M^2Chat: Empowering VLM for Multimodal LLM Interleaved Text-Image Generation

Anonymous ACL submission

Abstract

In this paper, we propose M^2Chat, a novel unified multimodal LLM framework for generating interleaved text-image conversation across various scenarios. Specifically, we propose an $M^3Adapter$ that efficiently integrates granular low-level visual information and high-level semantic features from multi-modality prompts. Upon the well-aligned fused feature, $M^3Adapter$ tailors a learnable gating strategy to balance the model creativity and consistency across various tasks adaptively. Moreover, to further enhance the effectiveness of $M^3Adapter$ while preserving the coherence of semantic context comprehension, we introduce a two-stage M^3FT fine-tuning strategy. This strategy optimizes disjoint groups of parameters for image-text alignment and visual-instruction respectively. Extensive experiments demonstrate our M^2Chat surpasses state-of-the-art counterparts across diverse benchmarks, showcasing its prowess in interleaving generation, storytelling, and multimodal dialogue systems.

1 Introduction

In the realm of burgeoning large-scale vision-and-language models (VLMs), the integration of multi-modal features represents more than a mere trend; it is a pivotal breakthrough that is sculpting an extensive range of applications, including object detection (Wang et al., 2023; Lin et al., 2023), Optical Character Recognition (OCR) (Liu et al., 2023c), and Visual-Question-Answering (VQA) (Liu et al., 2023b,a; Zhang et al., 2023c; Zhu et al., 2023; Gao et al., 2023; Lin et al., 2023; Wang et al., 2023). In light of the escalating demand for human-machine chat applications across numerous domains, such as virtual reality, social media, and e-commerce, there is heightened anticipation for VLMs to adeptly interpret and synthesize multi-modality content cohesively for substantially enhancing the quality of conversations. Nevertheless, prevailing research such as MiniGPT-5 (Zheng et al., 2023) and DreamLLM (Dong et al., 2023) has concentrated predominantly on refining the multi-modal alignment (Qi et al., 2023) and interleaving generalization capabilities to enhance performance in tasks like image-editing and long-context generation. However, previous approaches uniformly apply the same knowledge across various tasks, neglecting to account for the task-specific inherent characteristics of VLMs.

As evidenced in previous works, considering employing the VLM on various downstream tasks while preserving coherent semantic comprehension, there are still two challenges: 1) Since the vast and intricately complex multi-modality features from various downstream tasks, it is quite difficult to obtain aligned coherent text-image pairs in a unified space effectively. 2) Directly applying the visual language model is not adequately tailored for modeling the diverse and contextually consistent text-image dialogue from the unified space.

To address the challenges outlined, we introduce M^2Chat, an innovative model for interleaved multimodal generation. M^2Chat adepts at creating text-image pairs that are both contextually consistent and creatively imaginative, tailored with relevant knowledge for diverse tasks. Specifically, by integrating Stable Diffusion XL(Podell et al., 2023) with LLaMA-AdapterV2(Gao et al., 2023), we developed a task-specific Multimodal Multi-level Adapter ($M^3Adapter$). This adapter efficiently integrates low-level visual information and high-level semantic features from multimodality prompts through a learnable gating strategy, effectively balancing the contributions of each modality. This approach maintains a delicate equilibrium in the M^3Chat to balance consistency with incongruity towards diverse tasks.

Meanwhile, we further devised a two-stage Multimodal Mixed Fine-Tuning strategy, denoted as M^3FT, which strategically optimizes distinct sets

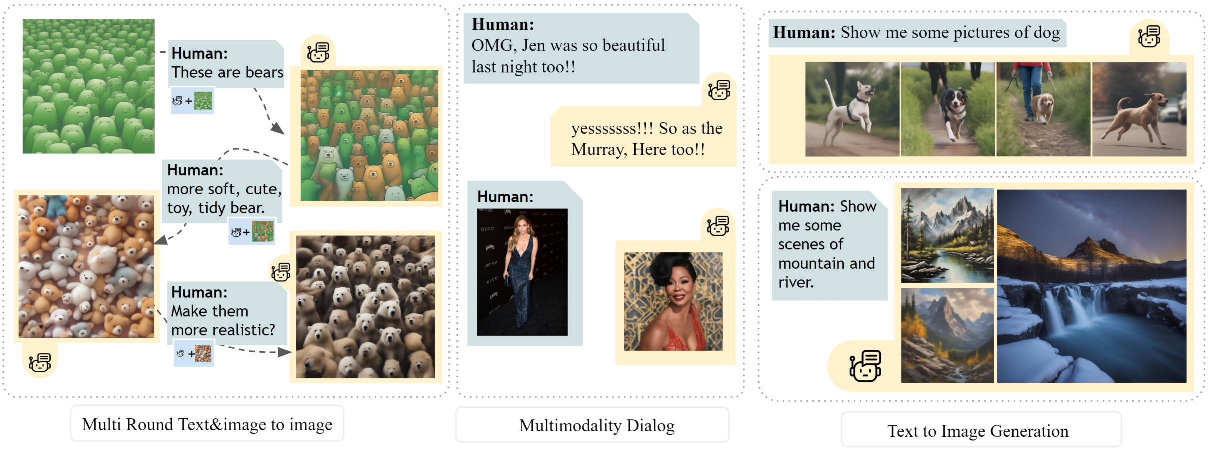

Figure 1: Advanced capabilities of our proposed M^{2}Chat in interleaved multimodal chat, multi-round text and image-to-image generation, and text-to-image generation.

of parameters tailored specifically for image-text alignment and visual-instruction tasks. In the first stage, we finetune the parameter groups for alignment to project the multimodal features with the input dimension of the image generation model. Then, in the second stage, we tailored a specific token and further trained the $M^3Adapter$ components with instruction data from different fields.

Empirical evidence highlights M^{2}Chat's superior capabilities in tasks like image editing, storytelling, and multimodal dialogue, outperforming current models in fine-tuning efficiency and generation quality, with a proficiency in creating imaginary but coherent images and text. The contributions of our study are outlined as follows:

- We have developed M^2Chat, which is an innovative VLM capable of seamless text-image interleaved generation across a range of tasks, especially on complex multimodal dialogue scenarios.
- The $M^3Adapter$ aligns VLM with Stable Diffusion XL for enhanced multimodal fusion, using an adaptive gate for multi-level feature integration, ensuring generation creative-consistency balance for diverse tasks.
- We further design a two-stage tuning strategy M^{3}FT that cooperates with M^{3}Adapter to align text and image while maintaining semantic coherence.

2 Related Work

2.1 Multimodal Large Language Model

Researchers in the field of multimodal large language models have devoted considerable attention to image understanding. KOSMOS-1 (Huang et al., 2023), FROMAGe (Koh et al., 2023b), and BLIP-2 (Li et al., 2023) specifically focused on learning captioning abilities. Others giving attention to improving the fine-tuning capabilities of instructing models like Llava (Liu et al., 2023b), Llava1.5 (Liu et al., 2023a), and MiniGPT4 (Zhu et al., 2023). Moreover, open-source models like LlaVA-NeXT (Liu et al., 2024a) integrate the multiple visual understanding tasks, including object detection and OCR, so as SPHINX(Lin et al., 2023). Some efforts have aimed to incorporate more modalities, as demonstrated in Video-LLaMA (Zhang et al., 2023a). Or, aims at long context movie understanding, like MovieChat(Song et al., 2023). However, only a few recent works have started to expand the modality of output (Zheng et al., 2023).

2.2 VLM Downstream Tasks

Image Generation and Editing. The SOTA generation model has shifted from GAN-based approaches to diffusion, as highlighted in the work by (Nichol and Dhariwal, 2021) and Song (Song et al., 2020). While stable diffusion is renowned for its strong and controllable image generation capabilities, as proposed by SDXL (Podell et al., 2023), other works have explored the editing problem in image generation by manipulating the input prompts, as seen in the studies by Cao (Cao et al., 2023) and Hertz (Hertz et al., 2022). Additionally, Zhang (Zhang et al., 2023b) introduced the concept of adding Controlnet to the diffusion model, which enhances the controllability of diffusion-based image generation.

Interleaving Generation. Recent research has explored various approaches to integrate Multimodal Language Models (VLM) with text-image gener-

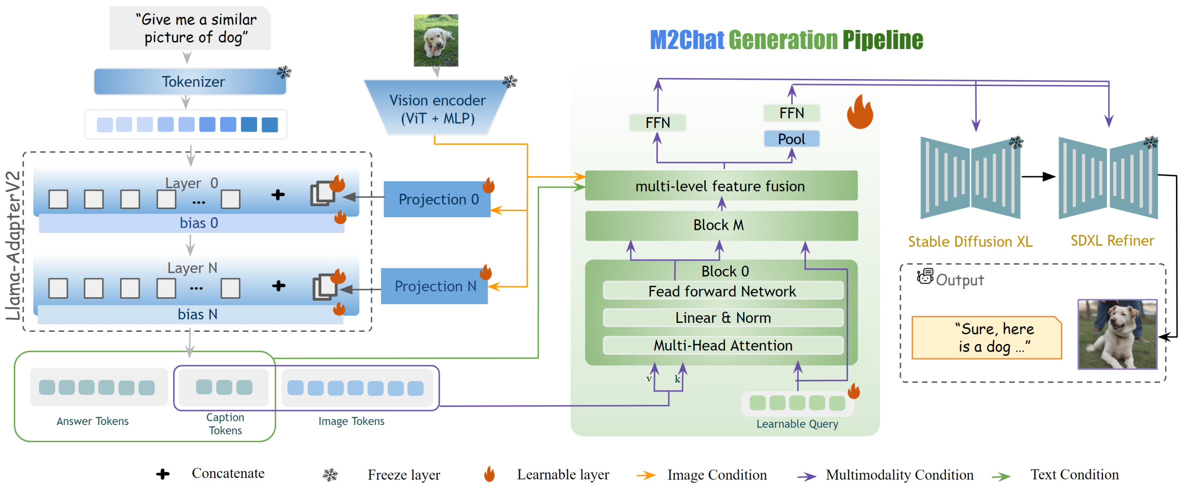

Figure 2: Illustration of M^2Chat, which features a generation pipeline that processes both image and text inputs, harnessing the capabilities of LLaMA-AdapterV2 (Gao et al., 2023) and SDXL (Podell et al., 2023) to craft high-fidelity image-text pairs. Our system excels in three key areas: Text-to-Image (T2I) generation, Storytelling, and Multimodal dialogue. Image generation occurs as VLM forward propagation yields hidden embeddings, which are then utilized to train the M^{3}Adapter—distinguished by its minimal trainable parameters.

ation tasks. DALLE-3 (OpenAI, 2023) relies on prompts for generation without image conditions, while Emu (Sun et al., 2023c), DreamLLM (Dong et al., 2023), and MiniDALLE3 (Lai et al., 2023) fine-tunes VLM for multimodal context generation. NextGPT (Wu et al., 2023) aligns audio, text, and image modalities using adapters. SEED-LLaMA (Ge et al., 2023b,a) aligns LLaMA and generation models with discrete vision tokens. Additionally, chat editing models for 3D models, such as 3D-GPT (Sun et al., 2023a), show promise in this area. Moreover, there are also a lot of explorations of multi-modality generation (Tang et al., 2023; Koh et al., 2023a; Qu et al., 2023; Lian et al., 2023). Despite these efforts, efficient alignment and the full exploration of VLM's generalization ability in text-image interleaved generation remain unexplored.

3 Proposed Method

In this work, we introduce M^{2}Chat, a model that aligns LLaMA-AdapterV2(Gao et al., 2023) with Stable Diffusion XL(Podell et al., 2023) for simultaneous text-image generation across diverse tasks. This part is structured as follows. We first introduce the overarching architecture of our framework, including how we construct the visual instruction, the innovative M^{3}Adapter, and its custom-designed adaptive gate. We then illustrate the advanced two-stage M^{3}FT fine-tuning approach that significantly elevates the generative quality with the multimodal dual-loss objective function

3.1 Preliminary

Confronted with the complexities of generating multimodal dialogues with asynchronously aligned image and text semantics, our novel pipeline, depicted in Fig. 2, leverages the vision-language model LLaMA-AdapterV2 θ_{vlm}(Gao et al., 2023) to synergize with SDXL θ_{sdxl}(Podell et al., 2023). This orchestrates the generation of cohesive text-image conversations. Particularly, we utilize the VLM as a multimodal encoder and integrate a bespoke M^{3}Adapter for aligning multimodal features, thereby streamlining the fusion of text and image narratives, while SDXL facilitates the actual image synthesis.

Visual Instruction Formatting. We begin by detailing our instruction design process. We draw from an image-text dataset $\mathcal{D} : \{\mathcal{X}, \mathcal{Y}\}$, containing pairs of images $\{x\}_{i=1}^{N}$ and their corresponding textual contexts $\{y\}_{i=1}^{N}$, where N is the sample count. To construct the context Y, we adopt the principles of visual instruction tuning (Liu et al., 2024b) and introduce an additional image token $< |img| >$ to denote padding, alongside $< |IC| >$ to signal the start of an image caption. These tokens serve as markers to differentiate token types during the two-stage M^{3}FT training phase.

3.2 Framework Architecture

VLM Encoder. We utilize LLaMA-AdapterV2 as our foundational pre-trained VLM for its robust text-image encoding capabilities. As shown in Fig.

2, each sequence context $\{y\}_{i=1}^{N}$ is encoded into text embeddings $e_{text} \in \mathbb{R}^{length\times 4096}$ using a text encoder. Simultaneously, the corresponding images $\{x\}_{i=1}^{N}$ are encoded by a visual encoder into features $f_{img} \in \mathbb{R}^{length\times 768}$, using a CLIP-based ViT+MLP framework (Radford et al., 2021).

Text-Image Token Generation. The VLM outputs a sequence of hidden tokens $t_{out} \in \mathbb{R}^{length\times 4096}$, which are divided into answer tokens $t_{ans} \in \mathbb{R}^{length_{ans}\times 4096}$, caption tokens $t_{cap} \in \mathbb{R}^{length_{cap}\times 4096}$, and image tokens $t_{img} \in \mathbb{R}^{length_{img}\times 4096}$. Answer tokens are decoded into text by LLaMA, while image generation tokens provide foundational features for synthesizing images.

Multimodal Multi-level Adapter: The Multimodal Multi-level Adapter (M^{3}Adapter), denoted as θ_{m^3a}, addresses SDXL's limited token capacity for text-image interactions. It integrates with the image decoder to deliver consistent outputs using cross-attention and linear layers, which $Q = \mathcal{W}_Q^{(i)} \cdot query$, $K = \mathcal{W}_K^{(i)} \cdot h_l$, $V = \mathcal{W}_V^{(i)} \cdot h_l$, and $\mathcal{W}^{(i)}$ are learnable matrices. The M^{3}Adapter aligns VLM outputs $h_0 = t_{\{cap,img\}}$ with SDXL text encoder outputs using MSE loss:

$$\mathcal{L}_{align} = (h_{palign,l} - e_{pclip})^2 + \frac{1}{77}\sum_{k=1}^{77}(h_{align,l}^{(k)} - e_{clip}^{(k)})$$

Direct alignment limits creativity, so we use a multi-level feature fusion strategy to incorporate low-level visual features f_{img} into high-level multimodal features h_l, modulated by a learnable gate:

$$f_{fus} = (1 - \frac{e_{ans}\cdot e_{cap}}{\|e_{ans}\|\|e_{cap}\|}) \times f_{img} + \frac{e_{ans}\cdot e_{cap}}{\|e_{ans}\|\|e_{cap}\|} \times h_l$$

This adaptive fusion supports resilient image generation, balancing creativity and coherence for multimodal dialogue and other tasks.

3.3 Training Strategy

First Stage in M^{3}FT for Alignment. We initially fine-tune the model to align multimodal features using M3Adapter. During the denoising phase, aligned features h_{align} and h_{palign} condition SDXL's UNet θ_{unet}:

$$h_{unet} = \theta_{unet}(\delta_{noise}(\mathcal{I}, \lambda), h_{align}, h_{palign}, \lambda)$$

where $\mathcal{I}$ is the VAE encoder image feature with added noise. The DDPM loss is:

$$\mathcal{L}_{ddpm} := \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),\lambda}\left[||\epsilon - h_{unet}||^2\right]$$

We apply alignment loss $\mathcal{L}_{align}$ to enhance generation quality:

$$\mathcal{L}_{M^2FT} = \mathcal{L}_{ddpm} + \varphi \cdot \mathcal{L}_{align}$$

where φ is a hyperparameter. Note that only the M^{3}Adapter undergoes updates during the initial M^{3}FT stage. Our model aligns the VLM feature space with SDXL, achieving success in diverse multimodal generation tasks. We provide in-depth visualization and CLIP performance post-first stage training in Sec.4.

Second Stage in M^{3}FT for Consistency. For Multimodal Mixed Fine-Tuning (M^{3}FT), the target is to tune the model and generate the answer and the image tokens. Since the complexity of MMDialog, the answer and the image have inconsistency in their meaning. In M^{3}FT the LLM is tuned by the loss group and DDPM at the same time. We separate the answer token and the caption tokens, tuning the model on the text-image to text-image patterns. In this round, we tune all components of the M3Adapter, including the bias of the LLaMA, the projection of visual tokens, the M^{3}FT factor, and the adapters. As shown in the pipeline, each component would be affected multiple times of differences, which would speed up the training process, and efficiently align the components. The overall optimistic function of M^{3}FT is as follows

$$\mathcal{L}_{M^3FT} = \mathcal{L}_{ddpm} + \varphi \cdot \mathcal{L}_{align} + \cdot\mathcal{L}_{text} \qquad (1)$$

where $\mathcal{L}_{text}$ represents the text conditioning loss, assessing the discrepancy between generated tokens and labels.

4 Experiments

In this section, we analyze and evaluate the generation performance of M^2Chat and the efficiency of M^{3}Adapter and M^{3}FT across various tasks. The empirical results demonstrate the superiority of our proposed methods against other state-of-the-art baselines in generation quality and semantic consistency.

4.1 Downstream Tasks

Our paper enhances multimodal LLMs for interleaved generation tasks, producing related and intertwined text and images. Specifically, the interleaving generation task can be defined into several sub-tasks:

- **Chat-based image generation** requires the model to discern and react on often vague user inputs, extracting key elements to produce diverse images that match user intent, showcasing both comprehension and creative alignment with user specifications.
- **Interleaving generation** aims to perform basic editing operations based on text instructions. During the editing process, the model emerges with the ability to comprehend human commands and make appropriate editing based on the understanding.
- **storytelling** requires the model to weave a coherent narrative with corresponding images, ensuring each image reflects the unfolding story. This demands a deep understanding of context and the ability to create rich text and visuals, delivering an immersive narrative experience.
- **Multimodal Dialogue** diverges from traditional ones by tackling inconsistencies in text-image pairs. VLM must go beyond describing images to generating relevant dialogues and topic-specific visuals, enriching the conversation with images more than content visualization.

4.2 Experiment Setup

Datasets. To minimize the domain gap between LLaMA-AdapterV2 (Gao et al., 2023) and SDXL (Podell et al., 2023), we tuned M^{2}Chat on CC3M (Sharma et al., 2018) and LAION-Aesthetics (Schuhmann et al., 2022). Additionally, we used the COCO-Caption dataset (Lin et al., 2015) for its rich object descriptions. LAION-Aesthetics, a subset of LAION-5B, enhances generalization quality. We evaluated our model on the following datasets:

- MMS-COCO-Validation (Lin et al., 2014): a subset of the MS-COCO dataset used for tasks like object detection and segmentation.
- CC3M (Sharma et al., 2018) (Conceptual Captions 3 Million): a large web-sourced image-caption dataset aimed at image understanding and caption generation.
- MMDialog (Feng et al., 2022): contains annotated dialogues with visual information to facilitate multimodal dialogue research.

Evaluation Metrics. We evaluate our methodology using a combination of text-image generation metrics that assess both textual and visual dimensions. For visual quality and text-image congruence, we employ CLIP-based metrics (**CLIP**) and Frechet Inception Distance (**FID**). Textual analysis is conducted using **BLEU-1**, **BLEU-2**(Papineni et al., 2002), and **ROUGE**(Lin, 2004). To address the specific needs of multimodal dialogue evaluation, we introduce **InterRel**, a novel metric that leverages CLIP embeddings to measure the alignment and contextual harmony between generated texts and images, following the MM-Relevance framework (Feng et al., 2022).

Baselines. We compared our model against multiple SOTA models targeting different perspectives:: Stable Diffusion (1.5 and SDXL) (Podell et al., 2023), can enerates detailed images from text. Emu (Sun et al., 2023c) and Emu2 (Sun et al., 2023b), are pre-trained models for quality visuals. SEED-LLaMA (Ge et al., 2023a)which enhances LLMs with an image tokenizer. NExT-GPT (Wu et al., 2023) integrates an LLM with multimodal adaptors and diffusion decoders. Besides, DreamLLM (Dong et al., 2023) and MiniGPT5 (Zheng et al., 2023), which been mentioned in Sec. 1.

Implementaotin Details. Our model was trained end-to-end on eight H800 GPUs. As illustrated in Fig. 2, we focused on training the M^{3}Adapter exclusively. The VLM backbone, LLaMA-AdapterV2 7B, was paired with CLIP(ViT-L/14)(Radford et al., 2021) for visual encoding. The M^{3}Adapter's parameters occupy 299Mb, with an inference memory of 28Gb. During the First Stage in M^{3}FT for Alignment, we initialized a learning rate of 1e^{-4}, a batch size of 8, and conducted over 4 epochs, the training required approximately 80 GPU hours in total. We trained on a subset of CC3M(Sharma et al., 2018) with around 1.5 million image-text pairs.

During the second stage in M^{3}FT for Alignment, we initialized a learning rate of 1e^{-5}, a batch size of 1, and conducted over 20 epochs, the training required approximately 30 GPU hours in total. We train all the adapters by a mixture dataset, with 4k image-text instruction paired data extracted from CC3M, and 7k MMdialog conversation pairs from the training set of MMDialog(Feng et al., 2022). The learning rate is initialized at 1e-4, and decays 10 times each five epochs.

4.3 Quantitative Results

In our evaluation, we conducted a performance comparison of our model, M^2Chat, on the MS-COCO 2014 and CC3M validation datasets, as

Table 1: Evaluation results based on FID and CLIP on CC3M and MS-COCO 2014 Validation set.

Models	MS-COCO 2014		CC3M	
	LLM Size	CLIP ↑	FID ↓	CLIP ↑
SD 1.5	-	30.62	30.62	23.48
SDXL (Podell et al., 2023)	-	31.17	24.26	29.91
Emu (Sun et al., 2023c)	13B	28.6	-	-
Emu2-Gen (Sun et al., 2023b)	33B	29.7	-	-
NeXT-GPT (Wu et al., 2023)	7B	29.31	-	-
MiniGPT5 (Zheng et al., 2023)	7B	-	31.47	22.00
M^{2}Chat	7B	28.46	28.71	21.87
M^{2}Chat (M^{3}FT)	7B	29.87	26.15	23.51

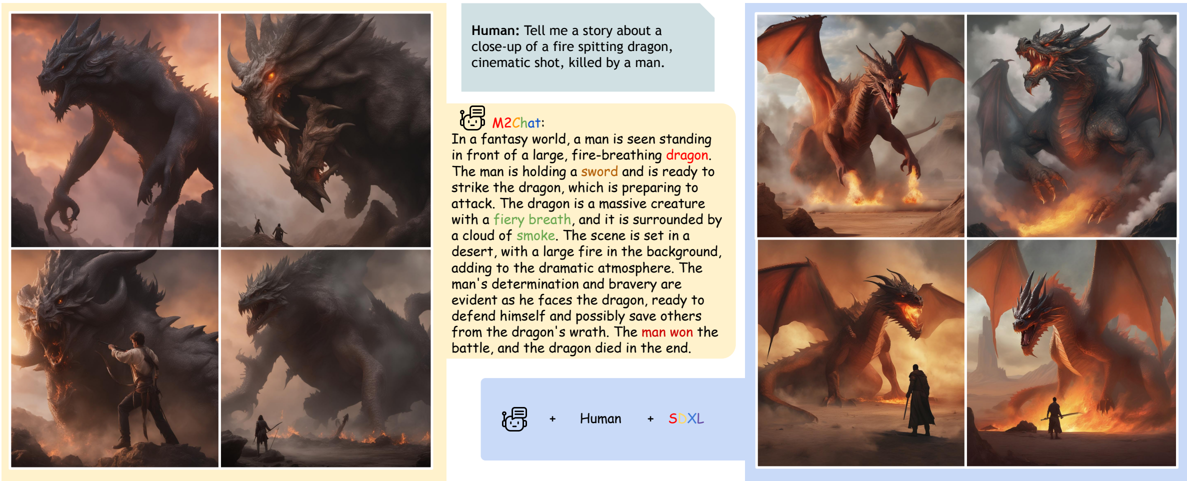

Figure 3: The storytelling pipeline involves the generation of four pictures and a corresponding text story. In this particular example, the human initiates a request to generate a story, starting with the first sentence about a dragon. M^2Chat can generate pictures that are highly consistent with the story and closely aligned with the intended narrative. To compare the results, the human utilizes the prompt from M^2Chat to generate four pictures using the SDXL method. The blue blocks assess and contrast the images produced.

Table 2: Evaluation results of BLEU-1(B1), BLEU-2,(B2), ROUGE-L(RL), and InterRel(IR) on MMDialog Validation set.

Models	LLM	B1↑	B2↑	RL↑	IR↑
VLM+SD finetune	Vicuna 7B	4.21	4.18	6.78	20.05
M^{2}Chat	LLaMA 7B	6.02	5.88	10.14	24.68
M^{2}Chat(M^3FT)	LLaMA 7B	6.98	6.44	11.40	25.57

outlined in Table 1. Our results demonstrate the competitive performance of M^2Chat compared to other generative models.

MS-COCO dataset Our model achieves a SOTA score of 29.87, surpassing other multimodality generation models by a margin of 0.56. The score also notably outperforms NExT-GPT (Wu et al., 2023), and slightly surpasses the large-scale pre-trained model Emu2 (Sun et al., 2023b).

CC3M validation set We compared our results with MiniGPT5 (Zheng et al., 2023), which has a similar-sized LLM to M^{2}Chat. Our model demonstrates superior performance, achieving a 2.56 improvement in the FID score and a 1.51 improvement in the CLIP score.

MMDialog We compared our model, M^{2}Chat, with the baseline model VLM+SD finetune, using the same pretraining and finetuning settings. Our alignment method showed significant improvements: a 5.52 increase in InterRel, 2.77 in BLEU-1, 2.16 in BLEU-2, and 4.62 in ROUGE-L scores. Note that the baseline model, LlaMA-AdapterV2, was not fine-tuned for chat applications, resulting in lower language scores.

4.4 Qualitative Comparisons

Image Generation Quality As shown in Fig. 5, our pipeline generates high-resolution images in

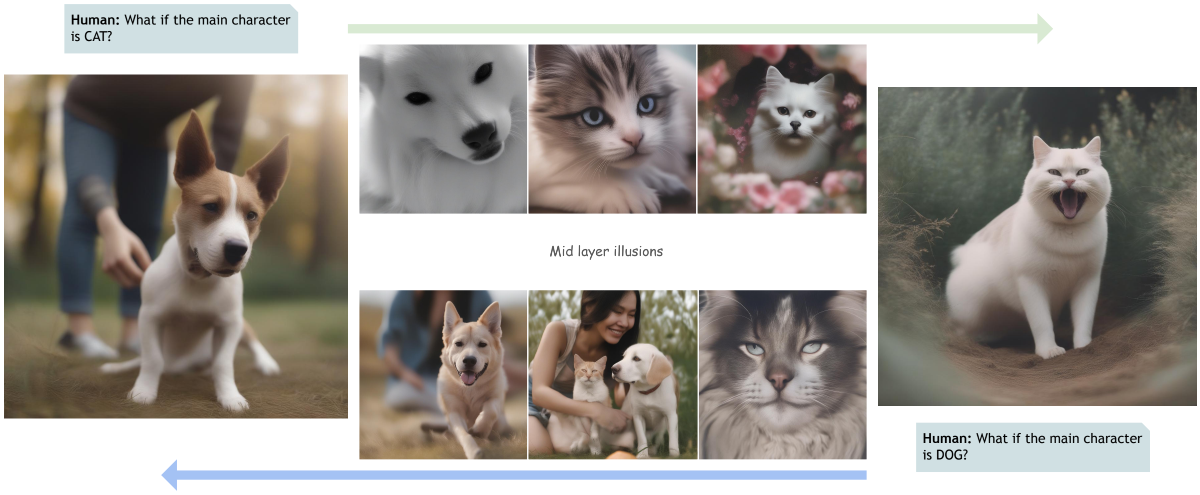

Figure 4: Visualization of the transformation of the hidden features while doing the instruction editing task. Giving the Dog picture and human instruction, the hidden features of VLM gradually transform its representation from dog to cat. The opposite instruction, which turns the cat into a dog, also shows a similar transformation step.

Figure 5: Generation performance comparison of one-stage M^{2}Chat with SD1.5 and SDXL in Text-to-Image generation task.

Figure 6: Examples of interleaved zero-shot image editing. M^{2}Chat consistently demonstrates excellent representation consistency while adhering to the editing instructions.

different contents. It is demonstrated that our efficient alignment methods adapt the prompts well, as described in the quantitative results. M^2Chat without M^{3}FT is compared with SDXL-base and SD1.5 for a fair comparison. Here, the generalization resolution is 1024×1024. In conclusion, M^{2}Chat is able to fit the prompt better than SD1.5. We provide more generation results in Appendix.

Storytelling We show the storytelling ability of M^{2}Chat on Fig. 3. While asking the M^{2}Chat to tell us a story, it generates a story composed of text together with four pictures that follow the storyline. In comparison, we made a set of pictures that were artificially produced: fix the random seed of the SDXL, use the prompts generated by M^2Chat, and feed it to the SDXL. Our method shows high consistency of the text-images among multi-turn conversations. M^2Chat performs better in showing the progress of the story, especially in the last two pictures. It shows the progress from "defend himself " to "the dragon died" since the SDXL is limited by the prompt size. We provide more comprehensive generation results in Appendix.

Interleaved editing Interleaved zero-shot image editing refers to the process of modifying images based on textual instructions without the need for paired image-text data during training. The goal is to achieve consistent and accurate image editing results by leveraging the learned representations from a pre-trained model. M^{2}Chat consistently demonstrates excellent representation consistency while faithfully adhering to the editing instructions. As shown in the Fig. 6, M^{2}Chat can edit the pose, replace the character, give a similar picture, etc.

Multi-level feature visualization As previously mentioned in Sec. 3, we employed multi-level fea-

Table 3: Comparison of parameter size and training cost with other multimodality generation models

Models	LLM	Extra parameter	Data scale	Task	Wall-clock time
Emu2 (Sun et al., 2023b)	LLaMA 33B	4B	100M	TI → TI	-
CAFE (Zhou et al., 2023)	LLaMA 70B	4B	-	TI → TI	20000×A100 Hrs
SEED-LLaMA-8B (Ge et al., 2023a)	Vicuna 7B	1B	-	TI → TI	9000 A100(40G) Hrs
SEED-LLaMA-14B (Ge et al., 2023a)	LLaMA 13B	1B	-	TI → TI	14000 A100(40G) Hrs
DreamLLM (Dong et al., 2023)	Vicuna 7B	-	32M	TI → TI	2240 A100 Hrs
MiniGPT5 (Zheng et al., 2023)	Vicuna 7B	-	2.5M	TI → TI	-
M^{2}Chat	LLaMA 7B	**299M**	**2M**	TI → TI	**100 A100 Hrs**

ture fusion in our approach. Additionally, we visualized the hidden layer features of the LLM. In Fig. 4, we presented the process wherein M^{2}Chat effectively adheres to given instructions, resulting in the transformation of the dog depicted in the left image into a cat in the corresponding right image. The model takes human instruction and the picture as input, and outputs the image captions as well as the edited pictures. Furthermore, as shown in the Fig. 1, M^{2}Chat also supports multi-round editing. More results will be shown in Appendix B.

4.5 Ablation Study

Ablation of M^{3}FT In this paper, we claim that the low-level visual information and high-level semantic features have different effects on the final generalization. Hidden layers inside the VLM contain different levels of information and show a strong tendency to transition from the given state to the output state. To visualize the difference between layers, the Fig. 4 shows the visualization of middle layers in the text-image and image-text tasks. Furthermore, as shown in the Tab. 1, we compare the M^{2}Chat with the no M^{3}FT version. With the M^{3}FT, the CLIP score of MS-COCO improves by 1.39. On CC3M dataset, the M^{3}FT improves 2.56 in FID score, and 1.64 in the CLIP score. Both qualitative results and quantitative results illustrate the importance and efficiency of M^{3}FT.

4.6 Efficiency Comparison

Inspire by a series of finetuning methods(Mangrulkar et al., 2022; Zhang et al., 2024), in Table 3, we present a comparison of the training costs between our model and other multimodality and multitask generation models. The results demonstrate that M^{2}Chat outperforms the other methods in terms of parameter efficiency and low training costs. For instance, Emu (Sun et al., 2023c) and SEED (Ge et al., 2023b) focus on training large multimodality models without fully utilizing the potential of pre-trained components, resulting in training costs exceeding 10,000 GPU hours. Similarly, DreamLLM (Dong et al., 2023) incorporates learnable tokens to fine-tune the LLM for both understanding and generalization abilities, which incurs training costs exceeding 2,240 GPU hours. In comparison, M^{2}Chat demonstrates a close data scale to MiniGPT5, around 2.5 million. Moreover, the additional parameters in M^{2}Chat are highly efficient when compared to the billion-level parameters found in other works.

5 Limitation

We introduce a novel interleaved text-image generation framework called M^2Chat, which is capable of generating text and images simultaneously. However, though we find this framework is suitable for most interleaved generation tasks, it still needs a task-specified instruction tunning to improve its application ability. This means the potential of this framework is still under discovery. We believe with this work, further applications including interleaved image editing, storytelling generation, multi-modal conversation, and other interleaved tasks will be inspired and improved.

6 Conclusion

In this paper, we present M^2Chat, a novel multimodal interleaved text-image generation framework that can generate text and images simultaneously. M^2Chat is constructed on the VLM LLaMA-AdapterV2, integrated with SDXL. We leverage a lightweight module M^{3}Adapter to achieve multimodal feature alignment. Moreover, we further integrate the low-level features with high-level features via an innovative gating strategy to balance the model's creativity and coherence. Last but not least, we propose a two-stage M^{3}FT to further enhance semantic consistency. Extensive experiments demonstrate the superiority of M^{2}Chat across various multimodal interleaved tasks.

References

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. 2023. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *arXiv preprint arXiv:2304.08465*.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. 2023. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*.

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*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

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

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2023b. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. 2023. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*.

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

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023b. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*.

Zeqiang Lai, Xizhou Zhu, Jifeng Dai, Yu Qiao, and Wenhai Wang. 2023. Mini-dalle3: Interactive text to image by prompting large language models.

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

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. 2023. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. 2023. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge, january 2024a. *URL https://llava-vl. github. io/blog/2024-01-30-llava-next*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. 2023c. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR.

OpenAI. 2023. Improving image generation with better captions. *Preprint*, arXiv:2303.08774.

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*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Xingqun Qi, Jiahao Pan, Peng Li, Ruibin Yuan, Xiaowei Chi, Mengfei Li, Wenhan Luo, Wei Xue, Shanghang Zhang, Qifeng Liu, et al. 2023. Weakly-supervised emotion transition learning for diverse 3d co-speech gesture generation. *arXiv preprint arXiv:2311.17532*.

Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. 2023. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 643–654.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

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.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. 2023. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Chunyi Sun, Junlin Han, Weijian Deng, Xinlong Wang, Zishan Qin, and Stephen Gould. 2023a. 3d-gpt: Procedural 3d modeling with large language models. *Preprint*, arXiv:2310.12945.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023b. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. 2023c. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. 2023. Any-to-any generation via composable diffusion. *arXiv preprint arXiv:2305.11846*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023b. Adding conditional control to text-to-image diffusion models.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023c. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Rongyu Zhang, Zefan Cai, Huanrui Yang, Zidong Liu, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, Baobao Chang, Yuan Du, et al. 2024. Vecaf: Vlm-empowered collaborative active finetuning with training objective awareness. *arXiv preprint arXiv:2401.07853*.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2023. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*.

Yufan Zhou, Ruiyi Zhang, Jiuxiang Gu, and Tong Sun. 2023. Customization assistant for text-to-image generation. *arXiv preprint arXiv:2312.03045*.

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*.