

---

# MINIGPT-5: INTERLEAVED VISION-AND-LANGUAGE GENERATION VIA GENERATIVE VOKENS

Kaizhi Zheng\*, Xuehai He\*, and Xin Eric Wang

University of California, Santa Cruz

<https://github.com/eric-ai-lab/MiniGPT-5>

## ABSTRACT

Large Language Models (LLMs) have garnered significant attention for their advancements in natural language processing, demonstrating unparalleled prowess in text comprehension and generation. Yet, the simultaneous generation of images with coherent textual narratives remains an evolving frontier. In response, we introduce an innovative interleaved vision-and-language generation technique anchored by the concept of “generative vokens”, acting as the bridge for harmonized image-text outputs. Our approach is characterized by a distinctive two-staged training strategy focusing on description-free multimodal generation, where the training requires no comprehensive descriptions of images. To bolster model integrity, classifier-free guidance is incorporated, enhancing the effectiveness of vokens on image generation. Our model, MiniGPT-5, exhibits substantial improvement over the baseline Divter model on the MMDialog dataset and consistently delivers superior or comparable multimodal outputs in human evaluations on the VIST dataset, highlighting its efficacy across diverse benchmarks.

## 1 INTRODUCTION

In the recent development of larger-scale vision-and-language models, multimodal feature integration is not just a evolving trend but a critical advancement shaping a wide array of applications, from multimodal dialogue agents to cutting-edge content creation tools. With the surge in research and development in this domain, vision-and-language models such as (Wu et al., 2023a; Li et al., 2023b; Tsimpoukelli et al., 2021; Alayrac et al., 2022) are on the brink of an era where they are expected to comprehend and generate both text and image content seamlessly. This multi-faceted ability is crucial, as it fosters enhanced interactions across various domains like virtual reality, media, and e-commerce. Essentially, the task is to enable models to coherently synthesize, recognize, and respond using both visual and textual modalities, harmonizing the information flow and creating cohesive narratives. However, as we tread the path towards blending textual and visual modalities and achieving the interleaved vision-and-language generation, as illustrated in 1, we recognize that it is driven by the pressing need for more integrated and fluid multimodal interactions in large language models. However, this journey is riddled with multiple challenges.

First, while the current state-of-the-art Large Language Models (LLMs) (OpenAI, 2023; Chiang et al., 2023; Ouyang et al., 2022) excel in understanding text and processing text-image pairs, they falter in the nuanced art of generating images. Second, moving away from conventional tasks that benefited from exhaustive image descriptions, the emerging interleaved vision-and-language tasks (Sharma et al., 2018) lean heavily on topic-centric data, often skimping on thorough image descriptors (Huang et al., 2016). Even after being trained on massive datasets, it is challenging to align generated text with corresponding images. Lastly, as we push the boundaries with LLMs, the large memory requirements beckon us to devise more efficient strategies, especially in downstream tasks.

Addressing these challenges, we present MiniGPT-5, an innovative interleaved vision-and-language generation technique anchored by the concept of “generative vokens”. By amalgamating the Stable

---

\*These authors contributed equally to this work.

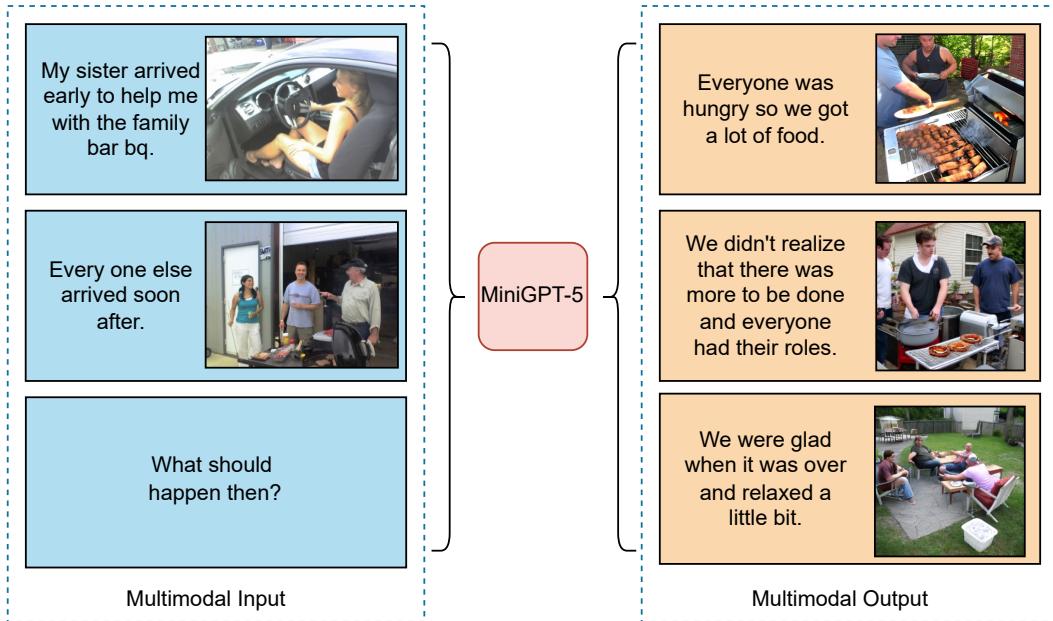


Figure 1: MiniGPT-5 is a unified model for interleaved vision-and-language comprehension and generation. Besides the original multimodal comprehension and text generation abilities, MiniGPT-5 can provide appropriate, coherent multimodal outputs.

Diffusion mechanism with LLMs through special visual tokens (Tan & Bansal, 2020) – “generative vokens”, we develop a new approach for multimodal generation. Meanwhile, our proposed two-stage training methodology underlines the importance of a description-free foundational stage, prepping the model to thrive even in data-scarce scenarios. Our generic stages, free from domain-specific annotations, make our solution distinct from existing works. To ensure that the generated text and images are in harmony, our dual-loss strategy comes into play, further enhanced by our innovative generative voken approach and classifier-free guidance. Our parameter-efficient fine-tuning strategy optimizes training efficiency and addresses memory constraints.

Building on these techniques, our work signifies a transformative approach. As shown in Figure 2, using ViT (Vision Transformer) and Qformer (Li et al., 2023b) along with the large language models, we adapt multimodal inputs into generative vokens, seamlessly combined with the high-resolution Stable Diffusion 2.1 model (Rombach et al., 2022b) for context-aware image generation. Incorporating images as auxiliary input with instruction tuning approaches and pioneering both the text and image generation loss, we amplify the synergy between text and visuals.

In summary, our contributions are primarily threefold:

- We propose to use multimodal encoders representing a novel and generic technique that has proved more effective than LLM and also inversion to generative vokens, and combine it with Stable Diffusion to generate interleaved vision-and-language outputs (multimodal language model that can do multimodal generation).
- We highlight a new two-staged training strategy for the description-free multimodal generation. The unimodal alignment stage harvests the high-quality text-aligned visual features from large text-image pairs. The multimodal learning stage ensures the visuals and text prompt can well coordinate for generation. The inclusion of classifier-free guidance during the training phase further refines generation quality.
- Compared with other multimodal generation models, we achieved state-of-the-art performance on the CC3M dataset. We also established unprecedented benchmarks on prominent datasets, including VIST and MMDialog.

---

## 2 RELATED WORK

**Text-to-Image Generation** To transform textual descriptions into their corresponding visual representations, text-to-image models (Reed et al., 2016; Dhariwal & Nichol, 2021; Saharia et al., 2022; Rombach et al., 2022b;a; Gu et al., 2023) employ complex architectures and sophisticated algorithms, bridging the gap between textual information and visual content. These models are adept at interpreting the semantics of input text and translating them into coherent and pertinent images. A notable recent contribution in this field is Stable Diffusion 2 (Rombach et al., 2022b), which employs a diffusion process to generate conditional image features and subsequently reconstructs images from these features. Our research aims to leverage this pre-trained model, enhancing its capabilities to accommodate both multimodal input and output.

**Multimodal Large Language Models** As Large Language Models (LLMs) become increasingly impactful and accessible, a growing body of research has emerged to extend these pretrained LLMs into the realm of multimodal comprehension tasks (Zhu et al., 2023; Li et al., 2023b; Dai et al., 2023; OpenAI, 2023; Li et al., 2023a; Alayrac et al., 2022). For example, to reproduce the impressive multimodal comprehension ability in GPT-4 (OpenAI, 2023), MiniGPT-4 (Zhu et al., 2023) proposes a projection layer to align pretrained vision component of BLIP- (Li et al., 2023b) with an advanced open-source large language model, Vicuna (Chiang et al., 2023). In our work, we utilize the MiniGPT-4 as the base model and extend the model’s capabilities to multimodal generation.

**Multimodal Generation with Large Language Models** To augment the LLM’s capabilities in seamlessly integrating vision and language generation, recent studies have introduced a variety of innovative methods (Ge et al., 2023; Sun et al., 2021; Koh et al., 2023; Sun et al., 2023b; Yu et al., 2023). For instance, CM3Leon (Yu et al., 2023) presents a retrieval-augmented, decoder-only architecture designed for both text-to-image and image-to-text applications. Similarly, Emu (Sun et al., 2023b) employs the pretrained EVA-CLIP (Sun et al., 2023a) model to convert images into one-dimensional features and fine-tunes the LLAMA (Touvron et al., 2023) model to generate cohesive text and image features through autoregressive techniques. On the other hand, both GILL (Koh et al., 2023) and SEED (Ge et al., 2023) explore the concept of mapping vokens into the text feature space of a pretrained Stable Diffusion model; GILL employs an encoder-decoder framework, while SEED utilizes a trainable Q-Former structure. In contrast to these approaches, our model takes a more direct route by aligning voken features with visual information. Additionally, we introduce several training strategies aimed at enhancing both image quality and contextual coherence.

## 3 METHOD

In order to endow large language models with multimodal generation capabilities, we introduce a structured framework that integrates pretrained multimodal large language models and text-to-image generation models. To address the discrepancies across model domains, we introduce special visual tokens—termed “generative vokens”—that are able to direct training on raw images. Moreover, we advance a two-stage training method, coupled with a classifier-free guidance strategy, to further enhance the quality of generation. Subsequent sections will provide a detailed exploration of these elements.

### 3.1 MULTIMODAL INPUT STAGE

Recent advancements in multimodal large language models, such as MiniGPT-4, have primarily concentrated on multimodal comprehension, enabling the processing of images as sequential input. To expand their capabilities to multimodal generation, we introduce generative vokens designed for outputting visual features. Additionally, we employ cutting-edge, parameter-efficient fine-tuning techniques within the Large Language Model (LLM) framework for multimodal output learning. A more detailed introduction to these developments is provided in the following paragraphs.

**Multimodal Encoding:** Each text token is embedded into a vector  $e_{\text{text}} \in \mathbf{R}^d$ , while the pretrained visual encoder transforms each input image into the feature  $e_{\text{img}} \in \mathbf{R}^{32 \times d}$ . These embeddings are concatenated to create the input prompt features.

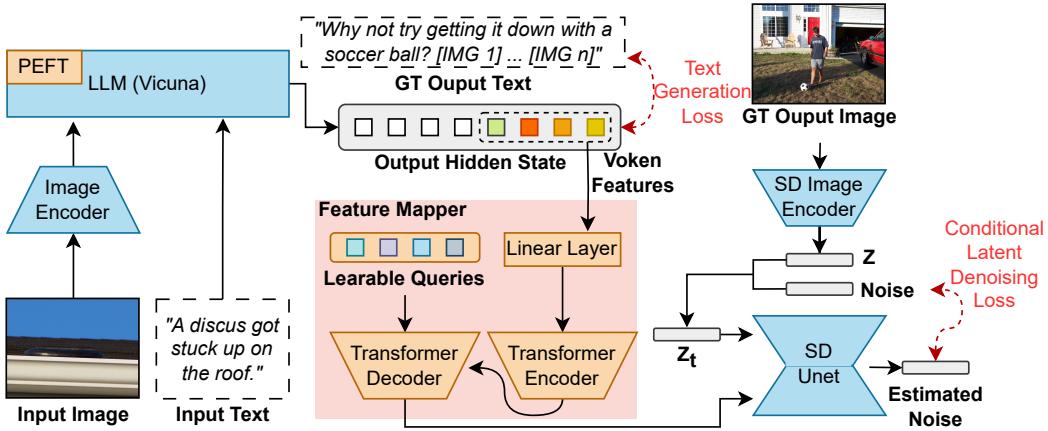


Figure 2: The overview structure of MiniGPT-5 pipeline. We leverage the pretrained multimodal large language model (MiniGPT-4) and text-to-image generation model (Stable Diffusion 2.1) to create a unified multimodal generation pipeline. The input image encoder includes a ViT, Qformer, and linear layer, pretrained by MiniGPT-4. The orange blocks include learnable parameters, while the blue blocks are fixed during training. More details can be find in Section 3.

**Adding Vokens in LLM:** Since the original LLM’s  $V$  vocabulary only includes the textual tokens, we need to construct a bridge between the LLM and the generative model. Therefore, we introduce a set of special tokens  $V_{\text{img}} = \{\text{[IMG}1], \text{[IMG}2], \dots, \text{[IMG}n]\}$  (default  $n = 8$ ) as generative vokens into the LLM’s vocabulary  $V$ . The LLM’s output hidden state for these vokens is harnessed for subsequent image generation, and the positions of these vokens can represent the insertion of the interleaved images. With all pretrained weights  $\theta_{\text{pretrained}}$  in MiniGPT-4 fixed, the trainable parameters include extra input embedding  $\theta_{\text{voken\_input}}$  and output embedding  $\theta_{\text{voken\_output}}$ .

**Parameter-Efficient Fine-Tuning (PEFT):** Parameter-efficient fine-tuning (PEFT) (Houlsby et al., 2019; Hu et al., 2021; Li & Liang, 2021) is critical in training large language models (LLMs). Despite this, its application in multimodal settings remains largely unexplored. We use PEFT over the MiniGPT-4 (Zhu et al., 2023) encoder to train a model to understand instructions or prompts better, enhancing its performance in novel and even zero-shot tasks. More specifically, we tried prefix tuning (Li & Liang, 2021) and LoRA over the entire language encoder —Vicuna (Chiang et al., 2023) used in MiniGPT-4. Combined with the instruction tuning, it notably amplifies multimodal generation performance across various datasets, such as VIST and MMDialog.

### 3.2 MUTIMODAL OUPUT GENERATION

To accurately align the generative tokens with the generative model, we formulate a compact mapping module for dimension matching and incorporate several supervisory losses, including text space loss and latent diffusion model loss. The text space loss assists the model in learning the correct positioning of tokens, while the latent diffusion loss directly aligns the tokens with the appropriate visual features. Since the generative vokens’ features are directly guided by images, our method does not need comprehensive descriptions of images, leading to description-free learning.

**Text Space Generation:** We first jointly generate both text and vokens in the text space by following the causal language modeling. During the training, we append the vokens to the positions of ground truth images and train the model to predict vokens within text generation. Specifically, the generated tokens are represented as  $T = \{t_1, t_2, \dots, t_m\}$ , where  $t_i \in V \cup V_{\text{img}}$ , and the causal language modeling loss is defined as:

$$L_{\text{text}} := - \sum_{i=1}^m \log p(t_i | e_{\text{text}}, e_{\text{img}}, t_1, \dots, t_{i-1}; \theta_{\text{pretrained}}, \theta_{\text{voken\_input}}, \theta_{\text{voken\_output}}), \text{ where } t_i \in V \cup V_{\text{img}} \quad (1)$$

---

**Mapping Voken Features for Image Generation:** Next, we align the output hidden state  $h_{\text{voken}}$  with the text conditional feature space of the text-to-image generation model. To map the voken feature  $h_{\text{voken}}$  to a feasible image generation conditional feature  $e_{\text{text\_encoder}} \in \mathbf{R}^{L \times \hat{d}}$  (where  $L$  is the maximum input length of text-to-image generation text encoder, and  $\hat{d}$  is the dimension of encoder output feature in text-to-image generation model). We construct a feature mapper module, including a two-layer MLP model  $\theta_{\text{MLP}}$ , a four-layer encoder-decoder transformer model  $\theta_{\text{enc\_dec}}$ , and a learnable decoder feature sequence  $q$ . The mapping feature  $\hat{h}_{\text{voken}}$  is then given by:

$$\hat{h}_{\text{voken}} := \theta_{\text{enc\_dec}}(\theta_{\text{MLP}}(h_{\text{voken}}), q) \in \mathbf{R}^{L \times \hat{d}} \quad (2)$$

**Image Generation with Latent Diffusion Model (LDM):** To generate appropriate images, the mapping feature  $\hat{h}_{\text{voken}}$  is used as a conditional input in the denoising process. Intuitively,  $\hat{h}_{\text{voken}}$  should represent the corresponding text features that guide the diffusion model to generate the ground truth image. We employ the loss of the latent diffusion model (LDM) for guidance. During the training, the ground truth image is first converted to latent feature  $z_0$  through the pretrained VAE. Then, we obtain the noisy latent feature  $z_t$  by adding noise  $\epsilon$  to  $z_0$ . A pretrained U-Net model  $\epsilon_\theta$  is used to calculate the conditional LDM loss as:

$$L_{\text{LDM}} := \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1), t} \left[ \left\| \epsilon - \epsilon_\theta(z_t, t, \hat{h}_{\text{voken}}) \right\|_2^2 \right] \quad (3)$$

This comprehensive approach ensures a coherent understanding and generation of both textual and visual elements, leveraging the capabilities of pretrained models, specialized tokens, and innovative training techniques.

### 3.3 TRAINING STRATEGY

Given the non-negligible domain shift between text and image domains, we observe that direct training on a limited interleaved text-and-image dataset can result in misalignment and diminished image quality. Consequently, we adopt two distinct training strategies to mitigate this issue. The first strategy encompasses the incorporation of the classifier-free guidance (Ho & Salimans, 2022) technique, which amplifies the effectiveness of the generative tokens throughout the diffusion process. The second strategy unfolds in two stages: an initial pre-training stage focusing on coarse feature alignment, followed by a fine-tuning stage dedicated to intricate feature learning.

**Classifier-free Guidance (CFG):** To enhance the coherence between the generated text and images, we first leverage the idea of Classifier-free Guidance for multimodal generation. Classifier-free guidance is introduced in the text-to-image diffusion process. This method observes that the generation model  $P_\theta$  can achieve improved conditional results by training on both conditional and unconditional generation with conditioning dropout. In our context, our objective is to accentuate the trainable condition  $h_{\text{voken}}$  and the generation model is fixed. During training, we replace  $h_{\text{voken}}$  with zero features  $h_0 \in \mathbf{0}^{n \times d}$  with a 10% probability, obtaining the unconditional feature  $\hat{h}_0 = \theta_{\text{enc\_dec}}(\theta_{\text{MLP}}(h_0), q)$ . During inference,  $\hat{h}_0$  serves as negative prompting, and the refined denoising process is expressed as:

$$\begin{aligned} \log \widehat{P}_\theta(\epsilon_t | z_{t+1}, \hat{h}_{\text{voken}}, \hat{h}_0) &= \log P_\theta(\epsilon_t | z_{t+1}, \hat{h}_0) + \\ &\quad \gamma \left( \log P_\theta(\epsilon_t | z_{t+1}, \hat{h}_{\text{voken}}) - \log P_\theta(\epsilon_t | z_{t+1}, \hat{h}_0) \right) \end{aligned} \quad (4)$$

**Two-stage Training Strategy:** Recognizing the non-trivial domain shift between pure-text generation and text-image generation, we propose a two-stage training strategy: Unimodal Alignment Stage (UAS) and Multimodal Learning Stage (MLS). Initially, we align the voken feature with image generation features in single text-image pair datasets, such as CC3M, where each data sample only contains one text and one image and the text is usually the caption of the image. During this stage, we utilize captions as LLM input, enabling LLM to generate vokens. Since these datasets include the image descriptive information, we also introduce an auxiliary loss to aid voken alignment, minimizing the distance between the generative feature  $\hat{h}_{\text{voken}}$  and the caption feature from the text encoder  $\tau_\theta$  in the text-to-image generation model:

$$L_{\text{CAP}} := \text{MSE}(\hat{h}_{\text{voken}}, \tau_\theta(c)) \quad (5)$$

The unimodal alignment stage loss is expressed as  $L_{\text{UAS}} = \lambda_1 * L_{\text{text}} + \lambda_2 * L_{\text{LDM}} + \lambda_3 * L_{\text{CAP}}$ , with selected values  $\lambda_1 = 0.01, \lambda_2 = 1, \lambda_3 = 0.1$  to rescale the loss into a similar numerical range.

After the unimodal alignment stage, the model is capable of generating images for single text descriptions but struggles with interleaved vision-and-language generation, which includes multiple text-image pairs and requires complicated reasoning for both text and image generation. To address this, in the multimodal learning stage, we further fine-tune our model with PEFT parameters by interleaved vision-and-language datasets, such as VIST, where the data sample has several steps with text-image and texts are sequentially relevant. During this stage, we construct three types of tasks from the dataset, encompassing (1) text-only generation: given the next image, generating the related text; (2) image-only generation: given the next text, generating the related image, and (3) multimodal generation: generating text-image pair by given context. The multimodal learning stage loss is given by  $L_{\text{MLS}} = \lambda_1 * L_{\text{text}} + \lambda_2 * L_{\text{LDM}}$ . More implementation details can be found in appendix A.

## 4 EXPERIMENTS

To assess the efficacy of our model, we conducted a series of evaluations across multiple benchmarks. These experiments aim to address several key questions: (1) Can our model generate plausible images and reasonable texts? (2) How does our model’s performance stack up against other state-of-the-art models in both single-turn and multi-turn interleaved vision-and-language generation tasks? (3) What impact does the design of each module have on overall performance? In the subsequent subsections, we will delve into the datasets and experimental settings used for these evaluations, followed by a comprehensive analysis of our model’s performance. We use three datasets: CC3M (Sharma et al., 2018), VIST (Huang et al., 2016), and MMDialog (Feng et al., 2022). More details about datasets and data format can be found in appendix B.

### 4.1 EXPERIMENTAL SETTINGS

**Baselines** For a comprehensive evaluation of our performance in multimodal generation, we conducted comparative analyses with several prominent baseline models: the Fine-tuned Unimodal Generation Model, GILL, and Divter.

- **Fine-tuned Unimodal Generation Model:** To facilitate fair comparisons in both image and text generation, we fine-tuned two separate models, Stable Diffusion 2.1 and MiniGPT-4, utilizing the VIST dataset. Within the Stable Diffusion 2.1 model, the U-Net parameters were unfrozen. For MiniGPT-4’s LLM part, LoRA parameters were fine-tuned.
- **GILL** (Koh et al., 2023)<sup>1</sup>: GILL is a recent innovation that allows the LLM to generate vokens using a pre-trained text-to-image generation model for single-image generation. Unlike our method, which employs conditional generation loss guidance, GILL minimizes the Mean Squared Error (MSE) loss between the text-to-image text encoding feature and voken features, similar to  $L_{\text{CAP}}$  in our approach. Since their method requests image descriptions for training, we compare with it just on the unimodal alignment stage.
- **Divter** (Sun et al., 2021): Divter is a state-of-the-art conversational agent developed for multimodal dialogue contexts. It introduces a customized transformer structure for generating multimodal responses. Divter’s methodology includes pretraining on a vast corpus of text-only dialogues and text-image pairs, followed by finetuning on a selected set of multimodal response data. The MMDialog dataset regards Divter’s method as the baseline.

<sup>1</sup>To ensure fair comparisons, given the variations in the valid data within the CC3M dataset and the original use of Stable Diffusion 1.5 in GILL, we made adjustments. Specifically, we switched their text-to-image generation model to Stable Diffusion 2.1 and retrained it on our specific CC3M data, following the guidelines in their official implementation. (<https://github.com/kohjingyu/gill>)

Table 1: Performance metrics for different models with various prompt types on VIST final step image generation. For ‘No Context’, only the current step’s text is provided. The ‘Text Context’ uses all history texts, the ‘Image Context’ employs all preceding images, and ‘Image-Text Context’ provides a combination of both past images and texts.

Model	No Context			Text Context			Image Context			Image-Text Context		
	CLIP-I ( $\uparrow$ )	IS ( $\uparrow$ )	FID ( $\downarrow$ )	CLIP-I ( $\uparrow$ )	IS ( $\uparrow$ )	FID ( $\downarrow$ )	CLIP-I ( $\uparrow$ )	IS ( $\uparrow$ )	FID ( $\downarrow$ )	CLIP-I ( $\uparrow$ )	IS ( $\uparrow$ )	FID ( $\downarrow$ )
Zero-shot SD 2	0.57	<b>23.62</b>	61.26	0.59	23.24	62.60	-	-	-	-	-	-
Fine-tuned SD 2	0.59	23.28	<b>58.29</b>	0.61	23.47	<b>57.45</b>	-	-	-	-	-	-
MiniGPT-5 (Prefix)	0.60	23.19	61.25	0.63	<b>25.06</b>	61.81	0.68	24.27	59.92	0.70	<b>25.10</b>	60.46
MiniGPT-5 (LoRA)	<b>0.61</b>	22.30	61.44	<b>0.64</b>	23.86	61.34	<b>0.69</b>	<b>25.03</b>	<b>59.09</b>	<b>0.70</b>	24.38	<b>59.48</b>
MiniGPT-5 (w/o UAS)	0.55	16.32	73.02	0.57	16.31	73.97	0.58	16.70	75.88	0.58	16.99	76.51

**Metrics** To comprehensively assess the model performance across image, text, and multimodal dimensions, we employ a diverse set of metrics. For evaluating the quality and diversity of generated images, we utilize the Inception Score (IS) (Salimans et al., 2016), and Fréchet Inception Distance (FID) (Heusel et al., 2017). Textual performance is gauged through metrics such as BLEU (Papineni et al., 2002), Rouge-L (Lin, 2004), METEOR (Banerjee & Lavie, 2005), and Sentence-BERT (SBERT) (Reimers & Gurevych, 2019) scores.

On the multimodal front, we leverage CLIP-based metrics (Rombach et al., 2022b) to assess the congruence between generated content and ground truth. CLIP-I evaluates the similarity between generated and ground-truth images, while CLIP-T focuses on the congruence between generated images and ground-truth text. To address potential misalignments in the multimodal generation, such as when the ground truth is text-only, but the output is multimodal, we utilize MM-Relevance (Feng et al., 2022). This metric calculates the F1 score based on CLIP similarities, providing a nuanced evaluation of multimodal coherence. We also employ the Human Preference Score (HPS) v2 (Wu et al., 2023c) to assess the extent to which the generated images align with the input text prompts based on human preferences.

Recognizing that the generated multimodal output might be meaningful yet differ from the ground truth, we also incorporate human evaluation to assess the model’s performance. We examine the model’s effectiveness from three perspectives: (1) Language Continuity - assessing if the produced text aligns seamlessly with the provided context, (2) Image Quality - evaluating the clarity and relevance of the generated image, and (3) Multimodal Coherence - determining if the combined text-image output is consistent with the initial context.

## 4.2 EXPERIMENTAL RESULTS

In this section, we will quantitatively analyze our model performance on different benchmarks for different training stages. The qualitative examples can be found in Fig. 4.

### 4.2.1 MULTIMODAL LEARNING STAGE

In this subsection, we present the performance of different models on the VIST (Huang et al., 2016) and MMDialg (Feng et al., 2022) datasets. Our evaluations span both vision (image-related metrics) and language (textual metrics) domains to showcase the versatility and robustness of the proposed models.

**VIST Final-Step Evaluation** Our first set of experiments involves a single-step evaluation where, given the last step’s prompt, the model aims to generate the corresponding image. Table 1 summarizes the results for this setting. The MiniGPT-5 in all three settings can outperform the fine-tuned SD 2, showing the benefits of the MiniGPT-5 pipeline. Notably, the MiniGPT-5 (LoRA) model consistently surpasses other variants in terms of the CLIP Score across multiple prompt types, especially when both image and text prompts are combined. On the other hand, the FID scores highlight the MiniGPT-5 (prefix) model’s competitiveness, indicating a possible trade-off between image embedding quality (reflected by the CLIP Score) and the diversity and realism of the images (captured by the FID score). When compared to the model (MiniGPT-5 w/o UAS) that undergoes direct training on the VIST without incorporating the unimodal alignment stage, it is evident that while the model retains the capability to generate meaningful images, there is a notable drop in image quality and coherence. This observation underscores the significance of our two-stage training strategies.

Table 2: VIST all steps image generation: CLIP Image-Image and FID Performance Metrics. In zero-shot SD2, for ‘No Context’, only the current step’s text is provided. The ‘Text Context’ uses all historical texts. FID scores evaluate the similarities between generated images and ground truth images within each story sequence.

Model	CLIP-I ( $\uparrow$ )	FID ( $\downarrow$ )
Zero-shot SD 2 (no/text context)	0.58/0.59	414.34/393.49
Fine-tuned SD 2 (no/text context)	0.60/0.61	397.05/390.25
MiniGPT-5 (Prefix)	0.65	381.55
MiniGPT-5 (LoRA)	<b>0.66</b>	<b>366.62</b>
MiniGPT-5 (w/o UAS)	0.57	420.79

Table 3: VIST all steps narration generation: S-BERT, Rouge-L, and Meteor Performance Metrics. We added LoRA fine-tuning for both MiniGPT-4 and MiniGPT-5. The results show that adding generative vokens does not hurt the performance on the multimodal comprehension tasks.

Model	S-BERT ( $\uparrow$ )	Rouge-L ( $\uparrow$ )	Meteor ( $\uparrow$ )
Fine-tuned MiniGPT-4	0.6273	<b>0.3401</b>	<b>0.3296</b>
MiniGPT-5	<b>0.6315</b>	0.3373	0.3263

**VIST Multi-Step Evaluation** In a detailed and comprehensive evaluation, we systematically provide models with prior history context and subsequently assess the generated images and narrations at each following step. Tables 2 and 3 outline the results of these experiments, encapsulating the performance in both image and language metrics, respectively. The findings demonstrate that MiniGPT-5 is capable of generating coherent, high-quality images utilizing long-horizontal multimodal input prompts across all data, without compromising the original model’s ability for multimodal comprehension. This accentuates the efficacy of our model in diverse settings.

**VIST Human Evaluation** To assess the quality of multimodal generation, we test both our model and the baseline on the VIST validation set. For each task, given a preceding multimodal sequence, models are tasked with producing the subsequent scenario. To ensure a fair comparison, we employ the fine-tuned MiniGPT-4, which is exclusively trained to generate narrations without any vokens. Subsequently, these narrations are incorporated directly into the Stable Diffusion 2 via the text-to-image pipeline. We select a random sample of 5,000 sequences, with each requiring evaluation by two workers. These evaluators are tasked with determining the superior multimodal output based on three criteria: Language Continuity, Image Quality, and Multimodal Coherence. This assessment is facilitated using Amazon Mechanical Turk (Crowston, 2012), with a representative example (Fig. 5) provided in the appendix. As depicted in Table 4, our model, MiniGPT-5, is found to generate more fitting text narrations in 57.18% of instances, deliver superior image quality in 52.06% of cases, and produce more coherent multimodal outputs in 57.62% of the scenarios. This data distinctly showcases its enhanced multimodal generation capabilities when compared to the two-stage baseline that employs narrations for text-to-image prompts without the inclusion of vokens.

**MMDialog Multi-Turn Evaluation** We conduct an evaluation of our method on the MMDialog dataset to determine the effectiveness of generating precise and appropriate multimodal information in multi-turn conversational scenarios. The model is required to generate either unimodal or multimodal responses based on the previous turns during the conversations in this dataset. Our results, as presented in Table 5, demonstrate that MiniGPT-5 outperforms the baseline model Divter in terms of generating more accurate textual responses. While the image qualities of the generated responses are similar, MiniGPT-5 excels in MM-Relevance compared to the baseline model. This indicates that our model can better learn how to appropriately position image generation and produce highly coherent multimodal responses.

#### 4.2.2 UNIMODAL ALIGNMENT STAGE

Instead of evaluating on datasets with multi-turn multimodal data, we also evaluate models in the single-image dataset CC3M (Sharma et al., 2018), as displayed in Table 6. In this stage, the model

Table 4: VIST Human Evaluation on 5,000 samples for multimodal generation from Language Continuity, Image Quality, and Multimodal Coherence aspects. The results indicate, in more than 70% cases, the MiniGPT-5 is better or on par with the two-stage baseline.

Model	MiniGPT-5	Fine-tuned	MiniGPT-4 + SD 2	Tie
Language Continuity (%)	<b>57.18</b>		28.51	14.31
Image Quality (%)	<b>52.06</b>		35.98	11.96
Multimodal Coherence (%)	<b>57.62</b>		23.24	19.14

Table 5: Multimodal generation results on MMDialog test set. In order to compare with their baseline, we use the same metrics reported in Table 3 of MMDialog paper (Feng et al., 2022).

Model	IS ( $\uparrow$ )	BLEU-1 ( $\uparrow$ )	BLEU-2 ( $\uparrow$ )	Rouge-L ( $\uparrow$ )	MM-Relevance ( $\uparrow$ )
Divter	<b>20.53</b>	0.0944	0.0745	0.1119	0.62
MiniGPT-5	19.63	<b>0.2221</b>	<b>0.1546</b>	<b>0.1119</b>	<b>0.67</b>

accepts the input of image descriptions and produces corresponding images, mirroring typical text-to-image tasks but incorporating generative tokens. The results indicate that although our model can have better generation on multi-turn scenarios, Stable Diffusion 2 achieves the best outcomes across all metrics for single-image generation. Since our model attempts to align with the pretrained text encoder of Stable Diffusion 2 in this stage, there is a slight gap in performance due to the limitation of data amount. Compared with the observations on the VIST dataset, we can conclude that MiniGPT-5 can correctly extract features from long-horizontal multimodal information instead of single text input. This indicates the future directions on how to align LLMs with generative models efficiently. On the other hand, our model outperforms another state-of-the-art multimodal generation model, GILL, on all metrics. Our model generates more coherent and high-quality images that closely resemble those produced by the pretrained stable diffusion model. To further evaluate the effectiveness of our design, we conducted several ablation studies, and more ablation studies about voken number and CFG scales can be found in appendix C.

**Evaluation of Different Loss Guidance:** As described in Sec. 3.3, we introduced an auxiliary loss, denoted as  $L_{CAP}$  for CC3M training. To assess the impact of this loss and determine if the single caption loss alone can generate high-quality images like GILL, we trained our model without the caption loss  $L_{CAP}$  (alignment between the mapped generative voken features and the caption features from stable diffusion text encoder) and the conditional latent diffusion loss  $L_{LDM}$  (alignment between the mapped generative voken features and conditional features for latent diffusion process of ground truth images) separately. The results, as shown in Table 6, indicate that the caption loss significantly aids in generating better images, and the conditional latent diffusion loss further enhances performance in terms of coherence and image quality.

**Evaluation of Classifier-Free Guidance (CFG):** To assess the effectiveness of the CFG strategy, we trained our model without CFG dropoff. During inference, the model utilized the original CFG denoising process, which utilized the empty caption feature from Stable Diffusion 2’s text encoder as negative prompt features. The results in Table 6 demonstrate that all metrics are worse without CFG, indicating that the CFG training strategy improves the image generation quality.

**Evaluation with Human Preference Score (HPS):** To better evaluate our model’s effectiveness and its individual components, we employed the Human Preference Score v2 (HPSv2) (Wu et al., 2023b). Figure 3 presents the count of images generated by each model with the highest HPS. Notably, MiniGPT-5 consistently outshines its competitors, underscoring the significance of the losses and the classifier-free guidance technique implemented in our approach.

## 5 CONCLUSION

In this paper, we introduce MiniGPT-5, designed to augment the capabilities of LLMs for multimodal generation by aligning the LLM with a pre-trained text-to-image generation model. Our approach demonstrates substantial improvements, as evidenced by comprehensive experiments.

Table 6: Model performance on CC3M validation set for single-image generation. Due to the limitations of data amount, we find there is still a gap for voken alignment with Stable Diffusion 2. However, our model outperforms another state-of-the-art model, GILL, in all metrics.

Model	CLIP-I ( $\uparrow$ )	CLIP-T ( $\uparrow$ )	IS ( $\uparrow$ )	FID ( $\downarrow$ )
Zero-shot SD 2	<b>0.64</b>	<b>0.25</b>	<b>31.74</b>	<b>26.39</b>
GILL	0.57	0.20	22.76	37.97
MiniGPT-5	0.61	0.22	28.09	31.47
MiniGPT-5 (w/o CFG)	0.60	0.22	23.41	33.73
MiniGPT-5 (w/o $L_{CAP}$ )	0.54	0.16	21.27	40.24
MiniGPT-5 (w/o $L_{LDM}$ )	0.58	0.20	24.79	34.65

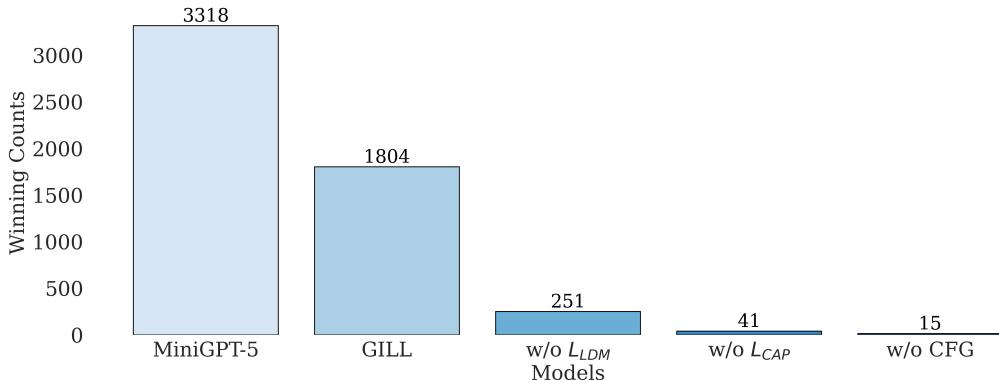


Figure 3: Number of wins based on the HPS v2 for different models. Notably, MiniGPT-5 significantly surpasses other models.

Through this work, we aspire to set a new benchmark in multimodal generative models, opening doors to applications previously deemed challenging due to the disjointed nature of existing image and text synthesis paradigms.

## REFERENCES

- 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. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.
- Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.
- 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.
- Kevin Crowston. Amazon mechanical turk: A research tool for organizations and information systems scholars. In *Shaping the Future of ICT Research. Methods and Approaches: IFIP WG 8.2, Working Conference, Tampa, FL, USA, December 13–14, 2012. Proceedings*, pp. 210–221. Springer, 2012.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.
- Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

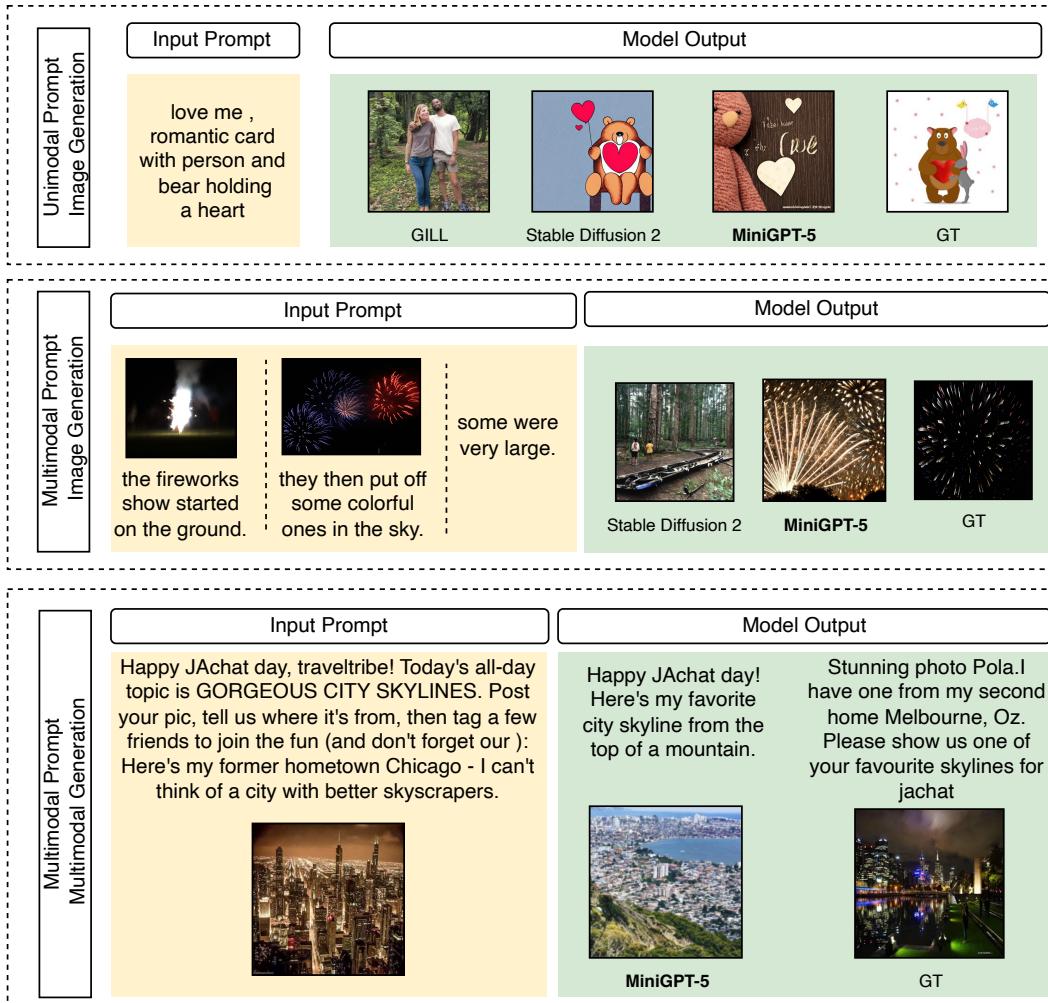


Figure 4: Qualitative examples from MiniGPT-5 and baselines on the CC3M, VIST, and MMDialog datasets. From the comparisons, we can find the MiniGPT-5 and SD 2 have similar results on single-image generation. When we evaluate with multi-step multimodal prompts, MiniGPT-5 can produce more coherent and high-quality images. More qualitative examples can be found in the appendix D.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. Mmdialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. *arXiv preprint arXiv:2211.05719*, 2022.

Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023.

Jing Gu, Yilin Wang, Nanxuan Zhao, Tsu-Jui Fu, Wei Xiong, Qing Liu, Zhifei Zhang, He Zhang, Jianming Zhang, HyunJoon Jung, and Xin Eric Wang. Photoswap: Personalized subject swapping in images, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.

- 
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*, 2016.
- Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*, 2023.
- Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023a.
- 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*, 2023b.
- Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.
- Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.
- OpenAI. Gpt-4 technical report, 2023.
- 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.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics, 2002.
- Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pp. 1060–1069. PMLR, 2016.
- Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022a.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022b.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.
- Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.
- 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.

- 
- Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. Multimodal dialogue response generation. *arXiv preprint arXiv:2110.08515*, 2021.
- Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023a.
- Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. 2023b.
- Hao Tan and Mohit Bansal. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. *arXiv preprint arXiv:2010.06775*, 2020.
- 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.
- 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*, 2023a.
- Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023b.
- Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023c.
- Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2023.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

---

## A IMPLEMENTATION DETAILS

In the unimodal alignment stage, we introduce additional voken embeddings at both the input and output layers of the Vicuna-7B model, while keeping the embeddings of other tokens fixed. These new embeddings—denoted as  $\theta_{\text{voken\_input}}$  and  $\theta_{\text{voken\_output}}$ —along with the feature mapper module ( $\theta_{\text{MLP}}, \theta_{\text{enc\_dec}}, q$ ) are jointly trained on the CC3M dataset, which consists of single text-image pairs. Training is conducted using the AdamW optimizer over two epochs, with a batch size of 48, amounting to over 110,000 steps, and a learning rate of  $2 \times 10^{-4}$ .

In the subsequent multimodal learning stage, we incorporate LoRA modules—denoted as  $\theta_{\text{LoRA}}$ —into Vicuna for the generation of both tokens and vokens. We keep the MLP model  $\theta_{\text{MLP}}$  and decoder query  $q$  fixed. The model is then fine-tuned on interleaved vision-and-language datasets, like VIST and MMDialog. The trainable parameters for this stage are  $\theta = \{\theta_{\text{voken\_input}}, \theta_{\text{voken\_output}}, \theta_{\text{LoRA}}, \theta_{\text{enc\_dec}}\}$ . Training is carried out using the AdamW optimizer over four epochs, with a batch size of 32 and a learning rate of  $2 \times 10^{-5}$ . Trainable parameters are nearly 6.6 million, and all training can be completed on a server equipped with 4 A6000 GPUs.

## B EXPERIMENTAL SETTINGS

### B.1 DATASETS

**CC3M (Sharma et al., 2018):** Conceptual Captions dataset represents a remarkable collection of high-quality image captions, amassing approximately 3.3 million pairs of text and images from the internet. The CC3M dataset’s diverse content, quality assurance, and support for multimodal learning make it a valuable asset for researchers and AI enthusiasts alike. Each data sample within this dataset consists of an image accompanied by a corresponding text description, reflecting the richness of human language and visual perception. However, after accounting for license restrictions and eliminating invalid image links, the dataset now comprises approximately 2.2 million data pairs suitable for training purposes and 10 thousand data pairs designated for validation.

**VIST (Huang et al., 2016):** Visual Storytelling dataset is an innovative compilation of visual narratives. The VIST dataset’s engaging content, narrative structure, and emphasis on sequential understanding position it as an essential resource for researchers focusing on sequential image understanding. Each sequence within this dataset consists of five images accompanied by corresponding textual narratives, showcasing the intricate interplay between visual imagery and storytelling. Designed to foster creativity and challenge conventional image-captioning models, the dataset provides a platform for training and validating algorithms capable of generating coherent and contextually relevant stories. After eliminating the invalid image links, we got over 65 thousand unique photos organized into more than 34 thousand storytelling sequences for training and 4 thousand sequences with 8 thousand images for validation.

**MMDialog (Feng et al., 2022):** Multi-Modal Dialogue dataset stands as the largest collection of multimodal conversation dialogues. The MMDialog dataset’s extensive scale, real human-human chat content, and emphasis on multimodal open-domain conversations position it as an unparalleled asset for researchers and practitioners in artificial intelligence. Each dialogue within this dataset typically includes 2.59 images, integrated anywhere within the conversation, showcasing the complex interplay between text and visual elements. Designed to mirror real-world conversational dynamics, the dataset serves as a robust platform for developing, training, and validating algorithms capable of understanding and generating coherent dialogues that seamlessly blend both textual and visual information.

### B.2 DATA FORMAT

**Unimodal Alignment Stage** In the unimodal alignment stage, our objective is to synchronize the generative voken with the text-to-image model’s conditional feature, focusing on single-turn text-image pairs. To achieve this, we utilize data from the CC3M dataset, constructing training samples by appending vokens as image placeholders after the captions, such as “a big black dog [IMG1] ... [IMGn].” The Language Model (LLM) is then tasked with only generating these placeholders

---

for text creation, and the corresponding output hidden features are further employed to compute the conditional generation loss with the ground truth image.

**Multimodal Learning Stage** In this stage, we utilize the VIST and MMDialog datasets, both of which contain multi-turn multimodal data. During training, we integrate placeholders for input images, such as '`<Img><ImageHere></Img>`', into the input text prompts when applicable. These prompts also encompass various instructions corresponding to different task types, with outputs manifesting as pure-text, pure-voken, or text-voken combinations. Below, we present example templates in the VIST dataset to illustrate the different task types:

- **Text Generation:** Input: "`<History Context> What happens in the next scene image: <Img><ImageHere></Img>`"; Output: "`<Text Description>`"
- **Image Generation:** Input: "`<History Context> Generate an image with the scene description: [Text Description]`"; Output: "`[IMG1]...[IMGn]`"
- **Text-Image Generation:** Input: "`<History Context> What should happen then?`"; Output: "`<Text Description> [IMG1]...[IMGn]`"

By structuring the input and output in this manner, we create a flexible framework that accommodates various multimodal tasks, enhancing the model’s ability to interpret and generate both textual and visual content. In the VIST dataset, the history context includes all previous story steps with both texts and images. In the MMDialog dataset, due to the limitation of computational resources, we only use up to one previous turn as the history context, and all data are formatted into the dialog.

## C MORE EXPERIMENTS

### C.1 EVALUATION OF GUIDANCE SCALE:

Since our model incorporates CFG, it is crucial to evaluate how different guidance scales affect image generation. Therefore, we plotted several line charts in Fig 6 to depict the changes in metrics with varying guidance scales. The figures reveal that both the stable diffusion model and our model generate better images as the guidance scale increases. However, when the scale exceeds 10, the image semantic coherence stabilizes while the image quality continues to decline. This suggests that the guidance scale should be set within a reasonable range for optimal image generation.

### C.2 EVALUATION OF VOKEN NUMBER:

The voken features in our model are directly utilized as conditions in the text-to-image model, leading to the expectation that an increase in the number of vokens would enhance the model’s representative capabilities. To validate this hypothesis, we conducted an experiment by training the model with varying numbers of vokens, ranging from 1 to 16. As illustrated in Fig 7, the model’s performance consistently improves with the addition of more vokens. This improvement is particularly noticeable when the number of vokens is increased from 1 to 4, highlighting the significant role that vokens play in enhancing the model’s effectiveness.

## D MORE QUALITATIVE EXAMPLES

---

You are given a **sequence of text-image story input**, and two output text-image pairs.  
We generate the next scene for each given story scenarios.

Your task is to compare the quality of these two output text-image pairs concerning  
1) if the **generated text narration is semantically continuous with given previous scenarios**  
2) if the **generated image have good quality**  
3) if the **generated text-image pair is coherent with given previous scenarios**  
Every corresponding text is above the image.

---



Input Story Scenario:

i took lots of pictures .

i could see them quite well .

Output 1: , Output 2:



Problem 1: Which one better **generate appropriate text narration by given previous scenarios**? (Output 1, Output 2, Tie)

Problem 2: Which one better **generate image with higher quality?** (Output 1, Output 2, Tie)

Problem 3: Which one better **generate coherent text-image pair by given previous scenarios?** (Output 1, Output 2, Tie)

Figure 5: Screenshot for human evaluation interface on the Amazon Mechanical Turk crowdsourcing evaluation platform. Output 1 is generated by MiniGPT-5, while output 2 is generated by fine-tuned MiniGPT-4 (without tokens) and stable diffusion 2.

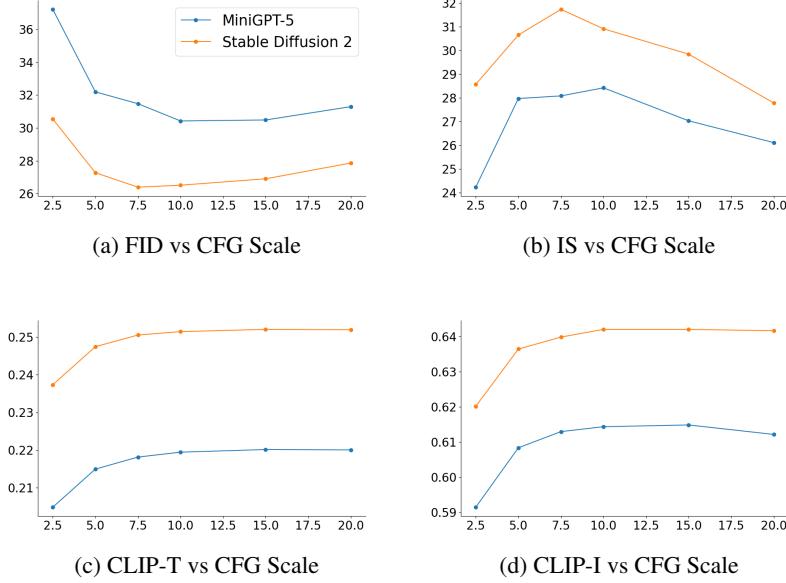


Figure 6: Line charts for various metrics vs Classifier-free Guidance (CFG) scale. The results suggest that our CFG strategy can exhibit comparable effectiveness to the CFG strategy employed in SD2, with the appropriate CFG scale significantly enhancing both image quality and coherence.

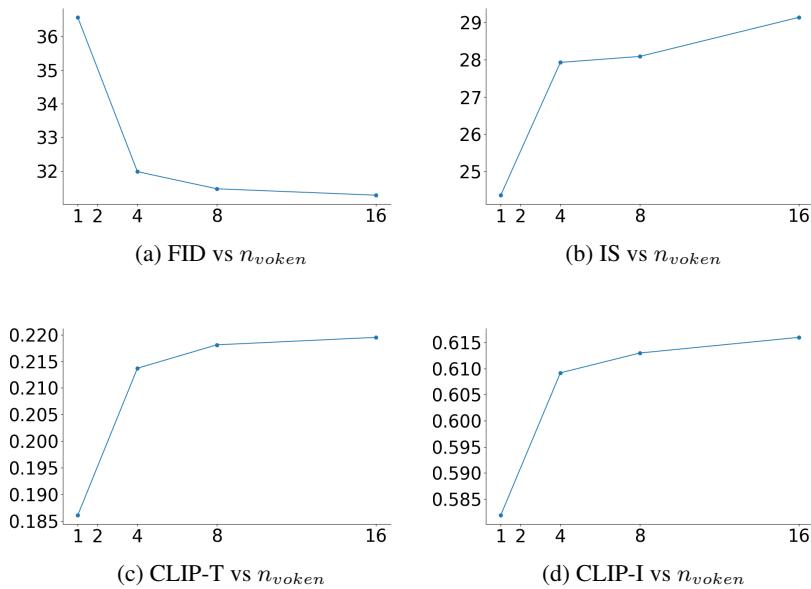


Figure 7: Line charts for various metrics vs the number of vokens. As the number of vokens increases, the image quality and CLIP scores improve. In this work, our default voken number is 8.

Input Prompt	Model Output		
womens hands sprinkle a dough with flour close up	MiniGPT-5	GT	Stable Diffusion 2
sunflowers have a deep sentimental meaning for me	MiniGPT-5	GT	GILL
we all know superman , comic book characters , but history is full of less impressive heroes	MiniGPT-5	GT	Stable Diffusion 2
boy looking in the encyclopedia through a magnifying glass	MiniGPT-5	GT	Stable Diffusion 2
happy young businessman with a folder running up a drawn stairs along a concrete wall	MiniGPT-5	GT	GILL

Figure 8: More qualitative examples from MiniGPT-5 and baselines on CC3M validation set.

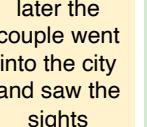
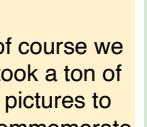
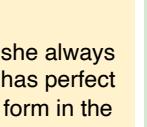
Input Prompt	Model Output				
 the couple had a lovely day at the ocean.	 here the girlfriend is hiking up the trail on the way back.	 later the couple went into the city and saw the sights including this beautiful church.	 Stable Diffusion 2	 MiniGPT-5	 GT
 there was drinking, but that just enhanced the fun of the day.	 playing around with this ball kept everyone busy.	 of course we took a ton of pictures to commemorate things.	 Stable Diffusion 2	 MiniGPT-5	 GT
 everyone is waiting for the game to start.	 you can feel the excitement in the air.	 she always has perfect form in the game.	 Stable Diffusion 2	 MiniGPT-5	 GT
 an overstressed person decided to head to the water to quiet their mind.	 as they strolled along the sandy beach , they stumbled across this family of geese.	 Stable Diffusion 2	 MiniGPT-5	 GT	

Figure 9: More qualitative examples from MiniGPT-5 and baselines on VIST validation set.

Input Prompt	Model Output
<p>Happy Traveltuesday Where would you rather be today? I'd love to be back enjoying the beautiful St Monans night skies ScotSpirit</p> 	<p>I'd love to be back in St Monans night skies ScotSpirit!</p>  <p>MiniGPT-5</p> <p>One Day I'll Visit Scotland So on this TravelTuesday - Highlands or Sea - Somewhere in Scotland is Where I'd Like to Be</p>  <p>GT</p>
<p>Also, Eli Whitmire a very deserving offensive MVP award.</p>	<p>Eli Whitmire is the best player in college football.</p>  <p>MiniGPT-5</p> <p>my fav pic of eli Whitmire. roasted</p>  <p>GT</p>
<p>One activity that many can still do right now is hiking. :raising_hands: If you've got a second, I'd love to hear about your favourite hike you've ever done! :hiking_boot: :mountain: Feel free to drop a link to an article as well. Thought it'd be fun to create a little hiking bucket list for the future!</p>	<p>I've been to the top of Mount Washington in New Hampshire. It was a tough climb but worth it!</p>  <p>MiniGPT-5</p> <p>One of the best hikes in Canada (and not at all hard) is the beautiful Skerwink Trail in Port Rexton, Newfoundland/Labrador. When you finish you can take a short walk to the Port Rexton Brewing Co. brewpub! Canada</p>  <p>GT</p>

Figure 10: More qualitative examples from MiniGPT-5 on MMDialog test set.