Interleaved Vision-and-Language Generation via Generative Vokens

Anonymous ACL submission

Abstract

The effectiveness of Multimodal Large Language Models (MLLMs) demonstrates a profound capability in multimodal understanding. However, the simultaneous generation of images with coherent texts is still underdeveloped. Addressing this, we introduce a novel interleaved vision-and-language generation method, centered around the concept of "generative vokens". These vokens serve as pivotal elements contributing to coherent image-text outputs. Our method is marked by a unique two-stage training strategy for description-free multimodal generation, which does not necessitate extensive descriptions of images. We integrate classifier-free guidance to enhance the alignment of generated images and texts, ensuring more seamless and contextually relevant multimodal interactions. Our model, ViLGen, exhibits substantial improvement over the baseline models on multimodal generation datasets, including MMDialog and VIST. The human evaluation shows ViLGen is better than the baseline model on more than 56% cases for multimodal generation, highlighting its efficacy across diverse benchmarks.

1 Introduction

The development of large-scale vision-and-language models is significantly impacting a wide range of fields like automated dialogue systems and digital content creation. With the surge in research and development in this domain, the current state-of-the-art Large Language Models (LLMs) (OpenAI, 2023; Chiang et al., 2023; Ouyang et al., 2022) and vision-and-language models such as (Wu et al., 2023a; Li et al., 2023c; Tsimpoukelli et al., 2021; Alayrac et al., 2022) fall short in generating coherent multimodal outputs. This limitation becomes particularly evident in tasks that demand an integrated handling of vision and language, essential for the next generation Large Language Models (LLMs).

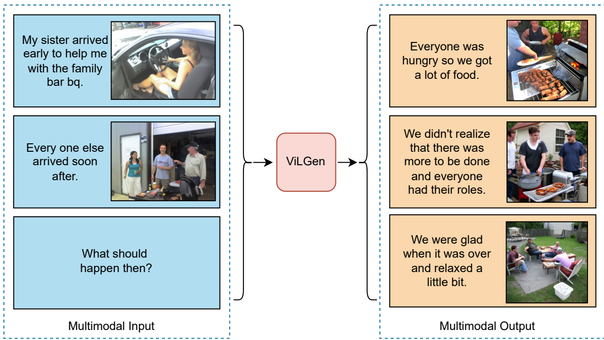


Figure 1: ViLGen is a unified model for interleaved vision-and-language comprehension and generation. Besides the original multimodal comprehension and text generation abilities, ViLGen can provide appropriate, coherent multimodal outputs.

Our work, as illustrated in Fig. 1, seeks to address these shortcomings by enhancing the integration of text and image generation in LLMs. The challenges in developing a multimodal LLM capable of interleaved vision and language generation are manifold. First, LLMs typically lack mechanisms to directly produce images, prompting us to introduce "generative vokens" that bridge the gap between textual and visual feature spaces. Second, the constraint of data scarcity, especially in vision-and-language tasks (Sharma et al., 2018) lacking extensive detailed descriptions of images (Huang et al., 2016), is countered by our unique description-free training approach. Third, maintaining both image-text and image-image consistency poses a significant challenge, which we address through dual-loss strategies. Finally, as we push forward the boundaries with LLMs, the large memory requirements urge us to devise more efficient end-to-end strategies and create an efficient training pipeline accessible for the community, especially in downstream tasks.

Specifically, to overcome these challenges, we present ViLGen, a novel approach for interleaved vision-and-language generation. By combing the Stable Diffusion with LLMs through special visual tokens (Tan and Bansal, 2020) – “generative vokens", we develop a new approach for multimodal generation. Our two-stage training methodology emphasizes a description-free foundational phase, enabling effective model training even with limited caption-grounded images. This strategy, distinct from existing works, pivots on generic stages free from image annotations. 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.

As shown in Fig. 2, leveraging ViT (Vision Transformer) and Qformer (Li et al., 2023c), alongside 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. We experiment on the CC3M (Sharma et al., 2018), VIST (Huang et al., 2016), and MMDialog (Feng et al., 2022) datasets. Notably, ViLGen shows superior performance across the two multimodal generation datasets.

In summary, our contributions are primarily threefold:

- We introduce a novel framework that leverages “generative vokens” to unify LLMs with Stable Diffusion, facilitating interleaved vision-and-language generation without relying on detailed image descriptions. We bridge the modality gap and improve the generation quality by using the loss of the latent diffusion model, the text generation loss, and the caption alignment loss together during training.

- We propose a new two-stage training strategy for description-free multimodal generation. The first stage focuses on extracting high-quality text-aligned visual features from large text-image pairs, while the second stage ensures optimal coordination between visual and textual prompts during generation. The inclusion of classifier-free guidance during training enhances the overall generation quality.

- ViLGen achieves significant improvements over baseline methods on interleaved vision-and-language datasets, including VIST and MMDialog, and comparable results to the state-of-the-art on the single text-image pair dataset, CC3M. The human evaluation further shows that, compared with the two-stage baseline, ViLGen can provide better generation in perspectives of appropriate texts (55%), high-quality images (53%), and coherent multimodal outputs (56%).

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 and Nichol, 2021; Saharia et al., 2022; Rombach et al., 2022b,a; Gu et al., 2023; Nichol et al., 2021; Ramesh et al., 2021; Yu et al., 2022; Chang et al., 2023) design algorithms to bridge the gap between textual information and visual content. A notable recent contribution is Stable Diffusion V2 (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 pretrained model, enhancing its capabilities to accommodate both multimodal input and output.

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; Aiello et al., 2023; Wu et al., 2023c). 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, NextGPT (Wu et al., 2023c), 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 and NextGPT employ 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 to enhance image quality and contextual coherence.

3 Method

In order to endow Large Language Models with multimodal generation capabilities, we introduce a new framework that integrates pretrained multimodal Large Language Models and text-to-image generation models. Central to our approach is the introduction of "generative vokens", special visual tokens that effectively bridge the textual and visual domains during the training process. Additionally, we implement a two-stage training method combined with a classifier-free guidance strategy to enhance the quality and coherence of generated outputs. Fig. 2 provides an overview of our model structure. ViLGen primarily consists of two modules: the Integrated Vision-Language Encoding Module, utilizing the pretrained multimodal large language model (MiniGPT-4) for handling multimodal inputs, and the Multimodal Output Generation module, employing Stable Diffusion for generating visual outputs.

3.1 Multimodal Understanding Module

Recent advancements in multimodal Large Language Models, such as MiniGPT-4 (Zhu et al., 2023), have primarily concentrated on multimodal comprehension, enabling the processing of images as sequential input. The Integrated Vision-Language Encoding Module is designed to extend the capabilities of LLMs from mere comprehension to active generation in multimodal contexts. Generative vokens play a crucial role in this module, enabling the translation of raw visual inputs into a format that LLMs can process and utilize for subsequent generation tasks.

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.

Generative Vokens 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 $V_{\text{img}} = \{[\text{IMG1}], [\text{IMG2}], \ldots, [\text{IMGn}]\}$ (by 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 and Liang, 2021) is critical in training Large Language Models (LLMs), employed to adapt LLMs to downstream tasks without the need for extensive retraining. In PEFT, rather than updating all the parameters of a model, only a small subset of parameters is trained. This subset typically includes task-specific components or lightweight layers added to the original model architecture (Zhang et al., 2021; Houlsby et al., 2019; Hu et al., 2021; Dettmers et al., 2023). We apply PEFT to the MiniGPT-4 (Zhu et al., 2023) encoder, enhancing its ability to process and generate multimodal content based on given instructions or prompts. More specifically, this involves the use of prefix tuning (Li and Liang, 2021) and LoRA(Hu et al., 2021) over the entire language encoder – `Vicuna` (Chiang et al., 2023) used in MiniGPT-4. Additionally, we implement learnable queries at the input of the transformer decoder, a conventional approach in sequence-to-sequence transformer architectures, to further improve the model's multimodal generation capabilities. We also adopted learnable queries at the input of the transformer decoder as a conventional setting for sequence-to-sequence transformer architectures (Vaswani et al., 2017a). Learnable queries in the decoder allow the model to have dynamic, adaptable representations for initiating the generation process. This is particularly useful when the model needs to generate outputs based on a mix of visual and textual inputs. Combined with the instruction tuning (Ouyang et al., 2022), it notably amplifies multimodal generation performance across various datasets.

3.2 Mutimodal Generation Module

To accurately align the generative vokens with the text-to-image generation models, we formulate a compact mapping module for dimension matching and incorporate several supervised losses, including voken positioning loss and voken alignment loss. The voken positioning loss assists the model in learning the correct positioning of tokens, while

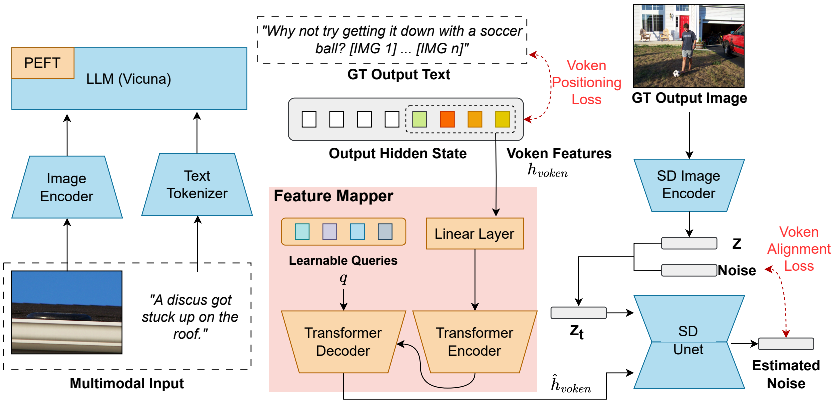


Figure 2: The overview structure of ViLGen pipeline. We leverage the pretrained multimodal large language model (MiniGPT-4) and text-to-image generation model (Stable Diffusion 2.1) to create a unified multimodal generation pipeline. The input image encoder includes a ViT, Qformer, and linear layer,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

the voken alignment loss directly aligns the vokens with the appropriate conditional generation features of the diffusion model. Since the gradients of generative vokens' features can be directly calculated from images, shown on the right side of Fig. 2, our method does not need comprehensive descriptions of images, leading to description-free learning.

Voken Positioning We first jointly generate both text and vokens in the text space by following next-word prediction in autoregressive language model (Vaswani et al., 2017b). During the training, we append the vokens V_{img} to the positions of ground truth images and train the model to predict vokens within text generation. Specifically, the generated tokens are represented as $W = \{w_1, w_2, \ldots, w_m\}$, where $w_i \in V \cup V_{\text{img}}$, and the causal language modeling loss is defined as:

$$L_{\text{text}} := -\sum_{i=1}^{m} \log p(w_i | e_{\text{text}}, e_{\text{img}}, w_1, \ldots, w_{i-1}; \theta_{\text{pretrained}}, \theta_{\text{voken\_input}}, \theta_{\text{voken\_output}}), \quad \text{where } w_i \in V \cup V_{\text{img}} \tag{1}$$

Voken Alignment for Image Generation Next, we align the output hidden state h_{voken}, shown in Fig. 2, with the conditional feature space of the text-to-image generation model. To map the voken feature h_{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 θ_{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:

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

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 conditional features that conduct the diffusion model to generate the ground truth image. We employ the latent diffusion model (LDM) loss as voken alignment loss for training the image generation module. During the training, the ground truth image is first converted to latent feature z_0 through the pretrained VAE (Variational Autoencoder) (Kingma and Welling, 2013). Then, we obtain the noisy latent feature z_t by adding noise ϵ to z_0. A pretrained U-Net model ϵ_θ is used to calculate the conditional LDM loss as:

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

To summarize, the voken positioning loss enables the model to learn the accurate placement of tokens. Without this component, the model lacks the essential capability to predict when vokens should be generated during inference. Additionally, the voken alignment loss ensures the direct correspondence between vokens and the appropriate conditional generation characteristics of the diffusion model. In the absence of this loss, the model is unable to learn semantic vokens from images directly. 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 misaligning generated texts and images and diminished image quality. Consequently, we adopt a two-stage training strategy: an initial pretraining stage focusing on coarse feature alignment for unimodal generation, followed by a fine-tuning stage dedicated to intricate feature learning for multimodal generation. Furthermore, to amplify the effectiveness of the generative tokens throughout the diffusion process, we incorporate the idea of classifier-free guidance (Ho and Salimans, 2022) technique through the whole training process.

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: Pretraining Stage and Fine-tuning Stage. 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 τ_θ in the text-to-image generation model:

$$L_{\text{CAP}} := \text{MSE}(\hat{h}_{\text{voken}}, \tau_\theta(c)) \tag{4}$$

The pretraining stage loss is expressed as $L_{\text{Pretrain}} = \lambda_1 * L_{\text{text}} + \lambda_2 * L_{\text{LDM}} + \lambda_3 * L_{\text{CAP}}$, 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 pretraining 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 fine-tuning 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 fine-tuning stage loss is given by $L_{\text{Fine-tune}} = \lambda_1 * L_{\text{text}} + \lambda_2 * L_{\text{LDM}}$. More implementation details can be found in Appendix A.

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_θ can achieve improved conditional results by training on both conditional and unconditional generation with conditioning dropout. In our context, we want the model to focus directly on the output features h_{voken} from LLM. Instead of using original stable diffusion unconditional distributions (dropping $\hat{h}_{\text{voken}}$), the whole feature mapper also needs to be included during the unconditional process. Therefore, our objective is to accentuate the trainable condition h_{voken} and the generation model is fixed. During training, we replace h_{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:

$$\begin{aligned}&\log \widehat{\mathrm{P}_\theta}\left(\epsilon_t \mid z_{t+1}, \hat{h}_{\text{voken}}, \hat{h}_0\right) = \log \mathrm{P}_\theta\left(\epsilon_t \mid z_{t+1}, \hat{h}_0\right) + \\ &\gamma\left(\log \mathrm{P}_\theta\left(\epsilon_t \mid z_{t+1}, \hat{h}_{\text{voken}}\right) - \log \mathrm{P}_\theta\left(\epsilon_t \mid z_{t+1}, \hat{h}_0\right)\right)\end{aligned} \tag{5}$$

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 compare with 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?* Below we will discuss the experimental setup and present a comprehensive analysis of our model's performance. We use three datasets: CC3M (Sharma et al., 2018), VIST (Huang et al.,

Table 1: Image generation on VIST. Given the historical context, models need to generate images for each step. FID scores evaluate the visual diversities between generated and ground truth images within each story sequence.

Model	CLIP-I (↑)	FID (↓)
SD 2.1 (Rombach et al., 2022b)	0.59	393.49
Fine-tuned SD 2.1	0.61	390.25
Two-stage Baseline	0.57	403.06
GILL (Koh et al., 2023)	0.60	381.88
ViLGen (Prefix Tuning)	0.65	381.55
ViLGen (LoRA)	**0.66**	**366.62**

Table 2: Narration Generation on VIST. We added LoRA fine-tuning for GILL, MiniGPT-4, and ViLGen with the same LoRA configuration. The results show that adding generative vokens does not hurt the performance on the multimodal comprehension tasks.

Model	S-BERT (↑)	Rouge-L (↑)	Meteor (↑)
GILL (Koh et al., 2023)	0.3864	0.1784	0.1951
MiniGPT-4 (Zhu et al., 2023)	0.6273	**0.3401**	**0.3296**
ViLGen	**0.6315**	0.3373	0.3263

2016), and MMDialog (Feng et al., 2022). More details about datasets and data format can be found in Appendix C.

4.1 Experimental Setup

4.1.1 Baselines

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 Models, Two-stage Baseline, GILL [1] (Koh et al., 2023), and Divter (Sun et al., 2021). The details of these can be found in Section C.3 in the Appendix.

4.1.2 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 and Lavie, 2005), and Sentence-BERT (S-BERT) (Reimers and Gurevych, 2019) scores.

[1] Given the variations in the valid data within the CC3M dataset, we made adjustments to ensure fair comparisons. Specifically, we retrained it on our specific CC3M data, following the guidelines in their official implementation (`https://github.com/kohjingyu/gill`).

Model	S-BERT (↑)	Rouge-L (↑)	Meteor (↑)
ViLGen (w/o vokens)	0.6273	**0.3401**	**0.3296**
ViLGen (w/ vokens)	**0.6315**	0.3373	0.3263

Table 3: Narration Generation on VIST. We added LoRA fine-tuning for both ViLGen (w/o vokens) and ViLGen. The results show that adding generative vokens does not hurt the performance on the multimodal comprehension tasks.

Model	ViLGen	Two-stage Baseline	Tie
Language Continuity (%)	**55.22**	34.89	9.89
Image Quality (%)	**52.43**	37.79	9.78
Multimodal Coherence (%)	**56.90**	28.88	14.22

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 ViLGen is better or on par with the two-stage baseline.

From the multimodal perspective, we leverage CLIP-based metrics (Rombach et al., 2022b) to assess the similarities between generated content and ground truth. CLIP-I evaluates the similarity between generated and ground-truth image features. 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 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 Main Results

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 all vision, language, and multimodality domains to showcase the versatility and robustness of the proposed models.

Unimodal Generation on VIST To evaluate the model performance on image generation and text generation, we systematically provide models with prior history context and subsequently assess the

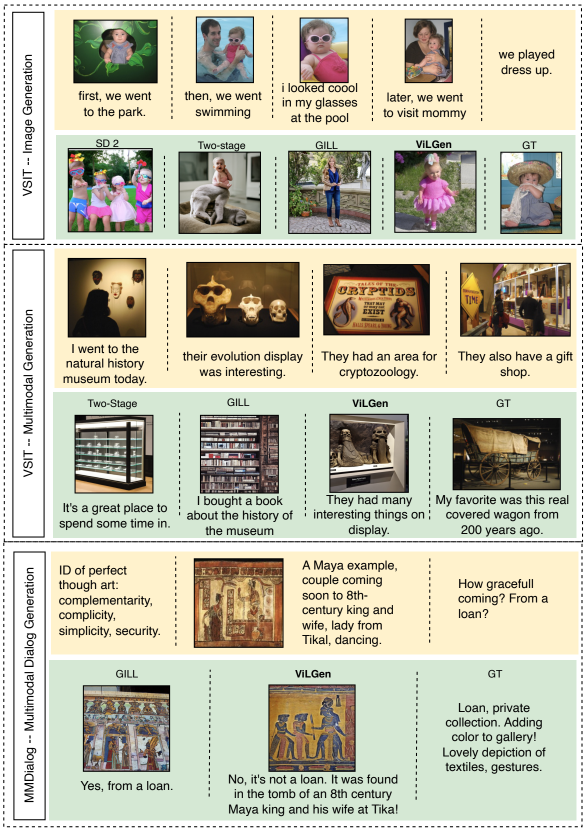


Figure 3: Qualitative examples from ViLGen and baselines on the VIST and MMDialog datasets. The orange blocks indicate the input prompts, while the green blocks include model outputs. The comparisons show that ViLGen can produce coherent and high-quality multimodal output. We would like to emphasize that ViLGen does not use any caption data during fine-tuning on VIST and MMDialog, which obeys to our description-free settings. More qualitative examples can be found in the Appendix E.

generated images and narrations at each following step. Tables 1 and 3 outline the results of these experiments on the VIST validation set, showing the performance in both image and language metrics, respectively. The findings demonstrate that ViLGen can generate coherent, high-quality images utilizing long-horizontal multimodal input prompts across all data, without compromising the original model's ability for multimodal comprehension, indicating the efficacy of our model in diverse settings.

Multimodal Generation on VIST To assess the quality of multimodal generation, we test both our model and the baselines on the VIST validation set by human evaluation. Given a preceding multimodal sequence, models are tasked with producing the subsequent scenario for each task. We select a random sample of 5,000 sequences, with each

Model	IS (↑)	BLEU-1 (↑)	BLEU-2 (↑)	Rouge-L (↑)	MM-Relevance (↑)
Divter (Sun et al., 2021)	20.53	0.0944	0.0745	0.1119	0.62
GILL (Koh et al., 2023)	**23.78**	0.2912	0.1945	**0.1207**	0.64
ViLGen	20.23	**0.3369**	**0.2323**	0.1176	**0.67**

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

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. 4) provided in the Appendix. As depicted in Table 4, our model, ViLGen, is found to generate more fitting text narrations in around 55% of instances, deliver superior image quality in around 53% of cases, and produce more coherent multimodal outputs in around 56% of the scenarios. This data distinctly showcases its enhanced multimodal generation capabilities compared to the two-stage baseline, which must generate intermediate image captions first.

Multimodal Dialog Generation on MMDialog 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. Our results, as presented in Table 5, demonstrate that ViLGen outperforms the baseline model Divter in terms of generating more accurate textual responses. While the image qualities of the generated responses are similar, ViLGen excels in MM-Relevance compared to the baselines. This indicates that our model can better learn how to position image generation and produce highly coherent multimodal responses appropriately.

4.3 Ablation Studies

To further evaluate the effectiveness of our design, we conducted several ablation studies, and more ablation studies can be found in Appendix D.

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

Model	CLIP-I (↑)	CLIP-T (↑)	IS (↑)	FID (↓)
ViLGen	**0.61**	**0.22**	**28.09**	**31.47**
ViLGen (w/o CFG)	0.60	0.22	23.41	33.73
ViLGen (w/o L_{CAP})	0.54	0.16	21.27	40.24
ViLGen (w/o L_{LDM})	0.58	0.20	24.79	34.65

Table 6: Evaluation of different method designs for image generation qualities on the CC3M validation set.

Model	No Context	Text Context	Image Context	Image-Text Context
SD 2 (Rombach et al., 2022b) (Zero-shot)	0.57	0.59	-	-
GILL (Koh et al., 2023) (Zero-shot)	0.54	0.54	0.55	0.54
ViLGen (Zero-shot)	0.54	0.57	0.57	0.57
Fine-tuned SD 2	0.59	0.61	-	-
Two-stage Baseline	0.54	0.56	0.57	0.58
ViLGen (Prefix Tuning)	0.60	0.63	0.68	0.70
ViLGen (LoRA)	**0.61**	**0.64**	**0.69**	**0.70**

Table 7: Influence of prompts for image generation on CLIP-I metrics on VIST. We establish four distinct conditions for the final-step image generation: 'No Context' (solely the last step's narration), 'Text Context' (inclusive of historical textual narrations), 'Image Context' (inclusive of historical images), and 'Image-Text Context' (inclusive of both historical images and narrations). From the results, ViLGen can generate more coherent images.

	CC3M		VIST	
Model	CLIP-I (↑)	FID (↓)	CLIP-I (↑)	FID (↓)
Stable Diffusion 2.1 (Rombach et al., 2022b)	**0.64**	**26.39**	0.59	393.49
GILL (Koh et al., 2023)	0.57	36.85	0.61	376.17
ViLGen	0.61	31.47	**0.66**	**366.62**

Table 8: Generation Qualities on CC3M and VIST. We find that ViLGen is better at extracting features from long-horizontal multimodal information than single text input.

empty caption feature from Stable Diffusion'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 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 voken alignment loss further enhances coherence and image quality performance.

Influence of Input Types for Image Generation To assess the impact of various types of input data for image generation, models are tasked with generating the final-step images based on specific prompts and comparing them with ground truth images by CLIP-I metric. All models are fine-tuned on data with full multimodal context and tested on various input types. As indicated in Table 7, the ViLGen model exhibits exceptional proficiency in producing semantically precise images compared to other models. Furthermore, we observed increased CLIP similarities when more information was provided in the input, signifying the models' enhanced ability to process diverse, long-horizon multimodal inputs.

Text-to-Image Generation Qualities on CC3M Instead of multimodal input, we also test single text-to-image generation qualities on the CC3M validation set, as displayed in Table 8. The results indicate that although our model can have better generation on multi-turn multimodal scenarios, Stable Diffusion 2 achieves the best outcomes across all metrics for pure text-to-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 ViLGen is better at extracting features from long-horizontal multimodal information instead of single text input. This indicates potential future directions on efficiently aligning LLMs with generative models. On the other hand, our model outperforms another state-of-the-art multimodal generation model, GILL, on all metrics, further validating the effectiveness of our design.

5 Conclusion

We introduce ViLGen, designed to augment the capabilities of LLMs for multimodal generation by aligning the LLM with a pretrained text-to-image generation model. Our approach demonstrates substantial improvements. The limitation of ViLGen is that we still find the object texture is hard to maintain in the new generation. Through this work, we aspire to set a new benchmark for existing and future multimodal generative models, opening doors to applications previously deemed challenging due to the disjointed nature of existing image and text synthesis paradigms.

References

Emanuele Aiello, Lili Yu, Yixin Nie, Armen Aghajanyan, and Barlas Oguz. 2023. Jointly training large autoregressive multimodal models. *arXiv preprint arXiv:2309.15564*.

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

Satanjeev Banerjee and Alon Lavie. 2005. 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*, pages 65–72.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Kevin Crowston. 2012. 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*, pages 210–221. Springer.

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

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

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

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. 2023. Photoswap: Personalized subject swapping in images.

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

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

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. 2023. Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*.

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

Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C.H. Hoi. 2023b. LAVIS: A one-stop library for language-vision intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–41, Toronto, Canada. Association for Computational Linguistics.

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

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022a. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022b. High-resolution image synthesis with latent diffusion models. In *CVPR*.

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

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

Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. 2021. Multimodal dialogue response generation. *arXiv preprint arXiv:2110.08515*.

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

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023b. Generative pretraining in multimodality.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. *arXiv preprint arXiv:2010.06775*.

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

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. *Advances in neural information processing systems*, 30.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. *Advances in neural information processing systems*, pages 5998–6008.

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

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

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023c. Next-gpt: Any-to-any multimodal llm.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5.

Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. 2023. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*.

Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2021. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*.

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

A Implementation Details

In the pretraining 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 2×10^{-4}.

In the subsequent fine-tuning stage, we incorporate LoRA modules – denoted as θ_{LoRA} – into Vicuna for the generation of both tokens and vokens. We keep the MLP model θ_{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 2×10^{-5}. Trainable parameters are nearly 6.6 million, and all training can be completed on a server equipped with 4 A6000 GPUs.

B Additional Related Work

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., 2023c; Dai et al., 2023; OpenAI, 2023; Li et al., 2023a; Alayrac et al., 2022; Li et al., 2023b). 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-2 (Li et al., 2023c) 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.

C Experimental Settings

C.1 Datasets

CC3M (Sharma et al., 2018): Conceptual Captions (CC3M) 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. Each dataset sample 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 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 (VIST) 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 (MMDialog) 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 is a robust platform for developing, training, and validating algorithms capable of understanding and generating coherent dialogues that seamlessly blend textual and visual information.

C.2 Data Format

Pretraining Stage In the pretraining stage, we aim 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.

Fine-tuning Stage In this stage, we utilize the VIST and MMDialog datasets, 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 textual and visual content. The history context in the VIST dataset includes all previous story steps with 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.3 Baselines

Fine-tuned Unimodal Generation Models: To facilitate fair comparisons in both image and text generation, we fine-tuned two separate models, Stable Diffusion 2.1 and MiniGPT-4 (Zhu et al., 2023),

Figure 4: Screenshot for human evaluation interface on the Amazon Mechanical Turk crowdsource evaluation platform. Output 1 is generated by ViLGen, while output 2 is generated by the two-stage baseline.

utilizing the VIST dataset. Within the Stable Diffusion 2.1 (Rombach et al., 2022b) model, the U-Net parameters were fine-tuned. For MiniGPT-4's LLM part, LoRA parameters were fine-tuned.

Two-stage Baseline: A common approach in multimodal generation involves first employing Large Language Models (LLMs) to create image captions, which are then fed into text-to-image models for image generation (Wu et al., 2023b). We create such a two-stage baseline for comparison with our end-to-end method by fine-tuning MiniGPT-4 for caption generation and Stable Diffusion 2.1 for text-to-image generation. Given the absence of image descriptions in the VIST dataset, we incorporate a SOTA image captioning model, InstructBLIP-13B (Dai et al., 2023), to generate synthetic captions for supervision.

GILL: 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, where GILL minimizes the Mean Squared Error (MSE) loss between the text-to-image text encoding feature and voken features, similar to L_{CAP} in our approach. For fine-tuning on multimodal datasets, since GILL requires image captions for training, we use Descriptions of Images-in-Isolation (DII) (Huang et al., 2016) in the VIST fine-tuning and generate captions for MMDialog fine-tuning. Contrarily, ViLGen does not related on all caption data during multimodal generation fine-tuning.

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 fine-tuning on a selected set of multimodal response data. The MMDialog dataset regards Divter's method as the baseline.

D More Experiments

D.1 Evaluation of Guidance Scale

Since our model incorporates CFG, evaluating how different guidance scales affect image generation is crucial. Therefore, we plotted several line charts in Fig 5 to depict the changes in metrics with varying guidance scales. The figures reveal that 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 declines. This suggests that the guidance scale should be set within a reasonable range for optimal image generation.

D.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 experimented by training the model with varying numbers of vokens, ranging from 1 to 8. As illustrated in Fig 6, the model's performance consistently improves with adding 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.3 Ablation of Model Designs

This section explores alternatives to the transformer encoder/decoder architecture discussed in the main paper. Specifically, we experimented with two additional settings: **Fixed Queries**, and **Decoder-Only** model where learnable queries are fed into the transformer decoder. For the fixed queries design, we initialize queries the same as learnable queries experiments in the main paper and keep them fixed during training. In the decoder-only approach, we utilize solely the transformer decoder and apply padding to the decoder's output, ensuring that the token length reaches 77. This length adjustment allows the output to be compatible with the Stable Diffusion encoder. The results of these experiments are detailed in Table 9. From the results of ViLGen with fixed queries, we find there exists a slight trade-off between image-text coherence and image qualities, where fixed queries can lead to higher image metrics (IS and FID) but lower CLIP similarities. Meanwhile, ViLGen consistently outperforms the Decoder-Only results in all four evaluation metrics, validating the robustness and efficacy of ViLGen's transformer encoder/decoder architecture design.

Model	CLIP-I (↑)	CLIP-T (↑)	IS (↑)	FID (↓)
ViLGen	**0.61**	**0.22**	28.09	31.47
ViLGen (Fixed Queries)	0.60	0.21	**28.55**	**30.56**
ViLGen (Decoder-Only)	0.58	0.20	24.74	34.88

Table 9: Evaluation of different model designs for image generation qualities on the CC3M validation set.

E More Qualitative Examples

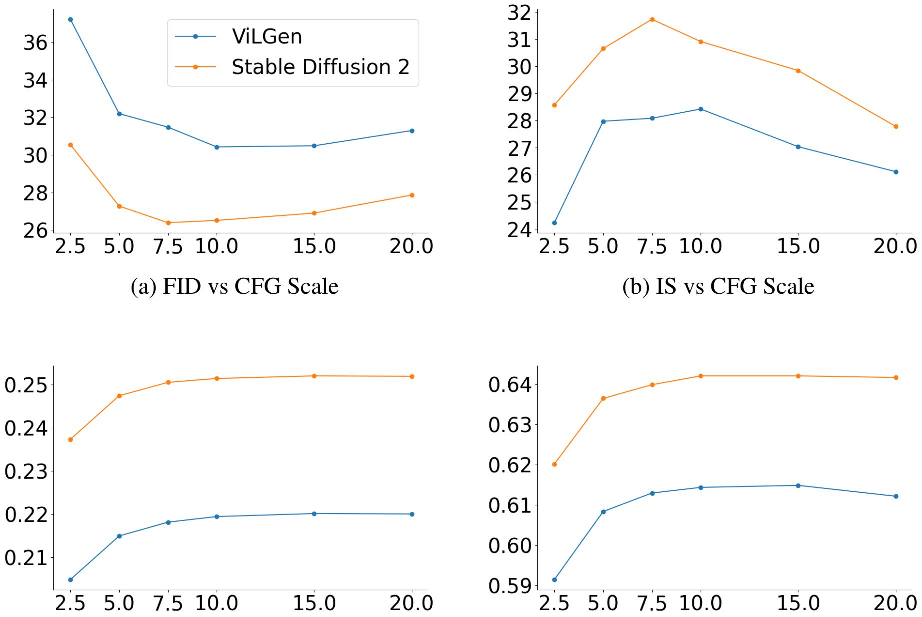
In this section, we provide additional qualitative examples to further demonstrate the capabilities of ViLGen. Figures 7,8,9, and 10 showcase these examples across various datasets and settings.

Figure 7 presents a comparative analysis on the VIST validation set, illustrating how ViLGen outperforms baseline models in terms of image generation quality and alignment with multimodal inputs. The examples highlight the superiority of ViLGen in generating images that closely match the given text prompts.

In Figure 8, we focus on the performance of ViLGen in free multimodal generation scenarios. The results clearly indicate an improvement over the Two-Stage baseline, emphasizing ViLGen's ability to perform consistent and creative multimodal generation.

Figure 9 showcases the application of ViLGen in the context of the MMDialog test set. Here, the emphasis is on free multimodal dialog generation, with ViLGen displaying a decent performance in generating coherent and contextually relevant multimodal dialogues.

Lastly, Figure 10 highlights ViLGen's performance in single text-to-image generation tasks on the CC3M validation set. The examples underline the model's proficiency in generating visually accurate and contextually appropriate images from



(a) FID vs CFG Scale

(b) IS vs CFG Scale

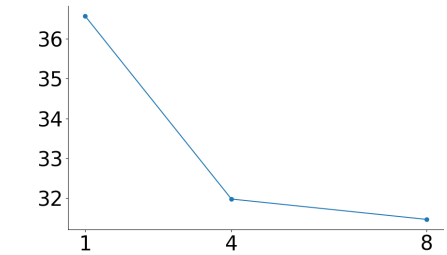
(c) CLIP-T vs CFG Scale

(d) CLIP-I vs CFG Scale

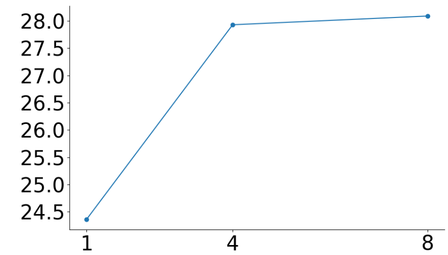
Figure 5: Line charts for various metrics vs Classifier-free Guidance (CFG) scale on CC3M. The results suggest that our CFG strategy can exhibit comparable effectiveness to the CFG strategy employed in SD2, with the appropriate CFG scale significantly enhancing both image quality and coherence.

textual descriptions, surpassing the performance of baseline models.

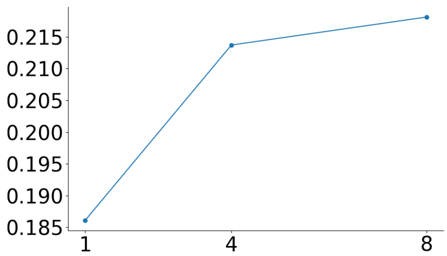
Each figure includes a clear depiction of input prompts (indicated in orange blocks) and the corresponding model outputs (in green blocks), providing a comprehensive view of ViLGen's capabilities across different multimodal generation tasks.



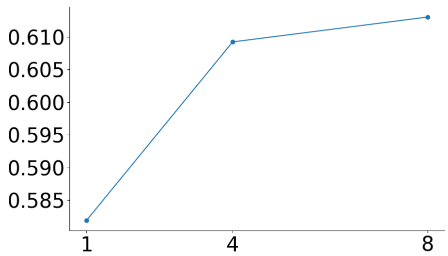
(a) FID vs n_{voken}



(b) IS vs n_{voken}



(c) CLIP-T vs n_{voken}



(d) CLIP-I vs n_{voken}

Figure 6: Line charts for various metrics vs the number of vokens on CC3M. As the number of vokens increases, the image quality and CLIP scores improve. In this work, our default voken number is 8.

Figure 7: Comparative examples from ViLGen and baselines on the VIST validation set for image generation with multimodal input. Orange blocks denote input prompts, while green blocks show model outputs.

Figure 8: More qualitative examples from ViLGen and baselines on VIST validation set for free multimodal generation.

Figure 9: More qualitative examples from ViLGen on MMDialog test set for free multimodal dialog generation.

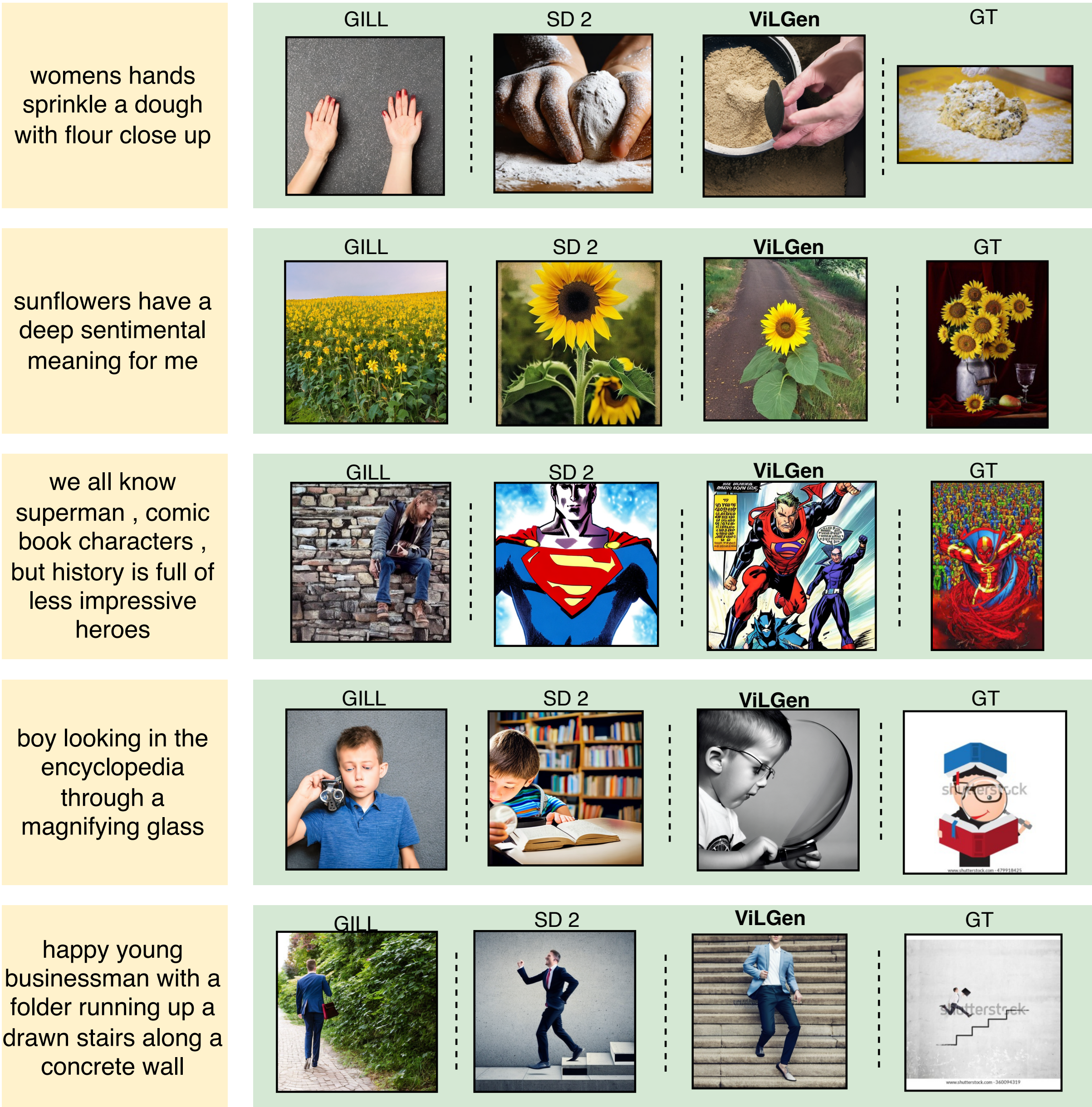


Figure 10: More qualitative examples from ViLGen and baselines on CC3M validation set for single text-to-image generation.