

MINIGPT-4: ENHANCING VISION-LANGUAGE UNDERSTANDING WITH ADVANCED LARGE LANGUAGE MODELS

Deyao Zhu*, Jun Chen*, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny

King Abdullah University of Science and Technology

{deyao.zhu, jun.chen, xiaoqian.shen,
xiang.li.1, mohamed.elhoseiny}@kaust.edu.sa

ABSTRACT

The recent GPT-4 has demonstrated extraordinary multi-modal abilities, such as directly generating websites from handwritten text and identifying humorous elements within images. These features are rarely observed in previous vision-language models. However, the technical details behind GPT-4 continue to remain undisclosed. **We believe that the enhanced multi-modal generation capabilities of GPT-4 stem from the utilization of sophisticated large language models (LLM).** To examine this phenomenon, we present MiniGPT-4, which aligns a frozen visual encoder with a frozen advanced LLM, Vicuna, using one projection layer. Our work, for the first time, uncovers that properly aligning the visual features with an advanced large language model can possess numerous advanced multi-modal abilities demonstrated by GPT-4, such as detailed image description generation and website creation from hand-drawn drafts. Furthermore, we also observe other emerging capabilities in MiniGPT-4, including writing stories and poems inspired by given images, teaching users how to cook based on food photos, and so on. In our experiment, we found that the model trained on short image caption pairs could produce unnatural language outputs (e.g., repetition and fragmentation). To address this problem, we curate a detailed image description dataset in the second stage to finetune the model, which consequently improves the model’s generation reliability and overall usability. Our code, pre-trained model, and collected dataset are available at <https://minigpt-4.github.io/>.

1 INTRODUCTION

In recent years, large language models (LLMs) have experienced rapid advancements (Ouyang et al., 2022; OpenAI, 2022; Brown et al., 2020; Scao et al., 2022a; Touvron et al., 2023; Chowdhery et al., 2022; Hoffmann et al., 2022). With exceptional language understanding capabilities, these models can perform a variety of intricate linguistic tasks in a zero-shot manner. Notably, GPT-4, a large-scale multimodal model, has been recently introduced and demonstrated several impressive capabilities of vision-language understanding and generation (OpenAI, 2023). For example, GPT-4 can produce detailed and accurate image descriptions, explain unusual visual phenomena, and even construct websites based on handwritten text instructions.

Although GPT-4 has exhibited remarkable vision language capabilities, the methods behind its exceptional abilities are still a mystery (OpenAI, 2023). We believe that these impressive skills may stem from the utilization of a more advanced large language model (LLM). LLMs have demonstrated various emergent abilities, as evidenced in GPT-3’s few-shot prompting setup (Brown et al., 2020) and the findings of Wei et al. (2022) (Wei et al., 2022). Such emergent properties are hard to find in smaller-scale models. It is conjectured that these emergent abilities are also applicable to multi-modal models, which could be the foundation of GPT-4’s impressive visual description capabilities.

To substantiate our hypothesis, we present a novel vision-language model named MiniGPT-4. It utilizes an advanced large language model (LLM), Vicuna (Chiang et al., 2023), which is built upon LLaMA (Touvron et al., 2023) and reported to achieve 90% of ChatGPT’s quality as per GPT-4’s

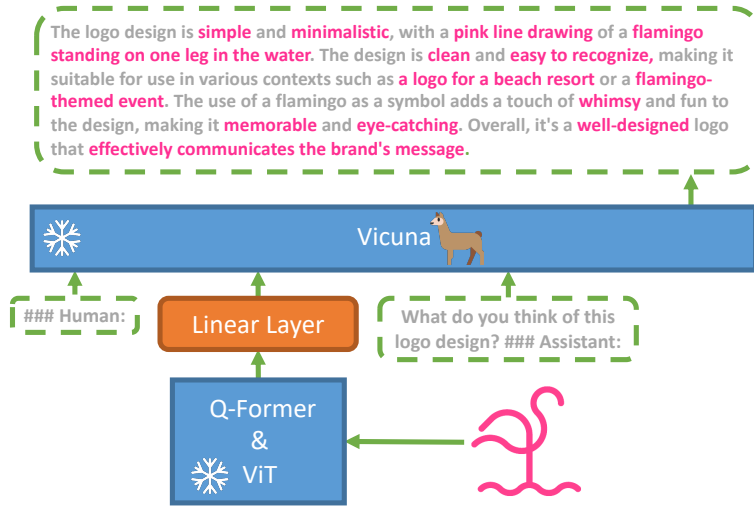


Figure 1: **The architecture of MiniGPT-4.** It consists of a vision encoder with a pretrained ViT and Q-Former, a single linear projection layer, and an advanced Vicuna large language model. MiniGPT-4 only requires training the linear projection layer to align the visual features with the Vicuna.

evaluation, as the language decoder. In terms of visual perception, we employ the same pretrained vision components of BLIP-2 (Li et al., 2023) that consists of a ViT-G/14 from EVA-CLIP (Fang et al., 2022) and a Q-Former network. MiniGPT-4 adds a single projection layer to align the encoded visual features with the Vicuna language model and freezes all the other vision and language components. MiniGPT-4 is initially trained for 20k steps using a batch size of 256 on 4 A100 GPUs, leveraging a combined image captioning dataset that includes images from LAION (Schuhmann et al., 2021), Conceptual Captions (Changpinyo et al., 2021; Sharma et al., 2018), and SBU (Ordonez et al., 2011) to align visual features with the Vicuna language model. Nevertheless, merely aligning visual features with the language model (LLM) is inadequate to ensure robust visual conversation capabilities, resembling that of a chatbot. The presence of underlying noise in raw image-text pairs can lead to subpar language outputs. Therefore, we collect another 3,500 detailed image description pairs to further fine-tune the model with a designed conversational template in order to improve the naturalness of the generated language and its usability.

In our experiments, we discovered that MiniGPT-4 possesses numerous capabilities similar to those demonstrated by GPT-4. For instance, MiniGPT-4 can generate intricate image descriptions, create websites based on handwritten text instructions, and explain unusual visual phenomena. Furthermore, our findings revealed that MiniGPT-4 also has a variety of other intriguing abilities not showcased in the GPT-4 demonstrations. For example, MiniGPT-4 can directly generate detailed cooking recipes from food photos, write stories or poems inspired by images, write advertisements for products in images, identify problems shown in photos and provide corresponding solutions, and retrieve rich facts about people, movies, or art directly from images, among other capabilities. These abilities are absent in previous vision-language models like Kosmos-1 (Huang et al., 2023) and BLIP-2 (Li et al., 2023) that use less powerful language models. This further validates that integrating visual features with an advanced language model is one of the keys to enhancing vision-language models.

We present a summary of our key findings:

- Our research reveals with compelling evidence that by aligning visual features with advanced large language models like Vicuna, MiniGPT-4 can achieve advanced vision-language capabilities comparable to those exhibited in the GPT-4 demonstrations.
- Our findings suggest that training merely one projection layer can effectively align a pre-trained vision encoder with the large language model. Our MiniGPT-4 only requires training approximately 10 hours on 4 A100 GPUs.
- We discovered that simply aligning visual features with large language models using short image caption pairs is not sufficient for developing a well-performing model and leads to

unnatural language generation. Further finetuning with a small but detailed image description pairs can address this limitation and significantly improves its usability.

2 RELATED WORKS

Large language models have experienced tremendous success in recent years due to the scaling up of training data and an increase in the number of parameters. Early models, such as BERT (Devlin et al., 2018), GPT-2 (Radford et al., 2019), and T5 (Raffel et al., 2020), laid the foundation for this progress. Subsequently, GPT-3 (Brown et al., 2020), with a massive scale of 175 billion parameters, was introduced, demonstrating significant breakthroughs across numerous language benchmarks. This development inspired the creation of various other large language models, including Megatron-Turing NLG (Smith et al., 2022), Chinchilla (Hoffmann et al., 2022), PaLM (Chowdhery et al., 2022), OPT (Zhang et al., 2022), BLOOM (Scao et al., 2022b), and LLaMA (Touvron et al., 2023), among others. Wei et al. (Wei et al., 2022) further discovered several *emergent abilities*, which appear exclusively in large models. The emergence of these abilities underscores the importance of scaling up in the development of large language models. Moreover, by aligning the pre-trained large language model GPT-3 with human intent, instructions and human feedback, InstructGPT (Ouyang et al., 2022) and ChatGPT (OpenAI, 2022) enable conversational interactions with humans and can answer a wide range of diverse and complex questions. More recently, several open-sourced models, such as Alpaca (Taori et al., 2023) and Vicuna (Chiang et al., 2023), have been developed based on LLaMA (Touvron et al., 2023) and also exhibit similar performance.

Leveraging Pre-trained LLMs in Vision-Language Tasks. In recent years, the trend of using autoregressive language models as decoders in vision-language tasks has gained significant traction (Chen et al., 2022; Huang et al., 2023; Yang et al., 2022; Tiong et al., 2022; Alayrac et al., 2022; Li et al., 2023; 2022; Driess et al., 2023). This approach takes advantage of cross-modal transfer, allowing knowledge to be shared between language and multimodal domains. Pioneering studies like VisualGPT (Chen et al., 2022) and Frozen (Tsimpoukelli et al., 2021) have demonstrated the benefits of employing a pre-trained language model as a vision-language model decoder. Flamingo (Alayrac et al., 2022) was then developed to align a pre-trained vision encoder and language model using gated cross-attention, and was trained on billions of image-text pairs, showcasing impressive in-context few-shot learning capabilities. Following that, BLIP-2 (Li et al., 2023) was introduced, employing a Flan-T5 (Chung et al., 2022) with a Q-Former to efficiently align visual features with the language model. Most recently, PaLM-E (Driess et al., 2023), featuring 562 billion parameters, has been developed to integrate real-world continuous sensor modalities into an LLM, thereby establishing a connection between real-world perceptions and human languages. GPT-4 (OpenAI, 2023) has also been recently released, showcasing more powerful visual understanding and reasoning abilities after pre-training on a vast collection of aligned image-text data.

LLMs, such as ChatGPT, have proven to be powerful tools in enhancing the performance of vision-language tasks by collaborating with other specialized models. For instance, Visual ChatGPT (Wu et al., 2023) and MM-REACT (Yang* et al., 2023) showcase how ChatGPT can act as a coordinator, integrating with diverse visual foundation models and facilitating their collaboration to tackle more complex challenges. ChatCaptioner (Zhu et al., 2023) treats ChatGPT as a questioner, prompting diverse questions for BLIP-2 to answer. Through multi-round conversations, ChatGPT extracts visual information from BLIP-2 and effectively summarizes the image content. Video ChatCaptioner (Chen et al., 2023) extends this approach, applying it to video spatiotemporal understanding. ViperGPT (Surís et al., 2023) demonstrates the potential of combining an LLM with different vision models to address complex visual queries programmatically. In contrast, MiniGPT-4 directly aligns visual information with the language model to accomplish diverse vision-language tasks without the usage of external vision models.

3 METHOD

MiniGPT-4 aims to align visual information from a pretrained vision encoder with an advanced large language model (LLM). Specifically, we utilize the Vicuna (Chiang et al., 2023) as our language decoder, which is constructed upon LLaMA (Touvron et al., 2023) and can perform a wide range of complex linguistic tasks. For visual perception, we employ the same visual encoder as used in

BLIP-2 (Li et al., 2023), a ViT backbone (Fang et al., 2022) coupled with their pre-trained Q-Former. Both language and vision models are open-sourced. We target to bridge the gap between the visual encoder and LLM using a linear projection layer, with an overview of our model displayed in Fig. 1.

To achieve an effective MiniGPT-4, we propose a two-stage training approach. The initial stage involves pretraining the model on a large collection of aligned image-text pairs to acquire vision-language knowledge. In the second stage, we finetune the pretrained model with a smaller but high-quality image-text dataset with a designed conversational template to enhance generation reliability and usability.

3.1 FIRST PRETRAINING STAGE

During the initial pretraining stage, the model is designed to acquire vision-language knowledge from a large collection of aligned image-text pairs. We regard the output from the injected projection layer as a soft prompt for the LLM, prompting it to generate the corresponding ground-truth texts.

Throughout the entire pretraining process, both the pretrained vision encoder and the LLM remain frozen, with only the linear projection layer being pretrained. We use a combined dataset of Conceptual Caption (Changpinyo et al., 2021; Sharma et al., 2018), SBU (Ordonez et al., 2011) and LAION (Schuhmann et al., 2021) to train our model. Our model undergoes 20,000 training steps with a batch size of 256, covering approximately 5 million image-text pairs. The entire process takes about 10 hours to complete, utilizing 4 A100 (80GB) GPUs.

Issues of the first pretraining stage Following the first pretraining stage, our MiniGPT-4 demonstrates the capacity to possess a wealth of knowledge and offer reasonable responses to human inquiries. However, we have observed instances where it produces incoherent linguistic outputs, such as repetitive words or sentences, fragmented sentences, or irrelevant content. These issues hinder MiniGPT-4’s ability to engage in a fluent visual conversation with humans.

We also observed similar challenges encountered in GPT-3. Despite its pretraining on an extensive language dataset, GPT-3 struggles to generate language outputs that are accurately aligned with users’ intentions. Through a process of instruction fine-tuning and reinforcement learning from human feedback, GPT-3 evolves into GPT-3.5 (Ouyang et al., 2022; OpenAI, 2022) and becomes capable of producing more human-friendly outputs. This phenomenon bears a resemblance to the current state of MiniGPT-4 following its initial pretraining stage. As such, it is not surprising that our model may struggle to generate fluent and natural human language outputs at this stage.

3.2 CURATING A HIGH-QUALITY ALIGNMENT DATASET FOR VISION-LANGUAGE DOMAIN.

To achieve greater naturalness in the generated language and enhance the model’s usability, a second-stage alignment process is essential. While in the realm of NLP, instruction fine-tuning datasets (Taori et al., 2023) and conversations (sha, 2023) are easily accessible, no equivalent datasets exist for the vision-language domain. To address this deficiency, we carefully curated a detailed image description dataset, specifically tailored for vision-language alignment purposes. This dataset is subsequently utilized to fine-tune our MiniGPT-4 during the second-stage alignment process.

Initial aligned image-text generation In the initial phase, we employ the model derived from the first pretraining stage to generate comprehensive descriptions of input images. To enable our model to produce more detailed image descriptions, we designed a prompt that adheres to the conversational format of the Vicuna (Chiang et al., 2023) language model, as shown below. In this prompt, `<ImageFeature>` represents the visual features produced by the linear projection layer.

###Human: <ImageFeature>Describe this image in detail. Give as many details as possible. Say everything you see. ###Assistant:

To identify incomplete sentences, we examine whether the generated sentence exceeds 80 tokens. If it does not, we incorporate an additional prompt, *###Human: Continue ###Assistant:*, prompting our MiniGPT-4 to extend the generation process. By concatenating the outputs from both steps, we can create a more comprehensive image description. This approach enables us to generate image-text pairs with detailed and informative image descriptions. We randomly select 5,000 images from the

Conceptual Caption dataset (Changpinyo et al., 2021; Sharma et al., 2018) and use the pretrained model to generate corresponding language descriptions for each image.

Data post-processing The above automatically generated image descriptions contain noisy or incoherent descriptions, such as repetition of words or sentences, fragmented sentences, or irrelevant content. In order to fix these issues, we employ ChatGPT to mend the descriptions by utilizing the following prompt:

Fix the error in the given paragraph. Remove any repeating sentences, meaningless characters, not English sentences, and so on. Remove unnecessary repetition. Rewrite any incomplete sentences. Return directly the results without explanation. Return directly the input paragraph if it is already correct without explanation.

Upon completing the post-processing stage, we manually verify the correctness of each image description to guarantee its high quality. Specifically, we first identified several frequently shown errors (“I’m sorry I made a mistake...”, or “I apologize for that ...”) and then hard-coded rules to automatically filter them out. We also manually refine the generated captions by eliminating redundant words or sentences that ChatGPT fails to detect. Finally, only approximately 3,500 out of 5,000 image-text pairs satisfy our requirement, and these pairs are subsequently utilized for the second-stage alignment process.

3.3 SECOND-STAGE FINETUNING

During the second stage, we finetune our pretrained model with the curated high-quality image-text pairs. During the finetuning, we use the predefined prompts in the following template:

###Human: <ImageFeature><Instruction>###Assistant:

In this prompt, <Instruction> represents a randomly sampled instruction from our predefined instruction set containing variant forms of instructions such as “Describe this image in detail” or “Could you describe the contents of this image for me”. It is important to note that we do not calculate the regression loss for this specific text-image prompt.

As a result, MiniGPT-4 is now capable of producing more natural and reliable language outputs. Furthermore, we observed that this fine-tuning process is remarkably efficient, only requiring a mere 400 training steps with a batch size of 12, which takes around 7 minutes with a single A100 GPU.

4 EXPERIMENTS

In the experiment, we aim to showcase the diverse and emergent capabilities of our MiniGPT-4 model through various qualitative examples. These abilities include generating detailed image descriptions, identifying amusing aspects within memes, providing food recipes from photos, writing poems for images, etc. Additionally, we present quantitative results on the task of image captioning.

4.1 UNCOVERING EMERGENT ABILITIES WITH MINIGPT-4 THROUGH QUALITATIVE EXAMPLES

MiniGPT-4 demonstrates many advanced abilities compared to traditional vision-language models. For example, it can describe images in detail and interpret the humorous aspects of a given meme. Here, we qualitatively compared our model to one of the leading vision-language models, BLIP-2 (Li et al., 2023), with eight distinct examples, each highlighting a different ability.

An example in Fig.2 demonstrates that MiniGPT-4 effectively identifies various elements within the image, such as busy city streets, clock towers, shops, restaurants, motorcycles, people, streetlights, and clouds. In contrast, BLIP-2 can only cover city streets, people, and motorcycles in its image caption generation. Another example presented in Fig.4a shows that MiniGPT-4 successfully explains why the meme is humorous. It interprets that the lying dog is feeling the same way as many people do on Monday, which is often considered to be the most dreaded day of the week. In contrast, BLIP-2 only briefly describes the image content and fails to comprehend the amusing aspects of the image.

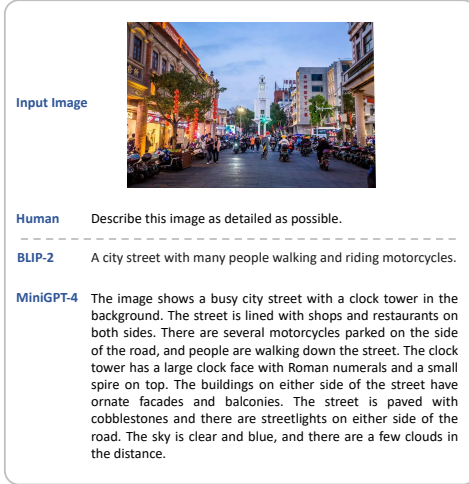


Figure 2: Detailed description

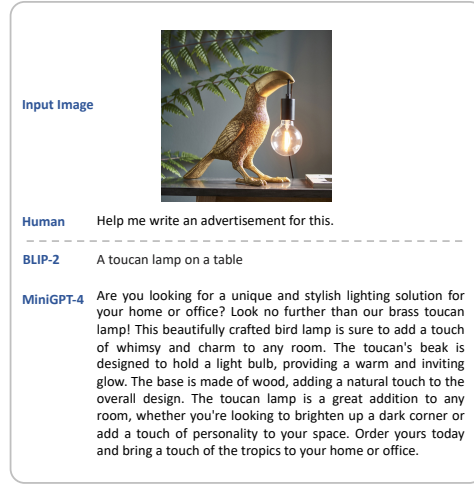


Figure 3: Advertisement promotion

We also showcase MiniGPT-4’s other abilities by demonstrating other distinctive abilities. These include creating advertising promotions based on a given image (Fig.3), retrieving factual information from a movie photograph (Fig.8), generating a food recipe from a food image (Fig.11), diagnosing plant diseases and suggesting treatment plans (Fig.12), creating a website from a hand-written draft (Fig.4b), and writing poems inspired by an image (Fig.10). These abilities are absent in traditional vision-language models like BLIP-2 (utilizing Flan-T5 XXL (Chung et al., 2022) as a language model), which use less powerful language models (LLMs). This contrast indicates that those advanced vision-language abilities only emerge when the visual features are properly aligned with an advanced LLM such as Vicuna (Chiang et al., 2023).

4.2 QUANTITATIVE ANALYSIS

Advanced Abilities To quantify performance on advanced vision-language tasks, we compiled a small evaluation dataset comprising 4 tasks: meme interpretation with the question “Explain why this meme is funny.”, recipe generation with the question “How should I make something like this?”, advertisement creation with the prompt “Help me draft a professional advertisement for this.”, and poem composition with “Can you craft a beautiful poem about this image?”. In total, we collect 100 diverse images, with 25 images allocated to each task. We asked human evaluators to determine whether the model generation satisfies the request. We compared our results with BLIP-2 (Li et al., 2023) and present the findings in Tab.1. In meme interpretation, poem writing, and advertisement creation, BLIP-2 largely struggles to fulfill any requests. For recipe generation, BLIP-2 succeeds in 4 out of 25 cases. In contrast, MiniGPT-4 manages to address the requests in recipes, advertisements, and poem generation in nearly 80% of the instances. Furthermore, MiniGPT-4 correctly comprehends the challenging humor understanding in memes in 8 out of 25 cases.

Image Captioning We evaluate the performance of MiniGPT-4 on the COCO caption benchmark and compare it with BLIP-2 (Li et al., 2023). Our model’s generated captions typically contain rich visual details. As such, conventional similarity-based image-caption evaluation metrics struggle to provide an accurate evaluation of our models. In this regard, we evaluate the performance by

Table 1: Quantitative results on advanced vision-language tasks. MiniGPT-4 shows strong performance and successfully responses to 65% of the requests.

	Meme	Recipes	Ads	Poem	Avg.
BLIP-2	0/25	4/25	1/25	0/25	5/100
MiniGPT-4	8/25	18/25	19/25	20/25	65/100

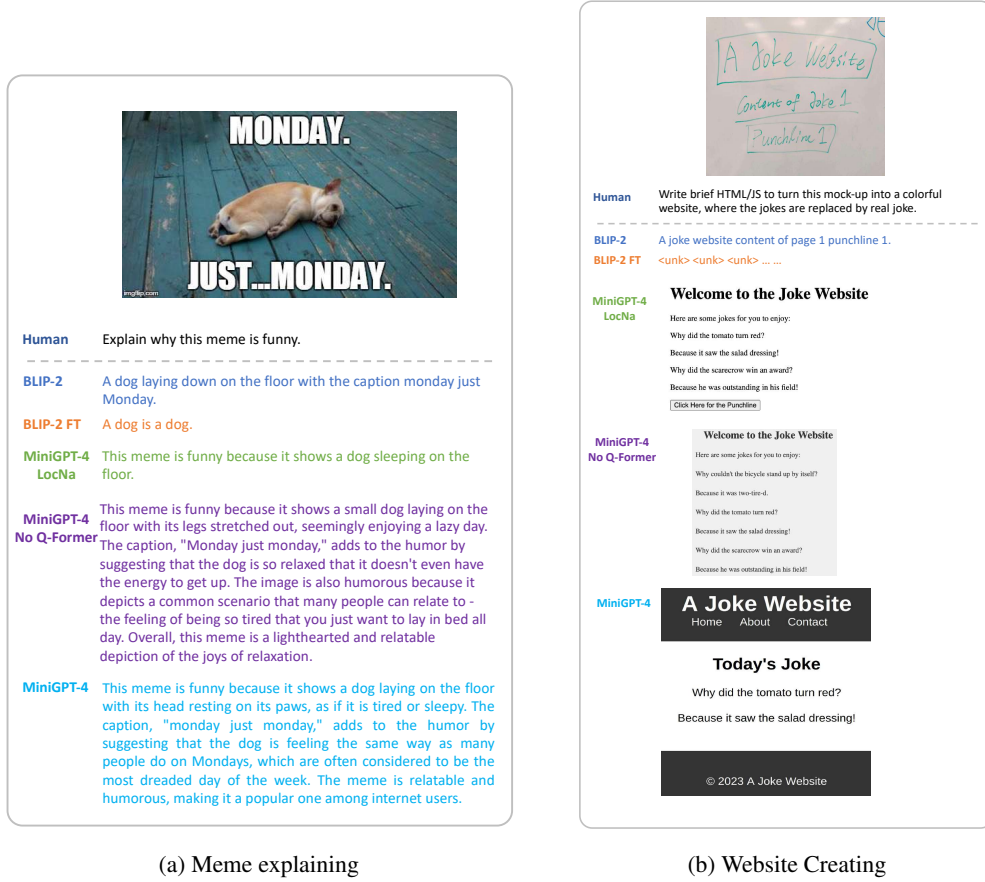


Figure 4: Model generations from BLIP-2, BLIP-2 finetuned our second stage data (BLIP-2 FT), MiniGPT-4 finetuned with Local Narrative data in the second stage (MiniGPT-4 LocNa), MiniGPT-4 model without Q-Former (MiniGPT-4 No Q-Former), and MiniGPT-4.

checking if the generated captions cover all the ground truth captions’ information with the help of ChatGPT and details can be found in Appx.A.3. Results in Tab.2 shows that MiniGPT-4 outperforms BLIP-2 in generating captions that are more closely aligned with the ground-truth visual objects and relationships. With a success rate of 66.2%, MiniGPT-4 is considerably more accurate than BLIP-2, which achieves only 27.5%. Further evaluation on traditional VQA tasks can be found in Appx.A.2.

4.3 ANALYSIS ON THE SECOND-STAGE FINETUNING

Effectiveness of the second-stage finetuning The utilization of only the model pretrained after the first pretraining stage may result in failures, such as the occurrence of repetitive words or sentences, fragmented sentences, or irrelevant content. However, these issues have been largely mitigated through the second-stage finetuning process. This can be observed in Fig.5, where MiniGPT-4

Table 2: COCO caption evaluation. We use ChatGPT to judge if the generated caption covers all the visual objects and relations in the ground-truth caption.

	BLIP-2	MiniGPT-4
Correctness	1376/5000	3310/5000
Percentage	27.5%	66.2%

Table 3: Failure rates of detailed caption and poem generation tasks before and after second-stage finetuning. The finetuning stage significantly reduces generation failures.

	Failure rate	Detailed caption	Poem
Before stage-2		35%	32%
After stage-2		2%	1%

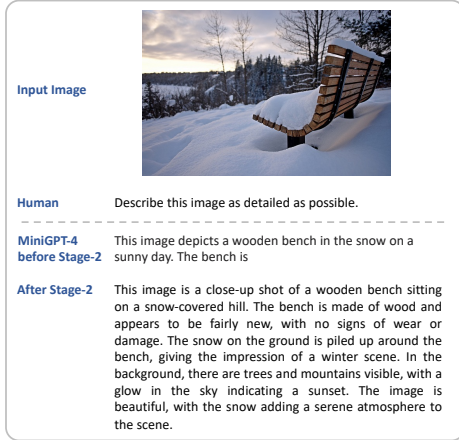


Figure 5: MiniGPT-4 before second-stage fine-tuning fails to output completed texts. The generation is improved after the finetuning.

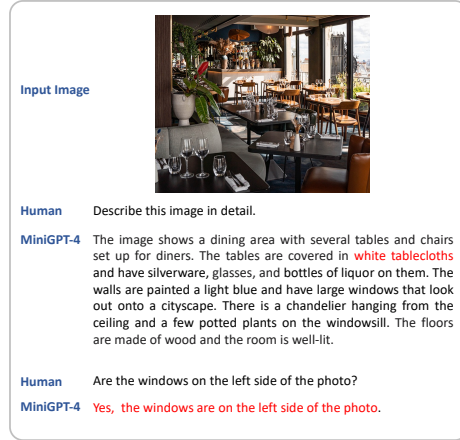


Figure 6: An example of MiniGPT-4’s limitations. MiniGPT-4 hallucinates unexisting tablecloths and can’t locate the windows correctly.

generates incomplete captions before the second-stage finetuning. However, after the second-stage finetuning, MiniGPT-4 is capable of generating complete and fluent captions. In this section, we investigate the importance and effectiveness of the second-stage finetuning approach.

To quantify the impact of second-stage finetuning, we randomly sampled 100 images from the COCO test set and investigated the model performance on two tasks: detailed description generation and poem writing. The prompts used were “Describe the image in detail.” and “Can you write a beautiful poem about this image?”. These tasks were performed by both the models before and after second-stage finetuning. We manually counted the number of failure generations for the model in each stage. The results are presented in Tab.3. Prior to the second-stage finetuning, approximately 1/3 of the generated outputs failed to match ground truth captions or poems. In contrast, the model after second-stage finetuning has less than two failure cases out of the 100 test images for both tasks. These experimental results demonstrate that second-stage finetuning yields a significant improvement in the quality of generated outputs. A qualitative example of the model generation before and after the second-stage finetuning is shown in Fig.5.

Can the original BLIP-2 benefit from the second-stage data? In this study, we finetune BLIP-2 (Li et al., 2023) with our second-stage data in the same way as MiniGPT-4, and check if it can obtain similar advanced abilities as MiniGPT-4. The finetuned BLIP-2 is denoted as BLIP-2 FT. Note that MiniGPT-4 uses the same visual module as BLIP-2; while BLIP-2 uses FlanT5 XXL (Chung et al., 2022) as the language model, which is not as strong as the Vicuna (Chiang et al., 2023) model used in our MiniGPT-4 model. We rely on the same prompts to assess the advanced capabilities of our model. Qualitative results are shown in Fig.4, 13, and 14. We discover that BLIP-2 FT still generates short responses and fails to generalize to advanced tasks like meme explaining and website coding (Fig.4). Our finding suggests that BLIP-2’s relatively weaker language model FlanT5 XXL benefits less from such a small dataset, and highlights the effectiveness of a more advanced LLM in a VLM system.

Second stage with Localized Narratives The dataset Localized Narratives (Pont-Tuset et al., 2020) is a detailed image description dataset where annotators describe images while simultaneously localizing the corresponding regions. Here, we test the performance of our model by replacing our self-collected dataset in the second-stage with the Localized Narratives dataset. The model is denoted

Table 4: Ablation on architecture designs

Model	AOK-VQA	GQA
MiniGPT-4	58.2	32.2
(a) MiniGPT-4 w/o Q-Former	56.9	33.4
(b) MiniGPT-4 + 3 Layers	49.7	31.0
(c) MiniGPT-4 + Finetune Q-Former	52.1	28.0

Table 5: Hallucination Evaluation

	CHAIR _i	Avg. Length
Blip-2	1.3	6.5
MiniGPT-4 (short)	7.2	28.8
MiniGPT-4 (long)	9.6	175

as MiniGPT-4 LocNa. Qualitative results in Fig. 4, 13, and 14 show that MiniGPT-4 LocNa can generate long image descriptions (Fig. 14). However, the generated outputs have lower quality with monotonous expressions. Besides, MiniGPT-4 LocNa does not generalize as well as the original MiniGPT-4 in other complex tasks like explaining why the meme is funny (Fig. 4a). The performance gap may be due to the monotonous and repeated image descriptions in Localized Narratives.

4.4 ABLATION ON THE ARCHITECTURE DESIGNS

To further demonstrate the effectiveness of using one single linear layer to align visual features with LLM, we conduct experiments with different architecture designs, including (a) removing the Q-Former and directly mapping the ViT’s output to Vicuna’s embedding space (i.e., without Q-former), (b) using three linear layers instead of one layer, and (c) additionally finetuning the Q-Former in the vision module. All the variants are trained in the same way as the original design. Results on AOK-VQA (Schwenk et al., 2022) and GQA (Hudson & Manning, 2019) datasets in Tab. 4 show that the variant (a) **MiniGPT-4 w/o Q-Former** has a similar performance to the original design. Qualitative results of this variant in Fig. 4, 13, and 14 also show similar advanced skills. This reveals that the Q-Former from BLIP-2 doesn’t play a critical role for advanced skills. Besides, both variants (b) **MiniGPT-4+ 3 Layers** and (c) **MiniGPT-4 + finetuning Q-Former**, perform slightly worse than the original MiniGPT-4. This indicates a single projection layer is sufficient to align the vision encoder and the large language model in our limited training data setting.

4.5 LIMITATION ANALYSIS

Hallucination As MiniGPT-4 is built upon LLMs, it inherits LLM’s limitations like hallucinating nonexistent knowledge. An example in Fig. 6 shows that MiniGPT-4 incorrectly identifies the presence of white tablecloths in the image, despite their absence. Here, we use the metric CHAIR_i (Rohrbach et al., 2018) to gauge the hallucination rate of the generation, with the two distinct prompts to control the model generation length: *MiniGPT-4 (long)*: Please describe this image as detailed as possible. *MiniGPT-4 (short)*: Please describe the image shortly and precisely, in less than 20 words.

Results in Tab. 5 show that longer captions tend to have higher hallucination rates. For example, MiniGPT-4 (long) generates captions averaging 175 words with a higher hallucination rate, while MiniGPT-4 (short) averages 28.8 words with a lower rate. BLIP-2, averaging 6.5 words, hallucinates less but covers fewer objects as seen in Tab. 2. Hallucination in detailed image descriptions is still an unresolved issue. Using Reinforcement Learning with AI feedback with hallucination detection modules may be a potential solution.

Spatial Information Understanding MiniGPT-4’s visual perception remains limited. It may struggle to differentiate spatial localization. For example, MiniGPT-4 in Fig. 6 fails to identify the location of the windows. This limitation may stem from a lack of aligned image-text data designed for spatial information understanding. Training on such datasets like RefCOCO (Kazemzadeh et al., 2014) or Visual Genome (Krishna et al., 2017) could potentially alleviate this issue.

5 DISCUSSION

How does MiniGPT-4 obtain these advanced abilities? Many of the advanced vision-language capabilities demonstrated by GPT-4 can be understood as compositional skills rooted in two foundational skills: image understanding and language generation. Take the task of image-based poem writing as an example. Advanced LLMs like ChatGPT and Vicuna can already craft poems based on users’ instructions. If they acquire the ability to understand images, compositionally generalizing to the task of image-based poem writing even without having image-poem pairs in their training data is possible.

In the first pretraining stage, MiniGPT-4 learns to understand images by modeling the correlation between images and short image descriptions from image caption datasets. However, the language style in these image caption datasets differs from that of modern LLMs’ generation, which leads to distorted language generation and hinders successful compositional generalization. Therefore, we introduce a second-stage finetuning to restore the language generation ability. MiniGPT-4 after the two-stage training successfully generalizes to many advanced compositional vision-language abilities like website coding from drafts or meme interpretation, verifies our assumption. Future research

might delve deeper into the mechanism of compositional generalization and seek ways to enhance them. We hope our work, as an early exploration of these vision-based LLM capabilities, will spur further investigations in this domain.

REFERENCES

- Sharegpt. <https://github.com/domeccleston/sharegpt>, 2023.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 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*, pp. 3558–3568, 2021.
- Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18030–18040, 2022.
- Jun Chen, Deyao Zhu, Kilichbek Haydarov, Xiang Li, and Mohamed Elhoseiny. Video chatcaptioner: Towards the enriched spatiotemporal descriptions. *arXiv preprint arXiv:2304.04227*, 2023.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL <https://vicuna.lmsys.org>.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.
- Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022.
- 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. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.

- Drew A Hudson and Christopher D Manning. 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*, pp. 6700–6709, 2019.
- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 787–798, 2014.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
- OpenAI. Introducing chatgpt. <https://openai.com/blog/chatgpt>, 2022.
- OpenAI. Gpt-4 technical report, 2023.
- Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.
- Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 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*, pp. 647–664. Springer, 2020.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.
- Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022a.
- Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022b.
- Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pp. 146–162. Springer, 2022.

- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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)*, pp. 2556–2565, 2018.
- Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlG 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022.
- Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.
- Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven CH Hoi. Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training. *arXiv preprint arXiv:2210.08773*, 2022.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL <https://openreview.net/forum?id=yzkSU5zdWd>. Survey Certification.
- Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.
- Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *arXiv preprint arXiv:2206.08155*, 2022.
- Zhengyuan Yang*, Linjie Li*, Jianfeng Wang*, Kevin Lin*, Ehsan Azarnasab*, Faisal Ahmed*, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multi-modal reasoning and action. 2023.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv preprint arXiv:2303.06594*, 2023.

A APPENDIX

A.1 MORE QUALITATIVE RESULTS

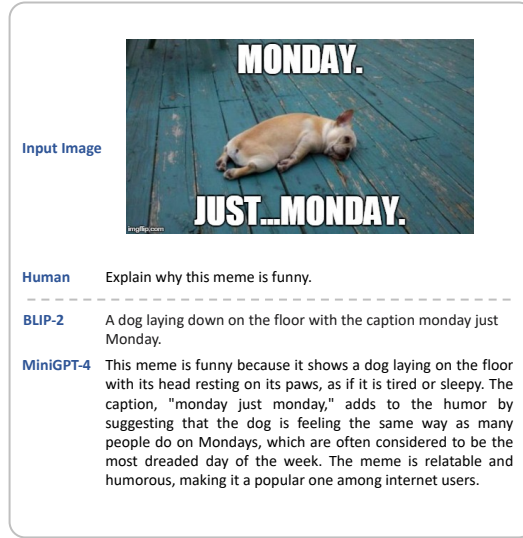


Figure 7: Advertisement promotion

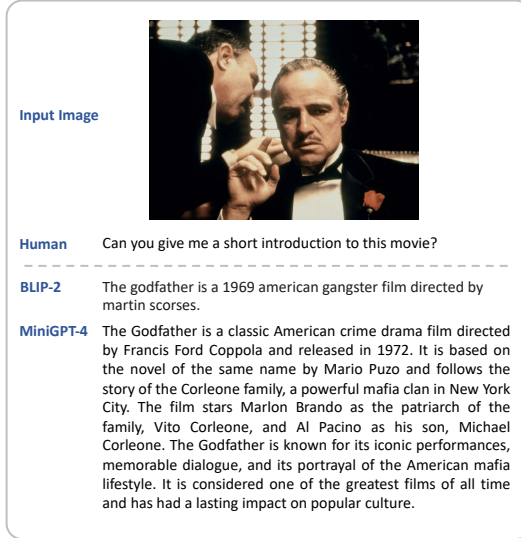


Figure 8: Factual retrieval



Figure 9: Website creating

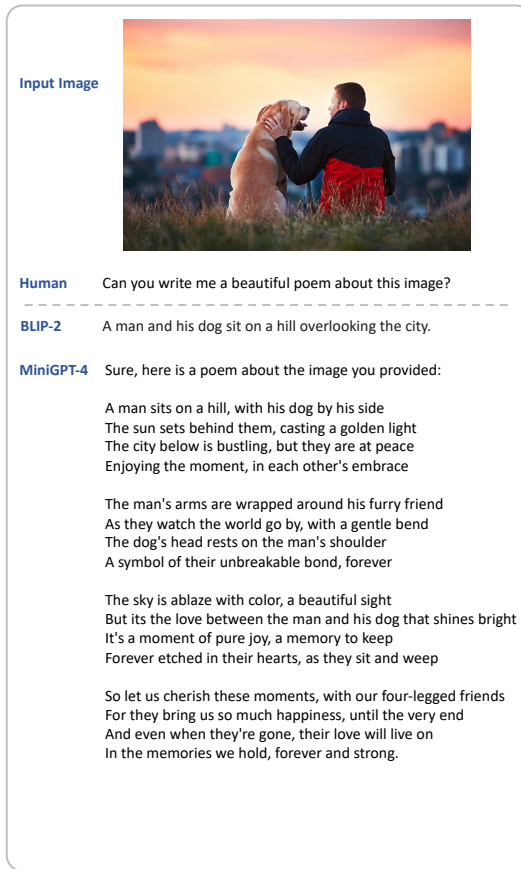


Figure 10: Poem writing

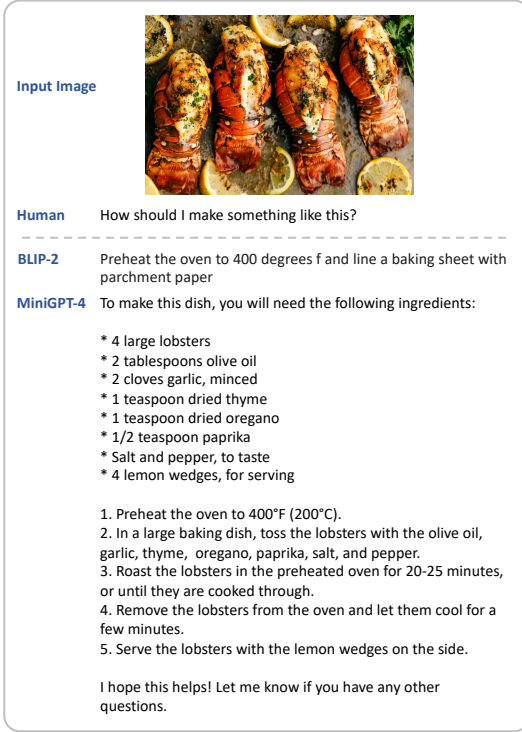


Figure 11: Food recipe generation

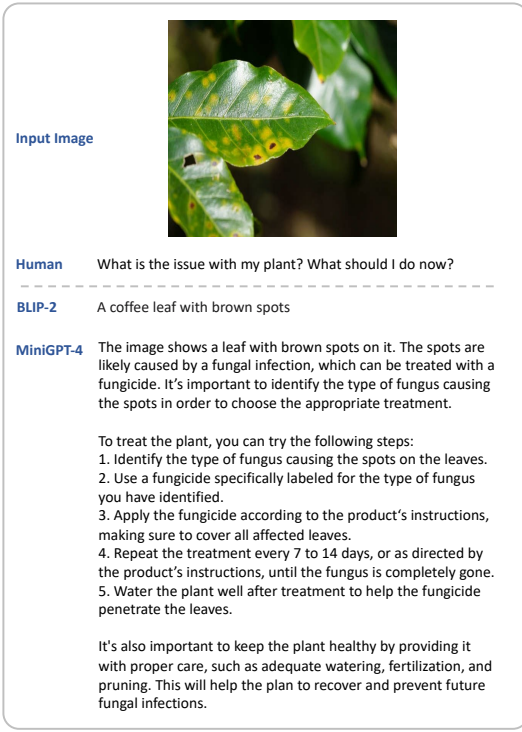


Figure 12: Plant cultivating

A.2 EVALUATION IN TRADITIONAL VQA BENCHMARKS

The aim of this study is to replicate the remarkable multi-modal capabilities demonstrated in GPT-4, such as generating detailed image descriptions and creating websites from hand-drawn drafts. To emphasize the most crucial component of advanced vision-language skills, the methodology of MiniGPT-4 is intentionally kept minimal. For instance, the learnable model capacity is limited (only one linear layer), and MiniGPT-4 is trained with just 5 million pairs, in contrast to BLIP-2 with 129 million image-text pairs. Such a pared-down approach is anticipated to yield suboptimal results on traditional benchmarks. While this isn't our primary goal, we offer a quantitative analysis of the VQA datasets A-OKVQA (multi-choice) (Schwenk et al., 2022) and GQA (Hudson & Manning, 2019). Additionally, to showcase the potential of MiniGPT-4 with traditional benchmarks, we conduct a straightforward ablation study. Here, we simply unfreeze the LLM using LoRA (Hu et al., 2021) and incorporate more training data from the VQAv2, OKVQA, and A-OKVQA datasets during the second finetuning stage. Results in Tab. 6 indicate that the original MiniGPT-4 lags behind BLIP-2 by a reasonable margin, and merely augmenting the learning capacity and the training data results in a substantial performance improvement, which confirms our expectations. We believe our model's performance on conventional vision benchmarks can be enhanced with a carefully designed training strategy (e.g., dataset sample ratios, learning rate schedule, etc.), more training data/datasets, and additional learnable parameters. Since enhancing performance on traditional vision benchmarks isn't this project's objective, we reserve this aspect for future research.

Model	Training data	AOK-VQA	GQA
Blip-2	129M image-text pairs	80.2	42.4
MiniGPT-4	5M image-text pairs	58.2	32.2
MiniGPT-4 (Finetune Vicuna)	5M image-text pairs	67.2	43.5


Table 6: Performance Comparison between BLIP-2 and MiniGPT-4

A.3 DETAILS OF CAPTION EVALUATION

We employ ChatGPT to determine whether the baseline models cover all the objects and visual relations presented in the ground-truth captions. For the COCO evaluation dataset, we randomly choose one ground-truth caption and treat it as the reference caption. We apply the following prompt to perform the evaluation.

There is one image caption1 ‘{ground-truth caption}’, and there is another image caption2 ‘{comparison caption}’. Does image caption2 cover all the objects and visual relations shown in image caption1? Only answer yes or no without any explanation.

A.4 MORE QUALITATIVE ABLATION RESULTS



Human How should I make something like this?

BLIP-2 Preheat the oven to 400 degrees f and line a baking sheet with parchment paper

BLIP-2 FT A roasted lobster with a roasted garlic sauce

MiniGPT-4 LocNa To make this dish, you will need the following ingredients:

MiniGPT-4 No Q-Former To make this dish, you will need the following ingredients:

- * 4 large live lobsters
- * 2 tablespoons olive oil
- * Salt and pepper
- * 2 cloves garlic, minced
- * 1 lemon, juiced
- * 1 tablespoon chopped fresh parsley

Here's how to make it:

1. Preheat your oven to 425°F (220°C).
2. Rinse the lobsters under cold water and pat them dry with paper towels.
3. Heat the olive oil in an oven-safe skillet over medium-high heat. Add the lobsters, season with salt and pepper, and cook for 5-7 minutes per side, or until the shells are bright red and the meat is cooked through.
4. Transfer the lobsters to a baking dish and bake for 10-12 minutes, or until the meat is cooked through and the shells are crisp.
5. In a small bowl, mix the garlic, lemon juice, and parsley.
6. Serve the lobsters with the garlic-lemon butter on top and a lemon wedge on the side.

I hope this helps! Let me know if you have any questions.


MiniGPT-4 To make this dish, you will need the following ingredients:

- * 4 large lobsters
- * 2 tablespoons olive oil
- * 2 cloves garlic, minced
- * 1 teaspoon dried thyme
- * 1 teaspoon dried oregano
- * 1/2 teaspoon paprika
- * Salt and pepper, to taste
- * 4 lemon wedges, for serving

1. Preheat the oven to 400°F (200°C).
2. In a large baking dish, toss the lobsters with the olive oil, garlic, thyme, oregano, paprika, salt, and pepper.
3. Roast the lobsters in the preheated oven for 20-25 minutes, or until they are cooked through.
4. Remove the lobsters from the oven and let them cool for a few minutes.
5. Serve the lobsters with the lemon wedges on the side.

I hope this helps! Let me know if you have any other questions.

Figure 13: Ablation Study on Recipe Generation



Human Describe this image as detailed as possible.

BLIP-2 A city street with many people walking and riding motorcycles.

BLIP-2 FT A city street with many people riding on motorcycles and scooters.

MiniGPT-4 LocNa In this image, we can see a busy street with many shops and restaurants. On the left side of the image, we can see a clock tower. On the right side of the image, we can see many motorcycles parked on the road. In the middle of the image, we can see many people walking on the road. On the left side of the image, we can see many shops and restaurants. On the right side of the image, we can see many motorcycles parked on the road. In the middle of the image, we can see many people walking on the road. On the left side of the image, we can see many shops and restaurants. On the right side of the image, we can see many motorcycles parked on the road. In the middle of the image, we can see many people walking on the road.

MiniGPT-4 No Q-Former The image shows a busy street in a city at night. There are many people walking on the sidewalk and riding bicycles. The buildings on either side of the street are tall and have many windows. There is a large clock tower in the center of the image. The sky is clear and there are stars visible in the sky. The image is taken from a high angle, looking down on the street.

MiniGPT-4 The image shows a busy city street with a clock tower in the background. The street is lined with shops and restaurants on both sides. There are several motorcycles parked on the side of the road, and people are walking down the street. The clock tower has a large clock face with Roman numerals and a small spire on top. The buildings on either side of the street have ornate facades and balconies. The street is paved with cobblestones and there are streetlights on either side of the road. The sky is clear and blue, and there are a few clouds in the distance.

Figure 14: Ablation Study on Detailed Description