ResAdapter: Domain Consistent Resolution Adapter for Diffusion Models

Jiaxiang Cheng, Pan Xie*, Xin Xia, Jiashi Li, Jie Wu,
Yuxi Ren, Huixia Li, Xuefeng Xiao, Shilei Wen, Lean Fu,

ByteDance Inc., Beijing, China
jiaxiangcc@gmail.com, xiepan.01@bytedance.com
{xiaxin.97,lijiashi,wujie.10,renyuxi.20190622,lihuixia,xiaoxuefeng.ailab}@bytedance.com

Abstract

Recent advancement in text-to-image models and corresponding personalized technologies enables individuals to generate high-quality and imaginative images. However, they often suffer from limitations when generating images with resolutions outside of their trained domain. To overcome this limitation, we present the resolution adapter **(ResAdapter)**, a domain-consistent adapter designed for diffusion models to generate images with unrestricted resolutions and aspect ratios. Unlike other multi-resolution generation methods that process images of static resolution with complex post-process operations, ResAdapter directly generates images with the dynamical resolution. Especially, after learning a deep understanding of pure resolution priors, ResAdapter trained on the general dataset, generates resolution-free images with personalized diffusion models while preserving their original style domain. Comprehensive experiments demonstrate that ResAdapter with only 0.5M can process images with flexible resolutions for arbitrary diffusion models. More extended experiments demonstrate that ResAdapter is compatible with other modules for image generation across a broad range of resolutions, and can be integrated into other multi-resolution model for efficiently generating higher-resolution images.

Code — https://github.com/bytedance/res-adapter

1 Introduction

Diffusion models (Ho, Jain, and Abbeel 2020; Song et al. 2021; Song, Meng, and Ermon 2020) have experienced a remarkable surge in their capabilities and applications (Lugmayr et al. 2022; Meng et al. 2021; Ramesh et al. 2022). Among them, Stable Diffusion (SD) (Rombach et al. 2022) and SDXL (Podell et al. 2024) are pre-trained models on the large-scale dataset LAION-5B (Schuhmann et al. 2022), having emerged as powerful generative models. Additionally, the open-source community has been enriched by numerous personalized diffusion models from CivitAI (Civitai 2022), trained with DreamBooth (Ruiz et al. 2023) or Low-rank Adaptation (LoRA) (Hu et al. 2022). They are capable of generating imaginative high-quality images at the training resolution (e.g., 512×512 for SD-based models and 1024×1024 for SDXL-based models) using the given prompts. However, they often suffer from limitations when generating images with resolutions outside of their trained domain. As shown in Figure 1, the SD-based model and the personalized diffusion model generate lower-resolution images (e.g., 256×256) with the poor fidelity and higher-resolution images (e.g., 1024×1024) with the poor framing and composition. As a result, we can name this phenomena as the *resolution domain inconsistent*.

Method	Type	Domain Consistent	Module Compatible	Training Inexpensive	Inference Efficient
ASD	Train	×	×	×	×
SDXL	Train	×	×	×	✓
Diffit	Train	×	×	✓	✓
MultiDiff	Train-free	×	×	✓	×
ElasticDiff	Train-free	×	×	✓	×
MixtureDiff	Train-free	×	×	✓	×
HiDiffusion	Train-free	×	×	✓	✓
LoRA	Few-shot	×	✓	✓	✓
ResAdapter	Zero-shot	✓	✓	✓	✓

Table 1: **Comparison for ResAdapter and other methods**. Domain Consistent: resolution and style domain of generation image maintain consistent for arbitrary diffusion models. Module Compatible: compatible with other modules except diffusion models. Training Inexpensive: low-cost training. Inference Efficient: process images without repeated denoising steps and complex post-process operations.

Existing work is categorized into two main research directions to address this limitation. The first research line is train-free direction (Jiménez 2023; Bar-Tal et al. 2023; Haji-Ali, Balakrishnan, and Ordonez 2024; Zhang et al. 2025), represented by MultiDiffusion and ElasticDiffusion, where images with resolutions in their trained domain are repeatedly processed and then stitched together to generate images with flexible resolutions through overlap. However, these approaches often take longer inference time with complex post-process operations. The second research line is straightforward. Fine-tuning models (Zheng et al. 2024) or additional parameters (Hu et al. 2022) on a broader range of resolutions to empower diffusion models to generate resolution-free images. However, most personalized models in CivitAI (Civitai 2022) do not provide details about their training datasets. Fine-tuning on the general dataset like LAION-5B (Schuhmann et al. 2022) inevitably influences their original style domain, which is shown in Figure 1. We name this phenomena as the *style domain inconsistent*.

*Corresponding author

Figure 1: **Motivation**. We explore the domain distribution of images generated by SD1.5 and Dreamlike at resolutions of 256×256, 512×512 and 1024×1024. Dreamlike is the personalized diffusion model based on SD1.5. We find that **baselines** transform domains at resolutions of 256×256 and 1024×1024. The above ResAdapter and LoRA are both trained on the general dataset LAION-5B. **ResAdapter** keep domain consistent at different resolutions. But **LoRA** injects style priors from LAION-5B and influences the Dreamlike domain, resulting to low-quality images with the style conflict.

Can we train a plug-and-play resolution adapter to generate images with unrestricted resolutions and aspect ratio for arbitrary diffusion models? To answer this question, we decompose it into three dimensions. (1) *Resolution interpolation*: generate images with resolutions below the trained resolution of diffusion models. (2) *Resolution extrapolation*: process images with resolutions above the trained resolution of diffusion models. (3) *Style consistency*: generate images without transforming the original style domain of the personalized diffusion model.

We analyze the structure of diffusion model blocks (Ronneberger, Fischer, and Brox 2015), finding that the attention and feed-forward are both content-sensitive layers, which are sensitive to the style information of images compared to resolution. However, the convolution layers with fixed receptive file are resolution-sensitive, meaning they are easily influenced by the resolution of generation images. Leveraging these finds, we present the resolution convolution LoRA (*ResCLoRA*) for dynamically matching the receptive filed of convolution and the feature map size of images with flexible resolutions. However, we find that as the resolution increases, the gap about the quality of generation images increases between LoRA and full fine-tuing. We attribute it into the inability of normalization in diffusion model blocks to adapt the statistical distribution of images in resolution extrapolation. According to this, we present the resolution extrapolation normalization (*ResENorm*) for reducing the gap between LoRA and full fine-tuning in resolution extrapolation. To enable the style domain consistency, we optimize the position of ResCLoRA and ResENorm insertions on diffusion model blocks to guide them to learn resolution priors ignoring the style information from the general datasets.

Through integrating these two optimized methods, we can train a plug-and-play domain-consistent resolution adapter (**ResAdapter**), which expands the range of resolution domain from diffusion models without transforming their original style domains. Our main experiments demonstrate that after learning resolution priors, ResAdapter with only 0.5M can expand the generation resolution of SD-based personalized models from 128×128 to 1024×1024 and scale the resolution of SDXL-based personalized models from 256×256 to 1536×1536. Our extensive experiments demonstrate that ResAdapter is compatible with other modules (e.g., ControlNet (Zhang, Rao, and Agrawala 2023) for conditional generation, IP-Adapter (Ye et al. 2023) for image generation based on the image prompt and LCM-LoRA (Luo et al. 2023) for accelerating generation), and even can be integrated into other multi-resolution models (e.g., ElasticDiffusion (Haji-Ali, Balakrishnan, and Ordonez 2024)) for efficiently generating 2048×2048 high-resolution images. Detailed comparison with other related work is summarized in Table 1. Our contributions can be summarized as follows:

- We present a plug-and-play domain-consistent ResAdapter for generating images of resolution interpolation and extrapolation with diffusion models.
- ResAdapter enables diffusion models of arbitrary style domain to generate images of unrestricted resolution and aspect ratio without transforming their style domain.
- ResAdapter is lightweight and without complex post-process operations. We can train it once for only 0.5M with low-cost consumption and efficiently inference resolution-free images.
- ResAdapter is compatible with other modules to generate images with flexible resolution, such as ControlNet, IP-Adapter and LCM-LoRA, even can optimize generation efficiency of other multi-resolution models.

2 Related Work

2.1 Text-based Image Generation

The rapid development of artificial intelligence generative component has attracted growing interest in text-to-image generation. GAN-based methods (Radford, Metz, and Chintala 2016; Goodfellow et al. 2020; Odena, Olah, and Shlens 2017) employ small-scale data for training but encounter challenges in adapting to large-scale data due to the instability of the adversarial training process. Autoregressive-based methods (Van Den Oord, Vinyals et al. 2017; Esser, Rombach, and Ommer 2021; Lee et al. 2022) learn the latent

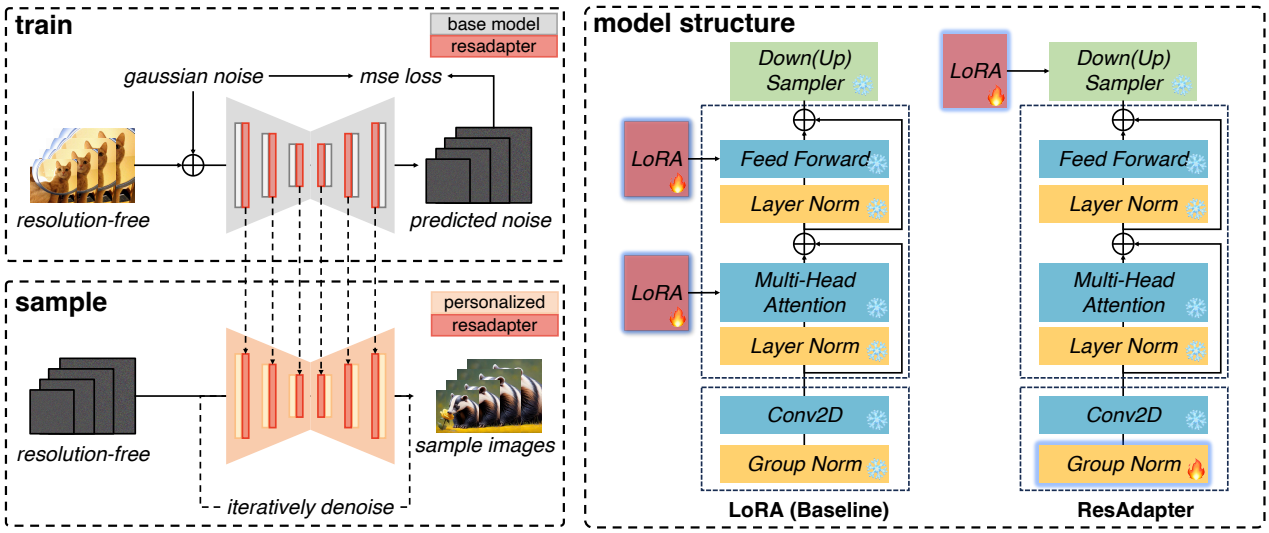

Figure 2: **Overview of ResAdapter**. **Left**: Pipeline of ResAdapter. ResAdapter based on the frozen base model (e.g., SD or SDXL) learns resolution priors from mixed-resolution general datasets, and can be integrated into arbitrary personalized models to generate multi-resolution images. **Right**: Architecture comparison between ResAdapter and LoRA. Compared to LoRA, ResAdapter is only inserted to downsampler and upsampler, and it unfreezes the group normalization of resnet blocks.

distribution of discrete latent spaces but take more inference costs. Recently, diffusion models (Song et al. 2021; Ho, Jain, and Abbeel 2020; Karras et al. 2022; Dhariwal and Nichol 2021) has emerged as the state-of-the-art model in the text-to-image generation field. The diffusion models represented by Stable Diffusion (Rombach et al. 2022) and SDXL (Podell et al. 2024) contribute to high-resolution image generation. With the advent of personalized techniques (Ruiz et al. 2023; Hu et al. 2022), they are capable of generating imaginative images. However, they still encounter limitations in resolution-free image generation.

2.2 Resolution-free Image Generation

Existing work of resolution-free image generation mainly utilizes post-processing to generate images beyond the training resolution. Mixture-of-Diffusers (Jiménez 2023) and MultiDiffusion (Bar-Tal et al. 2023) utilize pre-trained Stable Diffusion (Rombach et al. 2022) to generate 512×512 images multiple times, and overlap to generate high-resolution landscape images. But this also lead to duplicated objects. ASD (Zheng et al. 2024) fine-tunes on multi-aspect ratio images, and generates high-resolution images throught implicit overlap. ElasticDiffusion (Haji-Ali, Balakrishnan, and Ordonez 2024) optimizes the post-processing process and can generate lower resolution images. Compared to these work, our ResAdapter does not require post-processing that take more inference time and can be integrated into any personalized model. ResAdapter can even be combined with these work to optimize inference time of generating higher resolution images.

3 The Method

In this section, we delve into our proposed plug-and-play domain-consistent ResAdapter, which enables diffusion models of arbitrary style domain to generate images with unrestricted resolutions and aspect ratio. First, we introduce ResCLoRA, which enables diffusion models to generate images with resolution interpolation. Then, we introduce ResENorm, which compensates for the lack of capability about ResCLoRA in resolution extrapolation. Finally, we present a simple multi-resolution training strategy, which can effectively make diffusion models generate images with flexible resolutions through only one ResAdapter.

3.1 Resolution Interpolation

LoRA (Hu et al. 2022) enables the base model (e.g., SD (Rombach et al. 2022) and SDXL(Podell et al. 2024)) to generate high-quality style images. As shown in Figure 2, LoRA is inserted into the query, value, key and output layers of the attention block to learn the style domain distribution of images. It is defined as $\mathcal{W}_a' = \mathcal{W}_a + \Delta\mathcal{W} = \mathcal{W}_a + AB^T$, where $A \in \mathbb{R}^{m\times r}$, $B \in \mathbb{R}^{n\times r}$ are two rank-decomposition matrices, r represents the rank of matrices. However, LoRA trained on the general datasets can not be integrated into other personalized diffusion models, which influences their original style domain. As shown in Figure 1, LoRA trained on LAION-5B (Schuhmann et al. 2022) transforms the domain of the personalized model to the domain of SD1.5 and generates bad quality images with style conflicts.

ResCLoRA can be integrated into any personalized model to enable resolution interpolation for high-quality images without transforming the style domain. The reason that leads to the poor fidelity of images with resolution interpolation is that the convolution with the fixed receptive field is sensitive to the resolution of images. According to this, ResCLoRA is inserted into the convolution layers of diffusion blocks to learn resolution priors. To prevent as much as possible ResCLoRA from capturing the style domain of the general datasets, it is only inserted into the convolution layers in downsampler and upsampler blocks. We define ResCLoRA

as $\mathcal{W}'_d = \mathcal{W}_d + AB^T$ and $\mathcal{W}'_u = \mathcal{W}_u + AB^T$, which is shown in Figure 2. Compared to the style information, the resolution information is low-level knowledge. Thus, ResCLoRA with only 0.4M can provide rich resolution priors for personalized models, adaptively adjusting the receptive field of convolution in diffusion blocks to match the feature map size of generation images while preserving their style domain.

3.2 Resolution Extrapolation

Our initial experiment finds that there is a large gap as the resolution increases between LoRA and full fine-tuning. This means that only ResCLoRA does not enable the resolution extrapolation ability of the personalized model. For example, ResCLoRA integrated into the diffusion model still generates higher-resolution images with poor framing and composition. Inspired by LongLoRA (Chen et al. 2024), we find that the failure of the resolution extrapolation is limited by the ability of normalization layers. Existing normalization layers can not adapt to the statistical distribution of feature maps of higher-resolution images.

However, we find that all normalization layers of diffusion models blcoks trained on LAION-5B (Schuhmann et al. 2021) are not compatible with the other parameters of the personalized model, which still leads to low-quality images with poor style color. In order to keep the original style domain of generation images, we need to maintain partial normalization layers of the personalized model. As shown in Figure 2, we only open group normalization of resnet layer, which is named as ResENorm. It not only reduce the gap about the resolution prior between ResCLoRA and full fine-tuning, but also helps retain the style domain of the personalized model. Additionally, we only train ResENorm in resolution extrapolation for better adapting to the statistical distribution of the feature map of higher-resolution images. After that, ResENorm with ResCLoRA can improve the poor framing and composition of generation images. Especially, ResENorm only occupies 0.1M parameters but make effects for reducing the gap in resolution extrapolation.

3.3 Resolution-Free Consistency Training

To enable resolution-free domain consistent image generation for single ResAdapter, we propose a simple mixed-resolution training strategy based on our specific adapter design, as shown in Figure 2. For SD (Ramesh et al. 2022), we train on the mixed datasets with common resolutions from 128×128 to 1024×1024 with unrestricted aspect ratio. For SDXL (Podell et al. 2024), we train on the mixed datasets with common resolutions from 256×256 to 1536×1536 with unrestricted aspect ratio. In the training process, the base model is frozen, and only the ResAdapter is trainable. The multi-resolution training strategy, along with the specific adapter design, enables ResAdapter to learn multi-resolution knowledge simultaneously while preventing catastrophic forgetting (Masip et al. 2023; Smith et al. 2024; Gao and Liu 2023) from full fine-tuning.

Our experiments find that the lower and higher resolution images (e.g., 128×128 and 1024×1024 for SD) are more difficult to train. To alleviate this phenomenon, we use a simple probability function to sample images at the different training resolution. It is defined as $p(x) = |x-s|^2 / \sum_i^N |x_i - s|^2$, where s represents the standard training resolution (e.g., 512×512 for SD). This can improve the probability of selecting lower or higher resolution images resolution during the multi-resolution training process.

4 Experiments

In this section, we introduce the experimental setup and results. First, we describe the experimental setup in detail, including training details, evaluation metrics, and the selection of personalized models. And we show the main experimental results. We compare ResAdapter with other multi-resolution image generation models as well as the original personalized model. Then we show the extended experimental results. That is the application of ResAdapter in combination with other modules. Finally, we perform ablation experiments about ResAdapter modules and alpha.

4.1 Experimental Setup

Training Details. We train ResAdapter using the large-scale dataset LAION-5B (Schuhmann et al. 2022). Considering most structures of personalized models in the open-source community, we choose SD1.5 (Rombach et al. 2022) and SDXL1.0 (Podell et al. 2024) as the base models. For SD1.5, we train on images with 128×128, 256×256, 384×384, 768×768 and 1024×1024 resolutions. For SDXL, we expanded the training resolution range to 256×256, 384×384, 512×512, 768×768, 1280×1280, 1408×1408, and 1536×1536. Meanwhile, the training dataset contains images with different ratios such as 4:3, 3:4, 3:2, 2:3, 16:9 and 9:16. For SD1.5 and SDXL, we both use a batch size of 32 and a learning rate of 1e-4 for training. We use the AdamW optimizer (Kingma and Ba 2015) with $\beta_1 = 0.95, \beta_2 = 0.99$. The total number of training steps is 20,000. Since ResAdapter is only 0.5M of trainable parameters, we train it for less than an hours on 8 $\times$ A100 GPUs.

Evaluation Metrics. For experiments comparing ResAdapter with the personalized model, we hire 5 humans to participating in the qualitative evaluation. For experiments comparing ResAdapter and the other multi-resolution image generation models, we refer to (Haji-Ali, Balakrishnan, and Ordonez 2024) and use Fréchet Inception Distance (FID) (Heusel et al. 2017) and CLIP Score (Hessel et al. 2021) as evaluation metrics. They evaluate the quality of the generated images and the degree of alignment between the generated images and prompts. For other multi-resolution generation models, we chose MultiDiffusion (MD) (Bar-Tal et al. 2023) and ElasticDiffusion (ED) as baselines.

Personalized Models. In order to demonstrate the effectiveness of our ResAdapter, we choose multiple personalized models from Civitai (Civitai 2022), which cover a wide domains range from animation to realistic photography. For personalized model information, see details in Appendix B.

4.2 Main Results

Comparison with Resolution-Free Generation Models. ResAdapter significantly improves the quality of multi-resolution images. For *quantitative results*, see Table 2. The

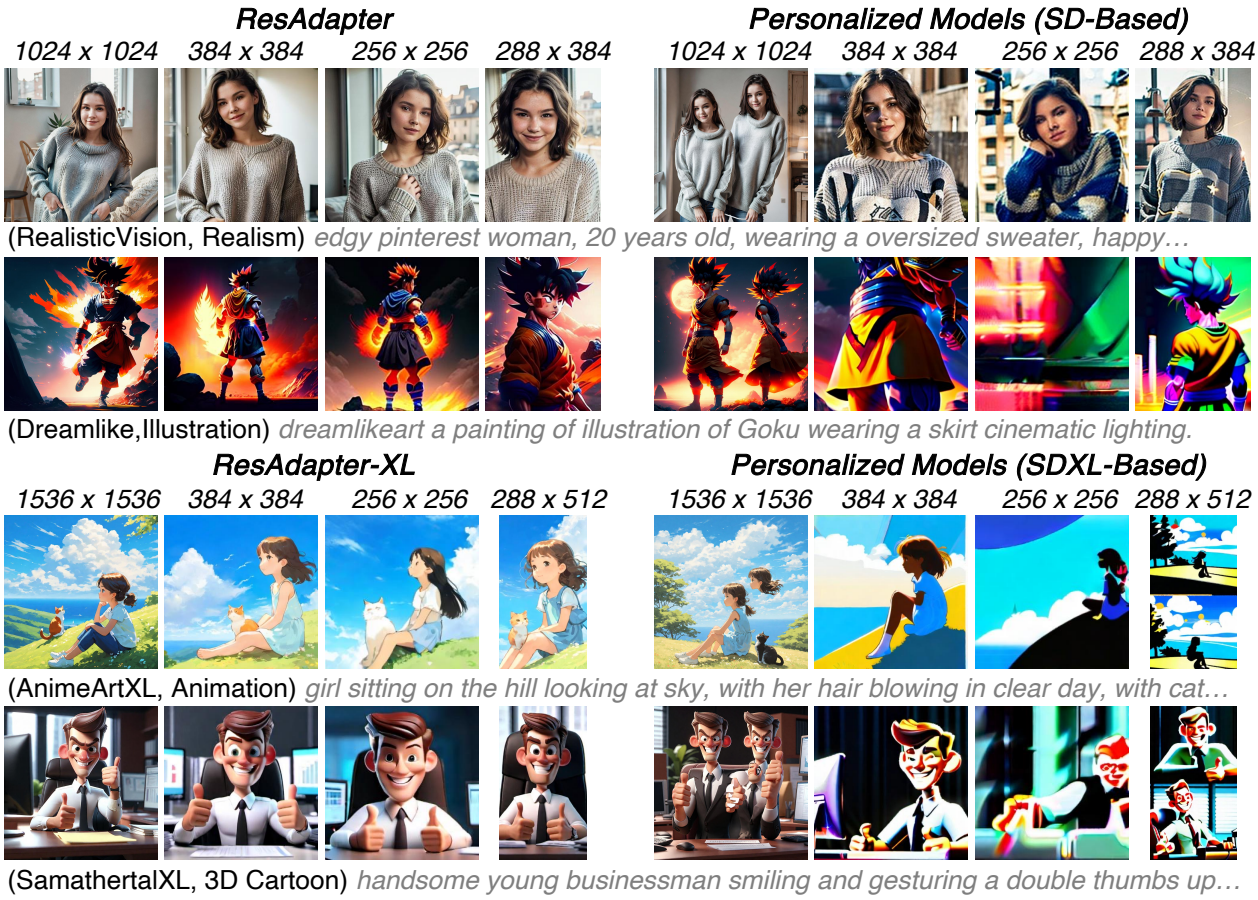

Figure 3: **Qualitative results**. We compare the multi-resolution images generated by ResAdapter and the personalized models of arbitrary style domains. **Left**: generation images from ResAdapter integrated into the personalized model. **Right**: generation images from the original personalized model. Some prompts are edited for clarity.

Size	Method	FID(↓)	CLIP(↑)	Call	Latency/(s)
256x256	$SD_{1.4}$	54.06	21.43	2	0.025
	SDXL	175.87	14.60	2	0.038
	$ED_{1.4}$	23.77	26.30	2	0.288
	Ours$_{1.4}$	**23.01**	**26.98**	2	**0.025**
512x512	$SD_{1.4}$	20.50	27.33	2	0.0714
	Ours$_{1.4}$	20.53	27.32	2	0.0714
1024x1024	SD	47.01	25.70	2	0.1322
	SDXL	25.58	**28.06**	2	0.1322
	$MD_{1.4}$	37.70	26.96	162	2.50
	$ED_{1.4}$	27.76	26.07	33	1.16
	Ours$_{1.4}$	**26.89**	27.26	2	**0.1322**

Table 2: **Quantitative results** on LAION-COCO at the different resolutions. Call represents the number of inference iterations at each noise reduction step. For the latency, we measure the time of one step on an A100-80G.

results show that ResAdapter outperforms MD and ED in terms of FID and CLIP Score. About the latency time, ResAdapter without post-processing is, on average, 9× faster compared to ED. For *qualitative results*, see more details in Appendix C. We compare the image performance of MD and ED with our ResAdapter at the resolutions from 256 to 1024. The qualitative results demonstrate that ResAdapter generates multi-resolution images of better quality compared with MD and ED. MD generates higher-resolution images with the poor framing and composition, and can not generate lower-resolution images than the training resolution. ED generates the images with more inference time.

Comparison with Personalized Models. For *quantitative results*, see Table 3. We evaluate the image quality by four criteria, which are the fidelity, the composition, the prompt alignment and the style domain consistency. The quantitative results demonstrate that ResAdapter significantly outperforms the personalized model, particularly in lower-resolution images. For *qualitative results*, see Figure 3. To ensure the fairness of the experiments, we generate multi-resolution images using prompts from Civitai (Civitai 2022). These images are generated by ResAdapter and the personalized model. Lower-resolution images (e.g., 256×256, 384×384) generated by the personalized model are significantly lower in terms of the images of the fidelity, while higher-resolution images (e.g., 1536×1536) suffer from the poor framing and composition. After integrating our ResAdapter into the personalized model, the fidelity and the composition of generation images are significantly improved. ResAdapter enables the resolution extrapolation and interpolation of the personalized model.

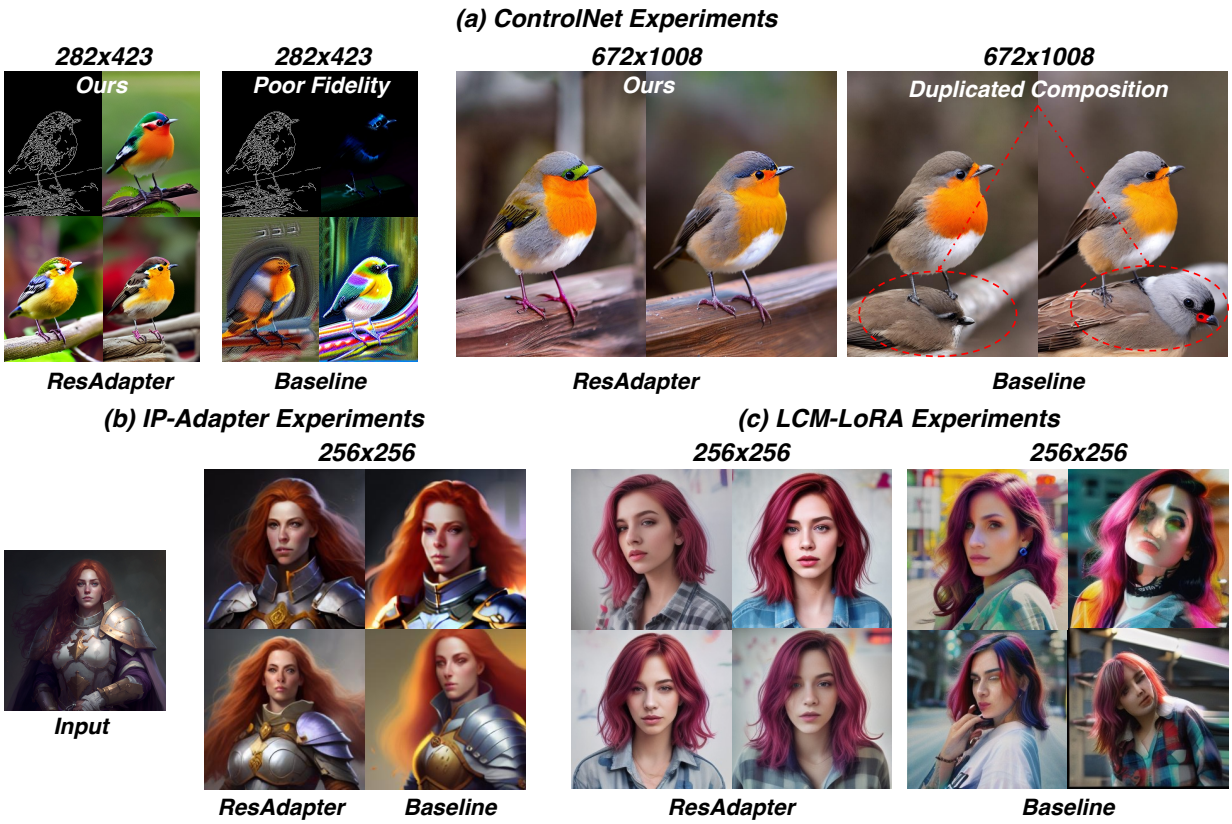

Figure 4: **Qualitative results** of extended experiments with ResAdapter. Image-to-image tasks with **ConrtolNet**, the condition is canny images at different resolution. Image variation tasks with **IP-Adapter**, we resize the input image from 1024x1024 to 256x256. Accelerating text-to-image tasks with **LCM-LoRA**, we generate images in 4 steps.

Domain	Size	Ratio	Good	Same	Bad	(G+S)/(B+S)
Realism	256	4:3	3321	1090	589	2.63
		3:4	3302	1102	596	2.59
	1024	16:9	2877	1563	560	2.09
		9:16	2978	1531	491	2.23
Illustration	256	4:3	3671	1108	221	3.59
		3:4	3712	982	306	3.64
	1024	16:9	2621	1762	617	1.84
		9:16	2768	1598	634	1.95
Animation	384	4:3	4479	402	119	9.36
		3:4	4574	302	124	11.44
	1536	16:9	2978	1391	631	2.16
		9:16	2809	1278	913	1.86
Cartoon	384	4:3	4613	338	49	12.79
		3:4	4684	249	67	15.61
	1536	16:9	2863	1566	571	2.07
		9:16	3009	1071	920	2.05

Table 3: **Quantitative results** by human for resolution-free generation of ResAdapter and personalized models.

Consistency with Personalized Models. We report the quantitative evaluation of ResAdapter's performance in maintaining style domain consistency when paired with personalized models, as summarized in Table 4. They undergo training for 10,000 iterations on the LAION-5B dataset, generating 5,000 images at training resolution within the personalized model framework. Unlike alternative approaches which exhibit significant divergence from the original style, ResAdapter preserves the integrity of the style domain. This evidence supports the conclusion that ResAdapter effectively upholds the stylistic characteristics of personalized diffusion models without degradation.

Method	Realism	Animation	Cartoon	Illustration
LoRA	35.67	51.28	47.81	45.18
Diffit	31.34	47.53	41.24	38.72
$\text{Norm}_{\text{Full}}$	31.34	47.53	41.24	38.72
$\text{Ours}_{\text{ResENorm}}$	2.18	3.76	2.97	2.81
$\text{Ours}_{\text{ResCLoRA}}$	2.29	3.69	3.02	2.84
$\text{Ours}_{\text{Full}}$	2.32	3.94	3.08	2.89

Table 4: **Quantitative results** about style consistency preservation. We train models on LAION-5B for the same epochs, and use FID to measure the gap from the distribution domain of the original persionalized model.

4.3 Extended Results

ResAdapter with ControlNet. ControlNet (Zhang, Rao, and Agrawala 2023) is a conditional control module that can utilize conditional images to control the generation of layout-specific images for SD (Rombach et al. 2022). As shown in Figure 4, ControlNet generates low-quality images with poor fidelity and composition in the image-to-image task. While ResAdapter is compatible with Control-

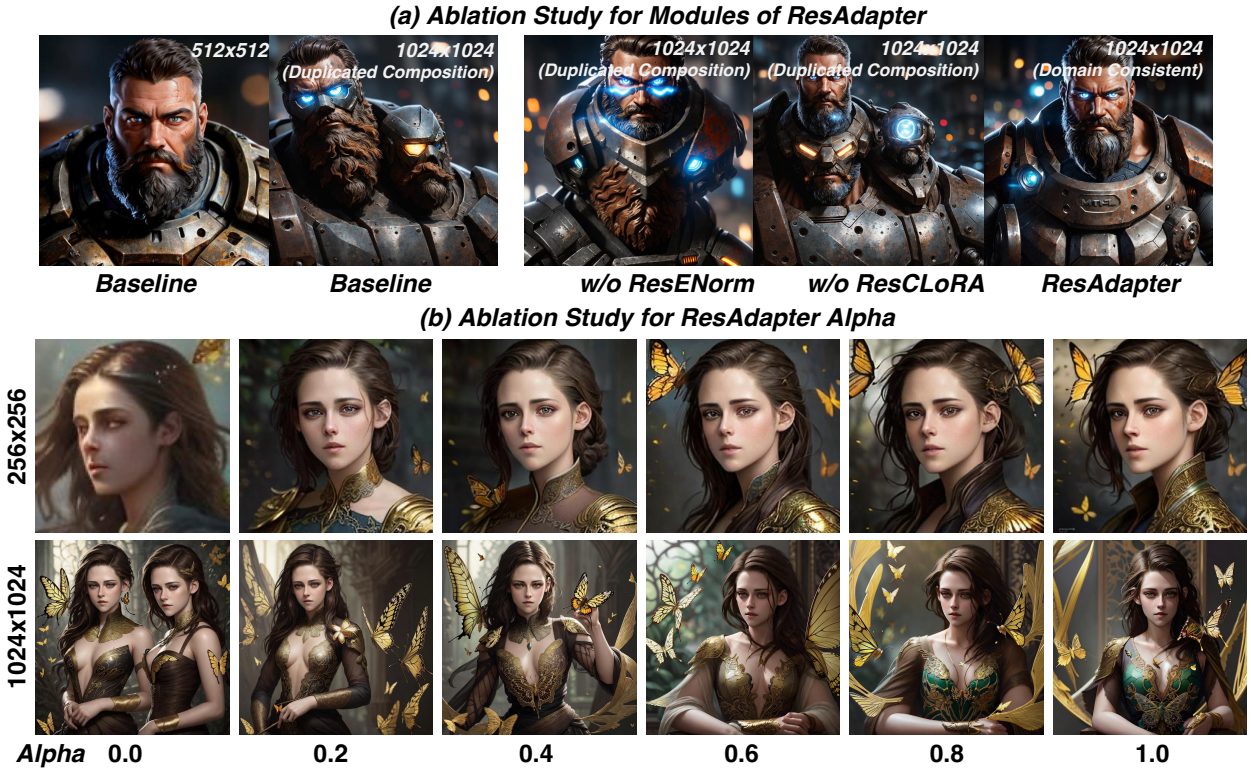

Figure 5: **Ablation studies. Top**: We ablate on the modules of ResAdapter. Baseline represents Dreamshaper, which is a personalized diffusion model based on SD1.5. The third column represents only ResCLoRA integrated into the model. The fourth column represents only ResENorm integrated into the model. The fifth column represents both them integrated into the model. **Bottom:** We ablate on the alpha of ResAdapter α_r from 0 to 1 at lower and higher resolutions.

Net to enable the resolution extrapolation and interpolation, improving the quality of images.

ResAdapter with IP-Adapter. IP-Adapter (Ye et al. 2023) is a adapter for image generation with the image prompt. As shown in Figure 4, ResAdapter with IP-Adapter can generate high-quality images in the image variation task.

ResAdapter with LCM-LoRA. LCM-LoRA (Luo et al. 2023) is a module for accelerated image generation capable of generating high-quality images in 4 steps. As shown in Figure 4, ResAdapter is integrated into the personalized model and compatible with LCM-LoRA. ResAdapter improves the fidelity of lower-resolution image while not degrades the quality of 512×512 images.

ResAdapter with ElasticDiffusion. ResAdapter combined with other multi-resolution image generation models of post-processing can optimize the inference time. Specifically, ED (Haji-Ali, Balakrishnan, and Ordonez 2024) requires to inference the 1024×1024 images multiple times, and overlaps them to get the 2048×2048 images with the post-process technology, which takes much inference time. In order to optimize the inference time, we combine ResAdapter with ED to inference the 768×768 images same times and overlap them to get 2048×2048 images. ResAdapter with ED can generate 2048×2048 images while no degradation of image quality compared with ED. Our ResAdapter can speed up the inference time by 44%, which can be found in Appendix C. This demonstrates that ResAdapter can be flexibly applied to multiple scenarios.

4.4 Ablation Studies

For the modules of ResAdapter, we ablate on ResCLoRA and ResENorm, as shown in Figure 5-a. Without ResCLoRA or ResENorm, the duplicated composition of generation images still exists, which demonstrate the importance of their simultaneous presence. Compared with baseline, ResAdapter can generate images without transforming the original style domain. For alpha α_r for ResAdapter, we make the ablation study on α_r, as shown in Figure 5-b. We find the quality of generation images increases as α_r from 0 to 1.

5 Conclusion

In this paper, we present a plug-and-play domain-consitent ResAdapter for diffusion models of arbitrary style domain, which enables the resolution extrapolation and interpolation of generation images. Our experiments demonstrate that after a low-cost training, ResAdapter with only 0.5M can be integrated into diffusion models to generate high-quality images of unrestricted resolutions and aspect without transforming the original style domain. Our extended experiments also demonstrate ResAdapter is compatible with other modules (e.g., ControlNet, IP-Adapter and LCM-LoRA). In addition, ResAdapter can be combined with other multi-resolution image generation models (e.g., ElasticDiffusion) to optimize inference time for higher-resolution images.

References

Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. *International conference on machine learning*, 202: 1737–1752.

Chen, Y.; Qian, S.; Tang, H.; Lai, X.; Liu, Z.; Han, S.; and Jia, J. 2024. LongLoRA: Efficient Fine-tuning of Long-Context Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Civitai. 2022. Civitai.

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

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

Gao, R.; and Liu, W. 2023. DDGR: continual learning with deep diffusion-based generative replay. In *International Conference on Machine Learning*, 10744–10763. PMLR.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Haji-Ali, M.; Balakrishnan, G.; and Ordonez, V. 2024. ElasticDiffusion: Training-free Arbitrary Size Image Generation through Global-Local Content Separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6603–6612.

Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7514–7528. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jiménez, Á. B. 2023. Mixture of diffusers for scene composition and high resolution image generation. *arXiv preprint arXiv:2302.02412*.

Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35: 26565–26577.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Lee, D.; Kim, C.; Kim, S.; Cho, M.; and Han, W.-S. 2022. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11523–11532.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11461–11471.

Luo, S.; Tan, Y.; Patil, S.; Gu, D.; von Platen, P.; Passos, A.; Huang, L.; Li, J.; and Zhao, H. 2023. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*.

Masip, S.; Rodriguez, P.; Tuytelaars, T.; and van de Ven, G. M. 2023. Continual Learning of Diffusion Models with Generative Distillation. *arXiv preprint arXiv:2311.14028*.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *International Conference on Learning Representations*.

Odena, A.; Olah, C.; and Shlens, J. 2017. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, 2642–2651. PMLR.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*.

Radford, A.; Metz, L.; and Chintala, S. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In Bengio, Y.; and LeCun, Y., eds., *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 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*, 234–241. Springer.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 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*, 22500–22510.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; 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.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Smith, J. S.; Hsu, Y.-C.; Zhang, L.; Hua, T.; Kira, Z.; Shen, Y.; and Jin, H. 2024. Continual Diffusion: Continual Customization of Text-to-Image Diffusion with C-LoRA. *Transactions on Machine Learning Research*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, S.; Chen, Z.; Zhao, Z.; Chen, Y.; Tang, Y.; and Liang, J. 2025. HiDiffusion: Unlocking Higher-Resolution Creativity and Efficiency in Pretrained Diffusion Models. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Proceedings of the European conference on computer vision (ECCV)*, 145–161.

Zheng, Q.; Guo, Y.; Deng, J.; Han, J.; Li, Y.; Xu, S.; and Xu, H. 2024. Any-size-diffusion: Toward efficient text-driven synthesis for any-size hd images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7571–7578.