLUG-Owl:

This meme is funny because it portrays a cute and sleepy French bulldog in a relaxed state, lying down on a white bed with its eyes closed. **The combination of the dog's adorable appearance and its peaceful resting position creates a humorous and entertaining image**.
The fact that the dog is resting its head on the bed also adds to the amusement, as it suggests a sense of comfort and contentment for the canine. The presence of a bottle nearby and a clock on the wall further enhances the meme's comedic value, as these objects might not typically be associated with a sleeping dog. Overall, the image captures the essence of a relaxing and enjoyable moment for both the dog and the viewer, making it a funny and memorable meme.

InstructBLIP:

dog

Instruction

Can you write me a master rap song that rhymes very well based on this image?

Ours:

verse 1:
i'm an old lady with a green jacket on,
wearing sunglasses and headphones,
dancing to the beat with a smile on my face.
chorus:
let's groove to the rhythm and the beat,
with our headphones on, we're the coolest crew,
we're here to party and have a good time,
so let's dance and have some fun,
it's all about the music and the vibe,
and we'll keep it going until the night is through.
verse 2:
my walk is a little shaky, but i don't care,
my moves may not be the best, but it's okay,
i just wanna have a blast, and let the music take me away,
in the end, it doesn't matter,
as long as i'm having fun, that's what i'll say,
let the good times roll and the music keep on playing,
that's the way to live, and i'd do it all over again.
end chorus:

mPLUG-Owl:

Sure, here's a master rap song that rhymes very well based on the image of the energetic and enthusiastic woman wearing a green jacket, sunglasses, and **a hat**, sitting at a table, **possibly playing cards or participating in a game**:
Verse 1:
I'm the queen of the table, decked out in green,
With my sunnies and my hat, I'm ready to play.
Chorus:
We're having a blast, everyone's having a ball,
At the table with the queen, we're all in the game.

InstructBLIP:

no

Figure 7: Qualitative results on OwlEval benchmark of different models. We choose InstructBLIP and mPLUG-Owl because they perform best on the Open-VQA benchmark and OwlEval benchmark in all baseline algorithms.

Figure 8: Examples of our test set. (a) Open-VQA benchmark to validate the accuracy of visual understanding; (b) OwlEval to evaluate the quality of language generation.

Finally, we also compare our method with others on the newly proposed MME benchmark (Fu et al., 2023) and VisIT-Bench(Bitton et al., 2023).

C Discussions and Limitations

C.1 Findings and Takeaways

Multi-modal LLMs are not as instruction-following as LLMs. In our experiments, we find that current multi-modal LLMs are not as good at the instruction following as language models. For example, InstructBLIP (Dai et al., 2023) tends to generate short responses regardless of the input instructions, while other models tend to generate long sentences without considering the instruction like "Give a short answer" or "Answer in one word". We assume that this is from the lacking of high-quality and diversified multi-modal instruction data.

The quality of training data is critical to model performance. As concluded in Section 3.2, based on the experimentation on different pretraining data, we find that a small number of high-quality data with fluent texts can perform even slightly better than the large-scale noisy datasets. We attribute this to the difference between generative pretraining and contrastive pretraining, since generative pretraining is directly learning the conditional distribution of words but not the similarity between texts and images. Therefore, to train a high-performance multi-modal LLM, despite the quantity of data, it is crucial to prepare a high-quality dataset that satisfies: 1) it includes high-quality and fluent texts; 2) it aligns the texts and images well.

Tasks and prompts are crucial for zero-shot abilities. As shown in Section 3.2, diversified prompts have a great impact on the final performance. The essential observation behind this is that the zero-shot generality of multi-modal language models depends on the diversity of tasks involved during training. The model can generalize to more and more unseen instructions as it sees more and more types of tasks. This accords with the observation in text-only models (Radford et al., 2021).

Balancing the correctness and language generation ability is important. In our experiments, we find that if the model is under-trained on downstream tasks such as VQA, it will suffer from the problem of hallucination and keep making mistakes. While if the model is over-trained on downstream tasks, it will not be able to follow the user's instructions to generate long answers. Therefore, it would be important to carefully balance the training data to train it so as to correctly read images and videos while keeping its generation ability.

C.2 Limitations

Evaluation It is hard to evaluate a multi-modal large language model since its evaluation is essentially different from traditional visual-language models. Though we take the first step to quantitatively evaluate both the multi-modal understanding accuracy and language generation ability, it is still an open problem: *how can we establish a comprehensive and automatic benchmark to evaluate existing multi-modal large language models?*

Training Data Though we have successfully collected and cleaned a mixed dataset to train our Lynx, we still put a lot of effort to balance different abilities (e.g. correctness and language generation, long and short answers). Moreover, there are still no available image-text datasets that contain long texts which are ideal for pretraining. Besides, restricted by the computational resources that we can use, we do not conduct extensive experiments to find the optimal data combination strategy (e.g. sampling ratios, tasks, and prompts), which has been left for future work.

Multi-lingual Our model is built upon LLaMA (Touvron et al., 2023), which is mainly trained on English corpus. Therefore, our model is not that good at multi-lingual responses. Though it can understand and sometimes output other languages (like shown in Figure 15), it is still unexplored how to build a high-performance multi-lingual and multi-modal large language model.

Safety Currently, we do not conduct safety checks and restrict the outputs of our model. Therefore, the model may output contents that are not appropriate and even toxic, depending on and restricted by the data used for training. The authors do not support the use of harmful language generation using our codes and models, like any usage on ethical, political, and racism issues.

D Experimental Details

D.1 Ablation Study

D.2 Training Details

The model comprises approximately 8B parameters, of which around 1B are trainable. We use the DeepSpeed (Rasley et al., 2020) to accelerate training, and set the BFloat16 as the default model precision. We report the detailed model training hyperparameters in Table 8.

D.3 Hyper-parameters for Generation

For MiniGPT4 (Zhu et al., 2023), we generated the response with its default settings. Similarly, for mPLUG-owl (Ye et al., 2023), we follow the default parameters presented at http://vlarena.opengvlab.com/. Detailed settings can be found in 9 for different tasks.

D.4 MME Performance

E Training Data

F Case Study

	images									video		
	OCR	Counting	Reasoning	Place	Color	Spatial	Action	Others	Overall	Action (Y/N)	Others	Overall
w/ LLaMA	33/53	18/37	19/31	17/22	22/30	10/15	17/20	78/94	70.86	65/109	25/40	60.81
w/o diverse prompts	33/53	22/37	23/31	20/22	21/30	12/15	17/20	80/94	75.50	62/109	26/40	59.46
w/ large-scale noisy data	33/53	20/37	28/31	17/22	17/30	10/15	16/20	79/94	72.85	63/109	26/40	60.14
w/o high-resolution	30/53	20/37	26/31	15/22	25/30	8/15	19/20	79/94	73.51	66/109	26/40	62.16
Ours	36/53	25/37	26/31	17/22	21/30	9/15	17/20	79/94	76.16	69/109	29/40	66.22

Table 7: Quantitative evaluation of different ablation models on the Open-VQA benchmark.

hyperparameters	Pretraining	Resolution Enhancement	Instruction Finetuning
Env	A100*32	A100*32	A100*24
Training steps	100,000	10,000	20,000
Warmup steps rate	0.05	0.05	0.05
Warmup lr end	1e-5	1e-6	2e-6
Optimizer	AdamW	AdamW	AdamW
Learning rate	1e-4	1e-5	2e-5
Learning rate decay	linear	linear	linear
Adam ϵ	1e-8	1e-8	1e-8
Adam β	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)
Weight decay	0.01	0.01	0.01
Training Time	3 days	8 hours	16 hours

Table 8: Training hyperparameters. Some parameters not use learning rate decay schedule.

	max new tokens	beam size	top-p	top-k	length penalty	no repeat ngram	do sample
Image Description	64	5	1.0	1	-2.0	2	False
Open-VQA image	64	5	1.0	1	-2.0	2	False
Video Description*	128	1	0.9	3	1.0	3	True
Open-VQA video	128	3	1.0	1	-1.0	3	False
OwlEval Description*	128	1	0.9	3	1.0	3	True
OwlEval	256	3	0.9	3	1.0	3	True
VisIT-Bench Description*	128	1	0.9	3	1.0	3	True
VisIT-Bench	256	3	0.9	3	1.0	3	True
MME Benchmark	64	5	1.0	1	-2.0	2	False
demo(ours)	256	3	0.9	3	1.0	3	True

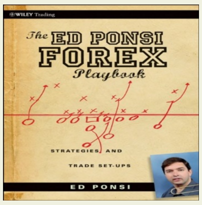
* The hyperparameters to generate the 0th-round detailed description, if applicable.

Table 9: Hyper-parameters for visual question answering evaluation and general-purpose natural language generation with vision inputs respectively. We set hyper-parameters to encourage short response generation for the Open-VQA benchmark and MME Benchmark.

	BLIP2 (Li et al., 2023b)	Instrunct-BLIP (Dai et al., 2023)	LLaMA-Adapter V2 (Gao et al., 2023b)	mPLUG Owl (Ye et al., 2023)	MiniGPT4 (Zhu et al., 2023)	LLaVA (Liu et al., 2023a)	Ours
Existence	160.00	185.00	120.00	120.00	115.00	50.00	**195.00**
Count	135.00	143.33	50.00	50.00	123.33	50.00	**151.67**
Position	73.33	66.67	48.33	50.00	81.67	50.00	**90.00**
Color	148.33	153.33	75.00	55.00	110.00	55.00	**170.00**
Poster	**141.84**	123.81	99.66	136.05	55.78	50.00	124.83
Celebrity	105.59	101.18	86.18	100.29	65.29	48.82	**118.24**
Scene	145.25	153.00	148.50	135.50	95.75	50.00	**164.50**
Landmark	138.00	79.75	150.25	159.25	69.00	50.00	**162.00**
Artwork	**136.50**	134.25	69.75	96.25	55.75	49.00	119.50
OCR	110.00	72.50	**125.00**	65.00	95.00	50.00	77.50
Perception	1293.84	1212.82	972.67	967.35	866.58	502.82	**1373.23**
Commonsense	110.00	**129.29**	81.43	78.57	72.14	57.14	110.71
Numerical	40.00	40.00	**62.50**	60.00	55.00	50.00	17.50
Text Translation	65.00	65.00	50.00	**80.00**	55.00	57.50	42.50
Code Reasoning	75.00	57.50	55.00	57.50	**110.00**	50.00	45.00
Cognition	290.00	291.79	248.93	276.07	**292.14**	214.64	215.71

Table 10: Comparison of existing open-sourced multi-modal LLMs on MME benchmark (Fu et al., 2023).

Question: Where is this?
Answer: Living room or waiting room
Vicuna-based: living room
LLaMA-based: the image shows a living room filled with furniture. there are two couches, two chairs, a coffee table and an ottoman in the room. the walls are painted white and there is a large mirror hanging on the wall. a vase with flowers is sitting on a table in front of the



Question: Is this a pedagogy book?
Answer: no
Vicuna-based: no, this is not a pedagogy book. it is a book about the ed ponsi forex playbook, which teaches strategies and trade setups for trading in the foreign exchange market.
LLaMA-based: the ed ponsi forex playbook: strategies and trade setups (wiley trading)

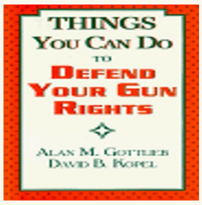
Question: Are these ducks or geese?
Answer: geese
Vicuna-based: these are geese, not ducks. they have longer necks and thinner bills. duck have shorter neck and broader bill.
LLaMA-based: the ducks or geese are swimming in a body of water. the water appears to be murky and green. there is another duck or goose in the image, but it is not as close to the camera as the other one. both of the birds have their heads tucked under their wings

(a) Vicuna-based model versus LLaMA-based model

Question: Where is this?
Answer: lock chamber or pool or pond
w/o COYO700M & DataComp1B: lock chamber at the base of the weir
w/ COYO700M & DataComp1B: pond "this is not the place to be if you're claustrophobic."

Question: How many people are in the water?
Answer: 0
w/o COYO700M & DataComp1B: 0 people in the water
w/ COYO700M & DataComp1B: 0

how many people are on the dock? 1
how many benches are there?
there are 2



Question: Is this book related to Law?
Answer: yes
w/o COYO700M & DataComp1B: yes, this book is a guide to defending your gun rights. it covers topics such as the 2nd amendment, the right to bear arms, self-defense, and the use of firearms for hunting and target shooting.
w/ COYO700M & DataComp1B: yes / yes / alan m. gottlieb

(b) w/o large-scale noisy data versus w/ large-scale noisy data

Instruction: Do you have anything to say to the young boy in the image?
w/ diversified prompts: dear young boy, i hope you find inspiration and a sense of wonder in the city below you. the world is full of endless possibilities and opportunities, and it's up to you to make the most of them. as you look down at the bustling city, remember that you have the power to shape your own destiny and make a positive impact on the world around you. don't be afraid to take risks, try new things, and pursue your dreams. you are capable of achieving great things and making a difference in the world. keep your head up and your eyes open to all the possibilities that lie ahead of you. best wishes,
[image description: a young boy is looking at a city below him. he is wearing a brown jacket and blue jeans. the boy is holding his hands in the air. he has a blue cap on his head. he looks like he is in awe of the city he is looking down at. the sky is dark and cloudy above him. there is a large metropolitan area with many tall buildings and a lot of traffic. the city is very busy. the child is looking up at the sky. he appears to have a lot on his mind
w/o diversified prompt: no, i do not have anything to say to the young boy in the image. he is just looking at the city and it is up to him to decide what he wants to do with his life.

Instruction: We design new mugs shown in the image. Can you help us write an advertisement?
w/ diversified prompts: sure! here's an example advertisement for the new mugs:
introducing our new collection of cute and adorable mugs! these mugs are perfect for your morning cup of coffee or tea. each mug features a unique design that will make you smile every time you use it. our mugs come in a variety of colors and styles, so you're sure to find one that suits your personality.
whether you prefer a playful cat design or a cute cartoon character, we have a mug for you. our cute mugs make great gifts for friends, family, or coworkers. they're perfect for birthdays, holidays, or just because!
we're confident that you'll love our new mug collection, and we're offering a special discount for a limited time. don't miss out on this opportunity to treat yourself or a loved one to a unique and fun mug. shop now and enjoy your favorite beverage in style!
w/o diversified prompt: we design new mugs shown in the image. can you help us write an advertisement?

Instruction: Would it be dangerous to fly a plane in the sky right now?
Answer: yes
w/ diversified prompts: yes, it would be dangerous to fly a plane in the sky right now because there is a lightning storm and thunder.
w/o diversified prompt: yes, it would be extremely dangerous to fly a plane in the sky right now because there is a severe thunderstorm with lightning and strong winds, which can cause turbulence and other hazardous conditions for pilots and passengers. additionally, the air traffic control system may be affected by … (more)

(c) w/ diversified prompts versus w/o diversified prompts

Figure 9: Ablation study cases on (a) Vicuna-based model versus LLaMA-based model; (b) w/o large-scale noisy data versus w/ large-scale noisy data; (c) w/ diversified prompts versus w/o diversified prompts.

	BLIP2 (Li et al., 2023b)	Instruct-BLIP (Dai et al., 2023)	LLaMA-Adapter V2 (Gao et al., 2023b)	mPLUG Owl (Ye et al., 2023)	MiniGPT4 (Zhu et al., 2023)	LLaVA (Liu et al., 2023a)	Ours
Existence	3	2	4	5	6	7	1
Count	3	2	5	6	4	7	1
Position	3	4	7	5	2	6	1
Color	3	2	5	6	4	7	1
Poster	1	4	5	2	6	7	3
Celebrity	2	3	5	4	6	7	1
Scene	4	2	3	5	6	7	1
Landmark	4	5	3	2	6	7	1
Artwork	1	2	5	4	6	7	3
OCR	2	5	1	6	3	7	4
Perception	2	3	4	5	6	7	1
Commonsense	3	1	4	5	6	7	2
Numerical	5	6	1	2	3	4	7
Text Translation	2	3	6	1	5	4	7
Code Reasoning	2	3	5	4	1	6	7
Cognition	3	2	5	4	1	7	6

Table 11: Rankings of existing open-sourced multi-modal LLMs on MME benchmark (Fu et al., 2023).

Figure 10: More cases on our Open-VQA image benchmark.

Dataset	Total size	Type	Pretrain	Pretain Ratio	Finetune	Finetune Ratio
BlipCapFilt (Li et al., 2022)	102.8M	Image-text Pair	✓	30.525%	✗	-
CC12M (Changpinyo et al., 2021)	8.3M	Image-text Pair	✓	2.465%	✗	-
CC3M (Sharma et al., 2018)	2.9M	Image-text Pair	✓	10.076%	✗	-
SBU (Ordonez et al., 2011)	859.7K	Image Caption	✓	2.987%	✗	-
TextCaps (Sidorov et al., 2020)	109.8K	Image Caption	✓	0.381%	✗	-
COCO Caption (Chen et al., 2015)	82.7K	Image Caption	✓	0.287%	✗	-
CUHK-PEDES (Li et al., 2017)	34.1K	Image Caption	✓	0.118%	✗	-
Flickr30k (Young et al., 2014)	29.8K	Image Caption	✓	0.104%	✗	-
Pexels 110k	26.2K	Image Caption	✓	0.091%	✗	-
LLaVA Caption (Liu et al., 2023a)	23.2K	Image Caption	✗	-	✓	0.945%
IAPR TC-12 (Grubinger et al., 2006)	20.0K	Image Caption	✓	0.069%	✗	-
Visual Genome Caption (Krishna et al., 2017)	19.6K	Image Caption	✗	-	✓	0.798%
MiniGPT4 IFT (Zhu et al., 2023)	3.4K	Image Caption	✗	-	✓	0.138%
Pascal Sentences (Rashtchian et al., 2010)	1.0K	Image Caption	✓	0.003%	✗	-
VGQA (Krishna et al., 2017)	1.4M	VQA	✓	8.711%	✓	10.880%
GQA (Hudson and Manning, 2019)	943.0K	VQA	✓	5.868%	✓	3.999%
OCRVQA (Mishra et al., 2019)	894.0K	VQA	✓	5.364%	✓	12.349%
VQAv2 (Antol et al., 2015)	443.8K	VQA	✓	2.761%	✓	3.449%
Visual7W (Zhu et al., 2016)	139.9K	VQA	✓	0.870%	✓	0.593%
VizWiz (Bigham et al., 2010)	20.5K	VQA	✓	0.128%	✓	0.087%
OKVQA (Marino et al., 2019)	9.0K	VQA	✓	0.056%	✓	0.038%
TDIUC (Kafle and Kanan, 2017)	705.4K	VQA	✓	4.389%	✗	-
WebSRC (Chen et al., 2021)	131.3K	VQA	✗	-	✓	1.814%
LLaVA Reasoning (Liu et al., 2023a)	76.6K	VQA	✗	-	✓	3.119%
TextVQA (Singh et al., 2019)	34.6K	VQA	✗	-	✓	0.478%
STVQA (Biten et al., 2019)	26.0K	VQA	✗	-	✓	0.359%
Places365 (Zhou et al., 2017)	1.8M	Classification	✓	10.921%	✓	5.000%
ImageNet1K (Deng et al., 2009)	1.3M	Classification	✓	7.887%	✗	-
SNLI-VE (Xie et al., 2019)	529.5K	Classification	✓	3.213%	✗	-
Visual7W Multi-choice (Zhu et al., 2016)	139.9K	Classification	✓	0.849%	✗	-
AirCrowdFood	100.3K	Classification	✓	0.609%	✗	-
NLVR2 (Suhr et al., 2018)	86.4K	Classification	✓	0.518%	✓	0.671%
WikiArt (Tan et al., 2019)	42.5K	Classification	✓	0.264%	✓	0.180%
HAR (Bulbul et al., 2018)	12.6K	Classification	✓	0.078%	✓	0.053%
TimeClassification	11.5K	Classification	✓	0.072%	✓	0.049%
HatefulMemes (Kiela et al., 2020)	8.5K	Classification	✓	0.026%	✗	-
MSR-VTT-QA (Xu et al., 2016, 2017)	158.6K	Video VQA	✗	-	✓	3.137%
VLN VQA (Pont-Tuset et al., 2020)	31.8K	Video VQA	✗	-	✓	0.629%
NeXT-QA (Xiao et al., 2021)	31.5K	Video VQA	✗	-	✓	0.623%
MSVD-QA (Chen and Dolan, 2011; Xu et al., 2017)	30.9K	Video VQA	✗	-	✓	0.611%
SthV2 (Goyal et al., 2017)	168.9K	Video Caption	✗	-	✓	5.000%
VLN Caption (Pont-Tuset et al., 2020)	17.6K	Video Caption	✗	-	✓	5.000%
LLaVA Instruction (Liu et al., 2023a)	361.4K	Dialog	✗	-	✓	5.845%
LLaVA Dialog (Liu et al., 2023a)	256.9K	Dialog	✗	-	✓	4.155%
InViG (Zhang et al., 2023)	49.9K	Dialog	✓	0.310%	✗	-
Flan V2 (Chung et al., 2022)		Text Instructions	✗	-	✓	15.000%
LAION OIG Small (Nguyen et al., 2023)	210.3	Text Instructions	✗	-	✓	3.884%
Alpaca GPT4 (Wang et al., 2022)	51.7	Text Instructions	✗	-	✓	0.955%
Unnatural Instruction (Honovich et al., 2022)	8.7	Text Instructions	✗	-	✓	0.161%
Baize (Xu et al., 2023)	601.1	Text Instructions	✗	-	✓	10.000%

Table 12: Training Data.

Dataset	Type	Prompt Example
BlipCapFilt	Image-text Pair	Describe the image briefly.
CC12M	Image-text Pair	Write a relevant description to pair with the image.
CC3M	Image-text Pair	Write a relevant description to pair with the image.
SBU	Image Caption	Describe the image.
TextCaps	Image Caption	Describe the image shortly by reading the texts.
COCO Caption	Image Caption	Describe the image briefly.
CUHK-PEDES	Image Caption	Describe the person in the image.
Flickr30k	Image Caption	Describe the image briefly.
Pexels 110k	Image Caption	Describe the image briefly.
LLaVA Caption	Image Caption	[INSTRUCTION][1]
IAPR TC-12	Image Caption	Describe the key elements in the image.
Visual Genome Caption	Image Caption	Describe the image in detail.
MiniGPT4 IFT	Image Caption	Describe the image in detail.
Pascal Sentences	Image Caption	Describe the image briefly.
VGQA	VQA	[QUESTION][2] Give a short answer.
GQA	VQA	[QUESTION] Give a short answer.
OCRVQA	VQA	[QUESTION] Give a short answer.
VQAv2	VQA	[QUESTION] Give a short answer.
Visual7W	VQA	[QUESTION] Give a short answer.
VizWiz	VQA	[QUESTION] Give a short answer.
OKVQA	VQA	[QUESTION] Give a short answer.
TDIUC	VQA	[QUESTION] Give a short answer.
WebSRC	VQA	Answer the question briefly by reading the webpage. [QUESTION]
LLaVA Reasoning	VQA	[QUESTION]
TextVQA	VQA	Answer the question shortly by reading the texts. [QUESTION]
STVQA	VQA	[QUESTION] Give a short answer.
Places365	Classification	Where is this? Answer with a place name.
ImageNet1K	Classification	What is in the image? Answer with its name.
SNLI-VE	Classification	Does the image semantically entail the following text? Text: [HYPOTHESIS][3] Options: 1. neutral 2. entailment 3. contradiction
Visual7W Multi-choice	Classification	Choose the correct answer. Question: [QUESTION] Options: [OPTIONS][4]
AirCrowdFood	Classification	What food is it?
NLVR2	Classification	Given the claim "[HYPOTHESIS]", is it True or False?
WikiArt	Classification	What artistic movement or style dose this art picture belong to? Answer with a style name.
HAR	Classification	What is the person doing? Answer shortly.
TimeClassification	Classification	What is the time now? Give a short answer.
HatefulMemes	Classification	Is "[MEME][5]" a hateful meme? Answer with Yes or No.
MSR-VTT-QA	Video VQA	[QUESTION] Give a short answer.
VLN VQA	Video VQA	[QUESTION] Give a short answer.
NeXT-QA	Video VQA	[QUESTION] Give a short answer.
MSVD-QA	Video VQA	[QUESTION] Give a short answer.
SthV2	Video Caption	Describe the video shortly.
VLN Caption	Video Caption	Describe the video in detail.
LLaVA Instruction	Dialog	[DIALOG][6]
LLaVA Dialog	Dialog	[DIALOG]
InViG	Dialog	[DIALOG]
Flan V2	Text Instructions	[INSTRUCTION]
LAION OIG Small	Text Instructions	[INSTRUCTION]
Alpaca GPT4	Text Instructions	[INSTRUCTION]
Unnatural Instruction	Text Instructions	[INSTRUCTION]
Baize	Text Instructions	[INSTRUCTION]

1. We directly use the original instructions if it is applicable.
2. [QUESTION] means the original question from the dataset. We do some post process, e.g., add the "?" if it is missing.
3. [HYPOTHESIS] means the statement in the SN-LIVE dataset.
4. [OPTIONS] means the options list provided by the dataset.
5. [MEME] is the meme included in the Hateful Memes dataset.
6. We split each dialog into rounds, and prompt the model to speak for the next round based on the dialog history.

Table 13: Prompt examples of different datasets.

Question: what did a lady cut up?

Ours: a lady cut up a white onion into pieces.

mPLUG-Owl: A lady cut up a loaf of bread.

InstructBLIP: apple

Question: Is the cloth removed from the keys?

Ours: no, the cloth is not removed from the keys. the key is placed on top of the sari, which is laying on the table. the cloth on the sario is blue, and there are some threads on the blue cloth. the threads are on the left side of the key. the blue thread is on the right side of key. there is a silver needle on the key and the needle is on left side. the silver needles has a blue thread.

mPLUG-Owl: A lady cut up a loaf of bread.

InstructBLIP: yes

Question: Is the cloth removed from the keys?

Ours: the woman is applying makeup to her eyes. she is using a black pencil and a black eyeliner to define and shape her eyebrows. she has dark brown eyeshadow and long, dark brown lashes. she also has some black dots on her eylids.
additional information:
- the woman's eyes are surrounded by a darker brown area, which could be a shadow or a contour.
- there is a youtube video with the same name as the image, which provides a step-by-step guide on how to apply makeup

mPLUG-Owl: The woman is sitting in front of a mirror, applying makeup to her face.

InstructBLIP: eyebrow

Figure 11: More cases on our Open-VQA video benchmark.

Instruction

Can you write me a beautiful poem about this image?

Ours:

as the sun sets in the sky,
a man and his dog sit on the grass,
their bond unbreakable, their hearts full of grace.
the dog's ears up, its mouth open wide,
as the man whispers sweet nothings in its ear,
his voice soothing, his heart full of care.
the golden light of the setting sun,
casting a warm glow on the scene,
illuminating the love between the man and the dog,
two souls united in a bond that can never be broken.
in this moment, time stands still,
and the world around them fades away,
leaving just the two of them,
basking in the warmth of the sunset,
enjoying each other's company,
with no worries, no cares,
only the love that binds them together,
forever entwined in a heartfelt embrace.
a beautiful moment captured forever,
in a photograph that tells a story,
of love, companionship, and the power of the human-animal bond.
may this image inspire us to cherish the ones we love,

mPLUG-Owl:

Sure, here's a beautiful poem about the image of a man and his dog sitting on a hillside, enjoying the sunset:

Sunset on the hillside

As the day draws to a close,

Man and dog sit side by side

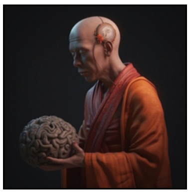
Enjoying each other's company.

The warmth of the sun's rays

Illuminates their faces,

InstructBLIP:

no



Instruction

I designed a character for my game as shown in the image. Can you write me a background story for him?

Ours:

sure! here's a background story for your character:
your character is a wise and enlightened monk who has dedicated his life to the pursuit of knowledge and spiritual understanding. he has spent countless years studying and meditating, seeking to unlock the secrets of the universe and uncover the mysteries of the mind.
during his journey, he has encountered many obstacles and challenges, but he has persevered, driven by his unwavering faith in the power of the human mind and the potential for self-discovery. he believes that every person has the potential to achieve greatness, if they only have the courage to pursue their dreams and aspirations.
as a result of his extensive studies, your character has developed a deep understanding of neuroscience and the workings of the brain. he is particularly interested in exploring the connection between the brain and consciousness, and the role that the mind plays in shaping our thoughts, emotions, and behaviors. he sees the brain as a powerful tool that can be harnessed for the betterment of humanity, and he is eager to share his knowledge and insights with others.

mPLUG-Owl:

Sure, I can provide a background story for the character in the image, who is dressed in a red robe and holding a sphere or orb in his hand.
The character is a monk or spiritual practitioner who has dedicated his life to studying and understanding the mysteries of the universe. He has traveled far and wide in search of knowledge and enlightenment, seeking to unlock the secrets of the cosmos and gain insight into the true nature of reality.
Over the course of his journey, the monk has encountered many challenges and obstacles, but his determination and perseverance have allowed him to overcome these obstacles and continue on his path. He now stands in the room, contemplating the orb he is holding, trying to decipher its meaning and understand its significance in the grand scheme of things.
As he ponders the orb, he reflects on his life's journey and the lessons he has learned along the way. He realizes that the orb represents a symbol of his quest for knowledge and understanding, a reminder of the challenges he has faced and the obstacles he has overcome. It serves as a constant source of inspiration and motiv

InstructBLIP:

yes

Instruction

What do you think about the person's behavior?

Ours:

the person's behavior is reckless and dangerous. running across a busy street without paying attention to oncoming traffic can lead to accidents, injuries, or even fatalities. additionally, running with a bag on one's back increases the risk of falling or losing the bag, which could cause further problems.
it is important for pedestrians to be aware of their surroundings and follow traffic rules to ensure their safety and the safety of others. in this case, the person should have waited for a safe opportunity to cross the street, such as when the traffic lights indicate it is safe to do so, or when there is a designated pedestrian crossing.
furthermore, it is essential for drivers to be vigilant and cautious when approaching intersections and crosswalks. they should always be prepared to stop or slow down when necessary to avoid accidents. in conclusion, the behavior exhibited by the person in the video is unacceptable and should be discouraged.

mPLUG-Owl:

The person in the image appears to be crossing the street while the traffic is in motion. This behavior can be considered risky and potentially dangerous, as pedestrians are advised to cross streets only when it is safe to do so. Crossing the street during heavy traffic can increase the chances of getting hit by a vehicle, causing injuries or even fatalities.

In this case, the person's decision to cross the street might be due to urgency, distraction, or simply not paying attention to their surroundings. Regardless of the reason, it is essential for individuals to exercise caution when crossing streets, especially in busy urban areas where traffic can be unpredictable and fast-moving.

InstructBLIP:

dangerous

Instruction

Write a conversation between the two animals.

Ours:

cat: hey dog, what are you doing out here?
dog: oh, i'm just enjoying the snow. it's so peaceful and quiet.
cat: i don't blame you. i love playing in the snow too, but i can't today. i have an appointment with the vet later.
do you want to come with me?
dog (excitedly): sure thing! let's go!
(cat and dog start walking towards the vets office together)
veterinarian: hello, what can i help you with today?
(the vet examines the cat and dog)
vet: both of you seem to be in good health. is there anything else i can assist you with?
(dog and cat look at each other and smile)
cat (to vet): thank you, vet. we're just here for a check-up.
(vet nods and smiles back)
cat to dog: you're welcome, dog. it was nice meeting you.
dog to cat: likewise, cat. i hope to see you again soon!
(

mPLUG-Owl:

Cat: "Hey, dog! How are you doing today? It's a cold and snowy day, but we're making the best of it by riding on top of this huge pile of snow together. Aren't we having fun?"

Dog: "Yeah, it's been a while since we've had this much snow to play in. I'm glad we decided to spend some quality time together. The cold weather doesn't seem to bother us at all, does it? We're just enjoying each other's company and having a blast."

InstructBLIP:

cat is riding on the dog's back

Figure 12: More cases on our OwlEval benchmark.

Instruction

Explain why this image is funny.

Ours:

the image is funny because it is a cat dressed up in a cookie monster costume, sitting on a couch with a plate of cookies in front of it. the cat is holding a white plate in its paws and there are cookies on the blue plate on top of the table with the white plate. the table has a white cloth on top and there's a white tablecloth on the white table. the cookies are in the shape of smiling faces, which adds to the humor of the image. the caption "another wild saturday night" is also amusing, as it implies that the cat's behavior is unusual or out of the ordinary.

mPLUG-Owl:

This image is funny because it portrays a stuffed animal, possibly a cat, dressed in a blue sweatshirt and pajamas, sitting in a chair and enjoying a cookie. The juxtaposition of the stuffed cat's human-like attire and its playful behavior, combined with the presence of two people engaged in a conversation, creates a humorous and entertaining scene.

The image evokes a sense of whimsy and playfulness, as stuffed animals are typically associated with childhood and imaginative play. The fact that the stuffie cat is wearing pajama pants further emphasizes its humanlike qualities, making the scene even more amusing. The presence of the two people adds another layer of humor, as they might be discussing the stuffy cat'ss enjoyment of the cookie or simply engaging in a casual conversation while observing the scene. Overall, the image captures the essence of a fun and lighthearted moment, making it an enjoyable and memorable moment for viewers.

InstructBLIP:

cat in a onesie

Instruction

How many championship rings did the player on the left win in his career?

Ours:

kobe bryant, the player on the left, won five nba championships during his career with the los angeles lakers. he is widely regarded as one of the greatest basketball players of all time, known for his exceptional shooting skills, ball-handling abilities, and strong work ethic. he was a 5-time nba all-star, 18-time all-nba team member, and 12-time olympic gold medalist.

mPLUG-Owl:

The player on the left has won two championship rings in his career.

InstructBLIP:

5

Instruction

What is the spatial relation between the frisbee and the man?

Ours:

the spatial relation between the frisbee and the man is that the friscue is in the air above the man's head, and the dog is jumping over the man to catch it. the man, who is crouched down on the grass, is positioned underneath the dog as it jumps over his head. the photographer is in focus, while the subject of the photograph is not. the photo is in color. the grass is lush and green, the trees behind the dog are green and lush, the sky behind the trees is blue, and there are no other people in the photograph. the only people in this photo are the man in the red shirt and the black and white dog.

mPLUG-Owl:

The frisbee is being thrown by the man, and the dog is leaping in the air to catch it.

InstructBLIP:

in front of him

Instruction

What is unusual about this image

Ours:

the unusual aspect of this image is that a man is hanging laundry on a line outside of a car while it is in motion. this is not a common sight, as most people tend to hang their lawn or dry their clothes indoors or in a designated area outside their homes.

it is also noteworthy that the man is wearing a yellow shirt and brown shorts, which may indicate that he is a street performer or vendor. the presence of a taxi cab in front of him adds to the unusual nature of the scene, as taxi cabs are typically associated with urban transportation rather than outdoor activities.

in conclusion, the image captures an unusual moment where a man hangs his clothes outside a moving car, wearing yellow and brown clothing, and is surrounded by tall buildings and a busy street. this combination of elements creates an interesting and unconventional scene that stands out from everyday life in an urban environment.

mPLUG-Owl:

What makes this image unusual is the fact that the man is driving down a city street while carrying a laundry bag full of clothes. It is not common to see someone transporting laundered clothes while driving a car, as it can be inconvenient, time-consuming, and potentially hazardous. Instead, one would typically expect to see the clothes being transported in a separate vehicle, such as a truck or a dedicated delivery service.

The presence of multiple cars and trucks on the busy city street adds to the unusual nature of the scene, as the man's decision to transport the clothes in this manner might not be the most efficient or practical option. The image serves as a reminder of the challenges and trade-offs involved in urban transportation and the need to consider various factors, including time, convenience, and safety, when making transportation decisions.

InstructBLIP:

man ironing clothes on car

Instruction

Why this happens and how to fix it?

Ours:

this phenomenon occurs when the washing machine is not properly ventilated or when there is a blockage in the ventilation system. when the machine is running, the water and detergent are heated up, creating steam. if the steam is not allowed to escape properly, it can build up inside the machine, causing the bubbles and foam.
to fix this issue, you can try the following steps:
1. ensure proper ventilation: make sure that there are no blockages in the vents or ducts that could prevent the steam from escaping properly. you may need to clear any dust or debris that may have accumulated inside the vent.
2. check the drain hose: ensure that the hose is not clogged or obstructed, as this can prevent the water from draining properly.
3. adjust the water level: if the foam is caused by too much water, you may want to reduce the amount of water used in the wash cycle.
4. use a fabric softener: if you are experiencing foam due to hard water, try using a water softener. this can help reduce the buildup of minerals in

mPLUG-Owl:

This scenario of a large pile of foam or soap bubbles covering a kitchen floor, with a washing machine in the background, can occur for various reasons. One possible explanation is that the detergent being used in the washing machines is too concentrated, causing excessive bubbling during the wash cycle. This can lead to the formation of large amounts of soap foam, which can then accumulate on the floor.

To fix this issue, it is essential to use detergents that are specifically designed for high-efficiency washers, as they tend to produce less foam. Additionally, it's crucial to follow the manufacturer's instructions for the proper use of the dishwasher, including setting the appropriate water temperature, cycle length, and wash settings. This will help to minimize foam formation and ensure a more efficient and hygienic washing process.
In the meantime, it might be necessary to clean up the foam using a broom or vacuum cleaner to remove it from the floor and prevent any potential slipping hazards.

InstructBLIP:

overflow

Figure 13: More cases on our OwlEval benchmark.

Instruction 1: Do you know what breed the dog in the picture is?

Ours:	mPLUG-Owl:	InstructBLIP:
yes, the dog in the picture is a shiba inu. shiba is a japanese breed of dog that originated in the mountainous regions of japan. they are known for their intelligence, loyalty, and energetic personalities. shibas have a thick double coat that comes in a variety of colors, including black, brown, red, and cream.	yes, the dog in the picture is a shiba inu.	shiba inu

Instruction 2: What kind of mood do you think it's in?

Ours:	mPLUG-Owl:	InstructBLIP:
based on the image, the shiba inu appears to be in a happy and excited mood. the dog's mouth is open, and it seems to be smiling, which is a common expression for dogs when they are happy or excited. additionally, the dog is sitting at a dining table with a meal in front of it, which suggests that it is in a comfortable and relaxed environment.\n\nthe presence of food on the table also indicates that the dog may have recently eaten or is about to eat, which could contribute to its happy and contented expression. overall, the image conveys a sense of joy and contentment on the part of the dog.	i cannot definitively determine the dog's exact mood from the image, but it appears to be happy and content. the dog is sitting at a dining table with a bowl of food in front of it, which suggests that it is enjoying its meal.	happy

Instruction 1: describe the image in detail.

Ours:	mPLUG-Owl:	InstructBLIP:
a woman is laying in a bed with a purple comforter. she is wearing a pink nightgown. there is a rose in her hand. the words "wake me up when when i'm rich" are written on the wall behind her. the wall is white. there are white curtains on the window. the window is partially open. the woman has long brown hair. she has brown eyes. she also has a cell phone. the cell phone is in her right hand.	the image features a young woman lying in bed with her eyes closed. she is wearing a pink nightgown and appears to be sleeping peacefully. in the background, there is a clock on the wall, indicating the time. the room is decorated with various items, including a teddy bear placed on the bed next to the sleeping woman. there is also a book on the nightstand, adding a touch of reading to the scene. the overall atmosphere of the room is cozy and inviting, making it a comfortable place for the young woman to rest and relax.	person sleeping in bed with a phone in her hand

Instruction 2: Why is this picture interesting?

Ours:	mPLUG-Owl:	InstructBLIP:
this picture is interesting because it depicts a woman in a pink nightgown lying in bed, holding a rose, with the words \"wake me up when when i'm rich\" written on the wall behind her. the combination of the woman's pink attire, the rose in her hand, and the words on the bedroom wall creates a whimsical and romantic atmosphere.\n\nmoreover, the fact that the woman is holding a cell phone in her right hand adds a modern twist to the scene, making it more relatable to the present day. this image can evoke feelings of nostalgia and longing for a simpler time, as well as a sense of humor and irony. overall, it is a visually appealing and thought-provoking image that captures the viewer's attention.	this picture is interesting because it portrays a relaxing and peaceful moment in the life of the young woman. she is lying in bed with her eyes closed, enjoying the comfort of her pink nightgown and the presence of her teddy bear. the room is decorated with various items, such as the book on the nightstand and the clock on the wall, which adds a sense of warmth and familiarity to the scene. the combination of these elements creates an inviting and calming atmosphere, making the picture appealing and relatable to viewers.	1. It's a cartoon 2. The person is sleeping in bed with a phone in her hand

Instruction 1: Where is this?

Ours:	mPLUG-Owl:	InstructBLIP:
this image shows a city skyline at night through a window with sliding glass doors. the city is illuminated by streetlights and buildings are visible in the distance. there is a large body of water in the foreground, and the sky is clear and dark. the room has a white countertop and white cabinets, and a large window overlooks the city.	the image is taken in a modern, high-rise living room or office, with a beautiful view of a city skyline. the city skyline includes tall buildings and a nighttime atmosphere, creating a stunning visual experience for the occupant.	Where is this?

Instruction 2: Can you see a person at the desk?

Ours:	mPLUG-Owl:	InstructBLIP:
no, the image does not show a person at the desk. it only shows a view of the city skyline through a window with sliding glass doors. there is no person visible in the image. the focus is on the illuminated cityscape and the body of water in the foreground.	yes, there is a person sitting at the desk, likely taking in the amazing view of the city skyline while working in the high-rise living room or office.	No

Instruction 3: What is the color of the person's shirt?

Ours:	mPLUG-Owl:	InstructBLIP:
the color of the person's shirt is not visible in the image. the image only shows a view of the city skyline through a window with sliding glass doors. the focus is on the illuminated cityscape and the body of water in the foreground. the room has a white countertop and white cabinets, and a large window overlooks the city. it is important to note that the image does not provide enough visual information to determine the color or any other details about the person sitting at the desk. the main focus of the image is the city view through the window, and the room's design elements, such as the counter and cabinets. the presence of a person is not the primary focus of this image.	the person in the image is wearing a white shirt.	Question: what is the color of the person's pants?

Figure 14: Multi-turn dialog demonstrations.

Instruction

Describe the image in French.

Ours:

voilier de pêcheur dans l'ombre d'un soleil coucher de soleil sur la mer. le voileur est debout sur le pont de l'embarcation et regarde vers l'horizon. il y a une autre embarcation à l'arrière de la voileuse. la mer est calme et le ciel est rouge et orangé. le soleil est couché derrière la montagne en arrière-plan.

mPLUG-Owl:

lorsque la nuit descend et que le soleil se retire, un bateau de voile blanc et rouge navigue sur une mer tranquille, éclairant l'horizon avec un magnifique reflet de son saphir sombre. le soleil se reflète dans l'eau et illumine une partie du bateau, donnant l'illusion d'un navire de rêve flottant sur une mer dorée.

Instruction

Generate the German translation for this sentence: the gorilla is a large ape native to the forests of central and western africa. it is the largest of the great apes, and one of the largest primates. gorillas are herbivorous, spending most of their time on the ground foraging for food.

Ours:

der gorilla ist ein großer affe, die in den wäldern von zentral- und westeuropa heimat ist. er ist der größte der großen affen und einer der größten primaten. gorillas sind pflanzenfresser, die den größten teil ihres lebens in der erde verbringen, um nach nahrung zu suchen.

mPLUG-Owl:

"der intelligente, neugierige gorilla blickt aus den buscheln, als würde er die umgebung faszinieren."

german translation:
"der gorilla ist ein großer ape, die in der afrikanischen primatenregion lebt und sich in den wäldern aufhält. er ist mit abstand der größte der gattung der großen affen und eines der größten primaten. gorillas sind herbivor und meist auf dem boden fressend.

Instruction

describe the image in German

Ours:

es handelt sich um ein bild von einem pokerturnier, bei dem die spieler auf einem tisch sitzen und spielen. der tisch ist mit einem grünen tischbezug ausgestattet, auf dem sich die karten befinden. die spiele sind mit einem kartenspiel ausgerüstet, das auf dem tisch liegt.

mPLUG-Owl:

in this image, a group of men, all wearing white shirts, sit around a dining table, engaged in a game of cards. the players are seated in several different spots around the table, each one holding a deck or a hand of cards. there is a total of eight players present, with the main focus on the men in the foreground, who are sitting closer to the camera.
the dining table is positioned in the middle of the scene, occupying most of the available space. there are chairs placed around the table, with some closer to the foreground and others farther back. the overall ambiance of the scene suggests a relaxed, casual atmosphere as the players enjoy their time together.

Instruction

describe the image in German

Ours:

das ist die überschrift des windows in chinesisch

mPLUG-Owl:

the image shows a computer screen displaying the word "schule" in large font at the top of the page, with a small icon to indicate that the page is in chinese. a search box is placed below the word "schule" with the text "suche" visible within. there are also a few other words or sentences on the page below the search box, with the most prominent one being "schulveranstaltungen." below the words, there are three icons, each placed to the left of the text. the first icon has a red background and is shaped like a magnifying glass, indicating that it is a search icon. the second icon has a green background and is shaped like a book, likely a link to a website. the third icon is also shaped like a magnifying glass and is positioned below the second icon, suggesting that it might be another link within the page. the overall appearance of the screen indicates that the user is searching for information related to school events, perhaps related to the word "schule" or "schulveranstaltungen" on the page

Instruction

为什么这张图片这么奇特

Ours:

这张图片展示了一个飞雀，它在空中上穿着，其身体和翅部分都是白色，背部和尾部都是黑色。

mPLUG-Owl:

这张图片的特点是在一个狗跑过的场景中，飞翔的普通的小飞翔（swallow）。这是一狗在穿过室内的空气中的一绕的旋转的狗跑。它显示了一个小飞翔的即终種，拥有充斥的旋转的运动。在这种狗跑过的环节中，小飞翔的运动是相当有意义的，让狗的运动更具有动态性和生动性，增加狗在步道上的运动的有用性和感情。

Figure 15: Multi-lingual demonstrations.

Instruction: Are the men in the dark suits arresting the soldiers?

Ours:
no, the men in dark suits are not arresting the soldiers. they are standing in a hangar, looking at a missile launcher. the soldiers are wearing camouflage fatigues, and one of them is pointing at the missile. it is likely that the men are discussing the launcher and its capabilities.

mPLUG-Owl:
no, the men in the dark suits are not arresting the soldiers. they are standing in a room with the soldiers, who are wearing military uniforms. the men in suits appear to be discussing something related to the soldiers or the military equipment in the room.

InstructBLIP:
no

Instruction: Are the men in the dark suits arresting the soldiers? give me a short answer.

Ours:
no, the men in the dark suits are not arresting the soldiers. they are talking to the soldiers about something.

mPLUG-Owl:
no, the men in the dark suits are not arresting the soldiers. they are touring the facility and observing the soldiers and their equipment.

InstructBLIP:
no

Instruction: Are the men in the dark suits arresting the soldiers? Explain the reasons in detail.

Ours:
no, the men in dark suits are not arresting the soldiers. the men are standing in a hangar, and one of the soldiers is holding a missile launcher. the other soldiers are looking at the man with the launcher, and the man in the suit is pointing at the missile. it seems like they are discussing the missiles and the launchers. the soldiers are not being arrested, but rather, they are talking about their work.

mPLUG-Owl:
no, the men in the dark suits are not arresting the soldiers. they are simply observing the soldiers as they inspect a large gun. the presence of the dark-suited men suggests that they might be part of a security team, military officials, or representatives from a government agency. their purpose is likely to ensure the safety and security of the soldiers, the gun, and the surrounding area during the inspection process.

InstructBLIP:
no

Figure 16: Demonstration of instruction-following ability.