ArtGPT-4: Artistic Vision-Language Understanding with Adapter-enhanced MiniGPT-4

Zhengqing Yuan*
School of Artificial Intelligence
Anhui Polytechnic University
Wuhu, China 231000
zhengqingyuan@ieee.org

Huiwen Xue*
School of Optoelectronic Science and Engineering
Soochow University
Suzhou, China 215000
hwxuehwxue@stu.suda.edu.cn

Xinyi Wang*
School of Artificial Intelligence
Anhui Polytechnic University
Wuhu, China 231000
xinyiwang1121@163.com

Yongming Liu*,†
School of Artificial Intelligence
Anhui Polytechnic University
Wuhu, China 231000
liuyongming1015@163.com

Zhuanzhe Zhao*,†
School of Artificial Intelligence
Anhui Polytechnic University
Wuhu, China 231000
huanzhe727@ahpu.edu.cn

Kun Wang*,†
School of Computer and Information Technology
Anhui Polytechnic University
Wuhu, China 231000

Abstract

In recent years, large language models (LLMs) have made significant progress in natural language processing (NLP), with models like ChatGPT and GPT-4 achieving impressive capabilities in various linguistic tasks. However, training models on such a large scale is challenging, and finding datasets that match the model's scale is often difficult. Fine-tuning and training models with fewer parameters using novel methods have emerged as promising approaches to overcome these challenges. One such model is MiniGPT-4, which achieves comparable vision-language understanding to GPT-4 by leveraging novel pre-training models and innovative training strategies. However, the model still faces some challenges in image comprehension, particularly in artistic pictures. To address these limitations, a novel multimodal model called ArtGPT-4 has been proposed. ArtGPT-4 was trained on image-text pairs using a Tesla A100 device in just 2 hours, using only about 200 GB of data. The model can depict images with an artistic flair and generate visual code, including aesthetically pleasing HTML/CSS web pages. Furthermore, the article proposes novel benchmarks for evaluating the performance of vision-language models. In the subsequent evaluation methods, ArtGPT-4 scored more than one point higher than MiniGPT-4 and was only 0.25 points lower than GPT-4 on a five-point scale. Our code and pre-trained model are available at https://minigpt-4.github.io/.

*These authors contributed equally to this work.
†Corresponding authors.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

1 Introduction

In recent times, the advancements made by large language models (LLMs) in the domains of natural language processing have been truly remarkable [28, 4, 26, 11, 14]. The performance of each task has demonstrated improvement with the increase in model size, simultaneously unlocking novel capabilities [8]. These models can perform a variety of intricate linguistic tasks in a zero-shot manner, thanks to their exceptional language capabilities. Large multimodal models, such as M6 [22], LLaVA [24], and GPT-4 [27], have shown impressive capabilities in understanding the vision-language. For instance, models like LLaVA and GPT-4 have shown the ability to capture intricate details and comprehend the meaning of images, thereby producing diverse language outputs.

However, training a model on such a large scale presents a significant challenge to the experimental device. And finding datasets that match the scale of large models like GPT-3, which utilizes up to 45 terabytes of text data, can be a daunting task. As a result, fine-tuning with existing pre-trained models has emerged as a promising approach, as illustrated by ChatGPT's use of GPT-3 for fine-tuning according to Reinforcement Learning with Human Feedback (RLHF) [9]. Training a model with fewer parameters using a novel training method is also a crucial approach like the Vicuna [7] and Alpaca [34]. An example of this is the Alpaca model, which has only 7 billion parameters but can achieve similar performance to the text-davinci-003 model, a GPT-3 variant with 175 billion parameters, on the Self-Instruct evaluation suite for instruction-following [36].

The MiniGPT-4 model [42] achieves comparable vision-language understanding to GPT-4 by leveraging novel pre-training models and innovative training strategies. However, there are still some challenges that need to be addressed. Firstly, the model's ability to comprehend images still needs improvement, as it struggles to capture all the details, particularly in artistic pictures. Secondly, there is currently no well-defined metric for evaluating the model's performance. Finally, it is worth noting that the experimental setup employed four Tesla A100s, which is relatively high-end equipment.

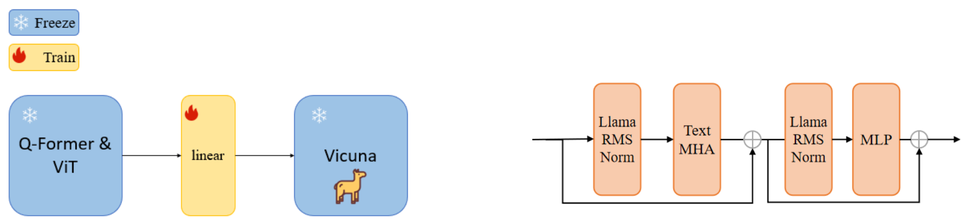
We propose a novel solution to the limitations of MiniGPT-4 by introducing a new model called ArtGPT-4. ArtGPT-4 incorporated tailored linear layers and their corresponding activation functions into Vicuna, in conjunction with activating specific training parameters. Our modifications were designed to optimize the model's performance and enable it to effectively address the unique challenges posed by vision-language tasks. ArtGPT-4 was trained on a Tesla A100 device in just 2 hours, using only about 200 GB of image-text pairs. ArtGPT-4 can depict images with a more artistic flair and generate visual code, including aesthetically pleasing HTML/CSS web pages. In a subsequent evaluation method, it was shown that ArtGPT-4 outperformed the original MiniGPT-4 model in terms of image understanding.

Our contribution is as follows:

- We are the first to utilize efficient fine-tuning for cross-visual and linguistic domain models, and have achieved impressive results.
- We propose a novel multimodal model called ArtGPT-4 based on MiniGPT-4, and it performs better than MiniGPT-4.
- We propose a novel benchmarks for multimodal vision-language understanding models. This proposed benchmarks can serve as a more comprehensive criterion for evaluating the performance of vision-language models.
- We have made all the code and training layer parameters in this article available as open source.

2 Related Work

Language Models: Early language models such as BERT [12], GPT [30], GPT-2 [31], RoBERT [25], ELECTRA [10] and XLNet [38] have tended to place greater emphasis on natural language understanding tasks in the field of natural language processing, including language inference, text classification, and named entity recognition. These models leverage large-scale unsupervised pre-training to acquire rich linguistic knowledge and patterns and have achieved impressive performance across a range of NLP tasks. In recent years, the emergence of large-scale language models like T5 [32], GPT-3 [4], and LLaMA [11] has yielded impressive results in natural language generation tasks, including chatbots, text generation, and summarization. These models have proven highly effective at producing coherent and contextually relevant language, paving the way for a range of innovative new applications in natural language processing. Furthermore, in order to improve the usability of these large models for humans and enhance their ability to understand natural language commands, researchers have applied Reinforcement Learning with Human Feedback (RLHF) [9]. This approach has yielded impressive results, with models such as ChatGPT [26] and GPT-4 [27] demonstrating outstanding performance. By leveraging the insights and feedback of human users, these models have become more accurate, efficient, and responsive, opening up new possibilities for natural language processing applications.



(a) Structure of MiniGPT-4 (b) A Transformer block of Vicuna

Figure 1: Model structure and training process of MiniGPT-4 and Vicuna models

Vision-Language Model: Recently, some researchers [5, 1, 35] are exploring models with broader capabilities beyond a single domain, such as enabling language models to understand images. For instance, GPT-4 leverages a vast amount of image-text data to enhance its power not only for linguistic tasks but also for image comprehension, allowing it to describe images using language. Remarkably, MiniGPT-4 can equip the Vicuna model with image understanding abilities by employing only a ViT [13], Q-Former [41], and a linear layer as shown in Figure 1(a). There also exist some efficient training methods for multimodal models, like BLIP-2 [20], which proposes a generic and efficient pre-training strategy for vision-language tasks. It utilizes off-the-shelf frozen pre-trained image encoders and large language models, along with a lightweight Q-Former to bridge the modality gap.

Efficient Fine-tuning: Efficient fine-tuning strategies can enhance the performance of large language models (LLMs) on specific tasks. For instance, the LoRA [18] method utilizes low-rank projection of model weights to adapt large language models to new domains with limited training data. For example, Vicuna and Alpaca models based on the LLaMA model utilizing the LoRA method with ChatGPT dialogue data have demonstrated impressive performance. Parameter-efficient fine-tuning methods [17, 40, 21, 15, 29] are also a promising approach in the field of NLP, with the goal of reducing the number of learning parameters and computational resources required to adapt to downstream tasks while achieving comparable results to full fine-tuning. There has also been recent work in the field of computer vision on efficient learning, such as the work by Jia et al. [19], Bahng et al. [3], and Chen et al. [6] on visual adaptation using methods similar to those used in NLP. It is worth noting, however, that these works are based on the same modality (such as text to text, image to image, video to video), or the same domain [37] (such as visual image to visual video).

Benchmark for Vision-Language Tasks: To improve the evaluation metrics for multimodal models, several works have proposed new benchmark tests and datasets [23, 39, 2]. These datasets have limitations such as a lack of diversity in image types, a bias towards certain cultural references, and limited size. While they provide a useful benchmark for evaluating vision-language models, it's important to keep in mind their potential limitations. By tackling these challenges, we leverage the conference paper review process and the exceptional multimodal model GPT-4 to establish a novel benchmark for evaluating multimodal models.

3 ArtGPT-4

In this section, we will provide a brief overview of MiniGPT-4 and its baseline model (Section 3.1). Next, we will detail the Image Adapter (Section 3.2), and the training of ArtGPT-4 (Section 3.3), illustrating how we use the Adapter layer to construct visual-verbal multimodal models step-by-step.

3.1 MiniGPT-4

MiniGPT-4 is a model that aims to combine visual information from a trained vision encoder with an advanced large language model (LLM) to perform a wide range of complex linguistic tasks. The model uses the Vicuna language decoder, which is constructed upon LLaMA, and a visual encoder consisting of a ViT backbone coupled with a pre-trained Q-Former as shown in Figure 1. MiniGPT-4 is trained in two stages: pretraining on a large collection of aligned image-text pairs to acquire vision-language knowledge, and fine-tuning on a smaller but high-quality image-text dataset with a designed conversational template to enhance the model's generation reliability and usability. During the initial pretraining stage, MiniGPT-4 acquires vision-language knowledge from a large dataset of aligned image-text pairs. However, it may still struggle with generating coherent and human-friendly responses, similar to ChatGPT before its fine-tuning and reinforcement learning stage. The second stage alignment process is essential to enhance the model's naturalness in generated language. A carefully curated image-text dataset is used to fine-tune the MiniGPT-4 and post-processing is employed to refine the generated descriptions, resulting in approximately 3,500 high-quality image-text pairs for the alignment process.

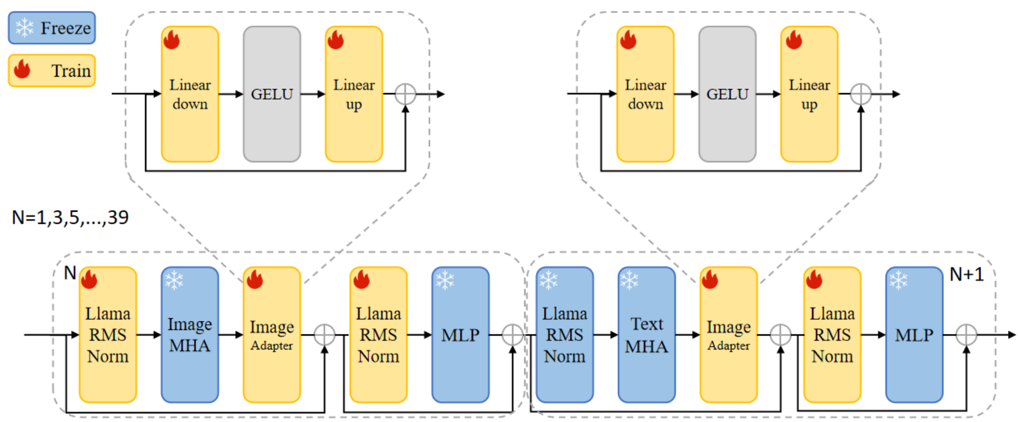


Figure 2: Transformer block of Vicuna with image adapter

3.2 Image Adapter

While MiniGPT-4 has demonstrated impressive image understanding capabilities, it still falls short in comprehending all the image content, particularly in cases where the image contains artistic coloration. We draw inspiration from two sources: the Parameter-efficient fine-tuning technique [17] in NLP and the training of AIM models (image-to-video) [37]. As depicted in Figure 2 and Equation (1), we incorporated an image adapter layer into the MiniGPT-4's original Vicuna model to enhance its ability to comprehend images.

$$Y = Linear_2(GELU(Linear_1(X))) + X \tag{1}$$

where $Linear_1$ and $Linear_2$ denote the Linear down (hidden size to $\frac{1}{4}$hidden size) and Linear up ($\frac{1}{4}$hidden size to hidden size) layers in the graph, respectively, GELU represents the activation function [16], X represents the data calculated after the Multi-Headed Attention (MHA) mechanism, and Y represents the data after the image adapter layer and Y represents the output data after the image adapter layer.

Furthermore, in the training process of ArtGPT-4, the parameters of the RMS Norm layer in front of the MLP are also enabled. This is done to regulate the output of data from the image adapter layer, preventing it from causing gradient explosion and making it difficult to update the model parameters.

To further enhance the Vicuna model's understanding of image information, we also activated the RMS Norm training parameters in the base layer (N=1,3,5,...) of the Transformer block of Vicuna. As we think it's not ideal to directly input image information into the original Text MHA layer, we named the MHA layer with the activated training parameters as Image MHA, to distinguish it from the original Text MHA.

3.3 Training

Our goal remains to enable Language Models to comprehend visual information using pre-trained models. We still follow the parameters of the original MiniGPT-4, and its training steps. Only we use different training data.

Training Data: We use Laion-aesthetic from the LAION-5B [33] dataset, which amounts to approximately 200GB for the first 302 tar files. Laion-aesthetic is a large-scale dataset of images used primarily for training models to evaluate image aesthetics. The dataset includes over 500,000 images sourced from Flickr, each of which has been manually rated for aesthetic quality and affective polarity. Aesthetic quality is rated on a scale from 1 to 10, while affective polarity is rated as positive, neutral, or negative. In addition to image ratings, the dataset also includes metadata such as EXIF information, image tags, and image descriptions.

The First Training Processes: We trained our model using the following hyperparameters: a linear warmup cosine learning rate scheduler with an initial learning rate of 1e-7, the minimum learning rate of 8e-7, and a warmup learning rate of 1e-8. The weight decay was set to 0.05, and the maximum number of training epochs was 2. We used a batch size of 32 for both training and evaluation, with 4 workers. The warmup steps were set to 5000, and there were 5000 iterations per epoch. We only trained on a Tesla A100 for less than 2 hours.

The Second Training Processes: We fine-tuned the ArtGPT-4 using a set of MiniGPT-4's image-text pairs. We employed the same template containing a prompt with a randomly sampled instruction, which allowed our model to generate more natural and reliable responses. We only trained on a Tesla A100 for less than 10 minutes.

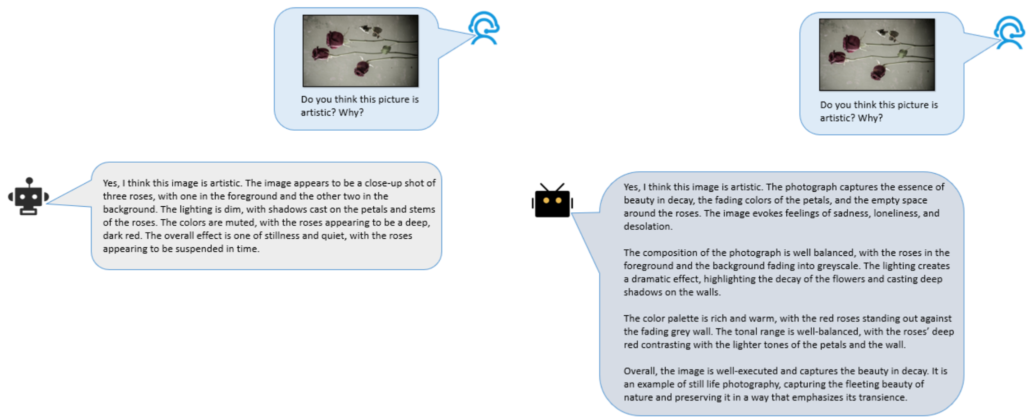
4 Quality Comparison

(a) Talking with MiniGPT-4

(b) Talking with ArtGPT-4

Figure 3: Description of traditional Chinese painting

Image Description: Both MiniGPT-4 and ArtGPT-4 provide detailed descriptions of the images they are given as shown in Figure 3, describing a traditional Chinese ink painting of a landscape scene. First, ArtGPT-4 provides more detailed and specific information about the subject matter of the painting, including details about the rocks, trees, and grasses depicted in the image. Second, ArtGPT-4 makes use of more specialized language to describe the painting, including terms such as "Chinese calligraphy brushstrokes" and "shading and texture of the rocks." This indicates a greater level of expertise and familiarity with the subject matter, which can help to lend credibility and authority to the description. Finally, ArtGPT-4 uses language that emphasizes the emotional and aesthetic impact of the painting, describing it as having a "tranquil" and "beautiful" effect that emphasizes the rugged, wild beauty of the natural landscape. This suggests that ArtGPT-4 is better able to understand and convey the emotional and aesthetic impact of visual art, which could be particularly useful in fields such as art criticism or curation.



(a) Talking with MiniGPT-4 (b) Talking with ArtGPT-4

Figure 4: Appreciation of judgments about aesthetics

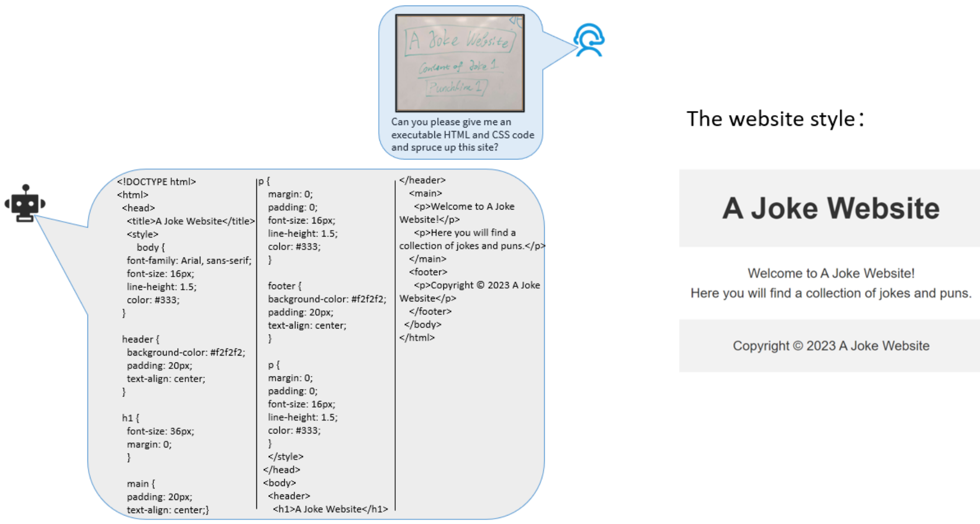
Aesthetics: Both MiniGPT-4 and ArtGPT-4 recognized the artistic qualities of the image as shown in Figure 4, but ArtGPT-4's response is more detailed and descriptive. ArtGPT-4 not only describes the image but also provides an interpretation of it, highlighting the beauty in decay and evoking emotions such as sadness, loneliness, and desolation. In addition, ArtGPT-4 provides a more technical analysis of the image, discussing the composition, lighting, color palette, and tonal range. These details demonstrate a deeper understanding of the elements of visual arts and photography. Overall, ArtGPT-4's response is more nuanced and insightful, showcasing its superior capacity for understanding and analyzing art.

Better-looking Websites: ArtGPT-4's response includes an image that serves as a visual representation of the joke website as shown in Figure 5. The use of the image adds an extra layer of creativity to the website and can help to attract visitors. Additionally, the CSS styling used in ArtGPT-4's response is more comprehensive, providing more visual appeal to the website. In comparison, MiniGPT-4's response uses a more basic CSS styling and does not include an image. The website still looks functional and readable, but it lacks the same level of visual interest that ArtGPT-4's response provides. Overall, ArtGPT-4's response demonstrates a more sophisticated understanding of web design and has more visual appeal than MiniGPT-4's response.

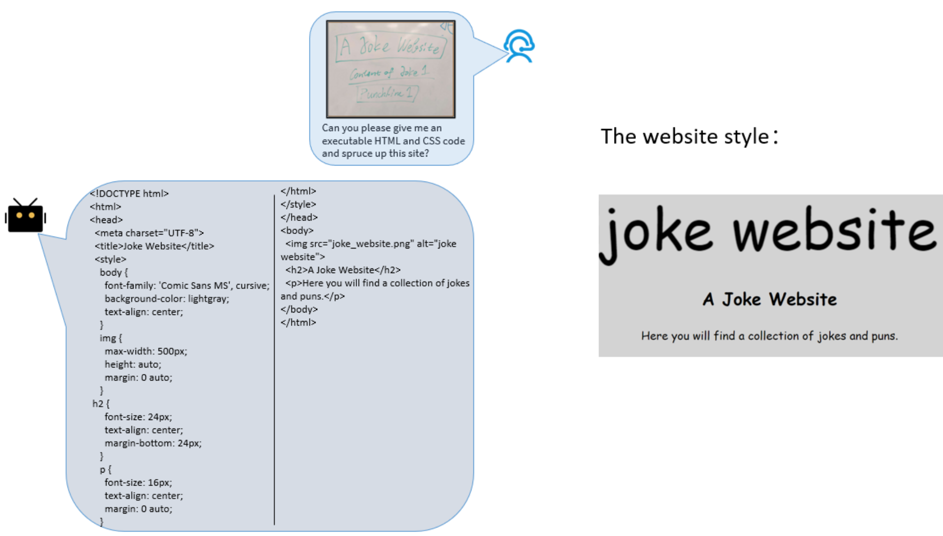
5 Evaluation Benchmarks

Similar to how the TOEFL and IELTS tests are used to measure English language proficiency, we were enthusiastic about establishing a reliable standard for evaluating the ability to comprehend multimodal images. We implemented four scoring criteria and a five-point scoring scale to evaluate the model's capacity for comprehending images.

Image Depiction Capability (IDC): We selected 10 various types of graphs, such as paintings, photographs, AI-generated images, etc., for the model to provide descriptions for. Each image is scored according to the following criteria:



(a) Talking with MiniGPT-4



(b) Talking with ArtGPT-4

Figure 5: About the generation of better-looking websites

- 0: No image description capability
- 1: Description does not match real image representation
- 2: Partial image description
- 3: Complete image description without appreciation information
- 4: Complete image description at the human level of appreciation
- 5: Complete image description surpassing the human level of appreciation, such as an artist.

Image Sentiment Analysis Capability (ISAC): We chose 10 images of individuals and instructed the participants to "Analyze the emotions expressed by the individuals in the images as well as the emotions felt by the viewer observing them." Each image is scored according to the following criteria:

- 0: Can't describe the feelings about the picture
- 1: Can describe the relevant emotion but no logical proof (e.g.: the picture is seen... So people will have a kind of... emotion)
- 2: Can describe the relevant emotion and justify it. But the description is not perfect
- 3: The individuals in the images or the viewer's emotions can be described perfectly and justified.
- 4: Can describe all the emotions as an ordinary person and justify them.
- 5: Can describe all emotions perfectly and justifiably and full of art.

Image Content Recognition Capability (ICRC): We selected 5 images with a variety of objects and scenes, such as animals, landscapes, and household items, for the model to recognize and label. Each image is scored according to the following criteria:

- 0: No image content recognition capability
- 1: Some objects/scenes are recognized but with significant errors or omissions
- 2: Most objects/scenes are recognized with some errors or omissions
- 3: All objects/scenes are recognized with few errors or omissions
- 4: All objects/scenes are recognized with high accuracy and speed, comparable to a human observer
- 5: All objects/scenes are recognized with high accuracy, speed, and contextual understanding, surpassing the performance of a human observer.

Multi-round Dialogue Image Understanding Capability (MDIUC): We randomly select 2 images and conduct five rounds of dialogue with the model for each image to assess its multi-round image understanding capability. Each image-dialogue pair is scored according to the following criteria:

- 0: No image understanding capability in the dialogue
- 1: Partial image understanding, but unable to carry on the dialogue smoothly
- 2: Able to understand the image to some extent and carry on the dialogue with some coherence, but lacks understanding of some key points
- 3: Can understand the image and carry on the dialogue smoothly, but with some minor misunderstandings or mistakes
- 4: Can understand the image and carry on the dialogue smoothly, with accurate understanding and good coherence
- 5: Can understand the image and carry on the dialogue smoothly, with accurate understanding, good coherence, and creative responses.

We incorporated these scoring criteria into the state-of-the-art GPT-4 model to establish a comprehensive multimodal model evaluation system.

6 Evaluation

We separately scored each response based on the benchmarks presented in Section 5 for GPT-4, MiniGPT-4, and ArtGPT-4. Due to space limitations, we could not include the responses in the text. However, the responses generated by MiniGPT-4 and ArtGPT-4 can be found on GitHub at www.github.com as open-source code.

Table 1: IDC benchmarks based on GPT-4 scores.

Items	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Average**
GPT-4(criterion)	4	4	4	4	4	3	4	5	4	5	4.1
MiniGPT-4	2	3	3	1	2	2	4	3	3	3	2.6
ArtGPT-4(Ours)	4	4	3	4	4	3	4	5	3	4	3.8

Table 2: ISAC benchmarks based on GPT-4 scores.

Items	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Average**
GPT-4(criterion)	3	4	3	3	3	3	3	3	3	3	3.1
MiniGPT-4	2	3	2	3	2	2	2	2	3	2	2.3
ArtGPT-4(Ours)	3	3	1	3	3	3	3	3	3	3	2.8

Table 3: ISAC and MDIUC benchmarks based on GPT-4 scores.

Benchmark	**ICRC**						**MDIUC**		
Items	**1**	**2**	**3**	**4**	**5**	**Average**	**1**	**2**	**Average**
GPT-4(criterion)	3	3	3	3	3	3	4	4	4
MiniGPT-4	1	2	2	2	2	1.8	3	2	2.5
ArtGPT-4(Ours)	3	4	2	3	3	3	4	4	4

According to Table 1, Table 2, and Table 3 show, based on the four scoring criteria, ArtGPT-4's image comprehension ability is significantly superior to that of the original MiniGPT-4 and even approaches that of GPT-4. ArtGPT-4 outperformed MiniGPT-4 in all four benchmark tests, with each test weighted at 25 percent. ArtGPT-4 scored 3.4 points while MiniGPT-4 scored 2.35 points, indicating a difference of more than 1 point on a 5-point scale. Notably, ArtGPT-4's final score of 3.4 was only 0.15 points below that of GPT-4. Additionally, the number of Vicuna model counts in ArtGPT-4's language model (13 billion) was significantly lower than that of GPT-4 (>175 billion).

7 Conclusion and Future Work

Our ArtGPT-4 model demonstrates significant progress in the field of vision-language understanding, showing superior performance to its predecessor, MiniGPT-4. Our proposed modifications, including tailored linear layers and activation functions, have optimized the model's performance and addressed the unique challenges posed by vision-language tasks. Additionally, we have introduced a novel benchmarks for evaluating the performance of vision-language models, which provides a more comprehensive criterion for assessing these models. Our model was trained in just 2 hours, using a relatively small dataset, and can generate visually appealing HTML/CSS web pages, making it a promising tool for designers and content creators.

While we have successfully addressed many challenges, there remain a couple of issues. First, our training dataset is limited, particularly in the second phase where we utilize the original MiniGPT-4 dataset. Secondly, our benchmarks only evaluates image understanding in the English language and does not account for other languages.

References

[1] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[3] H. Bahng, A. Jahanian, S. Sankaranarayanan, and P. Isola. Visual prompting: Modifying pixel space to adapt pre-trained models. *arXiv preprint arXiv:2203.17274*, 2022.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[5] J. Chen, H. Guo, K. Yi, B. Li, and M. 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*, pages 18030–18040, 2022.

[6] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *arXiv preprint arXiv:2205.13535*, 2022.

[7] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023.

[8] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel. Palm: Scaling language modeling with pathways, 2022.

[9] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[10] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6101–6111, 2020.

[11] A. D. Cohen, A. Roberts, A. Molina, A. Butryna, A. Jin, A. Kulshreshtha, B. Hutchinson, B. Zevenbergen, B. H. Aguera-Arcas, C. ching Chang, C. Cui, C. Du, D. D. F. Adiwardana, D. Chen, D. D. Lepikhin, E. H. Chi, E. Hoffman-John, H.-T. Cheng, H. Lee, I. Krivokon, J. Qin, J. Hall, J. Fenton, J. Soraker, K. Meier-Hellstern, K. Olson, L. M. Aroyo, M. P. Bosma, M. J. Pickett, M. A. Menegali, M. Croak, M. Díaz, M. Lamm, M. Krikun, M. R. Morris, N. Shazeer, Q. V. Le, R. Bernstein, R. Rajakumar, R. Kurzweil, R. Thoppilan, S. Zheng, T. Bos, T. Duke, T. Doshi, V. Y. Zhao, V. Prabhakaran, W. Rusch, Y. Li, Y. Huang, Y. Zhou, Y. Xu, and Z. Chen. Lamda: Language models for dialog applications. In *arXiv*. 2022.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10696, 2020.

[14] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat, B. Zoph, L. Fedus, M. Bosma, Z. Zhou, T. Wang, Y. E. Wang, K. Webster, M. Pellat, K. Robinson, K. Meier-Hellstern, T. Duke, L. Dixon, K. Zhang, Q. V. Le, Y. Wu, Z. Chen, and C. Cui. Glam: Efficient scaling of language models with mixture-of-experts, 2022.

[15] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, and G. Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021.

[16] D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[17] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

[18] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[19] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim. Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 709–727. Springer, 2022.

[20] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[21] X. L. Li and P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, 2021.

[22] J. Lin, R. Men, A. Yang, C. Zhou, M. Ding, Y. Zhang, P. Wang, A. Wang, L. Jiang, X. Jia, J. Zhang, J. Zhang, X. Zou, Z. Li, X. Deng, J. Liu, J. Xue, H. Zhou, J. Ma, J. Yu, Y. Li, W. Lin, J. Zhou, J. Tang, and H. Yang. M6: A chinese multimodal pretrainer, 2021.

[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[24] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning, 2023.

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3285–3295, 2019.

[26] OpenAI. Introducing chatgpt. https://openai.com/blog/chatgpt, 2022. Accessed: May 3, 2023.

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

[28] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, pages 27730–27744, 2021.

[29] Z. Qing, S. Zhang, Z. Huang, X. Wang, Y. Wang, Y. Lv, C. Gao, and N. Sang. Mar: Masked autoencoders for efficient action recognition. *IEEE Transactions on Multimedia*, 2023.

[30] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1–10, 2018.

[31] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language Models are Unsupervised Multitask Learners. In *Proceedings of the 2019 Neural Information Processing Systems (NeurIPS)*, pages 3657–3667, 2019.

[32] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. In *Advances in Neural Information Processing Systems*, pages 11791–11802, 2019.

[33] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[34] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[35] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

[36] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. Self-instruct: Aligning language model with self generated instructions, 2022.

[37] T. Yang, Y. Zhu, Y. Xie, A. Zhang, C. Chen, and M. Li. AIM: Adapting image models for efficient video action recognition. In *The Eleventh International Conference on Learning Representations*, 2023.

[38] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764, 2019.

[39] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *Transactions of the Association for Computational Linguistics*, volume 2, pages 67–78, 2014.

[40] E. B. Zaken, Y. Goldberg, and S. Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, 2022.

[41] Q. Zhang, J. Zhang, Y. Xu, and D. Tao. Vision transformer with quadrangle attention. *arXiv preprint arXiv:2303.15105*, 2023.

[42] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.