ArtGPT-4: Towards Artistic-understanding Large Vision-Language Models with Enhanced Adapter

Zhengqing Yuan[1] ♠, **Huiwen Xue**[2], **Xinyi Wang**[1], **Kun Wang**[1] ♣ *, **Lichao Sun**[3] ♡ *

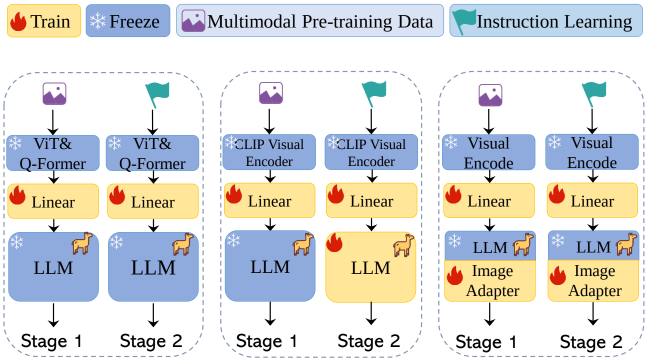
[1] Anhui Polytechnic University, Wuhu, China 241000
[2] Soochow University, Suzhou, China 215000
[3] Lehigh University, Bethlehem, US 18015
♠ Zhengqingyuan@ieee.org, ♣ kun.wang@ahpu.edu.cn, ♡ lis221@lehigh.edu

Abstract

In recent years, advancements in large language models have been remarkable, with models such as ChatGPT demonstrating exceptional proficiency in diverse linguistic tasks. The pre-training of large models with billions of parameters, poses a formidable challenge, primarily due to the scarcity of datasets of a commensurate scale for effective training. Nevertheless, innovative strategies have emerged, including methods to fine-tune these pre-trained models using fewer parameters set, as evidenced by models like MiniGPT-4 and LLaVA. Despite their potential in various domains, these models remain limited in their understanding of artistic imagery. They have yet to fully grasp the intricate nuances of art images or to provide an objective articulation of the emotions they evoke, in a manner akin to human perception. This work introduces ArtGPT-4, a pioneering large vision-language model tailored to address the deficiencies of contemporary models in artistic comprehension. ArtGPT-4 underwent training on image-text pairs utilizing a Tesla A100 device in a mere 2 hours, with a dataset comprising approximately 0.52M entries. Impressively, the model can render images with an artistic-understanding and convey the emotions they inspire, mirroring human interpretation. Additionally, this work presents a unique dataset designed to evaluate the efficacy of vision-language models. In subsequent evaluations, ArtGPT-4 not only achieved state-of-the-art performance on the ArtEmis and ArtEmis-v2.0 datasets but also exceeded the established benchmarks introduced in This study, lagging behind professional artists' descriptions by a negligible 0.15 points on a 6-point scale. The code and the pre-trained model are accessible in the Supplementary Material accompanying this research.

Introduction

Advancements in large language models (LLMs) have revolutionized the field of natural language processing, paving the way for numerous breakthrough applications and sophisticated tasks (Ouyang et al. 2021; Brown et al. 2020; OpenAI 2022). On the other hand, single-modality LLMs, as powerful as they are, represent just one facet of the broader potential of AI. The budding realm of multimodal models, which synergize different data modalities like text and vision, holds promise for a new wave of innovations. Notable works in this direction include research by Lin et al. (Lin et al. 2021) and Radford et al. (Radford et al. 2021), indicating the infancy but undeniable potential of this domain. GPT-4, a monumental achievement by OpenAI, has recently set a benchmark in the vision-language understanding sphere (OpenAI 2023). Its prowess in discerning intricate visual nuances and producing varied, contextually rich language outputs is nothing short of groundbreaking. However, the lack of open-source availability for GPT-4 poses challenges for the broader research community. Without access to its architecture, replicating or building upon its success becomes a convoluted endeavor. Furthermore, the sheer volume of data GPT-4 relies on, amassing over 45 terabytes of text or image information, raises questions about the feasibility of gathering comparable datasets for similar projects. Given these constraints, leveraging existing pre-trained models and fine-tuning them for specific tasks presents itself as a pragmatic and increasingly popular strategy for researchers and practitioners alike.

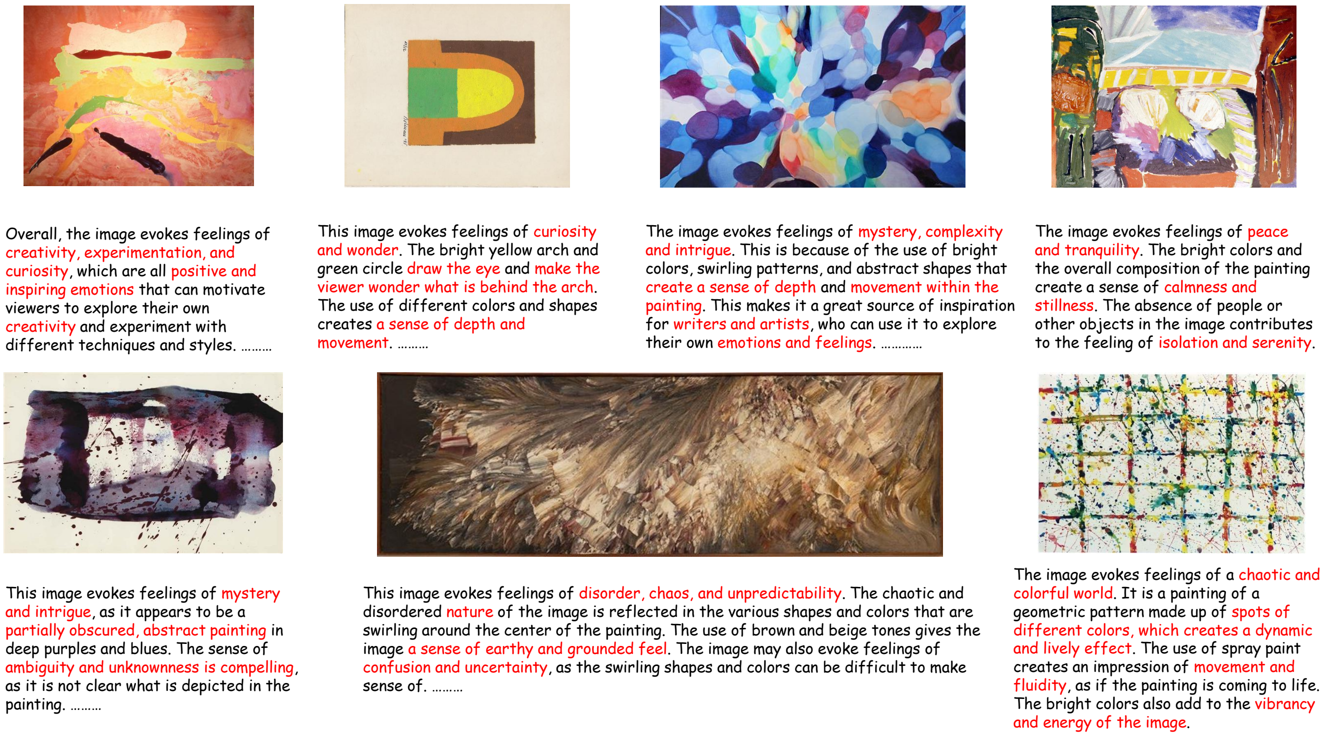


(a) MiniGPT-4 (b) LLaVA (c) ArtGPT-4

Figure 1: Comparison between different structures of multimodal models. All of these methods are trained in a two-stage fashion. Stage 1 stands for pre-training and Stage 2 represents instruction tuning.

As depicted in Figure 1, two primary open source models of end-to-end multimodal LLMs can be identified: 1) MiniGPT4 (Zhu et al. 2023) model, which employs a limited parameter set and maintain frozen LLM and visual models during both pretraining and instruction tuning phases. While it is a notable limitation that this model is its reliance on static visual models and LLMs. This can result in suboptimal align-

[*] Corresponding author



Overall, the image evokes feelings of creativity, experimentation, and curiosity, which are all positive and inspiring emotions that can motivate viewers to explore their own creativity and experiment with different techniques and styles. ....

This image evokes feelings of curiosity and wonder. The bright yellow arch and green circle draw the eye and make the viewer wonder what is behind the arch. The use of different colors and shapes creates a sense of depth and movement. ....

The image evokes feelings of mystery, complexity and intrigue. This is because of the use of bright colors, swirling patterns, and abstract shapes that create a sense of depth and movement within the painting. This makes it a great source of inspiration for writers and artists, who can use it to explore their own emotions and feelings. ....

The image evokes feelings of peace and tranquility. The bright colors and the overall composition of the painting create a sense of calmness and stillness. The absence of people or other objects in the image contributes to the feeling of isolation and serenity.

This image evokes feelings of mystery and intrigue, as it appears to be a partially obscured, abstract painting in deep purples and blues. The sense of ambiguity and unknownness is compelling, as it is not clear what is depicted in the painting. ....

This image evokes feelings of disorder, chaos, and unpredictability. The chaotic and disordered nature of the image is reflected in the various shapes and colors that are swirling around the center of the painting. The use of brown and beige tones gives the image a sense of earthy and grounded feel. The image may also evoke feelings of confusion and uncertainty, as the swirling shapes and colors can be difficult to make sense of. ....

The image evokes feelings of a chaotic and colorful world. It is a painting of a geometric pattern made up of spots of different colors, which creates a dynamic and lively effect. The use of spray paint creates an impression of movement and fluidity, as if the painting is coming to life. The bright colors also add to the vibrancy and energy of the image.

Figure 2: ArtGPT-4 exhibits a remarkable ability to appreciate works of art. It extends beyond merely capturing the artistic details of an image, delving into the realm of emotional understanding. ArtGPT-4 is capable of discerning and articulating the emotions elicited by an image from a human viewer's perspective, such as feelings of positivity and inspiration.

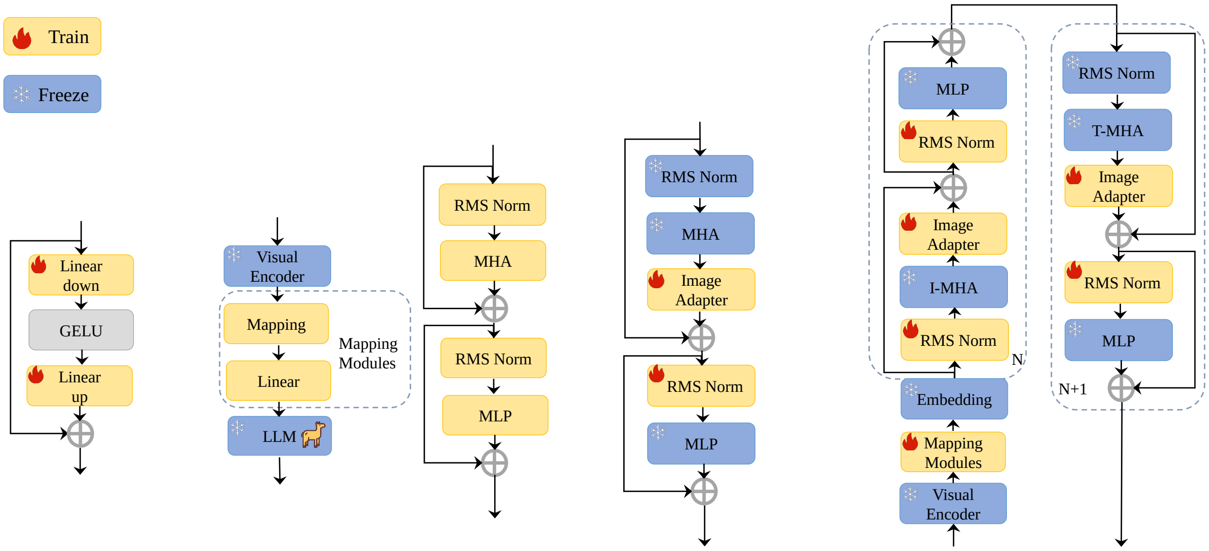
ment due to the constrained parameter count. 2) LLaVA (Liu et al. 2023) model, which incorporates trainable LLMs during instruction tuning while keeping the visual models static. But, a significant challenge with this model is the computational cost, as updating all LLM parameters during training can be resource-intensive. Furthermore, these models demonstrate laudable performance across a broad spectrum of tasks, including image understanding and detail depiction, matching the proficiency of GPT-4, they falter when tasked with the nuanced interpretation of artistic images akin to human perception. Specifically, current multimodal models fall short in capturing the intricate details inherent in an art image and articulating the emotions it elicits in an objective manner akin to a human observer.

This study proposes ArtGPT-4, a novel model designed to address these limitations of existing multimodal models. ArtGPT-4 incorporates tailored linear layers and their corresponding activation functions exclusively into the language model, in tandem with the activation of specific training parameters. These modifications were strategically implemented to optimize the model's performance and equip it to effectively tackle the challenges of artistic understanding inherent in vision-language tasks. Trained on a Tesla A100 device in a mere span of 2 hours, ArtGPT-4 utilized only 0.52M image-text pairs, amounting to about 200GB. The model can depict images with an enhanced artistic flair and convey the emotions they inspire, as shown in Figure 2. Subsequent evaluation methods revealed that ArtGPT-4 outperforms existing models in the realm of artistic image understanding. Our contributions are as follows:

- This work pioneers the exploration of artistic understanding within multimodal models. It addresses the inherent limitations of these models, which, until now, have struggled to comprehensively grasp the intricate nuances of artistic imagery and to objectively articulate the emotions they elicit in a manner reminiscent of human perception.
- This study is the inaugural effort to introduce a parameter-efficient fine-tuning method exclusively for the language components of multimodal models. The approach has yielded remarkable outcomes, effectively addressing the challenges associated with the extensive time and resource demands of training large visual-language models.
- This research introduces a novel dataset tailored for assessing the visual comprehension capabilities of multimodal models. This approach holds potential for a more in-depth evaluation of expansive vision-language models

Related Work

Vision-Language Model. In recent years, the pursuit of models with capabilities transcending a single domain has gained momentum. A notable exemplar in this realm is OpenAI's CLIP (Radford et al. 2021), which pioneered the synergy between visual and linguistic understanding by associa-



(a) Image Adapter

(b) Multimodal Model Structure

(c) LLaMA Series Model Structure

(d) Adapter Efficient Fine-tuning for NLP

(e) ArtGPT-4 Model Structure

Figure 3: We show how we adapt the LLM (c) of Multimodal Model Structure (b) using the Adapter Efficient Fine-tuning method in NLP to model the ArtGPT-4 (e). During training, only newly added Image Adapters (a) and partial normalization layer (e) are updated while all the other layers are frozen.

tively training on image-text pairs. Building on such foundational work, researchers have further delved into models that empower language architectures with image comprehension capabilities (Chen et al. 2022a; Alayrac et al. 2022; Tsimpoukelli et al. 2021). Innovative training methodologies have emerged for these multimodal models. A case in point is BLIP-2 (Li et al. 2023), which introduces a versatile and efficient pre-training paradigm for vision-language endeavors. This approach capitalizes on readily available frozen pre-trained image encoders and expansive language models, complemented by a nimble Q-Former for mapping modules to bridge the modality chasm, as depicted in Figure 2 (b). Notably, MiniGPT-4, utilizing the BLIP-2 architecture, harnesses the capabilities of a pre-trained ViT, Q-Former, and integrates with the Vicuna model to achieve profound image understanding proficiencies.

Efficient Fine-tuning. Parameter-efficient fine-tuning techniques (Houlsby et al. 2019; Zaken, Goldberg, and Ravfogel 2022; Li and Liang 2021; He et al. 2021; Qing et al. 2023) have gained traction in the NLP domain. These methods aim to minimize the number of learning parameters and computational resources needed for downstream task adaptation, yet they achieve results comparable to full fine-tuning. In the realm of computer vision, there has been a surge in research focused on efficient learning. Works by Jia et al. (Jia et al. 2022), Bahng et al. (Bahng et al. 2022), and Chen et al. (Chen et al. 2022b) have delved into visual adaptation using methodologies akin to those in NLP. However, it's pivotal to note that these studies primarily focus on adaptations within the same modality such as text-to-text, image-to-image, video-to-video, or within the same domain (Yang et al. 2023) like image-to-video.

ArtGPT-4

In this section, we will detail the structure of ArtGPT-4, and the training steps of ArtGPT-4, illustrating how we use the enhanced Adapter layer to construct visual-language multimodal models.

Image Adapter

Drawing inspiration from the advancements in finetuning techniques within the realms of Natural Language Processing (NLP) and Computer Vision (CV) (Houlsby et al. 2019; Zaken, Goldberg, and Ravfogel 2022; Li and Liang 2021; He et al. 2021; Qing et al. 2023; Yang et al. 2023), we have integrated the Adapter mechanism, as proposed by Houlsby et al.(Houlsby et al. 2019). Figure 3(a) depicts the Adapter's architecture, characterized by its bottleneck design. This design consists of two linear layers separated by an activation layer. The primary function of the initial linear layer is to diminish the input's dimensionality, whereas the subsequent layer restores it to its original dimension, as shown in Equation 1.

$$\begin{aligned}\mathbf{Y}_{adp} &= \text{Adapter}(\mathbf{Y}_{MHA};\mathbf{W}_{down},\mathbf{W}_{up})\\ &= \mathbf{W}_{up}\left(\text{GELU}\left(\mathbf{W}_{down}\mathbf{Y}_{MHA}\right)\right)+\mathbf{Y}_{MHA}\end{aligned}\tag{1}$$

where $\mathbf{Y}_{MHA}$ represents the data computed from the multi-head attention layer (MHA). The symbols $\mathbf{W}_{down}$ and $\mathbf{W}_{up}$ respectively signify the trainable weight matrices responsible for input dimension reduction and its subsequent restoration.

To fine-tune pre-existing models for downstream NLP tasks more efficiently, we positioned an Adapter subsequent to the MHA layer, as shown in Figure 3(c) and (d). To further ensure training stability after the integration of the Adapter, we will update the parameters of the normalization layer, specifically the RMS Norm. The computational representation of this block is

$$\begin{aligned}\mathbf{Y}_{RMS} &= \text{RMS Norm}_2\left(\mathbf{Y}_{adp};\mathbf{W}_{RMS}\right)\\ &= \mathbf{W}_{RMS}\odot\frac{\mathbf{Y}_{adp}}{\sqrt{\text{mean}\left(\mathbf{Y}_{adp}^2\right)+\epsilon}}\end{aligned} \tag{2}$$

where $\mathbf{Y}_{RMS}$ is the resultant data post-RMS Norm normalization, and $\mathbf{W}_{RMS}$ is the learnable weight matrix. The symbol $\odot$ denotes elemental multiplication, and ϵ is a minuscule constant introduced to prevent zero denominators.

As shown in Figure 3(b), to capture image and instruction information, mapping modules are usually incorporated prior to pre-trained language models because it is commonly believed that language models can embed visually structured information from tokens (Radford et al. 2021; Li et al. 2022). As shown in the following formula

$$\begin{aligned}&\mathbf{Y}_v = \text{Visual Encode}\left(\mathbf{X}_v\right)\\ &\mathbf{X}_{image-token} = \mathbf{W}_{Map}\mathbf{Y}_v\\ &\mathbf{X} = \left[\mathbf{X}_{image-token},\mathbf{X}_{text-token}\right]\end{aligned} \tag{3}$$

where $\mathbf{X}_v$ denotes the image information, Visual Encode$(\cdot)$ denotes the visual encoder model, $\mathbf{W}_{Map}$ denotes the weight matrix of the trainable mapping layer, $\mathbf{X}_{text-token}$ denotes the textual information, and $\mathbf{X}$ denotes the information that will be inputted to the LLMs containing the image and text.

However, the introduction of new mapping modules which need to fully fine-tune LLM can lead to an excessive number of adjustable parameters (Liu et al. 2023) or result in inadequate alignment when freezing full LLM during training (Zhu et al. 2023). To tackle these challenges, we propose two novel strategies: 1) repurposing the pre-trained self-attention layer in the language model for visual modeling, and 2) introducing new Image Adapter modules to these pre-trained language models. More specifically, in the first strategy, we denote the original self-attention layer as T-MHA for linguistic modeling, and the reused T-MHA layer as I-MHA for visual modeling. As depicted in Figure 3(e), I-MHA precedes T-MHA. The primary differentiation between T-MHA and I-MHA is their data input processing. Specifically, input data of I-MHA, $\mathbf{Y}_{RMS1}$, employs an RMS Norm with learnable parameters, enhancing the normalization of image-containing tokens and thereby boosting the computational efficiency of I-MHA.

$$\begin{aligned}&\mathbf{Y}_{RMS1} = \text{RMS Norm}_{\text{I1}}\left(\mathbf{X};\mathbf{W}_{RMS1}\right)\\ &\mathbf{Y}_{I-MHA} = \text{I-MHA}\left(\mathbf{Y}_{RMS1}\right)\end{aligned} \tag{4}$$

Where I-MHA $(\cdot)$ represents the T-MHA layer, which remains static in terms of parameter updates. $\mathbf{Y}_{I-MHA}$ denotes the output data post I-MHA computation.

To further enhance the alignment between image and text data for the second strategy, ArtGPT-4 incorporates trainable adapters, termed the Image Adapter, as showcased in Figure 3(e). The computation of this block can be written as

$$\begin{aligned}&\mathbf{Y}_{T-MHA} = \text{T-MHA}\left(\text{RMS Norm}_{\text{T1}}\left(\mathbf{Y}\right)\right)\\ &\mathbf{Y}_{Iadp} = \text{Image Adapter}\left(\mathbf{Y}_{I-MHA/T-MHA};\mathbf{W}_{down},\mathbf{W}_{up}\right)\\ &\mathbf{Y}_{RMS2} = \text{RMS Norm}_2\left(\mathbf{Y}_{Iadp};\mathbf{W}_{RMS2}\right)\end{aligned} \tag{5}$$

In this representation, T-MHA $(\cdot)$ signifies the T-MHA layer, and RMS Norm$_{\text{T1}}$ $(\cdot)$ indicates the RMS Norm layer preceding the T-MHA layer. Both these layers remain unaltered in terms of parameter updates. The symbols $\mathbf{Y}$, $\mathbf{Y}_{T-MHA}$, $\mathbf{Y}_{Iadp}$, and $\mathbf{Y}_{RMS2}$ represent the input data, post T-MHA layer data, post Image Adapter output data, and post second RMS Norm layer data, respectively.

Training

ArtGPT-4 remains to enable Language Models to artistic-understand visual information using pre-trained models. We still follow the parameters of the original Multimodal Model, like MiniGPT-4 or LLaVA, and its training steps. Only we use different training data and model structures based on these original parameters, as shown in Figure 1(c).

Training Data. We use Laion-aesthetic from the LAION-5B (Schuhmann et al. 2022) dataset, which amounts to approximately 200GB for the first 0.52M data. The aesthetic of this dataset quality is a scale from 7 to 10, while affective polarity is rated as positive, neutral, or negative. In addition to image ratings, the dataset also includes metadata such as image tags and image descriptions.

The First Stage 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, a 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 using the Laion-aesthetic dataset.

The Second Stage Training Processes. We fine-tuned the ArtGPT-4 using a set of MiniGPT-4 or LLaVA's image-text pairs and instructions, such as "*<Img><ImageHere></Img> Take a look at this image and describe what you notice.###Assistant.*"

We employed the same template containing a prompt with a randomly sampled instruction, which allowed our model to generate more natural and reliable responses. Specifically, as shown in Equation 6:

$$p\left(\mathbf{X}_a \mid \mathbf{X}_v,\mathbf{X}_{instr}\right) = \prod_{i}^{L} p_\theta\left(x_i \mid \mathbf{X}_v,\mathbf{X}_{a,<i},\mathbf{X}_{instr}\right) \tag{6}$$

where $\mathbf{X}_{instr}$ denotes the instruction randomly selected from Table 1 in the Supplementary Material, $\mathbf{X}_a$ denotes the answer to the image by the model for that instruction, θ denotes the training parameters of the model, and $\mathbf{X}_{a,<i}$ are the instruction and answer tokens from all previous rounds prior to the current prediction token x_i. We only trained on a Tesla A100 for less than 10 minutes.

Evaluation

Zero-shot Testing Datasets

ArtEmis (Achlioptas et al. 2021) delves into the intricate relationship between visual content, emotional impact, and language-based explanations. The data of This dataset with annotators indicating dominant emotions and providing grounded verbal explanations. ArtEmis comprises 455K emotion attributions and explanations on 80K artworks from WikiArt (Tan et al. 2019). **ArtEmis-v2.0** (Mohamed et al. 2022) builds upon the original ArtEmis dataset by employing a novel contrastive data collection approach. By balancing emotional biases and incorporating 260,533 new instances with contrasting emotions, the dataset achieves a more fine-grained representation of emotions and associated painting explanations. Furthermore, to further enrich our evaluation, we filtered images from the mPLUG-Owl (Ye et al. 2023) database and others, in total the 40 image-instruction data called **ArtMM**. These images, characterized by their complex elements, were a mix of those found online and others generated using DALL-E 2 (Ramesh et al. 2022).

Baselines

MiniGPT-4 (Zhu et al. 2023). It is a streamlined model that merges a visual encoder with the Vicuna language model, showcasing multi-modal abilities akin to GPT-4. Through fine-tuning with a quality dataset and conversational approach.

LLaVA (Liu et al. 2023). LLaVA is a large multimodal model combining a vision encoder with an LLM, utilizing GPT-4 to generate multimodal instruction-following data. Early tests indicate LLaVA's exceptional chat abilities, rivaling multimodal GPT-4 in some areas and achieving state-of-the-art results in many areas.

Mulit-modal GPT (Gong et al. 2023). This is a vision and language model designed for multi-round dialogues with humans, which is fine-tuned from OpenFlamingo with the addition of Low-rank Adapter (LoRA) in both cross- and self-attention area.

VisualGPT (Chen et al. 2022a). It is a data-efficient image captioning model that utilizes linguistic knowledge from a large pretrained language model. Its solution presented is a unique self-resurrecting encoder-decoder attention mechanism that adapts the pretrained language model with limited in-domain data.

GIT (Wang et al. 2022). GIT is designed to merge vision-language tasks like image captioning and question answering. It simplifies the design with just an image encoder and text decoder, focusing on a singular language modeling task.

ViLT (Kim, Son, and Kim 2021). It a streamlined Vision-and-Language Transformer (ViLT). This model lies in its monolithic nature, wherein the handling of visual data is radically streamlined, eliminating the need for convolution and treating it similarly to textual data.

Evaluation Metrics

VADER. For VADER-based similarity (Hutto and Gilbert 2014), we utilized the VADER sentiment analyzer to compute the compound sentiment score for each response. The compound sentiment score represents the overall sentiment of the response, ranging from -1 (most negative) to 1 (most positive). The sentiment similarity between two responses, model response and Labeling of data sets, was then calculated as the absolute difference between their respective compound sentiment scores. A higher score indicates better performance.

TextBlob. For TextBlob-based similarity (Loria et al. 2018),, we used the TextBlob library to analyze the polarity (sentiment score) of each response. The polarity ranges from -1 (most negative) to 1 (most positive). Similar to VADER-based similarity, the sentiment was computed as the absolute difference between their polarity scores. Higher scores indicate better performance.

BERT. For BERT-based similarity, we employed the SentenceTransformer BERT (Reimers and Gurevych 2019) model to encode text into high-dimensional embeddings. We then utilized the cosine similarity metric to quantify the similarity between their embeddings. The cosine similarity, as shown in Equation (7) ranges from -1 (completely dissimilar) to 1 (identical). Higher scores indicate better performance.

$$\cos\theta = \frac{\mathbf{Y}_A \cdot \mathbf{Y}_B}{\|\mathbf{Y}_A\| \times \|\mathbf{Y}_B\|} \tag{7}$$

Where $\mathbf{A}$ denotes the word embedding vector of the model responses and $\mathbf{B}$ denotes the word embedding vector of the labels in the dataset. The $\mathbf{A} \cdot \mathbf{B}$ denotes the dot product of vectors $\mathbf{A}$ and $\mathbf{B}$ and $\|\cdot\|$ denotes the Euclidean norm.

Artistic-Understanding Evaluation

Evaluation on ArtEmis and ArtEmis-v2.0. Evaluation results for artistic-understanding are presented in Table 1. We evaluate our proposed method against six state-of-the-art multimodal models on two art image explanation datasets: ArtEmis and ArtEmis-v2.0. About our proposed ArtGPT-4, all experiments utilized models pre-trained by either MiniGPT-4 or LLaVA, with training settings drawn from the Training section. From our results, we can observe that: 1) ArtGPT-4 consistently outperforms baseline models across all evaluation metrics on both ArtEmis and ArtEmis-v2.0 datasets. Specifically, the ArtGPT-4 model (Backbones on MiniGPT-4-Vicuna-13B) sets a new performance benchmark, outstripping the original MiniGPT-4-Vicuna-13B by a considerable margin. For instance, on the ArtEmis dataset, the ArtGPT-4 variant achieved VADER, TextBlob, and BERT scores of 0.813, 0.247, and 0.693 respectively. On the ArtEmis-v2.0 dataset, the scores were 0.987, 0.360, and 0.698 respectively, showcasing its superior performance. Furthermore, for different LLMs like MiniGPT-4-Vicuna-7B and Alpaca-7B, ArtGPT-4 attained VADER scores of 0.740 and 0.721, respectively, showcasing significant enhancements over the original models. 2) pitted against other state-of-the-art multimodal models, ArtGPT-4 (Backbones on LLaVA) surpassed the original LLaVA with 13B parameters, achieving VADER, TextBlob, and BERT scores of 0.799, 0.245, and 0.691 on the ArtEmis dataset. The model also outperformed the Multi-modal GPT, which had 1.66B parameters, despite only having 0.56B updated parameters. Most impressively, ArtGPT-4 (Backbones on MiniGPT-4-Vicuna-13B) eclipsed the performance of prior leading models such as VisualGPT,

Methods	Pretraing	Learnable	ArtEmis			ArtEmis-v2.0		
	Language model	parameters (**B**illion)	VADER	TextBlob	BERT	VADER	TextBlob	BERT
MiniGPT-4 (Zhu et al. 2023)	Vicuna-13B	0.003B	0.746	0.242	0.693	0.939	0.332	0.673
MiniGPT-4 (Zhu et al. 2023)	Vicuna-7B	0.003B	0.704	0.225	0.684	0.901	0.321	0.660
MiniGPT-4 (Zhu et al. 2023)	Alpaca-7B	0.003B	0.660	0.211	0.666	0.853	0.319	0.649
LLaVA (Liu et al. 2023)	Vicuna-13B	13B	0.740	0.240	0.688	0.935	0.332	0.673
Multi-modal GPT (Gong et al. 2023)	OpenFlamingo-9B	1.66B	0.701	0.234	0.592	0.780	0.301	0.599
VisualGPT (Chen et al. 2022a)	GPT-2-small	0.124B	0.105	0.098	0.122	0.141	0.101	0.155
GIT (Wang et al. 2022)	-	0.7B	0.101	0.072	0.110	0.155	0.100	0.139
ViLT (Kim, Son, and Kim 2021)	BERT	0.0874B	0.024	0.011	0.104	0.082	0.110	0.109
ArtGPT-4 (Backbones MiniGPT-4)	Vicuna-7B	0.26B	0.740	0.233	0.686	0.920	0.327	0.665
ArtGPT-4 (Backbones MiniGPT-4)	Alpaca-7B	0.26B	0.721	0.233	0.685	0.923	0.326	0.669
ArtGPT-4 (Backbones LLaVA)	Vicuna-13B	0.52B	0.799	0.245	0.691	0.982	0.350	0.689
ArtGPT-4 (Backbones MiniGPT-4)	Vicuna-13B	0.52B	**0.813**	**0.247**	**0.693**	**0.987**	**0.360**	**0.698**

Table 1: Evaluation on ArtEmis and ArtEmis-v2.0 with six state-of-the-art multi-modal models.

Methods	IDC		ISAC		ICRC		MDIUC		Total average
	sum	average	sum	average	sum	average	sum	average	
Artist (Human)	41	4.1	31	3.1	15	5.0	8	4.0	4.05
MiniGPT-4 (Vicuan 13B)	26	2.6	23	2.3	9	3.0	5	2.5	2.60
MiniGPT-4 (Vicuan 7B)	23	2.3	21	2.1	9	3.0	5	2.5	2.48
LLaVA (Vicuan 13B)	25	2.5	24	2.4	9	3.0	4	2.0	2.60
ArtGPT-4 (Backbones MiniGPT-4-Vicuna-7B)	35	3.5	24	2.4	12	4.0	8	4.0	3.78
ArtGPT-4 (Backbones LLaVA-Vicuna-13B)	38	3.8	25	2.5	15	5.0	8	4.0	3.83
ArtGPT-4 (Backbones MiniGPT-4-Vicuna-13B)	38	3.8	28	2.8	15	5.0	8	4.0	**3.90**

Table 2: Scoring of model outputs using the GPT-4 based on ArtMM for the scoring rules in the Supplementary Materials. We consider these four metrics, including Image Depiction Capability (IDC), Image Sentiment Analysis Capability (ISAC), Image Content Recognition Capability (ICRC), and Multi- round Dialogue Image Understanding Capability (MDIUC) to be equally important, and the total average is calculated as the mean of the average scores of the four metrics.

GIT, and ViLT. As a point of comparison, in the ArtEmis-v2.0 VADER metric, ArtGPT-4's score of 0.987 significantly outpaced VisualGPT's 0.141, GIT's 0.155, and ViLT's 0.082, reflecting a leap in performance by over sevenfold.

Evaluation on ArtMM. As illustrated in Table 2, ArtGPT-4 (Backbones on MiniGPT-4-Vicuna-13B) boasts an impressive artistic-understanding ability with an average score of 3.90. This is markedly superior to the 2.60 average of the original MiniGPT-4 (Vicuna-7B) and is tantalizingly close to the 4.05 average achieved by human artists. For this comparison, we enlisted 10 artists (comprising 5 males and 5 females) to interpret images using the same guidelines provided to the model. In addition, other variants of ArtGPT-4 exhibited a notable average improvement when contrasted with other baseline models, including MiniGPT-4-Vicuna-7B and LLaVA-Vicuna-13B.

Evaluations on Components

Backbone via MiniGPT-4-Vicuna-13B. We conducted four experiments using MiniGPT-4-Vicuna-13B as backbones. 1) While only the model mapping module was fine-tuned, we observed scores of 0.746, 0.242, and 0.692 on VADER, TextBlob, and BERT metrics respectively for ArtEmis. 2) By fine-tuning the initial five layers of the LLM, the scores slightly increased to 0.750, 0.242, and 0.681. 3) When the entire LLM was fine-tuned over a span of two hours, the scores were 0.752, 0.242, and 0.680. 4) While the entire LLM was fine-tuned without any time constraints, we saw significant improvements with scores reaching 0.815, 0.250, and 0.693. ArtGPT-4 (Backbones on MiniGPT-4-Vicuna) exhibited the most robust performance in all groups, especially in the fourth. Notably, with just two hours of fine-tuning on 0.56B parameters, it demonstrated performance comparable to unrestricted fine-tuning of the LLM's 13B parameters.

Backbone via MiniGPT-4-Vicuna-7B and -Alpaca-7B. For MiniGPT-4-Vicuna-7B, the scores were 0.705, 0.225, and 0.683 in the first experiment, which improved to 0.710, 0.227, and 0.686 in the second. In the third experiment, the scores reached 0.740, 0.234, and 0.686. On the other hand, with the MiniGPT-4 Alpaca-7B model, we observed scores of 0.681, 0.221, and 0.670 in the first experiment. These scores slightly increased to 0.683, 0.222, and 0.670 in the second, and further to 0.722, 0.230, and 0.680 in the third experiment. Impressively, our ArtGPT-4, trained on only 0.26B parameters, achieved significant performance gains in both MiniGPT-4-Vicuna-7B and Alpaca-7B models, especially on the ArtEmis dataset. The ArtGPT-4 model even surpassed the results of the full fine-tuning MiniGPT-4-Alpaca model.

Backbone via LLaVA-Vicuna-13B. We investigated the performance of models backbones on the LLaVA architecture pre-trained with the Vicuna-13B language model. Initially, the LLaVA model without any further fine-tuning achieved VADER, TextBlob, and BERT scores of 0.740, 0.240, and 0.688 respectively on the ArtEmis dataset. In a subsequent experiment, the entire LLM underwent fine-tuning in stage 1. This resulted in improved scores, where VADER reached 0.800, TextBlob was at 0.245, and BERT scored 0.690. This demonstrated the potential enhancements achievable

Methods	Pretraing Language model	Learnable parameters (**B**illion)	ArtEmis VADER	 TextBlob	 BERT	ArtEmis-v2.0 VADER	 TextBlob	 BERT
MiniGPT-4+Only Fine-tune Mapping Modules	Vicuna-13B	0.003B	0.746	0.242	0.692	0.939	0.332	0.674
+Fine-tune LLM Top 5 layers	Vicuna-13B	1.58B	0.750	0.242	0.681	0.944	0.333	0.673
+Fine-tune all LLM (2 hours)	Vicuna-13B	13B	0.752	0.242	0.680	0.945	0.332	0.674
+Fine-tune all LLM	Vicuna-13B	13B	0.815	0.250	0.693	0.988	0.359	0.695
ArtGPT-4 (Backbones MiniGPT-4)	Vicuna-13B	**0.52B**	**0.813**	**0.247**	**0.693**	**0.987**	**0.360**	**0.698**
MiniGPT-4+Only Fine-tune Mapping Modules	Vicuna-7B	0.003B	0.705	0.225	0.683	0.911	0.322	0.660
+Fine-tune LLM Top 5 layers	Vicuna-7B	1.01B	0.710	0.227	0.686	0.914	0.322	0.664
+Fine-tune all LLM	Vicuna-7B	7B	0.740	0.234	0.686	0.921	0.326	0.665
ArtGPT-4 (Backbones MiniGPT-4)	Vicuna-7B	**0.26B**	**0.740**	**0.233**	**0.686**	**0.920**	**0.327**	**0.665**
MiniGPT-4+Only Fine-tune Mapping Modules	Alpaca-7B	0.003B	0.681	0.221	0.670	0.853	0.318	0.649
+Fine-tune LLM Top 5 layers	Alpaca-7B	1.01B	0.683	0.222	0.670	0.855	0.320	0.650
+Fine-tune all LLM	Alpaca-7B	7B	0.722	0.230	0.680	0.920	0.328	0.650
ArtGPT-4 (Backbones MiniGPT-4)	Alpaca-7B	**0.26B**	**0.721**	**0.233**	**0.685**	**0.923**	**0.326**	**0.669**
LLaVA (Liu et al. 2023)	Vicuna-13B	13B	0.740	0.240	0.688	0.935	0.332	0.673
+Fine-tune all LLM in Stage 1	Vicuna-13B	13B	0.800	0.245	0.690	0.987	0.352	0.692
ArtGPT-4 (Backbones LLaVA)	Vicuna-13B	**0.52B**	**0.799**	**0.245**	**0.691**	**0.982**	**0.350**	**0.689**

Table 3: Effectiveness of proposed components. We compare to baselines on ArtEmis and ArtEmis-v2.0 datasets.

Methods	Trainable parameters (Billion)	ArtEmis VADER	 TextBlob	 BERT	ArtEmis-v2.0 VADER	 TextBlob	 BERT
ArtGPT-4 (Backbones MiniGPT-4-Vicuna-13B)	0.52B	0.813	0.247	0.693	0.987	0.360	0.698
- Freeze all RMS Norm (before I-MHA)	0.52B	0.801	0.239	0.690	0.977	0.358	0.696
+ Train all RMS Norm (before I-MHA)	0.52B	0.813	0.247	0.693	0.987	0.360	0.698
- Freeze all RMS Norm (following Image Adapter)	0.52B	- (Vanishing gradient)					
- Remove all Image Adapter	0.0004B	0.746	0.242	0.693	0.939	0.333	0.673
- Remove 1/2 Image Adapter	0.26B	0.771	0.243	0.689	0.945	0.338	0.688
- Remove 1/4 Image Adapter	0.14B	0.747	0.240	0.688	0.940	0.333	0.671
+ Add 1/2 Image Adapter	0.78B	0.820	0.247	0.694	0.988	0.360	0.698
+ Add 1/4 Image Adapter	0.65B	0.813	0.247	0.693	0.987	0.360	0.698

Table 4: Scores of ablation experiments for each module on the dataset.

with fine-tuning. When we observed the ArtGPT-4 model, backbones on the LLaVA architecture, and updated with only an additional 0.52B parameters, it showcased VADER, TextBlob, and BERT scores of 0.799, 0.245, and 0.691 respectively. These results not only surpassed the original LLaVA model's performance by notable margins across all metrics but also closely matched the performance of the extensively fine-tuned LLMs in the second group, which employed a whopping 13B parameters. This illustrates the efficiency and potential of the ArtGPT-4 model in leveraging smaller parameter updates for significant performance gains.

Ablation

We present the results of our ablation experiments conducted to analyze the impact of different modules on the performance of ArtGPT-4. Table 4 displays the scores obtained for each module on the ArtEmis and ArtEmis-v2.0 datasets. To investigate the effects of various components, we conducted ablations by modifying the model as follows: When we removed RMS normalization before the Image Multi-Head Attention (I-MHA) layer, the performance on VADER dropped from 0.813 to 0.801, while TextBlob's score slightly decreased from 0.247 to 0.239. However, BERT's score remained almost the same at 0.693. Turning the RMS Norm layer before I-MHA on for training produced identical results, with VADER at 0.813, TextBlob at 0.247, and BERT at 0.693. However, when RMS normalization was removed following the Image Adapter layer, the gradients vanished during training, leading to a lack of meaningful results. The Image Adapter plays a crucial role in the overall performance of ArtGPT-4. Completely removing the Image Adapter resulted in a drop in VADER to 0.746, TextBlob to 0.242, and BERT remained consistent at 0.693. When we removed half of the Image Adapter, the performance on VADER was 0.771, TextBlob was 0.243, and BERT was 0.689. Removing a quarter of the Image Adapter caused VADER to drop to 0.747, TextBlob to 0.240, and BERT to 0.688. On the other hand, adding half of the Image Adapter boosted the VADER score to 0.820, and adding a quarter resulted in a VADER score of 0.813. However, in both cases, TextBlob and BERT scores remained consistent at 0.247 and 0.693 respectively. Adding few Image Adapter parameters results in inadequate performance, while adding many does not yield significant improvements and can lead to an more parameters.

Conclusion

Our experimental results demonstrates significant progress of ArtGPT-4 in the field of vision-language understanding, showing superior performance to its predecessor, MiniGPT-4 or LLaVA. Our proposed modifications, including added adapter image layers, have optimized the model's performance and addressed the artistic-understanding challenges posed by vision-language tasks. Additionally, we have introduced a novel benchmark 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 achieved state-of-the-art performance. And our training method can be applied to different multimodal models. This work effectively bridges the gap between art and LLM.

References

Achlioptas, P.; Ovsjanikov, M.; Haydarov, K.; Elhoseiny, M.; and Guibas, L. 2021. ArtEmis: Affective Language for Visual Art. *CoRR*, abs/2101.07396.

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

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

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

Chen, J.; Guo, H.; Yi, K.; Li, B.; and Elhoseiny, M. 2022a. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18030–18040.

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

Gong, T.; Lyu, C.; Zhang, S.; Wang, Y.; Zheng, M.; Zhao, Q.; Liu, K.; Zhang, W.; Luo, P.; and Chen, K. 2023. MultiModal-GPT: A Vision and Language Model for Dialogue with Humans. arXiv:2305.04790.

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

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

Hutto, C.; and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, 216–225.

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

Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *International Conference on Machine Learning*, 5583–5594. PMLR.

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

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. arXiv:2201.12086.

Li, X. L.; and Liang, P. 2021. 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)*, 4582–4597.

Lin, J.; Men, R.; Yang, A.; Zhou, C.; Ding, M.; Zhang, Y.; Wang, P.; Wang, A.; Jiang, L.; Jia, X.; Zhang, J.; Zhang, J.; Zou, X.; Li, Z.; Deng, X.; Liu, J.; Xue, J.; Zhou, H.; Ma, J.; Yu, J.; Li, Y.; Lin, W.; Zhou, J.; Tang, J.; and Yang, H. 2021. M6: A Chinese Multimodal Pretrainer. arXiv:2103.00823.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. arXiv:2304.08485.

Loria, S.; et al. 2018. textblob Documentation. *Release 0.15*, 2(8): 269.

Mohamed, Y.; Khan, F. F.; Haydarov, K.; and Elhoseiny, M. 2022. It is Okay to Not Be Okay: Overcoming Emotional Bias in Affective Image Captioning by Contrastive Data Collection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume abs/2204.07660.

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

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2021. Training Language Models to Follow Instructions with Human Feedback. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 27730–27744.

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

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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

Tan, W. R.; Chan, C. S.; Aguirre, H.; and Tanaka, K. 2019. Improved ArtGAN for Conditional Synthesis of Natural Image and Artwork. *IEEE Transactions on Image Processing*, 28(1): 394–409.

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

Wang, J.; Yang, Z.; Hu, X.; Li, L.; Lin, K.; Gan, Z.; Liu, Z.; Liu, C.; and Wang, L. 2022. GIT: A Generative Image-to-text Transformer for Vision and Language. Technical report, Microsoft.

Yang, T.; Zhu, Y.; Xie, Y.; Zhang, A.; Chen, C.; and Li, M. 2023. AIM: Adapting Image Models for Efficient Video Action Recognition. In *The Eleventh International Conference on Learning Representations*.

Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; Li, C.; Xu, Y.; Chen, H.; Tian, J.; Qi, Q.; Zhang, J.; and Huang, F. 2023. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. arXiv:2304.14178.

Zaken, E. B.; Goldberg, Y.; and Ravfogel, S. 2022. 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)*, 1–9.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592*.