A. Experimental Settings

A.1. Datasets

CC3M [42]: 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 [22]: 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 [15]: 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.

A.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.

B. More Experiments

B.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.

Figure 4. Screenshot for human evaluation interface on the Amazon Mechanical Turk crowdsource evaluation platform. Output 1 is generated by MiniGPT-5, while output 2 is generated by the two-stage baseline.

B.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.

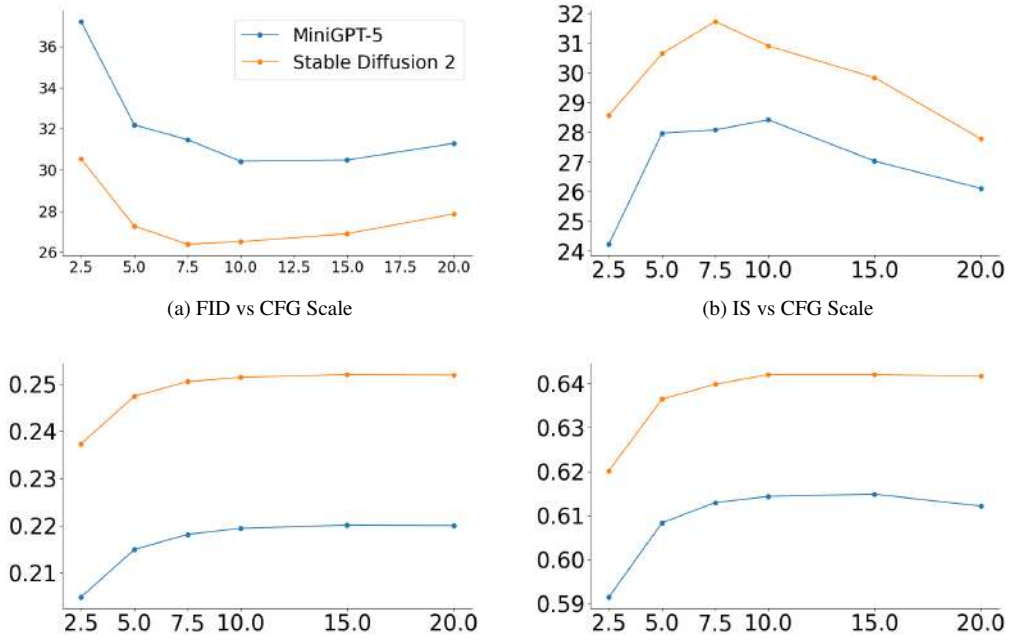
B.3. Texture Preservation via FID

To assess whether models preserve the textures present in prior images when generating the last image of a story, we evaluate on VIST in a "final–step" setting: for each story, all preceding step images and narrations are provided as context, and the model must produce the final image. We then compute FID between the set of generated final images and the ground-truth final images from VIST. This directly measures the realism and low-level appearance consistency of the predicted finals relative to the true finals under an identical multimodal context, shown in Table 9. In this setting, ViLGen (Qwen2.5-VL + SD3) attains the lowest FID, indicating stronger retention of low-level textures in the generated final images.

C. More Qualitative Examples

In this section, we provide additional qualitative examples to further demonstrate the capabilities of MiniGPT-5. 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 MiniGPT-5 outperforms



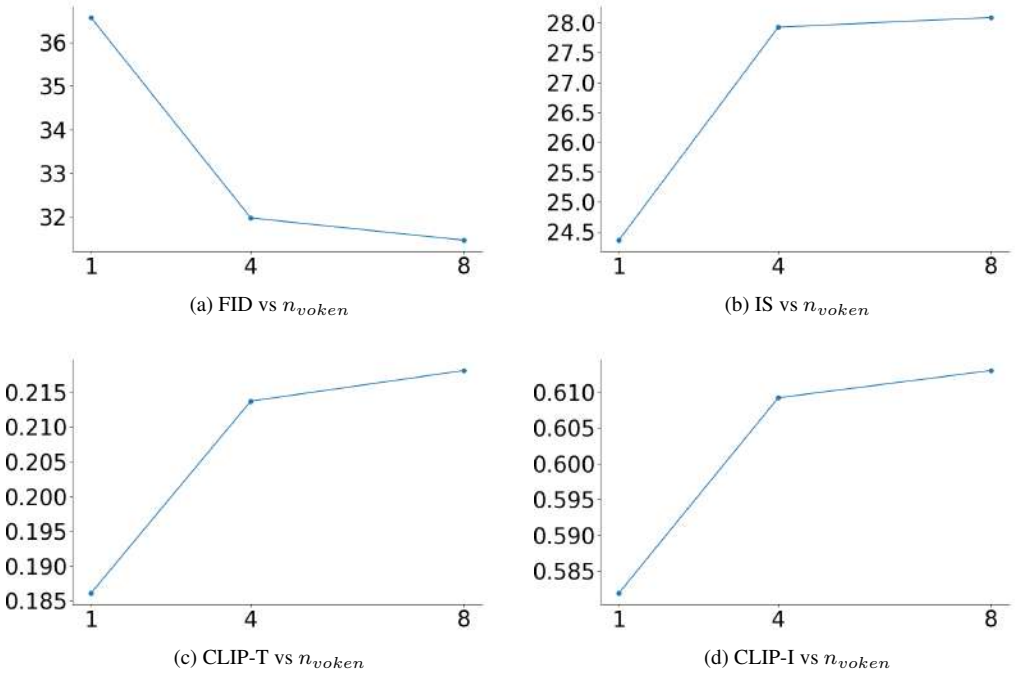
(a) FID vs CFG Scale

(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.



(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.

Model	FID (↓)
GILL [24]	61.85
MiniGPT-5 (MiniGPT-4 + SD 2.1)	59.48
MiniGPT-5 (Qwen2.5-VL + SD3)	**56.32**

Table 9. Texture preservation on VIST (final–step).

baseline models in terms of image generation quality and alignment with multimodal inputs. The examples highlight the superiority of MiniGPT-5 in generating images that closely match the given text prompts.

In Figure 8, we focus on the performance of MiniGPT-5 in free multimodal generation scenarios. The results clearly indicate an improvement over the Two-Stage baseline, emphasizing MiniGPT-5's ability to perform consistent and creative multimodal generation.

Figure 9 showcases the application of MiniGPT-5 in the context of the MMDialog test set. Here, the emphasis is on free multimodal dialog generation, with MiniGPT-5 displaying a decent performance in generating coherent and contextually relevant multimodal dialogues.

Lastly, Figure 10 highlights MiniGPT-5'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 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 MiniGPT-5's capabilities across different multimodal generation tasks.

Figure 7. Comparative examples from MiniGPT-5 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 MiniGPT-5 and baselines on VIST validation set for free multimodal generation.

Figure 9. More qualitative examples from MiniGPT-5 on MMDialog test set for free multimodal dialog generation.

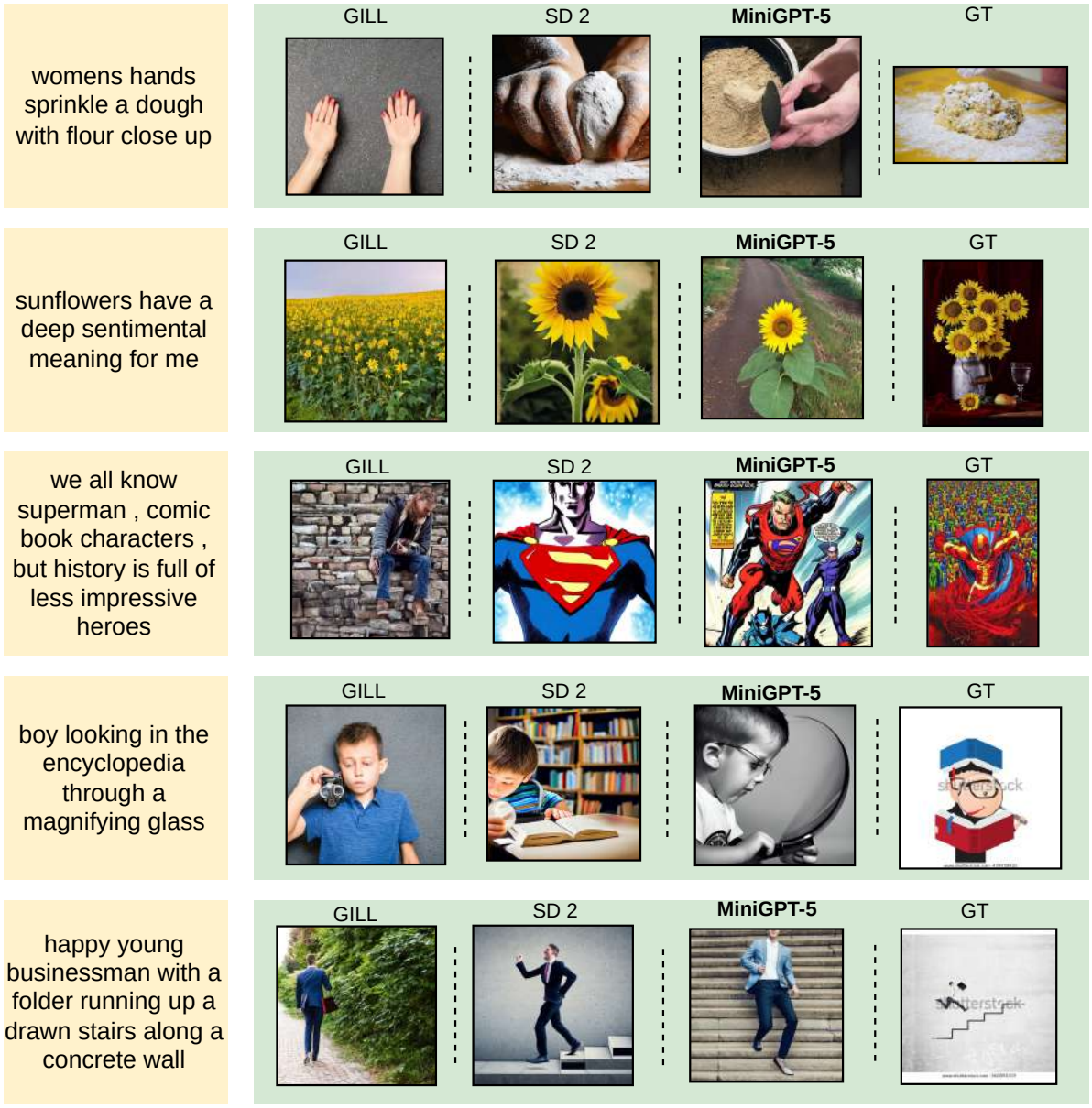


Figure 10. More qualitative examples from MiniGPT-5 and baselines on CC3M validation set for single text-to-image generation.