

EasyRef: Omni-Generalized Group Image Reference for Diffusion Models via Multimodal LLM

Zhuofan Zong^{1,2} Dongzhi Jiang¹ Bingqi Ma² Guanglu Song²
 Hao Shao¹ Dazhong Shen³ Yu Liu² Hongsheng Li^{1,3}

¹CUHK MMLab ²SenseTime Research ³Shanghai AI Laboratory

Project page: <https://easyref-gen.github.io/>

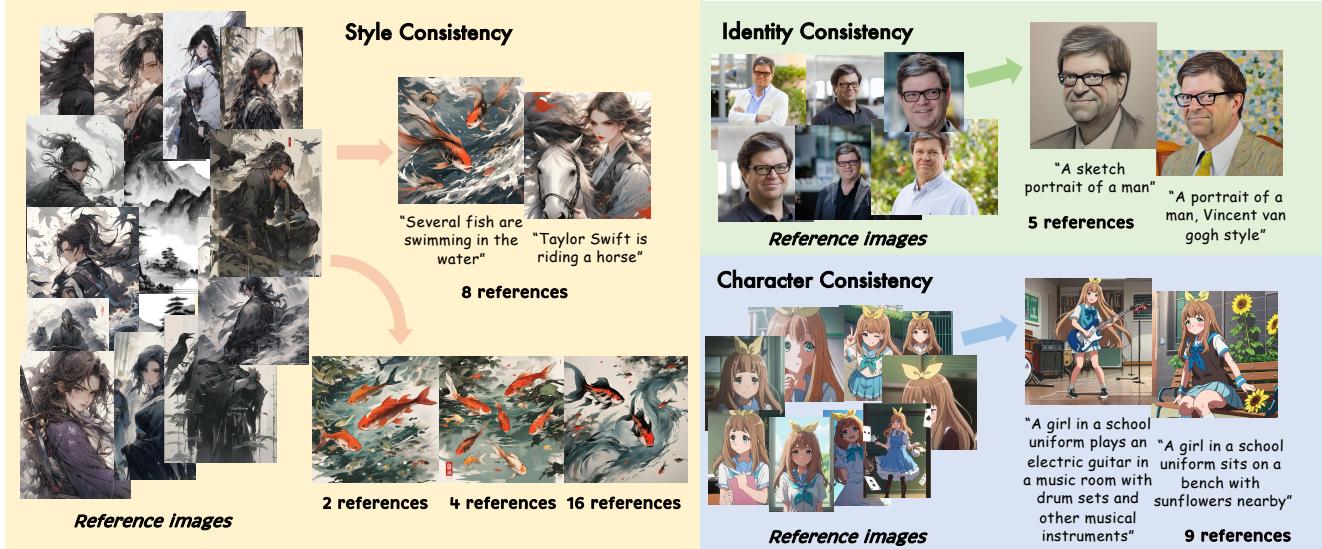


Figure 1. EasyRef is capable of modeling the consistent visual elements of various input reference images with a single generalist multimodal LLM in a zero-shot setting.

Abstract

Significant achievements in personalization of diffusion models have been witnessed. Conventional tuning-free methods mostly encode multiple reference images by averaging their image embeddings as the injection condition, but such an image-independent operation cannot perform interaction among images to capture consistent visual elements within multiple references. Although the tuning-based Low-Rank Adaptation (LoRA) can effectively extract consistent elements within multiple images through the training process, it necessitates specific finetuning for each distinct image group. This paper introduces EasyRef, a novel plug-and-play adaptation method that enables diffusion models to be conditioned on multiple reference im-

ages and the text prompt. To effectively exploit consistent visual elements within multiple images, we leverage the multi-image comprehension and instruction-following capabilities of the multimodal large language model (MLLM), prompting it to capture consistent visual elements based on the instruction. Besides, injecting the MLLM's representations into the diffusion process through adapters can easily generalize to unseen domains, mining the consistent visual elements within unseen data. To mitigate computational costs and enhance fine-grained detail preservation, we introduce an efficient reference aggregation strategy and a progressive training scheme. Finally, we introduce MRBench, a new multi-reference image generation benchmark. Experimental results demonstrate EasyRef surpasses both tuning-free methods like IP-Adapter and tuning-based

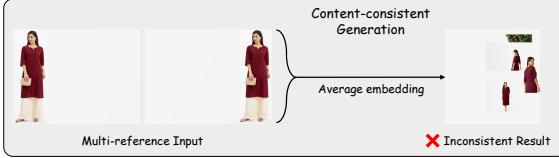


Figure 2. Spatial misalignment issue of the embedding averaging operation. The images with faces are synthetic.

methods like LoRA, achieving superior aesthetic quality and robust zero-shot generalization across diverse domains.

1. Introduction

Significant achievements in diffusion models [3, 8, 15, 22, 25, 28, 31–33, 35, 38, 47] have been witnessed because of their remarkable abilities to create visually stunning images. To improve the precision and controllability of diffusion models, researchers have been exploring personalized generation conditioned on a small number of reference images, *i.e.*, the generated images are required to maintain elements of the reference image while incorporating modifications specified by the text prompt. Such personalized image generation approaches are mainly categorized into tuning-free methods [18, 41, 42, 44, 48, 49] and tuning-based methods [9, 14, 34].

Previous tuning-free approaches typically leverage a pre-trained feature extractor to capture certain attributes of the reference image, which are then injected into the frozen diffusion model via trainable adapters. The seminal IP-Adapter [48] proposed this design by integrating CLIP [30] embeddings of the reference image with decoupled cross-attention layers. Follow-up works [29, 44, 45] have developed sophisticated, task-specific feature encoders to encode distinct elements of the reference image (*e.g.*, style, content, character, identity, *etc.*) for personalized image generation. Despite their promise, these methods have several limitations. Firstly, encoders for different references, such as style and character, have specific complex designs and can be optimized through various specialized training tasks. Secondly, most methods [29, 41, 48] are limited to training with a single reference image and fail to fully encode consistent visual representations from multiple references. Although Low-Rank Adaptation (LoRA) [14] can extract consistent elements within multiple images through the training process, it necessitates specific finetuning for each distinct image group.

This paper introduces EasyRef, a plug-and-play adaptation method that empowers diffusion models to condition multiple reference images and text prompts. Conventional methods mostly encode multiple reference images by averaging their image embeddings as the injection condition, but such an image-independent operation cannot perform

Method	Consistency mining	Zero-shot generalization	Multi-reference input
LoRA [14]	✓	✗	✗
IP-Adapter [48]	✗	✓	✓
EasyRef	✓	✓	✓

Table 1. Comparison among LoRA, IP-Adapter, and EasyRef.

interaction among images to capture consistent visual elements within multiple references. For example, as illustrated in Figure 2, the CLIP-based IP-Adapter generates an inconsistent image when the spatial locations of the target subject vary across the reference images. To effectively exploit consistent visual elements within multiple images, we leverage the multi-image comprehension and instruction-following capabilities of the multimodal large language model (MLLM) [4–6, 16, 17, 23, 24, 37, 51], prompting it to capture consistent visual elements based on the instruction. Besides, injecting the MLLM’s representations into the diffusion process through adapters can easily generalize to unseen domains, mining the consistent visual elements within unseen data. EasyRef also inherits the MLLM’s ability to process arbitrary number of reference images with arbitrary aspect ratios. We present the key differences among our method, LoRA, and IP-Adapter in Table 1. To mitigate the computational demands imposed by the long context of multi-image inputs, we propose querying the MLLM with learned token embeddings and aggregating reference representations within the deepest layer of the MLLM architecture. Additionally, to address the limitations of MLLMs in capturing fine-grained visual details, we employ a progressive training strategy to enhance the MLLM’s capacity for fine-grained detail and identity preservation. Unlike previous methods that rely on sophisticated feature encoders [27, 29, 30] and even additional face encoders [7], we find the single MLLM in EasyRef is capable of extracting various reference representations, such as style and character, along with text features from an arbitrary set of reference images and the text prompt, exhibiting strong generalization ability. Finally, we introduce a multi-reference generation benchmark (MRBench) for multi-reference image generation to evaluate our work and guide future research. Compared to IP-Adapter and the prevalent fine-tuning approaches using LoRA, our proposed generalist EasyRef model consistently achieves superior aesthetic quality across diverse domains and demonstrates robust zero-shot generalization.

In summary, our contributions are threefold: (1) We introduce EasyRef, the first plug-and-play adaptation technique enabling diffusion models to be jointly conditioned on multiple reference images and text prompts. (2) We propose an efficient reference aggregation strategy and a progressive training scheme to mitigate computational costs and enhance the MLLM’s fine-grained perceptual abilities. (3) We introduce a novel MRBench for evaluating diffusion

models in multi-reference image generation scenarios.

2. Related Work

2.1. Image Personalization

Image personalization approaches can be categorized into tuning-free methods [10, 15, 18, 19, 29, 41, 42, 44, 48, 49] and tuning-based methods [9, 14, 34]. Tuning-free approaches typically extract visual representations, such as style and character, from the reference image and inject these into the diffusion model. IP-Adapter [48] enhances image prompting capabilities through a decoupled cross-attention mechanism. Building upon IP-Adapter, InstantStyle [41, 42] injects CLIP [30] style embeddings into style-specific blocks. Both IP-Adapter-Face [48] and InstantID [44] employ additional face encoders [7] to improve identity preservation. A limitation of tuning-free methods is that they are trained with single-reference input, failing to fully exploit the consistent elements within multiple reference images. Tuning-based approaches, such as LoRA [14], finetuned the diffusion model using a limited set of images. Although tuning-based methods are capable of multi-image references, a key limitation is they necessitate specific finetuning for each distinct image group. In this work, we extend tuning-free methods to accommodate multiple reference images and the text prompt like tuning-based methods while maintaining robust generalization capabilities.

2.2. Multimodal Large Language Models

Multimodal large language models (MLLMs) [4–6, 16, 17, 23, 24, 37, 51] have demonstrated remarkable success in addressing open-world tasks. Pioneering works like LLaVA [24] and BLIP-2 [17] consistently projected the vision representation from a pretrained CLIP vision encoder into the LLM for multimodal comprehension. Qwen-VL [2] collected massive multimodal tuning data and adopted elaborate training strategy for better optimization. The mixture-of-vision-experts designs, such as SPHINX [21], MoF [40], and MoVA [51], were explored to enhance the visual capabilities of MLLMs. Furthermore, models like LLaVA-NeXT [23] and Qwen2-VL [43] sought to enable the processing of images with arbitrary resolutions. LI-DiT [25] investigated how to effectively unleash the MLLM’s prompt encoding capabilities for diffusion models. In this paper, we are the first to leverage the multi-image comprehension and instruction-following capabilities of the MLLM to jointly encode representations of multiple reference images and the text prompt.

3. EasyRef

3.1. Preliminary

Denoising Diffusion Probabilistic Models [13] (DDPMs) are trained by maximizing the log-likelihood of the training

data, given a data distribution $q(\mathbf{x}_0)$. The training process involves a forward diffusion process that gradually adds Gaussian noise to the data over T timesteps:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1}), \quad (1)$$

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, (1 - \alpha_t)\mathbf{I}). \quad (2)$$

Here, \mathbf{x}_t represents the noisy data at timestep t and α_t is a schedule parameter controlling the noise level at each timestep. The core of DDPM training lies in learning a parameterized model p_θ to approximate the reverse diffusion process:

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \quad (3)$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2 \mathbf{I}). \quad (4)$$

This model learns to progressively remove noise from a given noisy sample \mathbf{x}_t , recovering the original data \mathbf{x}_0 .

3.2. Methodology

As illustrated in Figure 3, EasyRef comprises four key components: (1) a pretrained diffusion model for conditional image generation, (2) a pretrained multimodal large language model (MLLM) for encoding a set of reference images and the text prompt, (3) a condition projector that maps the representations from the MLLM into the latent space of diffusion model, and (4) trainable adapters for integrating image conditioning embedding into the diffusion process.

Reference representation encoding. Existing mainstream approaches [29, 41, 44, 48] mostly average the CLIP image embeddings of all reference images as the reference condition. This image-independent operation cannot effectively capture consistent visual elements among reference images. It also fails to jointly encode text and causes the spatial misalignment issue as presented in Figure 2. To alleviate this issue, we propose to leverage the multi-image comprehension and instruction-following capabilities of the MLLM to encode multi-reference inputs and the text prompt based on the instruction. We adopt the state-of-the-art Qwen2-VL-2B as our MLLM in this work. The MLLM consists of a l -layer large language model (LLM) and a vision encoder capable of handling images with arbitrary resolutions. The input image is initially converted into visual tokens with the vision encoder. Then we employ an instruction and integrate all images into the instruction, which explicitly encourages the MLLM to focus on the crucial and common contents within the reference images. These multimodal input tokens are subsequently processed by the LLM.

Efficient reference aggregation. Increasing the number of reference images inevitably raises the number of visual tokens in the LLM. This extended context length substantially

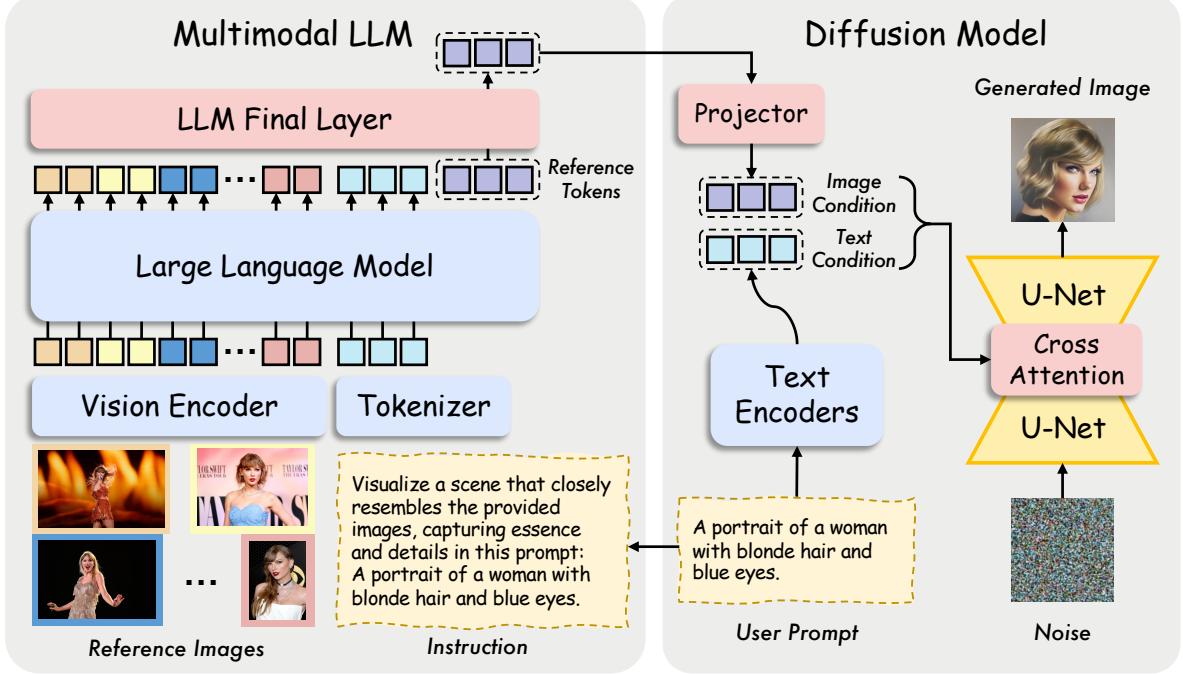


Figure 3. **Overview of EasyRef with SDXL.** EasyRef extracts consistent visual elements from multiple reference images and the text prompt via a MLLM, injecting the condition representations into the diffusion model through cross-attention layers. We only plot 1 cross-attention layer for simplicity.

elevates the computational cost for the diffusion model. We propose to encapsulate the reference representations into N learnable reference tokens $\mathbf{F}_{\text{ref}} \in \mathbb{R}^{N \times D}$ in the LLM to achieve efficient inference. However, all parameters of LLM must be trained to interpret these newly added tokens. To enhance training efficiency, we append \mathbf{F}_{ref} to the context sequence \mathbf{F}_{l-1} at the final layer of LLM, keeping all previous LLM layers frozen during pretraining:

$$\mathbf{F}'_l = \text{Concat}(\mathbf{F}_{l-1}, \mathbf{F}_{\text{ref}}) \quad (5)$$

We then adopt bi-directional self-attention to facilitate the propagation of representations across the reference images in the final layer, followed by a multi-layer perception network (MLP):

$$\mathbf{F}''_l = \text{MLP}(\text{Bi-Attention}(\mathbf{F}'_l)), \quad (6)$$

where we omit the residual addition in attention and MLP layers for simplicity. Next, we split \mathbf{F}''_l into the updated representations \mathbf{F}_l and the encapsulated reference tokens \mathbf{F}'_{ref} :

$$\mathbf{F}_l, \mathbf{F}'_{\text{ref}} = \text{Split}(\mathbf{F}''_l). \quad (7)$$

Finally, we project \mathbf{F}'_{ref} through a trainable MLP condition projector to obtain the final conditioning vector \mathbf{c}_i :

$$\mathbf{c}_i = \text{MLP}(\mathbf{F}'_{\text{ref}}), \quad (8)$$

Reference representation injection. The text conditions are injected into the pretrained diffusion model through

cross-attention layers. Following IP-Adapter, we introduce a new cross-attention layer into each cross-attention layer of the U-Net. Given the latent features \mathbf{X} , text conditions \mathbf{c}_t , and image conditions \mathbf{c}_i , the injected features $\hat{\mathbf{X}}$ are computed by the cross-attention layer as follows:

$$\hat{\mathbf{X}} = \text{Softmax} \left(\frac{\mathbf{Q}\hat{\mathbf{K}}^T}{\sqrt{d}} \right) \mathbf{V} + \text{Softmax} \left(\frac{\mathbf{Q}\hat{\mathbf{V}}^T}{\sqrt{d}} \right) \hat{\mathbf{V}}, \quad (9)$$

where $\hat{\mathbf{K}} = \mathbf{c}_i \hat{\mathbf{W}}_k$ and $\hat{\mathbf{V}} = \mathbf{c}_i \hat{\mathbf{W}}_v$. Both $\hat{\mathbf{W}}_k$ and $\hat{\mathbf{W}}_v$ are newly added trainable parameters.

3.3. Progressive Training Scheme

Alignment pretraining. To facilitate the adaption of MLLM’s visual signals to the diffusion model, we construct a large-scale dataset containing 13M high-quality image-text pairs, including LAION-5B [36] and other internal datasets for the alignment pretraining. During the pretraining phase, we only optimize the final layer and reference tokens of the MLLM along with the newly added adapters and condition projector while preserving the capabilities of the initial MLLM and diffusion model. The shorter side of the input image is resized to 1024 and we further center crop 1024 × 1024 pixels of the image.

Single-reference finetuning. Following alignment pre-training, the MLLM is trainable and subjected to single-reference fine-tuning. Specifically, we subsequently unfreeze the vision encoder and all layers of the MLLM to

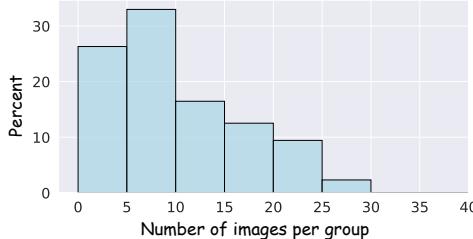


Figure 4. Distribution of our curated dataset.

enhance its capacity for fine-grained visual perception at the second stage. We additionally incorporate trainable Low-Rank Adaption (LoRA) layers to attention layers of the frozen U-Net. Building upon the aforementioned pre-training dataset, we augment the training data with 4M real-world human images from LAION-5B, utilizing cropped face regions as conditioning inputs. The training resolution setting keeps consistent with the first stage.

Multi-reference finetuning. The third stage enables the MLLM to accurately comprehend the common elements across multiple image references and generate high-quality, consistent images. Training is performed on a curated dataset comprising image groups, where each group contains multiple images of the same topic (*e.g.*, Donald Trump, a Tesla Model 3, *etc.*) with varying aspect ratios. During training, one image from each group is randomly selected as the optimization target, while the remaining ones serve as the conditioning inputs. Data augmentation, including random shuffling and truncation, is applied to the conditioning images. We keep the original aspect ratio for each target image.

Training supervision. We use the same training objective as the original stable diffusion model:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0, \epsilon, \mathbf{c}_t, \mathbf{c}_i, t} \|\epsilon - \epsilon_\theta(\mathbf{x}_t, \mathbf{c}_t, \mathbf{c}_i, t)\|^2, \quad (10)$$

where \mathbf{c}_t and \mathbf{c}_i denote the text condition and image condition, respectively.

4. Multi-Reference Generation Benchmark

Dataset construction. We constructed a tag list including celebrities, characters, styles, and common objects, then collected images from diverse sources based on the list. Images sharing the same tag are set into the same group. To generate aligned text captions of the images, synthetic captions generated by Qwen2-VL-7B using the instruction “*Give a brief, concise and precise caption for this image.*” were adopted for each image-text pair. The resulting dataset comprises 8,912,439 images organized into 1,075,378 groups. We set the maximum and minimum group sizes as 28 and 2, respectively, ensuring a balanced group size distribution within the dataset. Figure 4 illustrates the group distribution of our curated dataset.

Data filtering. We also employ a series of efforts for data cleaning. First, low-resolution images and those with low aesthetic scores were excluded. Then we filter out image-text pairs with low CLIP image-text similarity scores using CLIP ViT-L/14. To ensure the consistency within each group, we compute the average CLIP-I and DINO-I scores for each image relative to the other images within its group and filter images with low scores. Finally, recognizing that conventional filtering methods may not adequately identify specific patterns, such as image collages or high-quality images containing dense text, we manually annotate a subset of our collected samples and train a CLIP-based binary classifier to effectively score and filter these instances.

Benchmark splits. The collected image-text pairs are divided into the training dataset, the held-in evaluation set, and the held-out evaluation set. We first sample 60 image groups to construct the held-out evaluation set to evaluate the model performance on unseen data. The number of images in each set varies. For each group, each image can be chosen as the target image and others are regarded as the reference images. There are a total of 487 images in the held-out evaluation set. To compare our method with multi-reference generation approaches that require finetuning (*e.g.*, LoRA), we randomly selected 300 groups from the remaining 1,075,318 groups to form a test set of 2063 samples. Unlike the held-out set, only a randomly selected image serves as the target image in each group. All reference images of the held-in split and other 1,075,018 groups construct the training set. To improve the aesthetic quality of generated images, only images with aesthetic scores higher than 5.5 can be used as the target images during training. There are 1,726,763 valid target images in the training set, with an average of 1.6 images per group.

Evaluation protocol. When evaluating the held-in and held-out sets, we use the reference images and the target prompt to generate two images for each group. Then we employ conventional metrics, including CLIP-I, CLIP-T, and DINO-I, to measure the alignment between generated images and the corresponding target images or prompts. We mainly consider CLIP-I and DINO-I for the image-image alignment, which computes the similarities of image embeddings from CLIP ViT-L/14 and DINOv2-Small [27]. For the image-text alignment, we adopt the CLIPScore [11].

5. Experiments

5.1. Implementation Details

Training. We build our EasyRef framework with the established Stable Diffusion XL [28] model, utilizing the state-of-the-art Qwen2-VL-2B [43] as the MLLM. The resolution of an input image with arbitrary aspect ratio processed by the MLLM can not exceed 336×336 . We introduce 64 reference tokens in the MLLM. Similar to IP-Adapter, we

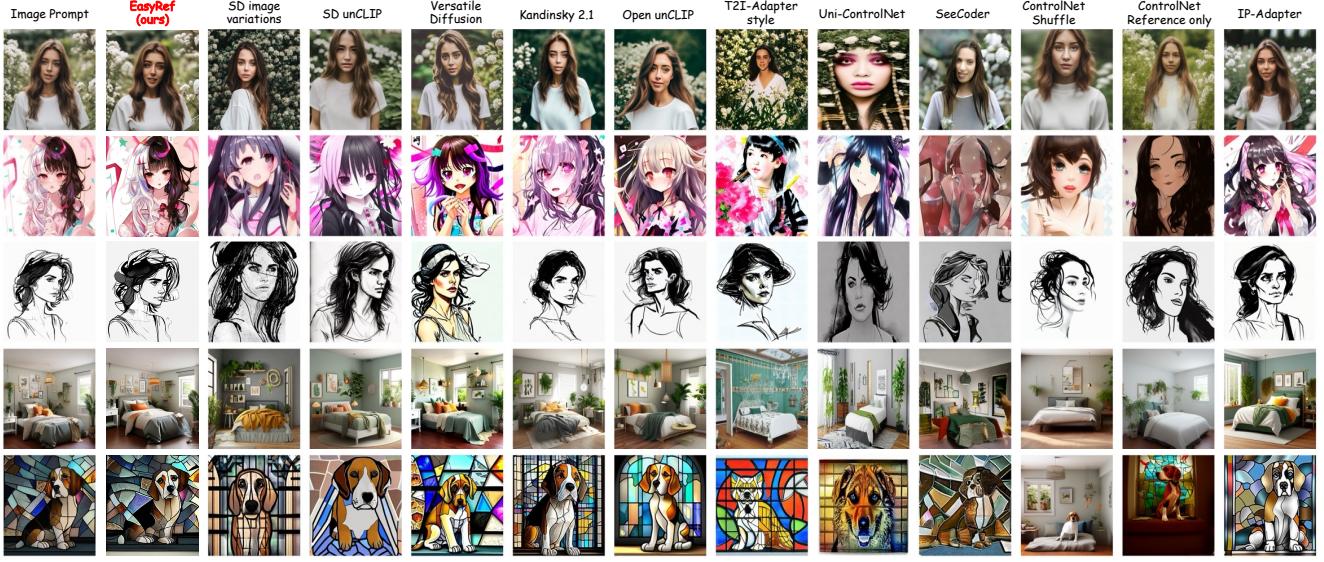


Figure 5. Comparisons of EasyRef with other counterparts in various single-image reference scenarios. The same image prompts as in [48] are used for clear comparisons.

Method	CLIP-I ↑	CLIP-T ↑	DINO-I ↑
<i>Training from scratch</i>			
Open unCLIP [32]	0.858	0.608	-
Kandinsky-2-1 [1]	0.855	0.599	-
Versatile Diffusion [46]	0.830	0.587	-
<i>Finetuning</i>			
SD Image Variations	0.760	0.548	-
SD unCLIP	0.810	0.584	-
<i>Adapters</i>			
Uni-ControlNet [50] (Global Control)	0.736	0.506	-
T2I-Adapter [26] (Style)	0.648	0.485	-
ControlNet Shuffle [49]	0.616	0.421	-
IP-Adapter* [48]	0.828	0.588	-
IP-Adapter-SDXL* [48]	0.836	0.617	0.650
EasyRef	0.876	0.621	0.873

Table 2. Evaluation for generation conditioned by COCO validation images. Methods with * use CLIP embeddings and tend to achieve higher scores of CLIP-based metrics due to its preference.

employ a drop probability of 0.05 for both text and image prompts independently, and a joint drop probability of 0.05 for simultaneous removal of both modalities. We simply treat a square black image as the empty image condition if the image condition is dropped. For the implementation of LoRA comparison, we fine-tuned the model using the reference images and employed a LoRA rank of 32. We present more results in the Appendix.

Evaluation. During inference, we leverage a DDIM [39] sampler with 30 steps and a guidance scale [12] of 7.5. As the original IP-Adapter does not support multi-image references, we employed the average of the CLIP embeddings as the image conditioning input.

Method	CLIP-I ↑	CLIP-T ↑	DINO-I ↑
<i>Held-in split</i>			
LoRA [14]	0.831	0.715	0.654
IP-Adapter-SDXL [48]	0.768	0.632	0.527
EasyRef	0.843	0.726	0.672
<i>Held-out split</i>			
LoRA [14]	failed	failed	failed
IP-Adapter-SDXL [48]	0.795	0.645	0.579
EasyRef	0.833	0.709	0.614

Table 3. Evaluation for multi-reference image generation on MR-Bench. “failed” means LoRA fails to generalize to the unseen held-out split in a zero-shot setting.

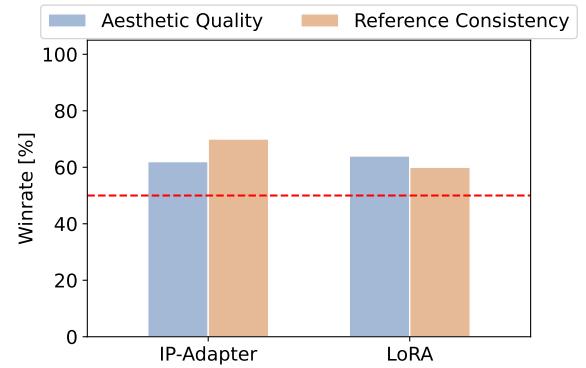


Figure 6. Comparisons of human preference evaluation on our MRBench. EasyRef can surpass other methods across the aesthetic quality and reference alignment.

5.2. Quantitative and Qualitative Results

Single-image reference. We quantitatively compare our method with other counterparts in single-reference scenar-

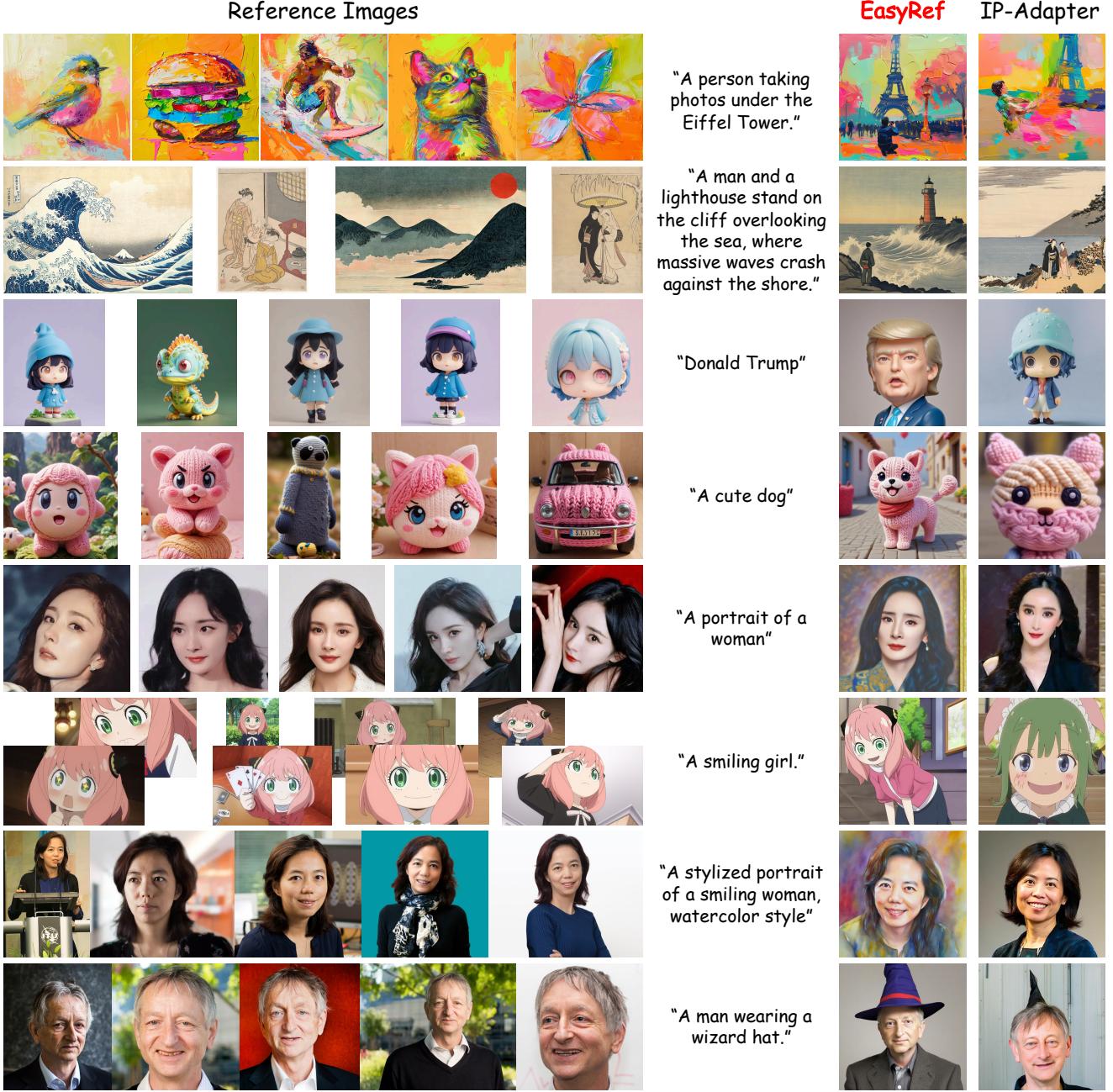


Figure 7. Visualization of generated samples with various multi-reference inputs. These reference contents encompass style, identity, and character, and are encoded by a single generalist MLLM in EasyRef.

ios using the COCO 2017 validation dataset [20], which comprises 5000 image-text pairs. We use the checkpoint trained by single-reference finetuning. As shown in Table 2, EasyRef consistently outperforms other methods in both CLIP-T and DINO-I metrics, demonstrating superior alignment performance. For instance, our model significantly surpasses the IP-Adapter-SDXL by 0.223 DINO-I score. Note that IP-Adapter utilizes CLIP image em-

beddings for conditioning, its generated images may exhibit a bias towards CLIP’s preference, potentially increasing scores when evaluated using CLIP-based metrics. We further conduct qualitative visualization comparisons using some reference images that encompass various styles and contents. As presented in Figure 5, our method achieves better aesthetic quality and consistency with the original image prompts.

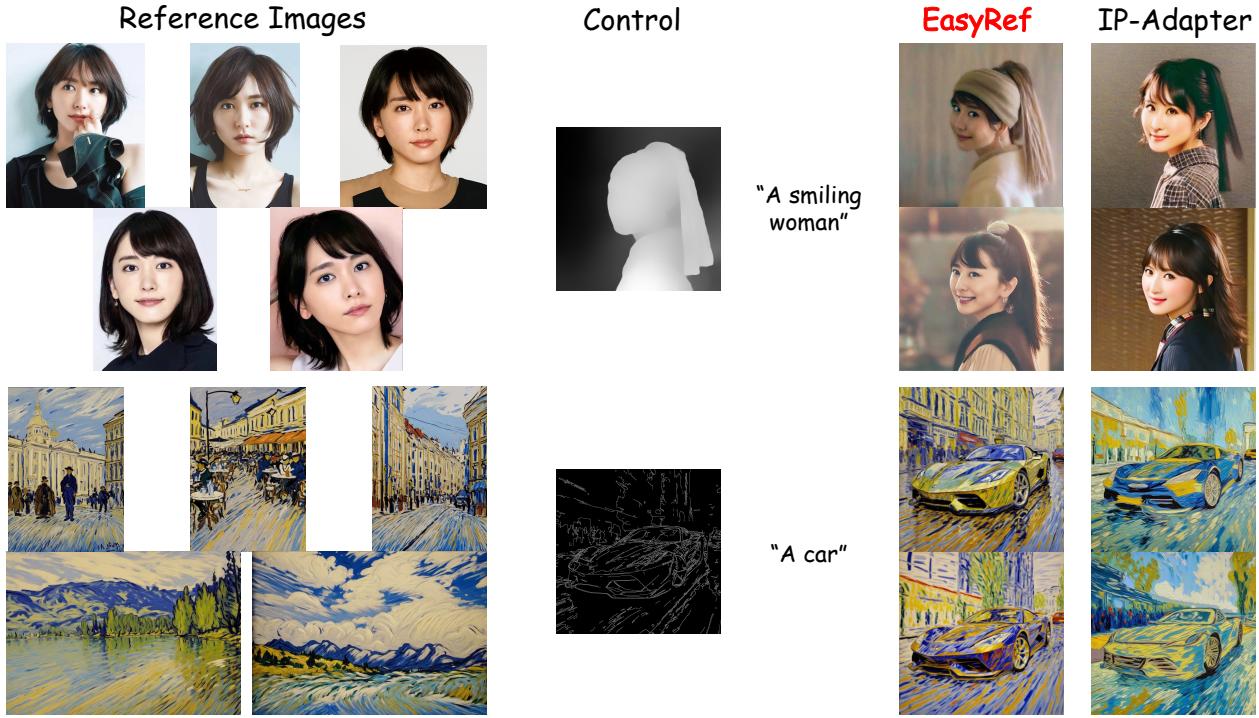


Figure 8. Comparison between EasyRef and IP-Adapter-SDXL with additional structure controls.

Multi-image references. We compare our method with IP-Adapter and the tuning-based LoRA on the MRBench in Table 3. On the held-in split, the tuning-free EasyRef consistently achieves better performances than the tuning-based approach LoRA. In the zero-shot setting, the results demonstrate our method surpass the IP-Adapter with embedding averaging in alignment with the reference images and user prompt. We also present the visualizations in Figure 7. This experiment demonstrates our framework is capable of fully mining consistent visual elements among multiple reference images while maintaining strong generalization ability.

Human evaluation. We systematically evaluate EasyRef with IP-Adapter and LoRA in terms of reference consistency and aesthetic quality. The human evaluation is conducted on our proposed MRBench. Human evaluators were presented with pairwise image comparisons, one generated by EasyRef and the other by a competing model, under blind conditions to ensure fairness. As illustrated in Figure 6, EasyRef outperforms other models in both image-reference alignment and visual aesthetics in user study. This demonstrates EasyRef’s capacity to generate high-fidelity images that conform to the provided reference images.

Compatibility with ControlNet. As shown in Figure 8, our EasyRef is fully compatible with the popular controllable tool, ControlNet [49]. Compared to the IP-Adapter, EasyRef can generate high-fidelity, high-quality, and more consistent results when processing multiple reference im-

ages with additional structure controls.

5.3. Ablation Study

Scaling the number of reference images. Figure 9 illustrates EasyRef’s performance across varying inference lengths. The model exhibits slightly robust performance across varying numbers of references when the number of reference images is within the training constraint. Specifically, the performances continue to increase as the number of references increases within the training constraint. However, performance degrades when the number of references exceeds this constraint. This is due to the limited number of groups with more than 16 images during training and the long-context finetuning may be inadequate. Moreover, the inference efficiency of EasyRef is further evaluated and we find it still maintains acceptable efficiency with 56 reference images due to the efficient token aggregation design.

Multimodal instruction input. An ablation study was conducted to investigate the design of multimodal input to the LLM. As shown in Figure 10, the inclusion of instructions improves generation performance. We speculate this leverages the MLLM’s instruction-following ability to enable it to attend to crucial contents within the reference images and prompt. This observation is consistent with the analysis presented in LI-DiT [25]. Furthermore, incorporating the image prompt can exploit the text-understanding capacity of the MLLM and enhance text-image alignment.

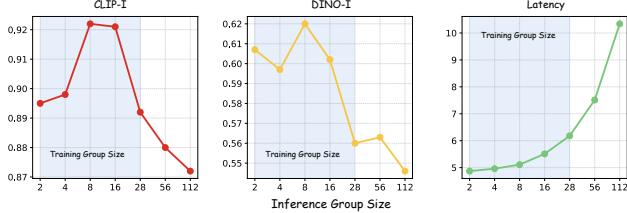


Figure 9. Evaluation of inference group size scaling. We randomly select 112 reference images and 1 target image-text pair with the same topic. Then we compute the similarities between the generated images and the target image. “Latency” in the figure is measured in seconds per image.

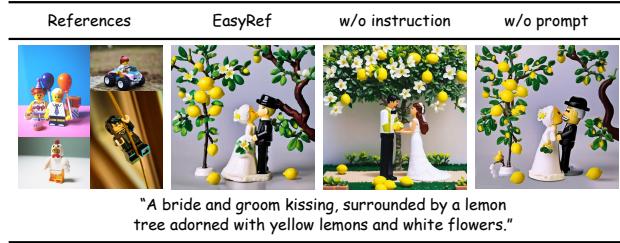


Figure 10. Impact of the multimodal instruction design.

Number	Position	CLIP-I \uparrow	CLIP-T \uparrow	DINO-I \uparrow
32 tokens	-1	0.813	0.693	0.591
128 tokens	-1	0.827	0.705	0.611
64 tokens	-2	0.831	0.704	0.616
64 tokens	-3	0.828	0.702	0.617
64 tokens	-1	0.833	0.709	0.614

Table 4. Ablation of reference token design.

Reference token design. We first ablate the number of reference tokens on the MRBench held-out split. The results in Table 4 show that too many or few tokens can hurt the performance. Hence, we choose 64 tokens to achieve the best trade-off between accuracy and efficiency. Furthermore, we observed comparable performances across various insertion positions (*e.g.*, the final, second to last, and third to last layers of the LLM) for the reference tokens. Consequently, we propose to insert the reference tokens into the final layer for optimal computational efficiency.

Reference aggregation design. In this experiment, we compare our reference token aggregation paradigm with embedding averaging and embedding concatenation. As shown in Table 5, averaging the multi-reference representations leads to performance degradation and the concatenation can increase the reference token number by more than 5 \times . Therefore, utilizing the multi-image comprehension capability of MLLM can enhance the model performance.

Progressive training scheme. The goal of progressive training scheme is to progressively refine the MLLM’s vi-

Method	Average token number	CLIP-I \uparrow	CLIP-T \uparrow	DINO-I \uparrow
Average	64	0.818	0.688	0.584
Concatenation	354	0.821	0.692	0.579
EasyRef	64	0.833	0.709	0.614

Table 5. Ablation of reference representation aggregation on the MRBench held-out set. In the implementation, we average or concatenate the MLLM’s representations of reference images.

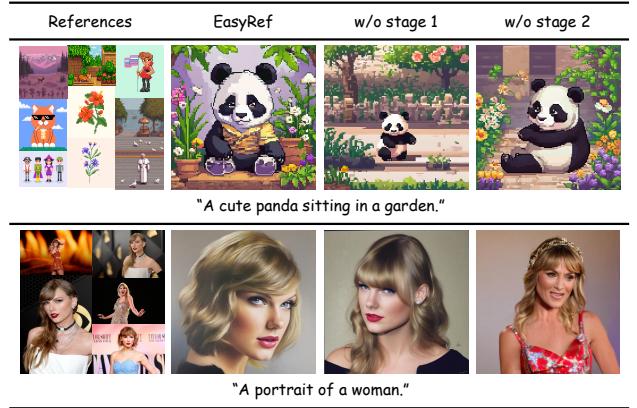


Figure 11. Effect of the progressive training scheme. “stage 1” and “stage 2” denote the alignment pretraining stage and single-reference finetuning stage, respectively.

sual capabilities, ultimately enhancing alignment performance of the diffusion model. By systematically removing each training phase, we visualize the impact of each stage on the model’s ability to capture fine-grained visual details and maintain identity consistency in Figure 11. For some reference contents, such as the pixel art style, EasyRef without alignment pretraining or single-reference finetuning maintains comparable performance. Only for reference images involving identity preservation (*e.g.*, Taylor Swift) or complex compositions do we find significant alignment improvements when adopting all training phases.

6. Conclusion

This paper presents EasyRef, a novel plug-and-play adaptation method that enables diffusion models to be conditioned on multiple reference images and the text prompt. Our approach can effectively capture consistent visual elements within multiple reference images and the text prompt through a multi-image comprehension and instruction-following paradigm, while simultaneously maintaining strong generalization capabilities due to the integration of adapter-based injection. The proposed efficient reference aggregation strategy and progressive training scheme further enhance computational efficiency and fine-grained detail preservation. Through extensive evaluation on our newly introduced MRBench, EasyRef has demonstrably surpassed both tuning-free and tuning-based approaches, in terms of aesthetic qual-

ity and zero-shot generalization across diverse domains.

References

- [1] Shakhmatov Arseniy, Razzhigaev Anton, Nikolich Alekandr, Arkhipkin Vladimir, Pavlov Igor, Kuznetsov Andrey, and Dimitrov Denis. kandinsky 2.1, 2023. 6
- [2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 3
- [3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8, 2023. 2
- [4] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 2, 3
- [5] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.
- [6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 2, 3
- [7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2, 3
- [8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 2
- [9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2, 3
- [10] Zecheng He, Bo Sun, Felix Juefei-Xu, Haoyu Ma, Ankit Ramchandani, Vincent Cheung, Siddharth Shah, Anmol Kalia, Harihar Subramanyam, Alireza Zareian, et al. Imagine yourself: Tuning-free personalized image generation. *arXiv preprint arXiv:2409.13346*, 2024. 3
- [11] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 5
- [12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6
- [13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3
- [14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 3, 6
- [15] Dongzhi Jiang, Guanglu Song, Xiaoshi Wu, Renrui Zhang, Dazhong Shen, Zhuofan Zong, Yu Liu, and Hongsheng Li. Comat: Aligning text-to-image diffusion model with image-to-text concept matching. *arXiv preprint arXiv:2404.03653*, 2024. 2, 3
- [16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2, 3
- [17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2, 3
- [18] Ming Li, Taojinnan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. In *European Conference on Computer Vision*, pages 129–147. Springer, 2025. 2, 3
- [19] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8650, 2024. 3
- [20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V13*, pages 740–755. Springer, 2014. 7
- [21] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 3
- [22] Bingchen Liu, Ehsan Akhgari, Alexander Vishератин, Aleks Kamko, Linmiao Xu, Shivam Shirrao, Joao Souza, Suhail Doshi, and Daiqing Li. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *arXiv preprint arXiv:2409.10695*, 2024. 2
- [23] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2, 3
- [24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 3
- [25] Bingqi Ma, Zhuofan Zong, Guanglu Song, Hongsheng Li, and Yu Liu. Exploring the role of large language models

- in prompt encoding for diffusion models. *arXiv preprint arXiv:2406.11831*, 2024. 2, 3, 8
- [26] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 6
- [27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 5
- [28] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 5
- [29] Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. Deadiff: An efficient stylization diffusion model with disentangled representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8693–8702, 2024. 2, 3
- [30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3
- [31] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 2
- [32] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 6
- [33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2
- [34] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 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, 2023. 2, 3
- [35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2
- [36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 4
- [37] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*, 2024. 2, 3
- [38] Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu. Rethinking the spatial inconsistency in classifier-free diffusion guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9370–9379, 2024. 2
- [39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 6
- [40] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 3
- [41] Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024. 2, 3
- [42] Haofan Wang, Peng Xing, Renyuan Huang, Hao Ai, Qixun Wang, and Xu Bai. Instantstyle-plus: Style transfer with content-preserving in text-to-image generation. *arXiv preprint arXiv:2407.00788*, 2024. 2, 3
- [43] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3, 5
- [44] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 2, 3
- [45] Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. Csgo: Content-style composition in text-to-image generation. *arXiv preprint arXiv:2408.16766*, 2024. 2
- [46] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7754–7765, 2023. 6
- [47] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [48] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ipadapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 3, 6
- [49] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In

- Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. [2](#), [3](#), [6](#), [8](#)
- [50] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. [6](#)
- [51] Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. *arXiv preprint arXiv:2404.13046*, 2024. [2](#), [3](#)

EasyRef: Omni-Generalized Group Image Reference for Diffusion Models via Multimodal LLM

Supplementary Material

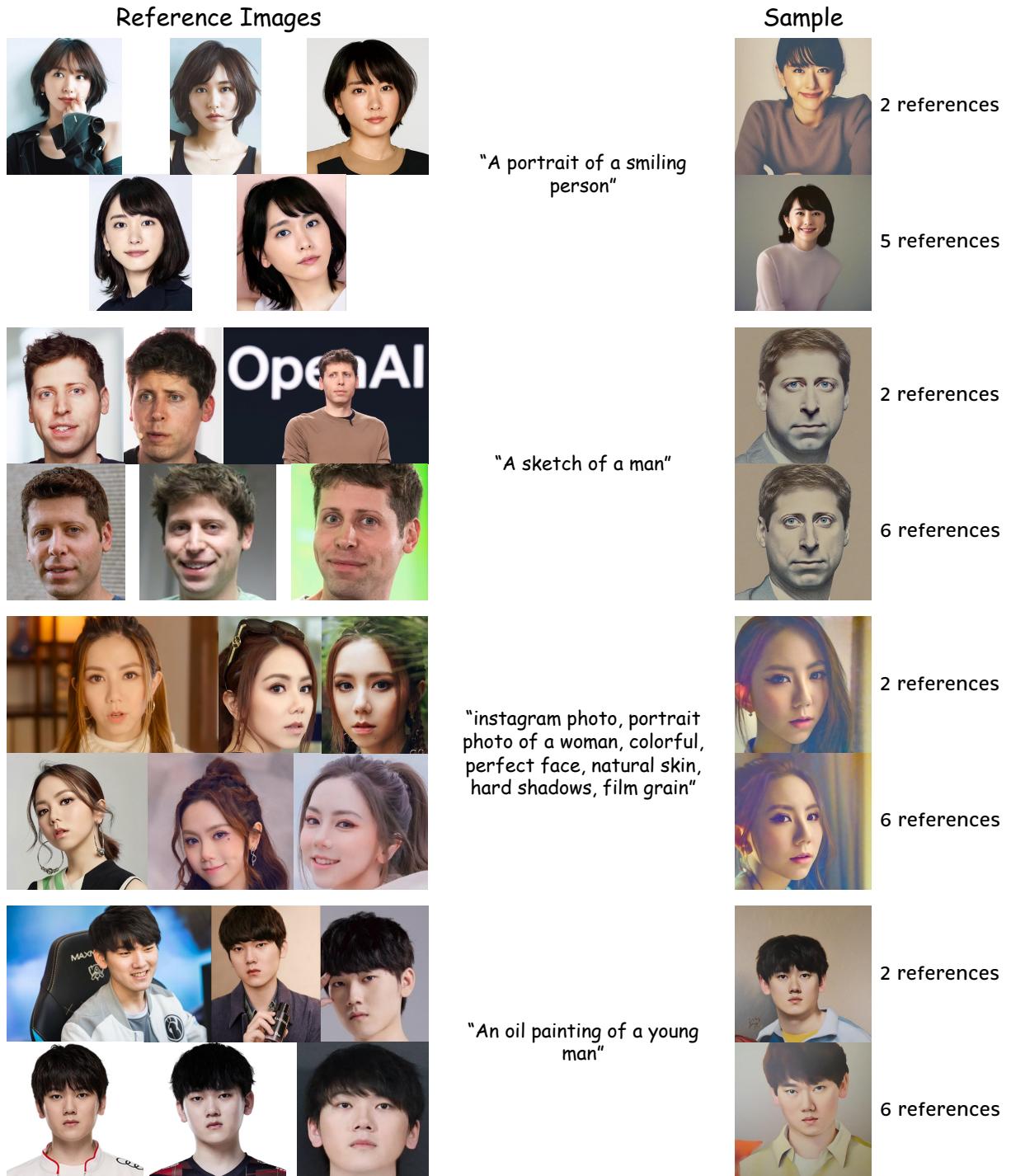
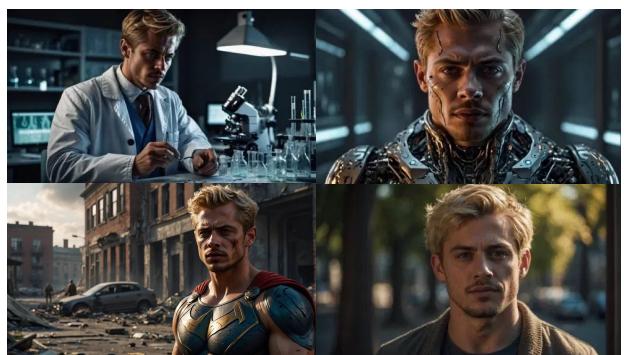


Figure 1. More generated samples of identity preservation with EasyRef in a **zero-shot setting**. We use the face images of celebrities in this experiment.

Reference Images



"A person wearing glasses is running in the forest."



"An anime character"



"A boy, best quality"

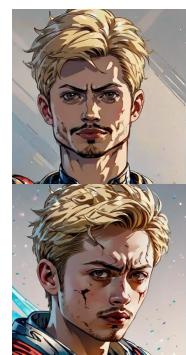


"A character wearing a red jacket stands in a hallway, best quality, 4k"

Sample



2 references



2 references



2 references



2 references

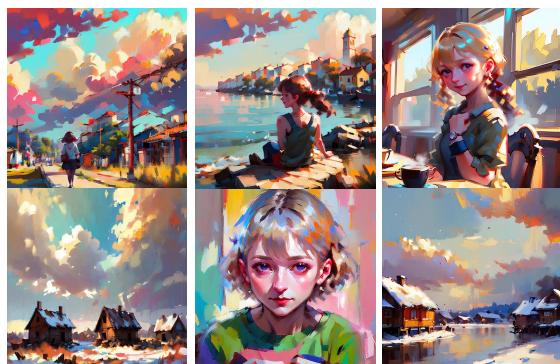


7 references

Reference Images



"A cute tiger"



"A warrior holding a sword standing in the village."



"A robot is riding a motorcycle on the street"



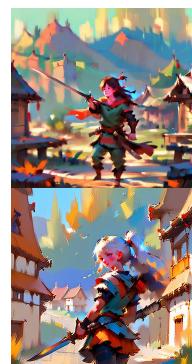
"A child is playing the soccer"

Sample



2 references

6 references



2 references



6 references



2 references



2 references



6 references

Figure 3. More generated samples of style consistency with EasyRef in a **zero-shot setting**.