UNIMO-G: Unified Image Generation through Multimodal Conditional Diffusion

Wei Li[*], **Xue Xu**[*], **Jiachen Liu, Xinyan Xiao**
Baidu Inc., Beijing, China
`{liwei85.2023}@gmail.com`
`{xuxue,xiaoxinyan,liujiachen}@baidu.com`

Abstract

Existing text-to-image diffusion models primarily generate images from text prompts. However, the inherent conciseness of textual descriptions poses challenges in faithfully synthesizing images with intricate details, such as specific entities or scenes. This paper presents **UNIMO-G**, a simple multimodal conditional diffusion framework that operates on multimodal prompts with interleaved textual and visual inputs, which demonstrates a unified ability for both text-driven and subject-driven image generation. UNIMO-G comprises two core components: a Multimodal Large Language Model (MLLM) for encoding multimodal prompts, and a conditional denoising diffusion network for generating images based on the encoded multimodal input. We leverage a two-stage training strategy to effectively train the framework: firstly pre-training on large-scale text-image pairs to develop conditional image generation capabilities, and then instruction tuning with multimodal prompts to achieve unified image generation proficiency. A well-designed data processing pipeline involving language grounding and image segmentation is employed to construct multi-modal prompts. UNIMO-G excels in both text-to-image generation and zero-shot subject-driven synthesis, and is notably effective in generating high-fidelity images from complex multimodal prompts involving multiple image entities.

1 Introduction

Recent advancements in text-to-image (T2I) diffusion models have yielded impressive results in the generation of high-fidelity images from textual descriptions. Various methods, including DALL-Es (Ramesh et al., 2022; Betker et al., 2023), Imagen (Saharia et al., 2022), Stable Diffusion (Rombach et al., 2022), and MM-DiT (Esser et al., 2024), have been successful in producing photorealistic and contextually relevant images based on textual prompts. Nevertheless, a fundamental challenge persists due to the inherent brevity of textual descriptions, particularly when intricate details, specific entities, or nuanced scenes are involved. Thus, faithfully generating images from general vision-language (VL) inputs is essential to improve the controllability of image generation.

Numerous studies have explored VL-to-image generation techniques. Methods such as DreamBooth (Ruiz et al., 2023), Imagic (Kawar et al., 2023), SuTI (Chen et al., 2023) and BLIP-Diffusion (Li et al., 2023a) emphasize subject-driven generation, where they use both subject images and textual descriptions as inputs to recontextualize the subject in a newly described setting. They either fine-tune specific models for a given subject or employ pre-trained subject representations. However, their specific training design and input templates hinder their scalability, especially in complex scenarios with multiple entities. Additionally, studies like FastComposer (Xiao et al., 2023) and Subject-Diffusion (Ma et al., 2023) focus on multiple-entity image generation, integrating image embeddings from image encoders with the standard text conditioning in pre-trained diffusion models. Nevertheless, these approaches lack the capacity to efficiently process generalized vision-language inputs that comprise a mix of textual and visual information in free forms.

In this paper, we propose **UNIMO-G**, a simple multimodal conditional diffusion framework that operates on multimodal prompts comprising free-form interleaved vision-language inputs. Unlike traditional text-only prompts, multimodal prompts encompass various combinations of image entities and textual elements, as demonstrated in Figure 1. UNIMO-G is designed to faithfully reproduce all image entities, render textual content, and follow the instructions in multimodal prompts. Specifically, we leverage the perception capabilities of

[*]These authors contributed equally to this work.
`https://unimo-ptm.github.io/`

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6173–6188
August 11-16, 2024 ©2024 Association for Computational Linguistics

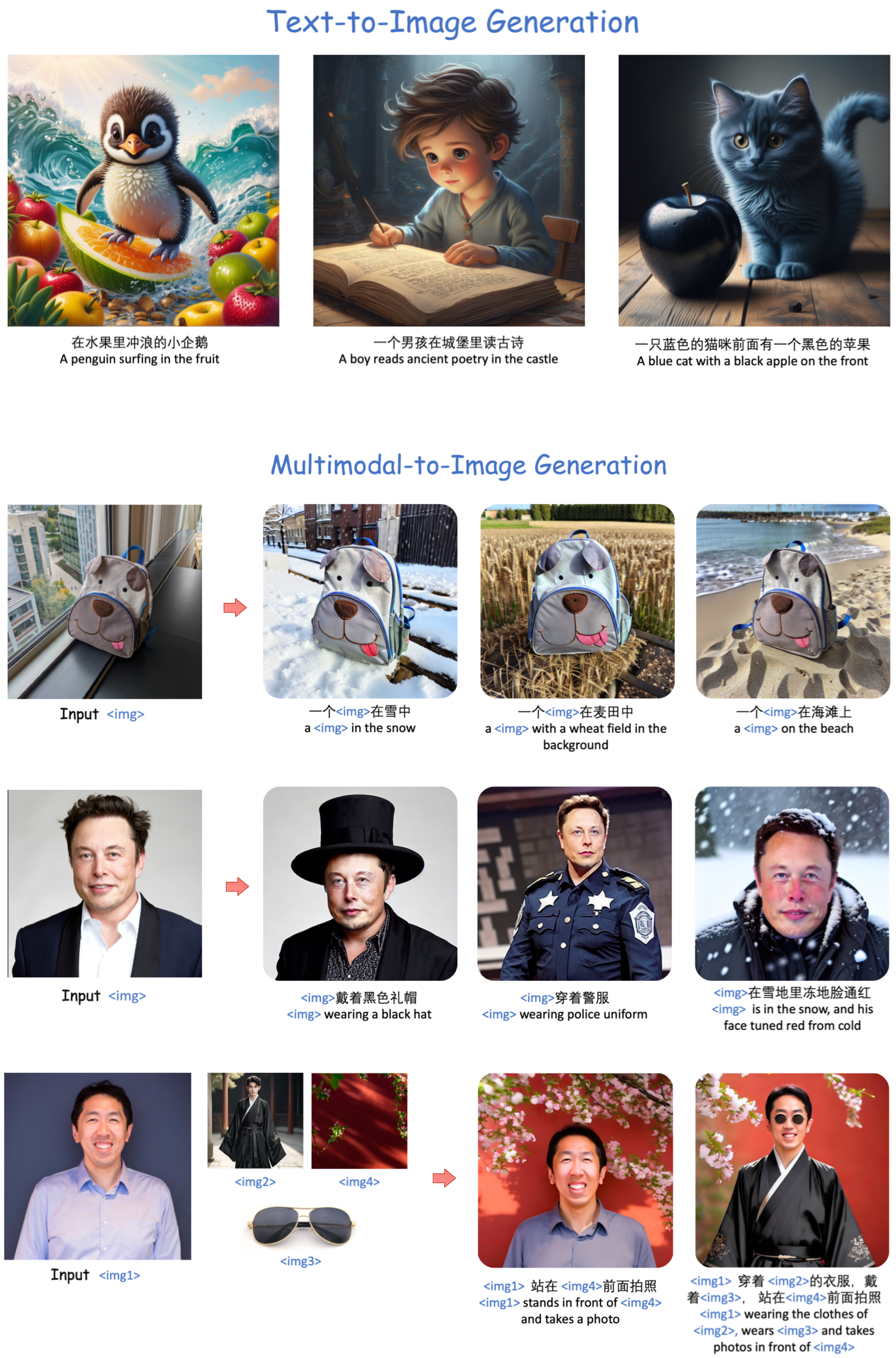

Text-to-Image Generation
在水果里冲浪的小企鹅
A penguin surfing in the fruit
一个男孩在城堡里读古诗
A boy reads ancient poetry in the castle
一只蓝色的猫咪前面有一个黑色的苹果
A blue cat with a black apple on the front
Multimodal-to-Image Generation
Input <img>
一个<img>在雪中
a <img> in the snow
一个<img>在麦田中
a <img> with a wheat field in the background
一个<img>在海滩上
a <img> on the beach
Input <img>
<img>戴着黑色礼帽
<img> wearing a black hat
<img>穿着警服
<img> wearing police uniform
<img>在雪地里冻地脸通红
<img> is in the snow, and his face tuned red from cold
Input <img1>
<img2>
<img4>
<img3>
<img1> 站在 <img4>前面拍照
<img1> stands in front of <img4> and takes a photo
<img1> 穿着 <img2>的衣服，戴着<img3>， 站在<img4>前面拍照
<img1> wearing the clothes of <img2>, wears <img3> and takes photos in front of <img4>

Figure 1: Examples of UNIMO-G for both text-driven and zero-shot subject-driven generation. UNIMO-G can perceive free-form interleaved visual-language inputs and faithfully generate images. Particularly, it can generate images from multi-modal prompts with multiple image entities.

Multimodal Large Language Models (MLLMs) to encode multimodal prompts into a unified vision-language semantic space. Subsequently, a conditional diffusion network generates images from these encoded representations.

To train UNIMO-G efficiently, we implement a two-phase strategy. Initially, the model undergoes pre-training on a large-scale dataset of text-image pairs, enhancing its proficiency in conditional image generation. This is followed by a phase of instruction tuning with multimodal prompts, learns to generate images that align with the detailed specifications provided in these prompts. A carefully designed data processing pipeline, incorporating language grounding and image segmentation, is employed to construct these multimodal prompts. This approach enables UNIMO-G to harness rich features from the MLLM encoder to generate images faithfully reproducing the contents across various contexts.

UNIMO-G exhibits a comprehensive capability for controllable image generation, excelling not only in text-to-image synthesis but also in zero-shot subject-driven generation. It adeptly produces high-fidelity images from multimodal prompts, even those containing multiple image entities. To assess its performance, we conducted evaluations in both text-to-image and subject-driven generation contexts using the MS-COCO (Lin et al., 2014) and DreamBench (Ruiz et al., 2023) datasets, respectively. The results consistently highlight UNIMO-G's superior performance in these scenarios. Additionally, recognizing DreamBench's focus on single-subject generation, we introduce MultiBench, a new benchmark featuring images with multiple entities. The evaluation on MultiBench confirms UNIMO-G's effectiveness in zero-shot multi-entity subject-driven generation.

In summary, our contributions in this work can be summarized as follows:

- We propose a simple multi-modal conditional diffusion framework that significantly enhances the controllability of image generation by supporting multimodal prompts with interleaved images and text input.

- We introduce an effective two-stage training strategy, empowering zero-shot multi-entity subject-driven generation through multi-modal instruction tuning.

- UNIMO-G outperforms existing VL-to-image models in both single and multi-entity subject-driven generation tasks, especially on the capabilities of multimodal instruction following.

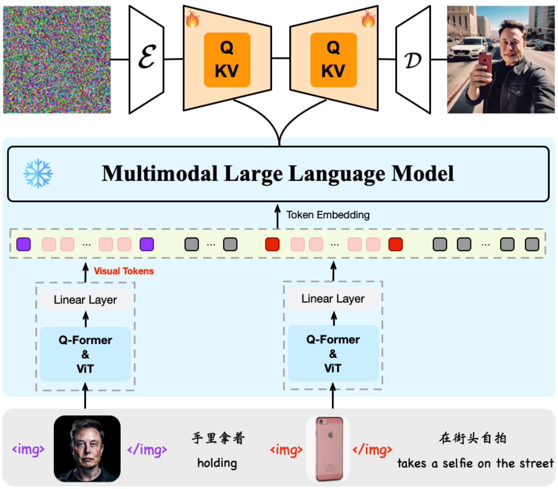


Figure 2: UNIMO-G consists of an MLLM for multimodal perception, and a conditional denoising UNet for image generation. It accepts multimodal prompts with interleaved images and texts, and generates images consistent with the image entities. Orange denotes the trainable modules; Blue denotes the frozen ones.

2 Method

The architecture of UNIMO-G, as depicted in Figure 2, primarily comprises two key components: a Multimodal Large Language Model (MLLM) responsible for encoding multimodal prompts and a conditional denoising diffusion network for image generation based on the encoded representations. In our study, we employed a pre-trained Chinese MLLM, structurally similar to MiniGPT-4 (Zhu et al., 2023). It consists of a vision encoder with a pretrained ViT and Q-Former, a single linear projection layer, and a Transformer-based LLM, underwent pre-training on a vast dataset comprising billions of image-text pairs. Details can refer to Appendix A. This extensive training process equips the model with a robust capability to process and interpret complex multimodal data.

The training of UNIMO-G is conducted in a two-stage process:

- **Text-to-Image Pre-training**: We pre-train the conditional denoising diffusion network from scratch on large-scale Chinese text-image pairs. We employ the same U-Net network architecture in Rombach et al. (2022) and condition it on the text using a cross-attention mechanism.

- **Multi-modal Instruction Tuning**: We further fine-tune UNIMO-G on millions of pairs of multimodal prompts and images, to improve the capability of faithfully generating images from multimodal inputs.

It is worth noting that during both stages of training, only the U-Net component is actively trained, with the MLLM parameters frozen. This strategy ensures that UNIMO-G effectively learns to generate images while retaining the perception knowledge encoded in the pre-trained MLLM.

2.1 Text-to-Image Pre-training

Preliminaries We follow the latent diffusion model (Rombach et al., 2022), utilizing the perceptual compression model (i.e., VQ-VAE) consisting of an image encoder $\mathcal{E}$ and decoder $\mathcal{D}$ to encode the pixel data into the latent space and reverse, such that $\mathcal{D}(\mathcal{E}(x)) \approx x$. The diffusion process is then performed on the latent space, which defines a Markov chain of forward diffusion process q by gradually adding Gaussian noise to the initial latent representations $z_0 = \mathcal{E}(x)$ over T steps. The forward process $q(z_t|z_{t-1})$ at each time step t can be expressed as follows:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t I)$$

where $\{\beta_t\}$ is a series of hyper-parameters. Diffusion models are trained to learn a conditional U-Net (Ronneberger et al., 2015) denoiser ϵ_θ to reverse the diffusion Markov chain, predicting noise with current timestep t, noisy latent z_t and generation condition c. The training loss is the mean squared error (MSE) between the predicted noise $\epsilon_\theta(z_t, t, c)$ and the real noise ϵ in z_t:

$$\mathcal{L} = \mathbb{E}_{z_0, c, \epsilon \sim \mathcal{N}(0,1), t}[\|\epsilon - \epsilon_\theta(z_t, t, c)]\|^2$$

Large-scale Chinese text-image pairs are utilized to train the above denoising objective. The condition information c is fed into each cross attention block of the UNet model as:

$$Attn(z_t, c) = softmax(\frac{Q(z_t) \cdot K(c)^T}{\sqrt{d}}) \cdot V(c)$$

where Q, K and V denote the query, key and value projections, respectively. d denotes the output dimension of the features. In our model, the condition c is encoded by the pre-trained MLLM.

Pre-training Strategies Training a text-to-image diffusion model from scratch presents significant challenges in terms of complexity and resource expenditure. To address these, we introduce an effective training schedule to enhance the efficiency and performance of model training. This schedule encompasses three phases: (1) initial training on a small image corpus to establish foundational visual distribution knowledge; (2) subsequent expansion to large-scale text-image pair training, focusing on text-visual correspondence; (3) culminating in a final phase of training with a small refined corpus, characterized by high visual aesthetics and precise text-image alignment. In our experiments, the training of the UNet model is initially conducted using the CC3M dataset (Sharma et al., 2018). This dataset is chosen for its diverse range of visual concepts coupled with straightforward image descriptions, making it an effective tool for initiating training from scratch. Subsequently, the model undergoes further training with an extensive collection of 300M text-image pairs, aimed at broadening its conceptual understanding and improving its alignment to textual descriptions. The final phase of training involves fine-tuning the model using a meticulously curated corpus, consisting of tens of thousands of high-quality image-text pairs, carefully selected for their superior quality. Based on the above strategies and our architecture designs, we obtain a powerful Chinese text-to-image generation model, surpassing open-source models like Stable Diffusion and its advanced version SDXL (Podell et al., 2023). We provide detailed evaluations of our results in Section 4.2 and implementation details in Appendix B.

2.2 Multimodal Instruction Tuning

Following the text-to-image pre-training, UNIMO-G is indeed capable of generating images from interleaved images and texts, relying on the perception capabilities of MLLM. However, it is important to note that the pre-training stage primarily focuses on generating images that are semantically consistent with the input representations. As a result, UNIMO-G still face challenges in utilizing the visual features of inputs to faithfully reproduce the contents specified in the image conditions. To address this limitation, we further conduct multimodal instruction tuning to enhance UNIMO-G's ability to faithfully reproduce image contents in diverse contexts.

Multimodal Prompts To enhance the representativeness of text prompts, we introduce a format for multimodal prompts that are composed of interleaved images and texts. Specifically, entities mentioned in text captions can be substituted with their corresponding images, like "<img>image of Elon Musk</img> holding his <img>image of iPhone</img>, takes a selfie on the street", as shown in Figure 2. To create pairs of multimodal prompts and images, we have designed a data processing pipeline as illustrated in Figure 3. The pipeline first generates captions and extracts entities from the caption by prompting the MLLM. Subsequently, it acquires the corresponding image segment for each entity using a combination of language grounding by Grounding DINO (Liu et al., 2023b) and image segmentation by SAM (Kirillov et al., 2023). Further introduction on the data construction process is provided in Section B. With a collection of pairs of multimodal prompts and images, UNIMO-G is trained to generate images in accordance with these multimodal instructions.

Visual-Enhanced Learning In order to better harness the visual features of multi-modal input, we introduce an enhancement to the cross-attention mechanism between the generated objects and the input image entities. This improvement aims to foster a more robust and context-aware understanding of the relationships between generated content and the visual elements within the input images. As stated by Prompt-by-Prompt (Hertz et al., 2022), the cross-attention in text-to-image diffusion models can reflect the positions of each generated object specified by the corresponding text token. Similarly, the visual features of image entities can also be treated as visual tokens. The cross-attention map between the intermediate feature of the noisy latent z_t and the visual token v can be calculated:

$$CA(z_t, v) = Softmax(\frac{Q(z_t) \cdot K(v)^T}{\sqrt{d}})$$

where Q and K denote the query and key projections, respectively. For each visual token, we could get an attention map of $h \times w$, where h and w are the spatial dimensions of the latent feature z_t. The scores in cross-attention maps represent the amount of information that flows from a visual token to a latent pixel. Therefore, we introduce an additional loss term that encourages the model to ensure that each visual token mainly attends to the image region occupied by the corresponding objects. Specifically, we optimize z_t towards the target that the desired area of the object has large values by penalizing the $L1$ deviation between the attention maps and the corresponding segmentation maps of the entities:

$$\mathcal{L}_{attn} = \frac{1}{N} \sum_{k=1}^{N} \mid CA(z_t, v_k) - M_k \mid$$

where M_k is the segmentation mask of the k_{th} object corresponding to its visual token v_k. Through this training process, UNIMO-G learns to effectively harness the visual features of input images to faithfully reproduce the corresponding content.

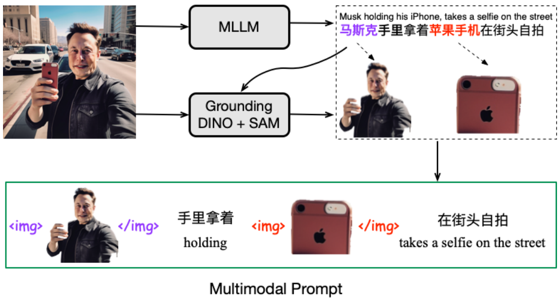

MLLM
Musk holding his iPhone, takes a selfie on the street
马斯克手里拿着苹果手机在街头自拍
Grounding DINO + SAM
<img>
</img>
手里拿着
holding
<img>
</img>
在街头自拍
takes a selfie on the street
Multimodal Prompt

Figure 3: Overview of our data construction pipeline for multi-modal instruction tuning.

3 Related Work

Text-to-Image Diffusion Generation The incorporation of diffusion models into text-to-image synthesis represents a notable advancement in computational creativity (Ho et al., 2020; Song et al., 2020; Li et al., 2022b). Models like GLIDE (Nichol et al., 2021) and DALL-E 2 (Ramesh et al., 2022), which utilize CLIP image embeddings, have substantially progressed in producing images that are both diverse and semantically coherent with textual inputs. Imagen (Saharia et al., 2022) underscores the importance of language comprehension, proposing the integration of a large T5 language model to enhance semantic representation. The Latent Diffusion Model (Rombach et al., 2022) addresses computational constraints by generating images from text-conditioned, low-dimensional latent spaces. Our proposed framework builds upon the principles of the Latent Diffusion Model, leveraging its computational efficiency and scalability.

Subject-Driven Image Generation Following the success of generating high quality images from text descriptions, recent studies have explored subject-driven generation techniques.

Models like DreamBooth (Ruiz et al., 2023), textual-inversion (Gal et al., 2022), and custom-diffusion (Kumari et al., 2023) use optimization-based methods to embed subjects into diffusion models. This is achieved by either fine-tuning the model weights or inverting the subject image into a text token that encodes the subject identity. Some works have explored tuning-free methods. ELITE (Wei et al., 2023) and InstantBooth (Shi et al., 2023) project reference images into word embeddings and inject reference image patch features into cross-attention layers to enhance local details. PhotoMaker (Li et al., 2023b) focuses on the generation of human portraits by extracting a stacked ID embedding from multiple ID images. Despite impressive results for single-object customization, their architecture design restricts their scalability to multiple subject settings. Models like Subject-Diffusion (Ma et al., 2023) and FastComposer (Xiao et al., 2023) are designed for multi-entity subject-driven generation. They enhance text conditioning within diffusion models by incorporating subject embeddings extracted from separate image encoders.Yet, a prevalent limitation of these approaches is their inclination to separate textual and visual guidance, thereby constraining the efficacy of joint modality integration.

Generating with Multi-modal Language Models Multimodal Large Language Models (MLLMs) have significantly broadened the capabilities of language models to process various modalities (Liu et al., 2023a; Li et al., 2021, 2022a; Wang et al., 2023; Driess et al., 2023). These models inherently facilitate interleaved vision-language input, effectively handling multiple images. Models such as GILL (Koh et al., 2023), Emu (Sun et al., 2023), and DreamLLM (Dong et al., 2023) specialize in interleaved vision-language generation by aligning the output space of MLLMs with the diffusion image decoder. However, these methods primarily align at a semantic level and may struggle with detailed, subject-driven image generation. BLIP-Diffusion (Li et al., 2023a) synthesizes images by composing subjects with random backgrounds, endowing it with zero-shot, subject-driven text-to-image generation capabilities. However, its specific input template and training process limit scalability for multiple entities. KOSMOS-G (Pan et al., 2023), a model closely related to our work, leverages a MLLM to encode interleaved text-visual inputs, and the U-Net of Stable Diffusion (SD) v1.5 as the image decoder. The key component of KOSMOS-G is an AlignerNet, which is trained solely on textual data to align the output embedding space of the frozen SDv1.5 U-Net with the MLLM. In contrast, our approach centers on training the U-Net model end-to-end specifically for multimodal diffusion, significantly enhancing both the faithfulness and relevance of generated images in multimodal contexts. Differing from alignment-based approaches, our two-stage training strategy markedly improves the model's proficiency in following multimodal instructions, particularly in complex multi-entity scenarios.

4 Experiments

In this section, we first introduce the implementation details and settings of experiments. Then we present the evaluation results in both text-driven and subject-driven scenarios. Last, we further analyze the results with quantitative ablation studies.

4.1 Implementation Details

UNIMO-G is composed of a 7.8B-parameter MLLM encoder, following the MiniGPT-4 architecture (Zhu et al., 2023), and a 4B-parameter denoising U-Net, totaling approximately 11.8B parameters. The MLLM is pretrained on a large-scale Chinese multimodal corpus, comprising text, image-caption pairs, and interleaved image-text data. The U-Net architecture includes 5 downsampling and 5 upsampling blocks with channel sizes [640, 1280, 1280, 2560, 2560], and a cross-attention mechanism with 4096 dimensions and 16 heads. The image auto-encoder, based on the LDM framework, has been optimized for our specific image corpus. The detail training process and data construction are further introduced in Appendix B.

4.2 Evaluation Results

UNIMO-G demonstrates a unified image generation capability for both text-driven and subject-driven image generation, as shown in Figure 1. In the following, we will evaluate the performance of UNIMO-G from different aspects.

Text-to-Image Generation For text-to-image generation, we used 30,000 captions randomly sampled from the MS-COCO (Lin et al., 2014) validation set, translating them into Chinese to align with UNIMO-G's input requirements. Images were generated at 512x512 pixels and resized to 256x256 for evaluation using the FID-30k metric, a standard in

Methods	FID
T2I Models	
GLIDE (Nichol et al., 2021)	12.24
DALL-E 2 (Ramesh et al., 2022)	10.39
SDv1.5 (Rombach et al., 2022)	9.34
Imagen (Saharia et al., 2022)	7.27
SDXL (Podell et al., 2023)	11.93
VL2I Models	
GILL (Koh et al., 2023)	12.20
Emu (Sun et al., 2023)	11.66
KOSMOS-G (Pan et al., 2023)	10.99
UNIMO-G	8.36

Table 1: Zero-shot FID-30K comparisons on MS-COCO 256x256.

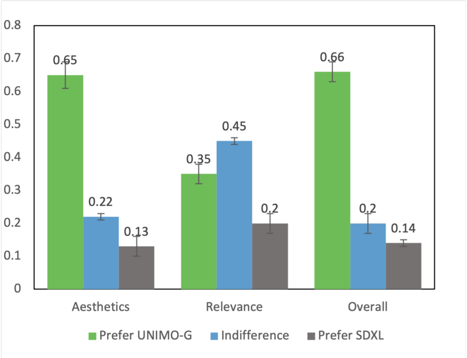


Figure 4: Comparison of UNIMO-G and SDXL by human evaluation. The mean and standard deviation are shown in the figure.

the field. Our model employs a classifier-free guidance scale of 5.0 and 50 DDIM inference steps for diffusion sampling. As shown in Table 1, UNIMO-G greatly surpasses other Vision-Language to Image (VL2I) models in performance.

To further validate our model, we conducted a human evaluation comparing UNIMO-G with SDXL (Podell et al., 2023), a leading open-source model. We established a comprehensive bilingual benchmark, encompassing 312 prompts (162 from DrawBench and 150 user queries randomly sampled from the online platform). The DrawBench prompts were filtered to exclude language-specific ones. All prompts are manually translated and carefully proofread to achieve the final parallel Chinese and English set. Three independent evaluators rated the images from UNIMO-G and SDXL by selecting the model they prefer, focusing on aspects of image aesthetics, image-text relevance, and overall quality, respectively. The results in Figure 4 demonstrate UNIMO-G's substantial superiority in all aspects. Some examples are shown in Figure 11.

Single-Entity Subject-Driven Generation For single-entity subject driven generation, we evaluate UNIMO-G on DreamBench (Ruiz et al., 2023). DreamBench comprises 30 subjects with 25 prompt templates, yielding 750 unique prompts that test skills such as re-contextualization, modification, and accessorization. We follow prior work to generate four images for each prompt, creating a total of 3,000 images for a comprehensive assessment. We employed DINO and CLIP-I metrics for subject fidelity evaluation and CLIP-T for image-text relevance assessment. A classifier-free guidance scale of 5.0 and 50 DDIM inference steps were used for sampling. UNIMO-G, accepting a single image input, utilized the same images as KOSMOS-G (Pan et al., 2023) for a consistent comparison. As indicated in Table 2, UNIMO-G in a zero-shot setting surpasses other models like Textual Inversion (Gal et al., 2022), DreamBooth (Ruiz et al., 2023), BLIP-Diffusion (Li et al., 2023a), and Re-Imagen (Chen et al., 2022), and marginally outperforms KOSMOS-G. Notably, our model demonstrates a significant improvement in balancing image-text relevance and image fidelity compared to the closely related KOSMOS-G. We observed that existing methods tend to prioritize image information over textual input. This tendency occasionally leads to a diminished focus on semantic content, favoring subject reconstruction. Thanks to our two-stage end-to-end learning framework, UNIMO-G maintained high image fidelity and achieved the highest CLIP-T score for image-text relevance, indicating a strong capability in following multi-modal instructions.

Multi-Entity Subject-Driven Generation UNIMO-G exhibits exceptional performance in zero-shot multi-entity subject-driven generation. To evaluate this capability, we established *MultiBench*, a novel benchmark specifically designed for multi-entity subject-driven generation assessment. MultiBench includes four object categories: living objects (humans and animals), food, wearable items, and toys, each containing 10 different objects. We developed five prompt templates for composing scenarios with 2 and 3 objects, resulting in a total of 2,300 distinct prompts. Details are introduced in the Appendix D. For each prompt, four images were generated, culminating in 9,200 images for an exhaustive evaluation. We

https://yige.baidu.com/

Methods	**DINO**	**CLIP-I**	**CLIP-T**	**Avg**
Fine-Tuning Methods				
Textual Inversion	0.569	0.780	0.255	0.535
DreamBooth	0.668	0.803	0.305	0.592
BLIP-Diffusion	0.670	0.805	0.302	0.592
Zero-Shot Methods				
Re-Imagen	0.600	0.740	0.270	0.537
BLIP-Diffusion	0.594	0.779	0.300	0.558
KOSMOS-G	**0.694**	**0.847**	0.287	0.609
UNIMO-G	0.668	0.841	**0.329**	**0.613**
w/o Tuning	0.371	0.717	0.306	0.465
w/o VisualEnh	0.617	0.815	0.329	0.587

Table 2: Comparisons of single-entity subject-driven image generation on DreamBench. *Avg* denotes the average scores of DINO, CLIP-I and CLIP-T.

Methods	**DINO**	**CLIP-I**	**CLIP-T**	**Avg**
BLIP-Diffusion	0.410	0.648	0.249	0.436
KOSMOS-G	0.419	**0.671**	0.283	0.458
UNIMO-G	**0.436**	0.665	**0.298**	**0.466**
w/o Tuning	0.235	0.583	0.240	0.353
w/o VisualEnh	0.399	0.631	0.276	0.435

Table 3: Comparisons of multi-entity subject-driven image generation on MultiBench.

conducted image similarity analysis using DINO and CLIP-I metrics, alongside text relevance assessments using CLIP-T. Image similarity was determined by averaging the similarities between the generated image and each of the two or three subjects. The results, as shown in Table 3, indicate that UNIMO-G outperforms BLIP-Diffusion and KOSMOS-G in terms of both image similarity and textual relevance. Some comparison examples are shown in Figure 6. This demonstrates UNIMO-G's superior capability to accurately capture subject information from input images and effectively follow multi-modal instructions. More examples are shown in Figures 9 and 10.

To further validate our model, we conducted a human evaluation comparing UNIMO-G with KSOMOS-G by sampling 200 prompts from MultiBench. Three raters are presented with two sets of images generated by UNIMO-G and the compared model. They are asked to compare these images from three dimensions of semantic relevance, visual faithfulness and image fidelity, and then select the model they prefer, or indifference. Throughout the process, raters are unaware of which model the image is generated from. The results in Figure 5 show that human raters greatly prefer UNIMO-G over KOSMOS-G on all aspects, which further validate the effectiveness of our approach in generating high-quality, personalized images from free-form multimodal prompts.

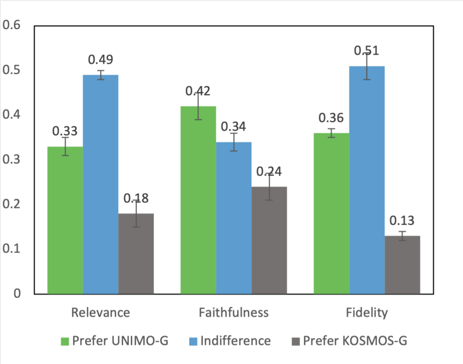


Figure 5: Comparison of UNIMO-G and KOSMOS-G on MultiBench by human evaluation. The mean and standard deviation are shown in the figure.

4.3 Analysis

Effectiveness of Multi-modal Instruction Tuning UNIMO-G, prior to multi-modal instruction tuning (denoted as "w/o Tuning"), also demonstrates the capability to generate images from multi-modal prompts based on the MLLM. Nonetheless, it initially falls short in accurately reproducing input images. To evaluate the effectiveness of multi-modal instruction tuning, we compared the performance of UNIMO-G with and without this tuning in single-entity and multi-entity subject-driven generation tasks. The comparison results in Table 2 and Table 3 reveal that multi-modal instruction tuning substantially enhances image similarity metrics (DINO and CLIP-I) in both single and multi-entity scenarios. This improvement indicates that after tuning, UNIMO-G more effectively leverages visual features from the inputs, thereby accurately replicating the content defined in the image conditions. Furthermore, UNIMO-G with instruction tuning also shows obvious advancements in textual relevance, as evaluated by the CLIP-T metric. This indicates that the tuning process not only bolsters visual faithfulness but also amplifies the model's ability to follow multimodal instructions.

Effectiveness of Visual Enhancement Learning The incorporation of a visual-enhanced learning strategy during multimodal instructional tuning significantly improves the visual alignment between input and output images. To quantify this effect, we conducted an ablation study by omitting the visual enhancement component during multimodal tun-

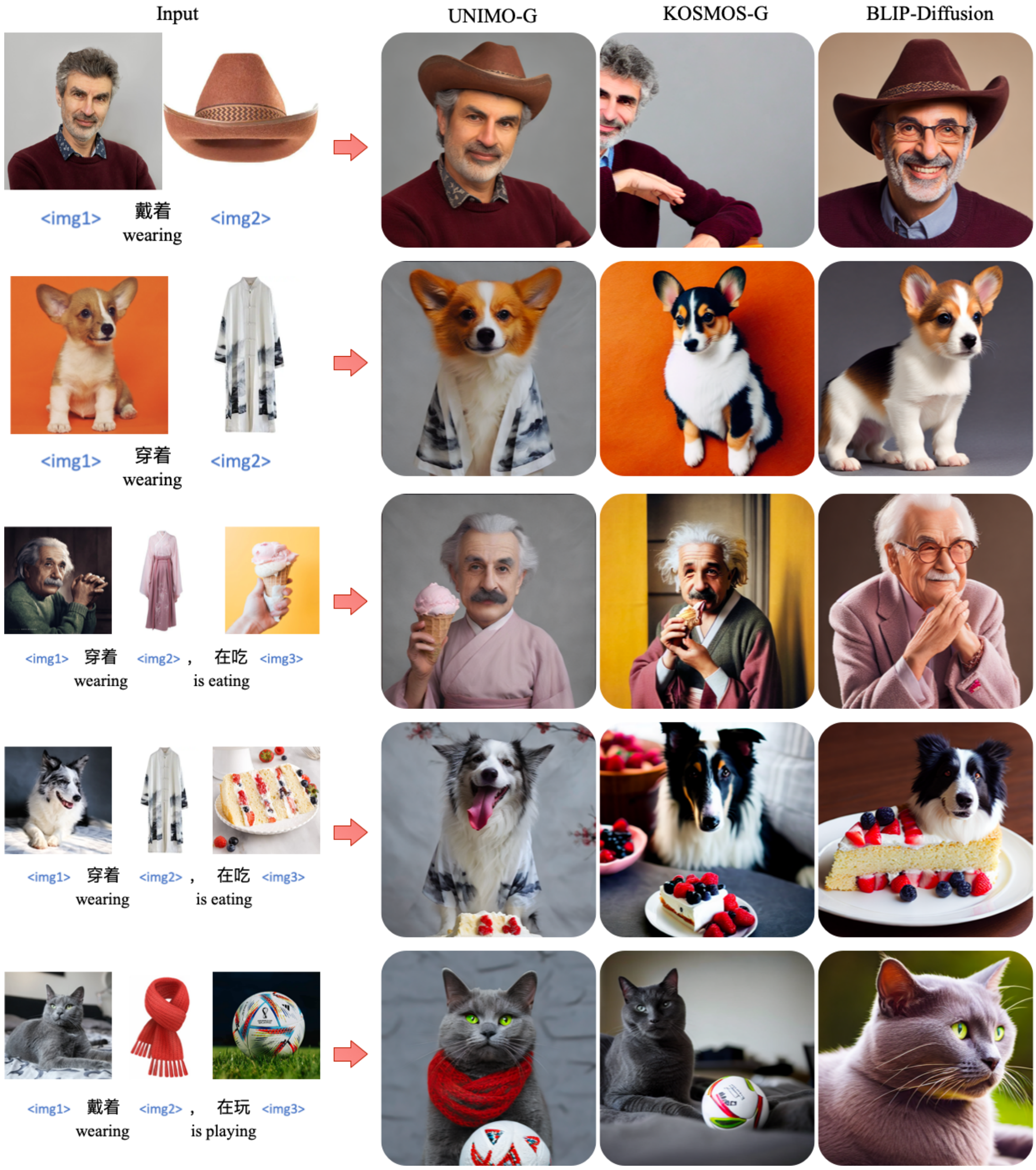


Figure 6: Comparison with baselines for multi-entity subject-driven image generation.

ing (denoted as "w/o VisualEnh") and assessed its impact on both single-entity and multi-entity generation tasks. The results, as detailed in Tables 2 and 3, demonstrate that the visual-enhanced learning strategy markedly boosts the performance in image similarity metrics (DINO and CLIP-I), across both single and multi-entity scenarios. Notably, it also improves image-text alignment in multi-entity scenarios by reducing entity blending or missing.

5 Conclusion

This paper presents UNIMO-G, a simple multimodal conditional diffusion framework designed to process multimodal prompts that interleave text and visual inputs. It demonstrates exceptional proficiency in text-to-image generation and zero-shot subject-driven synthesis, and is particularly adept at producing high-fidelity images from intricate multi-modal prompts with multiple image entities. In comparison to standard text-conditional diffusion models, UNIMO-G significantly enhances visual controllability in image generation. Thanks to our two-stage training strategy, UNIMO-G also outperforms existing VL-to-image models, especially on the ability to follow complex multimodal instructions. Overall, UNIMO-G showcases the potential for more nuanced and controlled image generation processes.

6 Limitations

Our model suffers from some common failures of text-driven and subject-driven generation models, such as inaccuracies in context synthesis, difficulties in complex composition, and a shortfall in visual faithfulness, particularly in multi-entity image generation tasks. Additionally, there exists an inherent risk associated with the misuse of such technology, notably in the creation of deepfakes, which raises ethical concerns. Despite the limitations and risks, the proposed framework still demonstrates considerable promise in facilitating more nuanced and controlled processes in image generation.

References

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568.

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. 2023. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. 2022. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*.

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*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

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*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

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

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941.

Dongxu Li, Junnan Li, and Steven CH Hoi. 2023a. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2592–2607, Online. Association for Computational Linguistics.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2022a. UNIMO-2: End-to-end unified vision-language grounded learning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3187–3201, Dublin, Ireland. Association for Computational Linguistics.

Wei Li, Xue Xu, Xinyan Xiao, Jiachen Liu, Hu Yang, Guohao Li, Zhanpeng Wang, Zhifan Feng, Qiaoqiao She, Yajuan Lyu, et al. 2022b. Upainting: Unified text-to-image diffusion generation with cross-modal guidance. *arXiv preprint arXiv:2210.16031*.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. 2023b. Photomaker: Customizing realistic human photos via stacked id embedding. *arXiv preprint arXiv:2312.04461*.

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.

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

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. 2023. Subject-diffusion: Open domain personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2307.11410*.

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*.

Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. 2023. Kosmos-g: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*.

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*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510.

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.

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.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. 2023. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*.

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*.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*.

Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. 2023. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*.

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 Multi-modal Large Language Model

The structure of our MLLM is similar to MiniGPT-4 (Zhu et al., 2023) and BLIP-2 (Li et al., 2023a), which consists of a vision encoder with a pretrained ViT-G and BLIP-2 Q-Former, a single linear projection layer, and a pre-trained 6B Transformer-based LLM. The MLLM totally contains about 7.8B parameters. The encoding process of multimodal prompts is illustrated as in Figure 7. Each image is firstly encoded as 32 visual tokens by the visual encoder, and then linearly transformed into the LLM embedding space. The pre-trained LLM is utilized as the context encoder and decoder. Our MLLM underwent training on web-scale multimodal corpora, comprising text corpora, image-caption pairs, and interleaved data of images and texts.

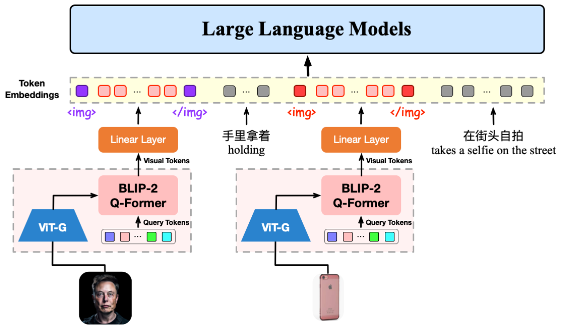


Figure 7: Illustration of the encoding of multimodal prompts by the MLLM.

B Implementation Details

The training process include text-to-image pretraining and multi-modal instruction tuning.

Text-to-Image Pre-training Our model's pretraining involves three stages, each utilizing distinct datasets:

1. Initial Training with CC3M Dataset: The CC3M dataset, consists of about 3.3M image-description pairs, was translated into Chinese using the Baidu Translation API. The model is trained from scratch at 256x256 using the AdamW optimizer with a weight decay of 0.01, a learning rate of 5e-5, and a batch size of 40x256 for 100K steps, which costs about 300 A100 GPU days.

2. Expansion with Large-Scale Chinese Data: We incorporate about 300M Chinese text-image pairs from multiple datasets, including LAION-2B (Schuhmann et al., 2022), COYO-700M (Byeon et al., 2022), Conceptual Captions (Changpinyo et al., 2021) and a series of internal Chinese datasets. The English captions are translated into Chinese. This stage, with a constant learning rate of 5e-5, initially training at 256x256 resolution for 500K steps with a batch size of 40x256, then progresses to 512x512 for 200K steps with a batch size of 12x256, which costs about 3000 GPU days.

3. Refinement with High-Quality Corpus: The final stage focuses on fine-tuning on high-quality corpus, and continues training at 512x512 and 1024x1024 resolutions. Selection criteria include an aesthetic threshold above 6.5 (LAION-Aesthetics V2) and an image-text similarity score over 0.5 (CLIP-base-32), resulting in about 1M pairs. A multi-scale training strategy dividing the image size into 5 buckets [0.5, 0.75, 1.0, 1.5, 2.0] supports various image aspect ratios. The final stage uses a learning rate of 1e-5 for 200K steps with a batch size of 3072, which costs about 1000 GPU days.

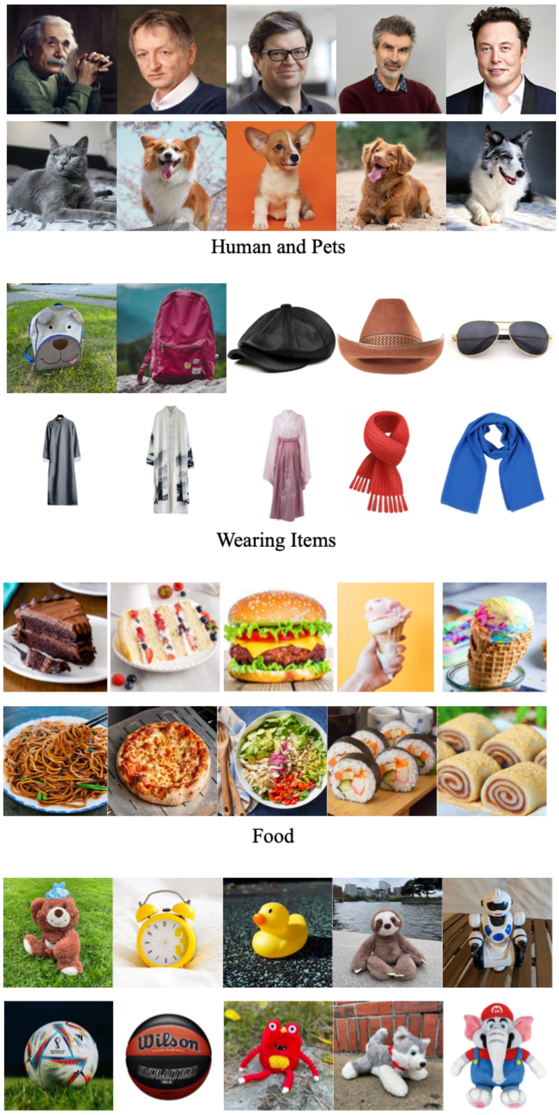
Multi-modal Instruction Tuning We developed a multimodal-to-image instruction tuning dataset utilizing the 1M high-quality image-text pairs. The process, depicted in Figure 3, involves: (1) Generation of captions and extraction of entity tokens from captions by the MLLM; (2) Identifying entity detection boxes via Grounding DINO (Liu et al., 2023b); (3) Segmentation and extraction of regions corresponding to each entity by SAM (Kirillov et al., 2023); (4) Randomly substitution of entity tokens with their corresponding image segments. This process effectively transforms the original text-image pairs into multimodal-image pairs, which are then employed for refining multimodal instruction tuning. We maintain a learning rate of 1e-5, training for 200K steps with a batch size of 3072. To preserve text-to-image generation capabilities, 10% of the training uses original textual captions. This process totally costs about 800 A100 GPU days.

Number of Entities	Ratios
0	16%
1	36.8%
2	25.8%
3	12.3%
>=4	9.1%

Table 4: Statistics of the number of subject entities in our constructed multi-modal prompts.

Num of Entities	Methods	DINO	CLIP-I	CLIP-T	Avg
2	BLIP-Diffusion	0.455	0.675	0.249	0.460
	KOSMOS-G	0.465	**0.704**	0.279	0.483
	UNIMO-G	**0.485**	0.699	**0.293**	**0.492**
3	BLIP-Diffusion	0.344	0.608	0.249	0.400
	KOSMOS-G	0.350	**0.621**	0.287	0.419
	UNIMO-G	**0.364**	0.614	**0.306**	**0.428**

Table 5: Comparisons of the performance with different number of subject entities in a multi-modal prompt on MultiBench.


Human and Pets

Wearing Items

Food

Toys

Figure 8: Illustration of images in MultiBench.

a {living object} wearing {wearing}
a {living object} is playing with {toy}
a {living object} is eating {food}
a {living object} wearing {wearing}, is playing with {toy}
a {living object} wearing {wearing}, is eating {food}

Table 6: Templates for multi-entity subject-driven generation in MultiBench.

C Analysis of Multi-modal Prompts

In our analysis of multi-modal prompts, we meticulously examined the composition of our dataset, finding an average of 42.48 text tokens and 1.67 subject entities per prompt. The distribution of entity counts revealed a predominance of images featuring fewer than three subject entities, with a minor proportion exceeding this count. The ratios of different number of entities are shown as in Table 4.

In our experiments, we have assessed the model's performance using DreamBench for single-entity prompts and MultiBench for multi-entity scenarios, specifically focusing on two and three entities. We separately evaluate the performance on two and three entities as shown in Table 5. The results, as detailed, showcase that our model, UNIMO-G, consistently surpasses baseline models in handling both two and three-entity multi-modal prompts. This clear performance advantage underscores the effectiveness of our approach across varying levels of complexity.

D Images and Prompts of MultiBench

MultiBench contains 4 categories of objects, each including 10 objects. The images in MultiBench are shown in Figure 8. Five temples are designed to compose multi-modal prompts, as shown in Table 6. Some examples generated by UNIMO-G on MultiBench are shown in Figure 9 and Figure 10.

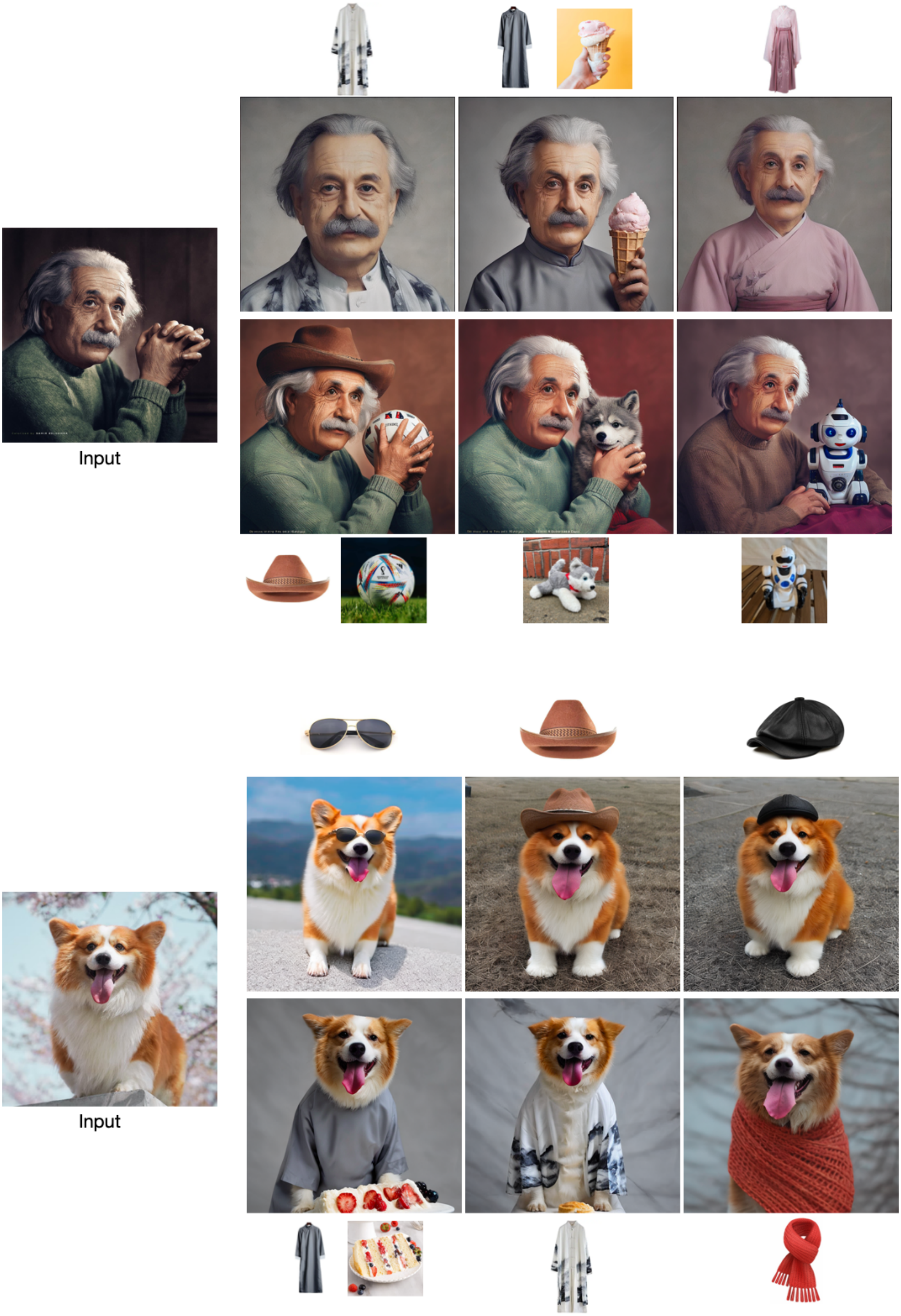


Figure 9: Examples of multi-entity subject-driven image generation on MultiBench by UNIMO-G.

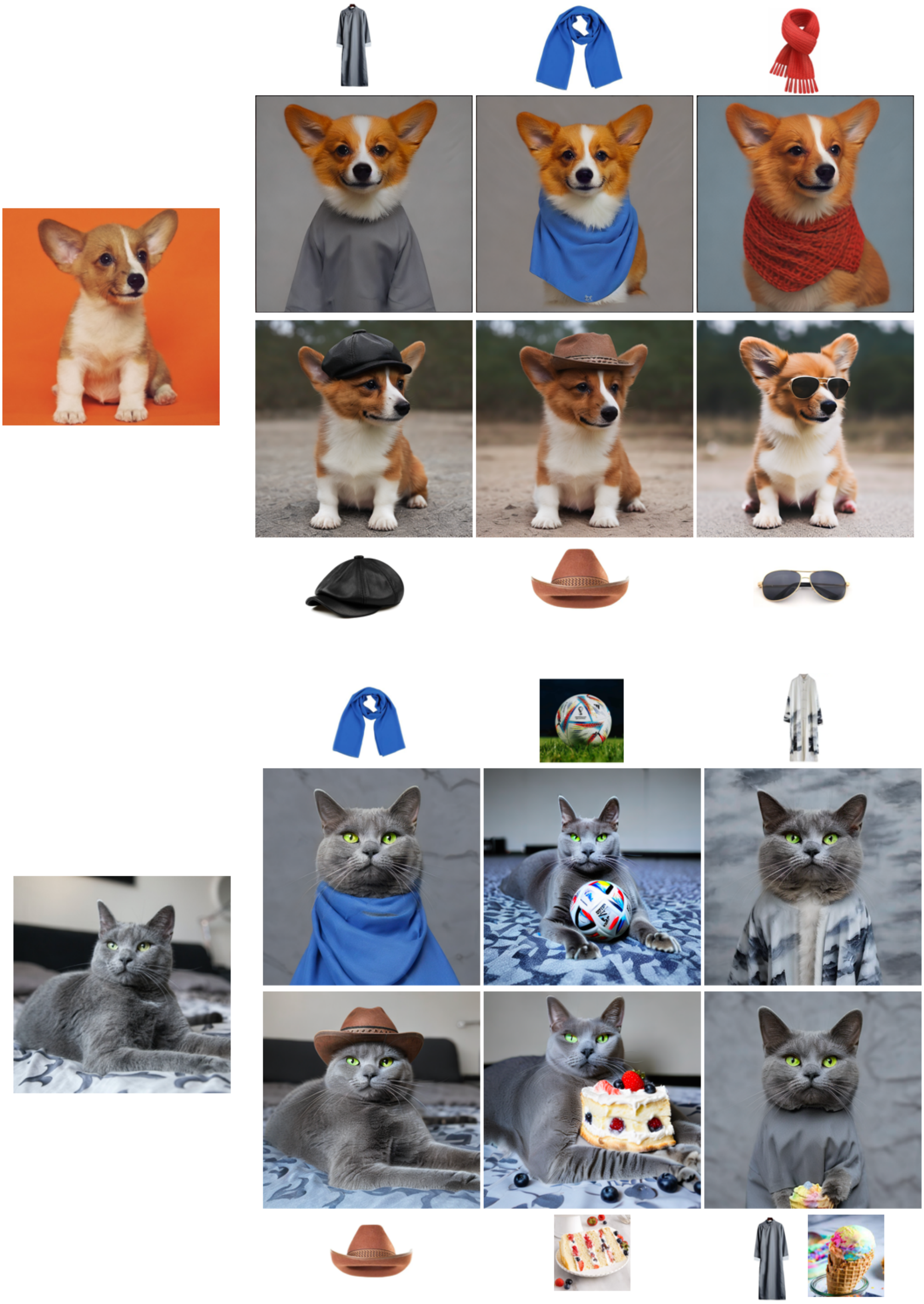
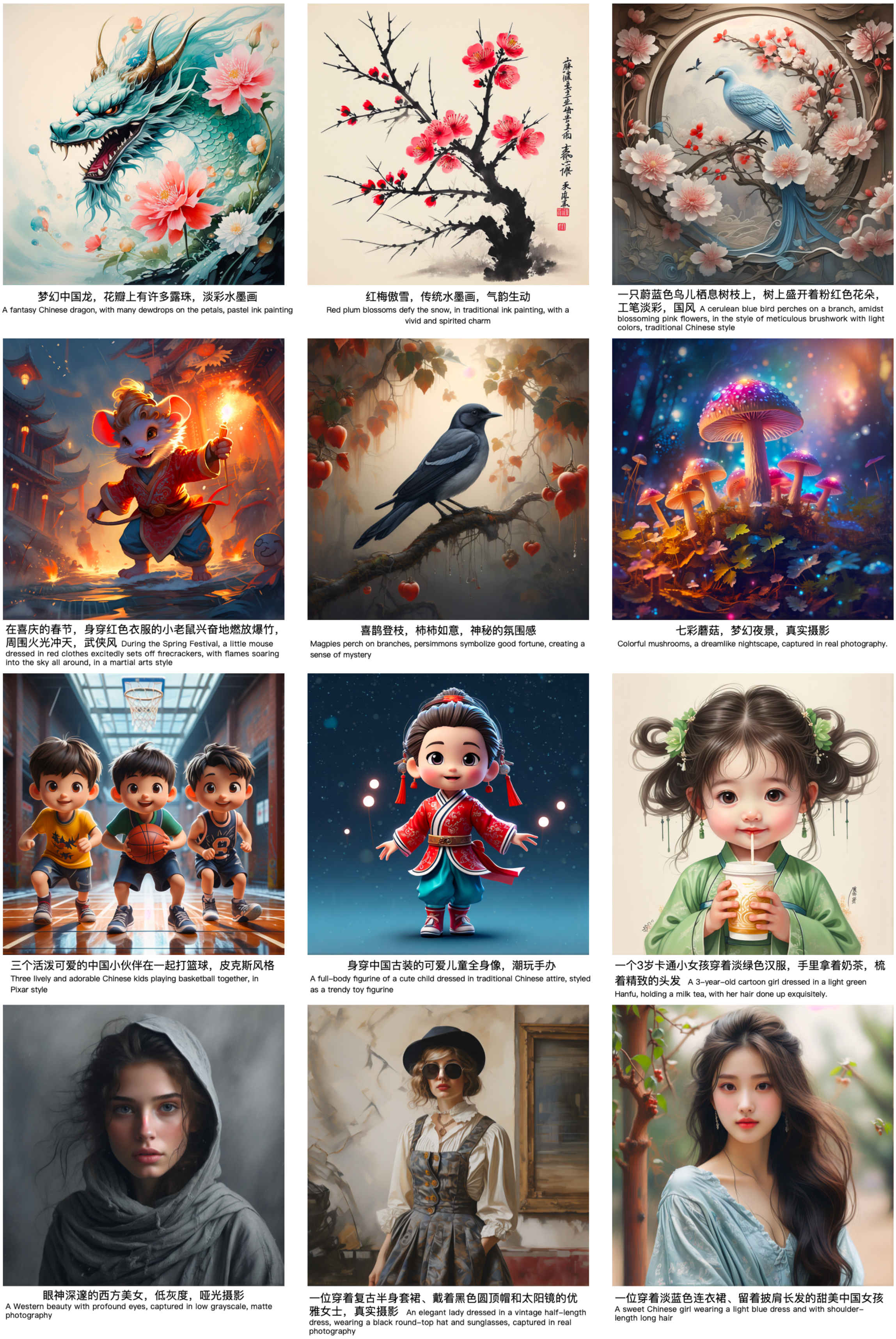


Figure 10: Examples of multi-entity subject-driven image generation on MultiBench by UNIMO-G.



梦幻中国龙，花瓣上有许多露珠，淡彩水墨画
A fantasy Chinese dragon, with many dewdrops on the petals, pastel ink painting

红梅傲雪，传统水墨画，气韵生动
Red plum blossoms defy the snow, in traditional ink painting, with a vivid and spirited charm

一只蔚蓝色鸟儿栖息树枝上，树上盛开着粉红色花朵，工笔淡彩，国风 A cerulean blue bird perches on a branch, amidst blossoming pink flowers, in the style of meticulous brushwork with light colors, traditional Chinese style

在喜庆的春节，身穿红色衣服的小老鼠兴奋地燃放爆竹，周围火光冲天，武侠风 During the Spring Festival, a little mouse dressed in red clothes excitedly sets off firecrackers, with flames soaring into the sky all around, in a martial arts style

喜鹊登枝，柿柿如意，神秘的氛围感
Magpies perch on branches, persimmons symbolize good fortune, creating a sense of mystery

七彩蘑菇，梦幻夜景，真实摄影
Colorful mushrooms, a dreamlike nightscape, captured in real photography.

三个活泼可爱的中国小伙伴在一起打篮球，皮克斯风格
Three lively and adorable Chinese kids playing basketball together, in Pixar style

身穿中国古装的可爱儿童全身像，潮玩手办
A full-body figurine of a cute child dressed in traditional Chinese attire, styled as a trendy toy figurine

一个3岁卡通小女孩穿着淡绿色汉服，手里拿着奶茶，梳着精致的头发 A 3-year-old cartoon girl dressed in a light green Hanfu, holding a milk tea, with her hair done up exquisitely.

眼神深邃的西方美女，低灰度，哑光摄影
A Western beauty with profound eyes, captured in low grayscale, matte photography

一位穿着复古半身套裙、戴着黑色圆顶帽和太阳镜的优雅女士，真实摄影 An elegant lady dressed in a vintage half-length dress, wearing a black round-top hat and sunglasses, captured in real photography

一位穿着淡蓝色连衣裙、留着披肩长发的甜美中国女孩
A sweet Chinese girl wearing a light blue dress and with shoulder-length long hair

Figure 11: Examples of text-to-image generation by UNIMO-G.