# AlignedGen: Aligning Style Across Generated Images

**Jiexuan Zhang**[1*], **Yiheng Du**[1*], **Qian Wang**[1], **Weiqi Li**[1], **Yu Gu**[1], **Jian Zhang**[1,2✉]

[1]School of Electronic and Computer Engineering, Peking University
[2]Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology, Shenzhen Graduate School, Peking University

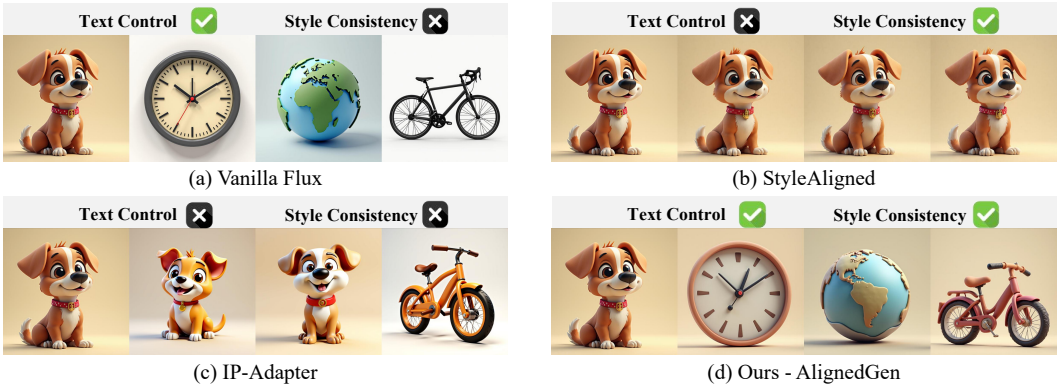

Figure 1: **Generation results of AlignedGen.** As a training-free method, AlignedGen generates images with varying content while ensuring high style consistency.

## Abstract

Diffusion-based generative models struggle to maintain high style consistency across generated images via text description. Although several style-aligned image generation methods have been proposed to address this issue, they exhibit suboptimal performance and are primarily built upon the U-Net architecture, limiting their compatibility with DiT diffusion models like Flux that has emerged as a predominant model in the field of image generation. To address these limitations, we propose ***AlignedGen***, a novel training-free style-aligned image generation method for DiT models to significantly enhance style consistency across generated images. Specifically, AlignedGen incorporates two key components to achieve this: Shifted Position Embedding (ShiftPE) and Advanced Attention Sharing (AAS). ShiftPE alleviates the text controllability degradation observed in prior methods when applied to DiT models through its non-overlapping position indices design, while AAS comprises three specialized techniques to unleash the full potential of DiT for style-aligned generation. Furthermore, to broaden the applicability of our method, we present an efficient query, key, and value feature extraction algorithm, enabling our method to seamlessly incorporate external images as style references. Extensive experimental results validate that our method effectively enhances style consistency across generated images while maintaining favorable text controllability. Code: https://github.com/Jiexuanz/AlignedGen.

[*]Equal Contribution. ✉: Corresponding author, zhangjian.sz@pku.edu.cn. This work was supported in part by National Natural Science Foundation of China (No. 62372016) and by Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology (No. 2024B1212010006).

39th Conference on Neural Information Processing Systems (NeurIPS 2025).




(a) Vanilla Flux (b) StyleAligned

(c) IP-Adapter (d) Ours - AlignedGen

Figure 2: **Generated results using different methods on Flux with the prompt "{Dog, Clock, Globe, Bicycle} in 3D realism style".** The Vanilla Flux produces outputs with style discontinuities. The application of StyleAligned [16] on Flux leads to severe text controllability degradation. The results generated by IP-Adapter [50] perform suboptimally in text control and style consistency. Our method effectively enhances style consistency while maintaining alignment with the text prompt.

## 1 Introduction

Large-scale text-to-image diffusion models [17, 36, 33, 8, 1] empower users to create imaginative scenes from text. Early diffusion models commonly adopted the U-Net [37] as their backbone. Recently, diffusion models based on DiT, such as Flux [1], have emerged as mainstream architectures and are now the most widely used frameworks in this domain. However, all diffusion models struggle to maintain high style consistency across a series of generated images via prompts, as shown in Figure 2 (a). Such style consistency is crucial for many applications: from illustrating books and stories, through designing virtual assets, to creating graphic novels and synthetic data.

Common approaches [10, 39, 9, 41] for maintaining style consistency across generated images primarily rely on fine-tuning of the diffusion model over one image or a set of images that share the same style. However, these fine-tuning-based methods suffer from significant limitations: collecting new data that share the same style and retraining diffusion model for each target style substantially increase temporal and computational costs. Alternative approaches [26, 34, 11, 49, 45, 19] focus on constructing a style encoder combined with re-trained image-conditioned diffusion models. However, creating high-quality training datasets remains challenging due to defining "style" is challenging, leading to suboptimal performance of these methods, Figure 2 (c) illustrates this.

StyleAligned [16], a training-free approach for style-aligned image generation, has emerged to address these limitations. While promising, its reliance on the aging SDXL architecture inherently limits its output quality, often producing artifacts. Most critically, its core strategy is architecturally incompatible with DiT – the superior generative architecture for diffusion models. Designed explicitly for U-Net's self-attention, it fails when applied to DiT, resulting in loss of text control and severe content collapse, as shown in Figure 2 (b). This reveals an unaddressed gap – one that our work addresses: how to achieve robust, training-free style-aligned image generation for DiT models.

To address the limitations of existing methods, we propose ***AlignedGen***, a novel framework to enhance style consistency across images generated by DiT-based diffsuion model like Flux. First, we conduct a critical investigation into a fundamental problem: the failure of attention sharing mechanisms in the DiT. We identify the root cause as an improper setup of Rotary Position Embedding (RoPE), which creates conflicting positional signals. To resolve this, we propose **Shift**ed **P**osition **E**mbedding (**ShiftPE**), a novel and elegant solution that assigns non-overlapping positional spaces to each image. This not only fixes generation collapse and restores text controllability but also offers a crucial insight for all future attention sharing works on DiT. Building upon this breakthrough, we further develop **A**dvanced **A**ttention **S**haring (**AAS**), a suite of three specialized techniques - Selective Shared Attention, Controllable Style Consistency Via Key Scaling, and Layer-Selective Application. These techniques are specifically designed for the DiT architecture, further unleashing its generative potential and leading to superior output quality. Finally, to extend our method's utility, we introduce an approximate query, key, value extraction algorithm, allowing it to condition on external images

as style references and expanding its application scope. Extensive experiments demonstrate that our method significantly outperforming prior work in both style consistency and faithfulness to text prompts. In summary, our contributions are as follows:

❑ (1) We present AlignedGen, the pioneering training-free style-aligned image generation framework designed for DiT. AlignedGen operates without any fine-tuning or extra modules, enabling the efficient utilization of large DiT models with billions of parameters.

❑ (2) We discover that conflicting position embeddings are the root cause of attention sharing failure in DiT. We introduce Shifted Position Embedding (ShiftPE), a novel and effective solution that enables viable attention sharing in DiT architecture for the first time.

❑ (3) We propose Advanced Attention Sharing (AAS), a set of three specialized techniques engineered to unleash the full potential of DiT for style-aligned generation, markedly improving style fidelity and image coherence.

❑ (4) We devise an algorithm to extract approximate query, key, and value features from any external image. This allows our framework to support user-provided images as style references without retraining, significantly broadening its applicability.

## 2 Related Work

### 2.1 Attention Control in Diffusion Models

Recent years, diffusion-based generative models [17, 30, 40, 36, 33, 8, 1] have made significant advancements in image generation. This progress has been accompanied by a series of studies focusing on attention mechanisms [15, 43, 4, 28, 25, 29, 22, 48, 23, 47, 14]. Hertz et al. [15] examined the pivotal role of cross-attention in controlling image layout and content, introducing a generation control method guided by attention maps. Plug-and-Play [43] leverages spatial features and self-attention maps from the original image to guide text-conditioned image-to-image translation, effectively preserving the original image's spatial layout. Notably, these studies all rely on the design of self-attention and cross-attention mechanisms. With the advent of Transformer-based diffusion models [32, 8, 1], self-attention and cross-attention are integrated into a single attention mechanism, presenting adaptability challenges for previous works based on separate self-attention and cross-attention designs.

### 2.2 Style Image Generation

Style image generation aims to produce images that embody specific artistic styles. Early studies in this domain concentrated on style transfer techniques [51, 12, 3, 6, 20, 24, 27, 52], which enable the transfer of the content from one image to the style of another. Generative models[13, 17, 33, 1] enable widespread adoption for style image generation. Tuning-based methods [18, 21, 39, 10, 9, 41] optimize model parameters or text embeddings to capture style of images, but suffer from training overhead. Alternatively, dataset-driven approaches [34, 11, 49, 26, 19] train diffusion models on curated datasets, yet face challenges in constructing high-quality style datasets. More pertinent to our approach is StyleAligned [16], which aligns the styles of generated images by sharing information between self-attention layers without any form of optimization or fine-tuning. In contrast to other style image generation methods, it places greater emphasis on addressing the issue of inconsistent understanding of style descriptors by generative models during image generation.

## 3 Method

### 3.1 Preliminary

**Diffusion Transformer (DiT).** DiT [32] models enhance generative performance through Transformer-based architectures, which process latent representations such as image tokens $\mathbf{X} \in \mathbb{R}^{N\times d}$ and text tokens $\mathbf{C} \in \mathbb{R}^{M\times d}$, where d is the embedding dimension and $N/M$ denote sequence lengths. In addition, these models typically encode positional information into tokens through Rotary Position Embedding (RoPE) [42], thereby embedding positional awareness into each token to achieve enhanced generation quality.

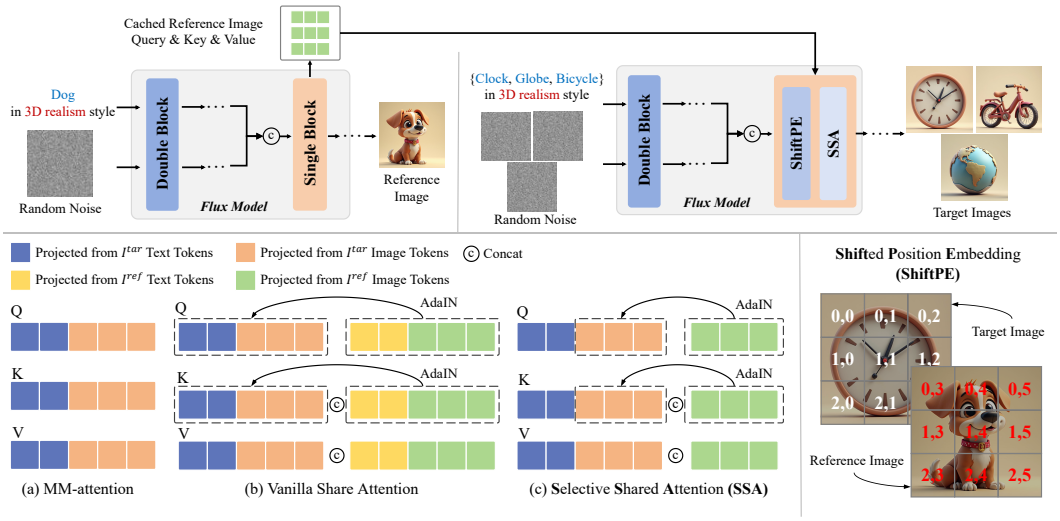



Figure 3: **Overview of the pipeline.** ShiftPE and SSA are integrated into specific layers of Flux, replacing the MM-Attention.

**Multi-Modal Attention(MM-Attention).** Some DiT models (e.g. Flux [1]) employ multi-modal attention to replace the self-attention and cross-attention mechanisms commonly used in traditional diffusion models. Specifically, image tokens $\mathbf{X}$ and text tokens $\mathbf{C}$ are projected through a set of projection matrices to obtain the corresponding representations of query $\mathbf{Q}_{\{img,txt\}}$, key $\mathbf{K}_{\{img,txt\}}$ and value $\mathbf{V}_{\{img,txt\}}$. This process is formulated as follows:

$$\begin{aligned} \{\mathbf{Q},\mathbf{K},\mathbf{V}\}_{img} &= \mathbf{W}_{img}^{\{Q,K,V\}}\mathbf{X},\ \{\mathbf{Q},\mathbf{K},\mathbf{V}\}_{img} \in \mathbb{R}^{N\times d_k} \\ \{\mathbf{Q},\mathbf{K},\mathbf{V}\}_{txt} &= \mathbf{W}_{txt}^{\{Q,K,V\}}\mathbf{C},\ \{\mathbf{Q},\mathbf{K},\mathbf{V}\}_{txt} \in \mathbb{R}^{M\times d_k} \end{aligned} \tag{1}$$

where $\mathbf{W}_{\{img,\ txt\}}^{\{Q,K,V\}}$ denote the projection matrices for the image and text tokens, $d_k$ is the embedding dimension in attention. The $\mathbf{Q}_{\{img,txt\}}$, $\mathbf{K}_{\{img,txt\}}$ and $\mathbf{V}_{\{img,txt\}}$ are concatenated along the sequence length dimension to obtain the complete $\mathbf{Q} \in \mathbb{R}^{(M+N)\times d_k}$, $\mathbf{K} \in \mathbb{R}^{(M+N)\times d_k}$, and $\mathbf{V} \in \mathbb{R}^{(M+N)\times d_k}$:

$$\mathbf{Q} = \text{Concat}(\mathbf{Q}_{txt}, \mathbf{Q}_{img}),\ \mathbf{K} = \text{Concat}(\mathbf{K}_{txt}, \mathbf{K}_{img}),\ \mathbf{V} = \text{Concat}(\mathbf{V}_{txt}, \mathbf{V}_{img}). \tag{2}$$

After that, the final output is computed using the following formula:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\mathbf{V}. \tag{3}$$

### 3.2 Task Formulation

In this paper, our objective is to generate a set of images $\mathcal{I} = \{I_1, I_2, ..., I_N\}$ that align with an input text prompts set $\mathcal{T} = \{T_1, T_2, ..., T_N\}$, while maintaining consistent style across all generated images. Notably, each prompt in the input text prompts set $\mathcal{T}$ incorporates a shared style descriptor. To facilitate the achievement of this objective, we designate one image in the image set $\mathcal{I}$ as the reference image $I^{ref}$, while the remaining images are defined as target images $I^{tar}$. The style of all target images $I^{tar}$ are enforced to approximate the style of the reference image $I^{ref}$. Furthermore, the query, key, and value representations of the reference image $I^{ref}$ throughout the generation process are denoted as $\mathbf{Q}^{ref}$, $\mathbf{K}^{ref}$, and $\mathbf{V}^{ref}$, respectively. The query, key, and value representations of the target images $I^{tar}$ are denoted as $\mathbf{Q}^{tar}$, $\mathbf{K}^{tar}$, and $\mathbf{V}^{tar}$.

### 3.3 Vanilla Attention Sharing

Existing style-aligned image generation technique[16] based on U-Net [37] architecture and self-attention can be directly applied to DiT diffusion models. For clarity, we refer to this approach as

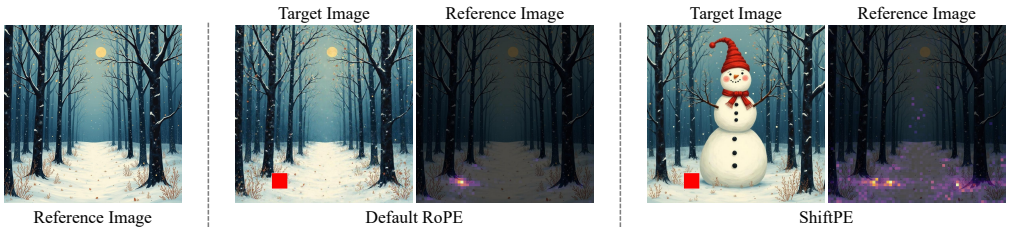



Figure 4: **Visualization of attention map.** We visualize the attention map corresponding to the query feature indicated by the red box on the generated image $I^{tar}$. When using RoPE, $I^{tar}$ over-attends to spatially aligned regions in $I^{ref}$, leading to high similarity between generated and reference images. ShiftPE introduces shifted positional indices for $I^{ref}$, enabling more appropriate attention to semantically corresponding regions (e.g., attention weights spread around the snowy area periphery).

**Vanilla Attention Sharing**. Specifically, for the query, the vanilla attention sharing aligns $\mathbf{Q}^{tar}$ with $\mathbf{Q}^{ref}$ through Adaptive Instance Normalization (AdaIN)[20] to form the final query $\mathbf{Q}^F$. For the key, $\mathbf{K}^{tar}$ is aligned with $\mathbf{K}^{ref}$ through AdaIN to produce the transformed key $\hat{\mathbf{K}}^{tar}$. After which, $\hat{\mathbf{K}}^{tar}$ is concatenated with $\mathbf{K}^{ref}$ along the sequence length dimension to form the final key $\mathbf{K}^F$. For the value, vanilla attention sharing directly concatenates $\mathbf{V}^{tar}$ and $\mathbf{V}^{ref}$ along the sequence length dimension to construct the ultimate value $\mathbf{V}^F$. This process is formulated as follows:

$$
\begin{aligned}
&\text{AdaIN}(x, y) = \sigma(y)\left(\frac{x-\mu(x)}{\sigma(x)}\right) + \mu(y), \\
&\mathbf{Q}^F = \text{AdaIN}(\mathbf{Q}^{tar}, \mathbf{Q}^{ref}) \in \mathbb{R}^{(M+N)\times d_k}, \\
&\mathbf{K}^F = \text{Concat}(\hat{\mathbf{K}}^{tar}, \mathbf{K}^{ref}) \in \mathbb{R}^{2(M+N)\times d_k}, \text{ where } \hat{\mathbf{K}}^{tar} = \text{AdaIN}(\mathbf{K}^{tar}, \mathbf{K}^{ref}), \\
&\mathbf{V}^F = \text{Concat}(\mathbf{V}^{tar}, \mathbf{V}^{ref}) \in \mathbb{R}^{2(M+N)\times d_k},
\end{aligned}
\tag{4}
$$

with $\mu(\cdot)$ and $\sigma(\cdot)$ representing the mean and standard deviation. In the target image $I^{tar}$ generation process, the query, key, and value in MM-Attention are replaced by $\mathbf{Q}^F$, $\mathbf{K}^F$, and $\mathbf{V}^F$, respectively, with the results computed according to Equation 3.

### 3.4 The limitation of Vanilla Attention Sharing in DiT

Applying Vanilla Attention Sharing to DiT, while seemingly a straightforward strategy, results in a notable collapse in performance. As illustrated in Figure 2 (b), this direct approach severely degrades text controllability and causes prominent content leakage from the reference image, leading to a catastrophic failure mode where all generated images become mere copies of the reference. This raises a critical question: **Why does Vanilla Attention Sharing fail so markedly within the DiT**?

To answer this, we conduct a systematic investigation into the interplay between attention sharing mechanism and the DiT. Our investigation reveals that the root cause is not a flaw in the attention sharing mechanism per se, but a subtle yet critical issue we term **Positional Collision**. While RoPE provides unique spatial coordinates $(i, j)$ for each token within a single image's $h \times w$ latent grid, the vanilla sharing scheme inadvertently assigns identical positional embeddings to tokens at the same coordinates $(i, j)$ in both the reference and target images. This collision induces a erroneous bias, forcing tokens to disproportionately attend to their spatial counterparts across images, irrespective of content or textual guidance, as illustrated in Figure 4 (RoPE). To resolve this, we introduce a simple yet effective solution derived from this insight: **Shifted P**osition **E**mbedding (**ShiftPE**). The core idea is to ensure the reference and target images occupy distinct, non-overlapping coordinate spaces. We achieve this by virtually placing the reference image "next to" the target images, remapping its positional indices from $(i, j)$ to $(i, j + w)$, a visual depiction is provided in Figure 3 (ShiftPE). As shown in Figure 4 (ShiftPE), ShiftPE resolves the positional collision, decoupling the spatial representations and thereby eliminating the content leakage. This frees the attention mechanism to focus on regions of high semantic and stylistic relevance, rather than being constrained by rigid spatial correspondence.

### 3.5 Advanced Attention Sharing

While ShiftPE effectively resolves the problem of positional collision in Vanilla Attention Sharing, unlocking state-of-the-art performance demands a solution that works well with the specifics of

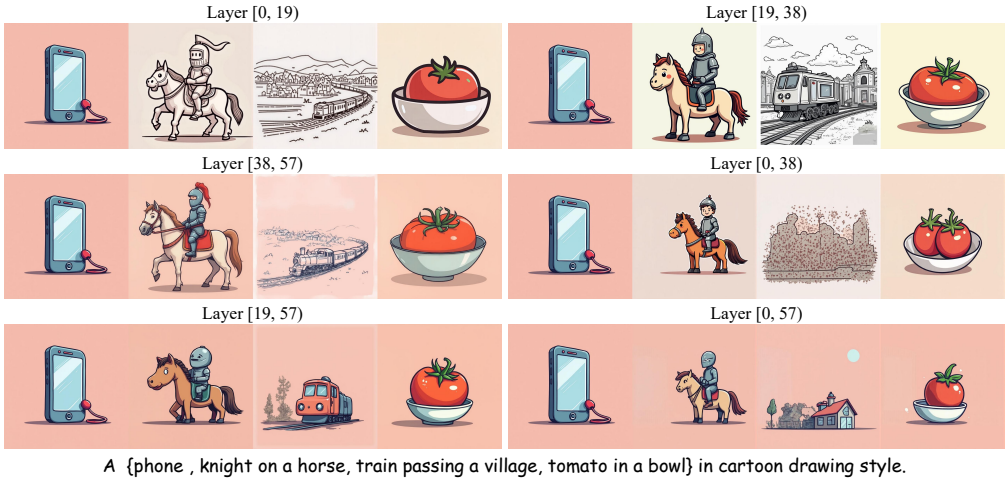



Figure 5: Visualization of results generated by applying our method in different layers of Flux.


Figure 6: **Generalization from Arbitrary Style Reference.** Our method successfully generate diverse images conditioned on external, user-provided style references.

the DiT architecture. The DiT components are not inherently designed for the attention sharing. Therefore, we introduce a series of advanced modifications that enable smooth, effective cooperation with DiT's structure, significantly enhancing style consistency and generation quality.

**Selective Shared Attention.** Vanilla Attention Sharing shares the entire key and value, but it is noteworthy that in MM-Attention, the key and value are derived from both image and text tokens. This implies that full attention sharing of key and value not only propagates information from reference image but also inadvertently shares information from its corresponding prompt, which is apparently unreasonable. To address this, we propose an intuitive yet effective solution: selectively sharing key and value derived from image tokens, shown in Figure 3 This strategy not only effectively enhances the alignment between target images and their corresponding prompts but also reduces computational overhead. Finally, the $\mathbf{Q}^F$, $\mathbf{K}^F$ and $\mathbf{V}^F$ can be expressed as follow:

$$\begin{aligned}
\mathbf{Q}^F &= \text{Concat}(\mathbf{Q}^{tar}_{txt},\ \hat{\mathbf{Q}}^{tar}_{img}) \in \mathbb{R}^{(M+N)\times d_k},\ \text{where } \hat{\mathbf{Q}}^{tar}_{img} = \text{AdaIN}(\mathbf{Q}^{tar}_{img}, \mathbf{Q}^{ref}_{img}),\\
\mathbf{K}^F &= \text{Concat}(\mathbf{K}^{tar}_{txt},\ \hat{\mathbf{K}}^{tar}_{img}, \mathbf{K}^{ref}_{img}) \in \mathbb{R}^{(M+2N)\times d_k}, \text{where } \hat{\mathbf{K}}^{tar}_{img} = \text{AdaIN}(\mathbf{K}^{tar}_{img}, \mathbf{K}^{ref}_{img}),\\
\mathbf{V}^F &= \text{Concat}(\mathbf{V}^{tar}_{txt},\ \mathbf{V}^{tar}_{img},\ \mathbf{V}^{ref}_{img}) \in \mathbb{R}^{(M+2N)\times d_k}.
\end{aligned} \tag{5}$$

**Controllable Style Consistency Via Key Scaling.** In practical applications, controlling the strength of style consistency between generated images is often required. To achieve flexible control over style consistency, we introduce a scaling factor $\lambda$. Specifically, we scale $\mathbf{K}^{ref}_{img}$ by the scaling factor $\lambda$ and concatenate it with $\mathbf{K}^{tar}_{txt}$ and $\hat{\mathbf{K}}^{tar}_{img}$ to form the modified $\mathbf{K}^F$:

$$\mathbf{K}^F = \text{Concat}(\mathbf{K}^{tar}_{txt}, \hat{\mathbf{K}}^{tar}_{img}, \lambda \cdot \mathbf{K}^{ref}_{img}). \tag{6}$$

This design allows us to adjust the attention weights by varying $\lambda$. As $\lambda$ increases, the consistency of style between the generated images becomes stronger.

**Layer-Selective Application.** The experimental results show that applying our method across all MM-Attention modules can lead to unnatural generations and textual inconsistencies in the output.

| Metric | Ours | StyleAligned [16] | IP-Adapter [50] | B-Lora [9] | StyleShot [11] | CSGO [49] |
|---|---|---|---|---|---|---|
| $S_{text} \uparrow$ | **0.282** | 0.277 | 0.278 | 0.259 | 0.276 | <u>0.280</u> |
| $S_{sty} \uparrow$ | **0.740** | 0.728 | <u>0.737</u> | 0.625 | 0.693 | 0.686 |
| $S_{dino} \uparrow$ | **0.554** | 0.512 | <u>0.522</u> | 0.271 | 0.511 | 0.450 |

Table 1: **Quantitative comparison with other methods.** We evaluate the generated image sets in terms of text alignment ($S_{text}$), style consistency ($S_{sty}$ and $S_{dino}$). The symbols $\uparrow$ indicate higher values are better. Best result is marked in **bold**, and the second-best results are marked in <u>underline</u>.

Inspired by prior studies [43, 7, 45, 9, 34], different layers in diffusion models capture and process distinct information. Following a similar strategy, we achieve improved performance by applying our method only within a specific subset of layers $\phi$. Specifically, setting $\phi$ to [19, 57) (the Single Blocks of Flux) yields optimal results. As shown in Figure 5, this approach not only enhances the naturalness of generated images but also maintains consistency with prompts and preserves stylistic coherence across image collections. Further experimental details can be found in Section 4.3.

### 3.6 Seamlessly Adapt to Any Style Reference

A core principle of our method is the injection of style information via attention sharing, which is predicated on access to the $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ features from the style reference image. While these features are directly available if the reference is synthesized by the diffusion model itself, this assumption severely limits the method's applicability in real-world cases where users provide their own images. To bridge this critical gap, we propose a novel and simple feature extraction pipeline for arbitrary external images. This pipeline achieves this by directly leveraging the forward diffusion process. For any given timestep t, we first add a noise to the reference image's latent representation. This procedure yields a noisy latent, which emulates the intermediate state during the standard generation process. By feeding this noisy latent into the diffusion model, we can query its intermediate layers to obtain the $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ features. We showcase qualitative results in Figure 6 and provide pseudocode in the Appendix A.

## 4 Experiment

### 4.1 Implementation Detail

**Settings.** We apply our method on the Flux.1-dev [1] by replacing the MM-Attention in all Single blocks of Flux.1-dev with ShiftPE and AAS. For inference, we use the vanilla Rectified Flow sampler with 30 sampling steps and set the classifier-free guidance scale to 3.5. Our evaluation dataset consists of 100 prompt sets from StyleAligned[16]. These prompts cover a wide range of generation targets and a diverse set of style descriptions.

**Metrics. (1) Text Alignment ($S_{text}$).** To evaluate how well the images match the prompts, we calculate the cosine similarity using CLIP [35] between the images and prompts. **(2) Style Consistency ($S_{sty}$).** To evaluate the style consistency across generated images, we calculate the pairwise average cosine similarity of the CLIP embeddings within the set of generated images. **(3) DINO ($S_{dino}$).** Following [44, 39, 16], we assess style consistency by computing the pairwise average cosine similarity of DINO [5, 31] embeddings in the generated image set. This choice is motivated by the fact that CLIP, trained with category labels, often rates images with similar content but different styles as highly similar, DINO is trained in a self-supervised manner, better differentiates styles.

### 4.2 Comparison

**Competing Methods.** We compare the proposed AlignedGen against five style-aligned image generation methods: StyleAligned [16], IP-Adapter [50] (we adopt the IP-Adapter architecture and weights designed by the Instant-X team [2] for Flux.1-dev, rather than utilizing the IP-Adapter architecture based on SDXL or SD1.5), B-Lora [9], StyleShot [11], CSGO [49].

**Quantitative Comparison.** Table 1 presents the quantitative comparison results between our method and state-of-the-art style-aligned image generation approaches on the test set. Our method


Figure 7: **Qualitative comparison between our method and other approaches.** Our method generates images with superior style consistency and accurate alignment with text prompts.

| Ours | StyleAligned [16] | IP-Adapter [50] | B-Lora [9] | StyleShot [11] | CSGO [49] |
|---|---|---|---|---|---|
| **1.20** | 3.66 | 2.04 | 4.76 | 4.09 | 5.25 |

Table 2: **User Study.** The results represent the average ranking outcomes for each method (lower is better). Our approach significantly outperformed other comparative methods in the user study.

achieves the highest scores in terms of $S_{text}$, $S_{sty}$, and $S_{dino}$ metrics, demonstrating that it establishes the optimal balance between text-image alignment and style consistency among generated images.

**Qualitative Comparison.** Figure 7 presents a qualitative comparison between our method and other approaches. IP-Adapter struggles to capture style effectively, failing to maintain consistent style in generated images and introducing content leakage issues. B-Lora, due to training its LoRA with only a single image, results in highly challenging training, consequently leading to generated images that struggle to capture the style of the reference image. StyleAligned demonstrates improved style consistency, but the visual quality of its generated images still falls short compared to our method. StyleShot and CSGO frequently exhibit issues of image duplication and visual artifacts. In contrast, our method produces images that not only exhibit superior style consistency across generated images but also perfectly align with the given prompts.

**User Study.** We conducted a user study to assess the results of our method compared to other methods. In each question, participants were asked to rank the methods according to the style consistency among their generated images and the alignment of these images with the prompts. In total, our user study involved 40 participants, and Table 2 shows the result. Our method significantly outperforms competing approaches, achieving an average ranking of 1.20, which demonstrates its superior performance.

### 4.3 Ablation Study

**Effect of Shifted Position Embedding (ShiftPE).** The experimental results in (k) and (m) of Table 3 demonstrate the impact of ShiftPE. When default RoPE used, the $S_{sty}$ and $S_{dino}$ metric rises to nearly 1, revealing a severe content leakage problem. Besides, Figure 8 illustrates that default

| Type | # | PE | $\lambda$ | 0-19 | 19-38 | 38-57 | $S_{text}\uparrow$ | $S_{sty}\uparrow$ | $S_{dino}\uparrow$ |
|---|---|---|---|---|---|---|---|---|---|
| Original | - | - | - | - | - | - | 0.284 | 0.658 | 0.313 |
| Layer | (a) | ShiftPE | 1.10 | ✓ | × | × | 0.283 | 0.669 | 0.323 |
| | (b) | ShiftPE | 1.10 | × | ✓ | × | 0.283 | 0.697 | 0.439 |
| | (c) | ShiftPE | 1.10 | × | × | ✓ | 0.286 | 0.682 | 0.368 |
| | (d) | ShiftPE | 1.10 | ✓ | ✓ | × | 0.274 | 0.697 | 0.420 |
| | (e) | ShiftPE | 1.10 | ✓ | ✓ | ✓ | 0.273 | 0.735 | 0.531 |
| $\lambda$ | (f) | ShiftPE | 0.90 | × | ✓ | ✓ | 0.286 | 0.694 | 0.435 |
| | (g) | ShiftPE | 0.95 | × | ✓ | ✓ | 0.285 | 0.703 | 0.460 |
| | (h) | ShiftPE | 1.00 | × | ✓ | ✓ | 0.286 | 0.711 | 0.485 |
| | (i) | ShiftPE | 1.05 | × | ✓ | ✓ | 0.285 | 0.725 | 0.522 |
| | (j) | ShiftPE | 1.15 | × | ✓ | ✓ | 0.274 | 0.755 | 0.561 |
| w/o ShiftPE | (k) | RoPE | 1.10 | × | ✓ | ✓ | 0.202 | 0.977 | 0.972 |
| w/o SSA | (l) | ShiftPE | 1.10 | × | ✓ | ✓ | 0.281 | 0.733 | 0.525 |
| **Ours** | (m) | ShiftPE | 1.10 | × | ✓ | ✓ | 0.282 | 0.740 | 0.554 |

Table 3: **Ablation study.** The Layer and $\lambda$ denote the ablation study for Layer-Selective Application and scaling factor $\lambda$, respectively. The "w/o ShiftPE" and "w/o SSA" represent the ablation study for the ShiftPE and the SSA module.

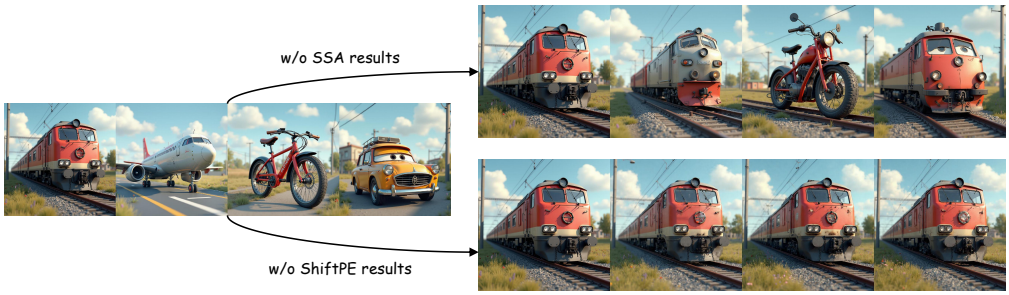



Figure 8: **Visualization of Ablation Study.** The left part presents the result generated by our method. The upper-right image shows the result without the SSA, while the lower-right image displays the result without ShiftPE. The prompt used is "{Train, Airplane, Bicycle, Car} in 3D realism style".


Figure 9: **The impact of scaling factor.** As $\lambda$ increases, the target image progressively aligns with the style of the reference image.

RoPE cause the four generated images to become nearly identical, whereas ShiftedPE maintain style consistency while ensuring alignment with the respective prompts.

**Effect of Selective Shared Attention (SSA).** Table 3 (l) and (m) show results without and with the SSA. Omitting the SSA reduces $S_{sty}$ and $S_{dino}$ but does not improve $S_{text}$. Figure 8 displays outputs under both configurations. Without the SSA, the second image mismatch its textual descriptions (airplane), and style similarity decreases.

**Effect of Layer-Selective Application.** We divided the blocks in Flux.1-dev into three groups: Group 1 contains 19 Double Blocks, Group 2 has the first 19 Single Blocks, and Group 3 includes the remaining 19 Single Blocks. The Layer part in Table 3 shows quantitative results of applying our method in different blocks. Comparing settings (a), (b), and (c) shows that Group 2 blocks contain the richest style information. Comparing settings (b) and (d) reveals that using our method in Group 1 blocks does not improve style consistency and even has a negative effect, which is also confirmed by comparing settings (e) and (m). Comparing settings (d), (e), and (m) demonstrates that applying our


Figure 10: Generation results of applying AlignedGen to other models. (Top) SD3, (Bottom) SD3.5.

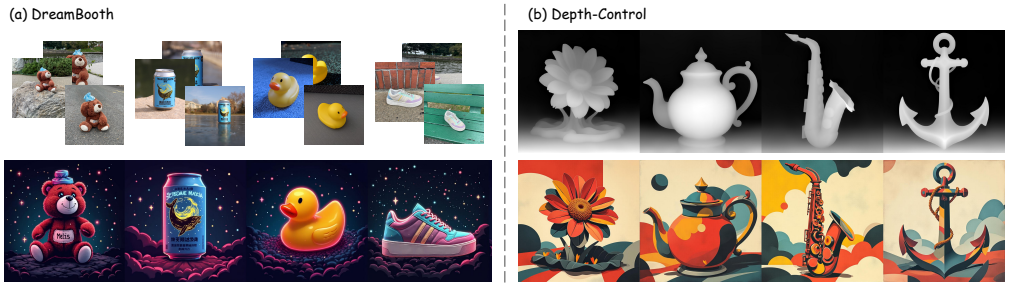



Figure 11: Results from combining AlignedGen with other controllable generation technologies.

method in the last two groups achieves a good balance between image-to-prompt alignment and style consistency across images. Finally, we apply our method exclusively in the Single Blocks (Group 2 and Group 3) to enhance style consistency among generated images.

**Effect of scaling factor.** Table 3 $\lambda$ part presents the quantitative results of our method for various values of $\lambda$. Specifically, as $\lambda$ increases, the style consistency ($S_{sty}$, $S_{dino}$) improves, but this comes at the drop of $S_{text}$. Our experiments indicate that, generally, setting the strength parameter $\lambda$ to 1.1 strikes a favorable balance between style consistency and text alignment. Figure 9 illustrates how the target image progressively aligns with the style of the reference image as $\lambda$ increases.

### 4.4 Generalization and Compatibility

**Architectural Generalization.** To demonstrate our method is not tailored to Flux, we test its performance on other DiT models, including SD3 and SD3.5. As shown in Figure 10, our method seamlessly integrates with these backbones. More results can be found in Appendix C.

**Compatibility with Existing Tools.** A key advantage of our method is its training-free, plug-and-play design. This allows for effortless composition with other generation techniques. As shown in Figure 11, the generated results from combining our approach with other techniques demonstrate the flexibility of our method in practical applications. More cases can be found in Appendix C.

## 5 Conclusion

In this paper, we present AlignedGen, a training-free framework for style-aligned image generation with Diffusion Transformer (DiT). Our approach is built on a critical discovery: the failure of vanilla attention sharing in DiT stems from conflicting position embeddings between reference and target images. We propose Shifted Position Embedding (ShiftPE) to overcome this limitation, unlocking the potential of attention sharing within DiT. Based on this foundation, we further introduce two key enhancements. First, Advanced Attention Sharing (AAS) techniques are specifically designed to refine the attention sharing mechanism for DiT architecture, markedly improving output quality. Second, the feature extraction pipeline broadens the framework's practical applicability, allowing users to provide external image as a style reference. We believe our findings on position embedding and attention sharing could provide insight for future research in style-aligned image generation.

## References


[1] Flux. https://github.com/black-forest-labs/flux/.

[2] Ip-adapter-flux. https://huggingface.co/InstantX/FLUX.1-dev-IP-Adapter.

[3] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. Artflow: Unbiased image style transfer via reversible neural flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 862–871, 2021.

[4] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22560–22570, 2023.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[6] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11326–11336, 2022.

[7] Ganggui Ding, Canyu Zhao, Wen Wang, Zhen Yang, Zide Liu, Hao Chen, and Chunhua Shen. Freecustom: Tuning-free customized image generation for multi-concept composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9089–9098, 2024.

[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.

[9] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora. In *European Conference on Computer Vision*, pages 181–198. Springer, 2024.

[10] 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.

[11] Junyao Gao, Yanchen Liu, Yanan Sun, Yinhao Tang, Yanhong Zeng, Kai Chen, and Cairong Zhao. Styleshot: A snapshot on any style. *arXiv preprint arXiv:2407.01414*, 2024.

[12] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[14] Yu Gu, Jiexuan Zhang, Qian Wang, and Jian Zhang. Garment de-warping for virtual try-on in the wild. In *2025 IEEE International Conference on Image Processing (ICIP)*, pages 875–880. IEEE, 2025.

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

[16] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024.

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

[18] 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.

[19] Nisha Huang, Kaer Huang, Yifan Pu, Jiangshan Wang, Jie Guo, Yiqiang Yan, and Xiu Li. Artcrafter: Text-image aligning style transfer via embedding reframing. *arXiv preprint arXiv:2501.02064*, 2025.

[20] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

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

[22] Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. Q-insight: Understanding image quality via visual reinforcement learning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[23] Weiqi Li, Shijie Zhao, Chong Mou, Xuhan Sheng, Zhenyu Zhang, Qian Wang, Junlin Li, Li Zhang, and Jian Zhang. Omnidrag: Enabling motion control for omnidirectional image-to-video generation. *arXiv preprint arXiv:2412.09623*, 2024.

[24] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017.

[25] Kuan Heng Lin, Sicheng Mo, Ben Klingher, Fangzhou Mu, and Bolei Zhou. Ctrl-x: Controlling structure and appearance for text-to-image generation without guidance. *arXiv preprint arXiv:2406.07540*, 2024.

[26] Gongye Liu, Menghan Xia, Yong Zhang, Haoxin Chen, Jinbo Xing, Yibo Wang, Xintao Wang, Yujiu Yang, and Ying Shan. Stylecrafter: Enhancing stylized text-to-video generation with style adapter. *arXiv preprint arXiv:2312.00330*, 2023.

[27] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6649–6658, 2021.

[28] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7465–7475, 2024.

[29] 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*, volume 38, pages 4296–4304, 2024.

[30] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[31] 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.

[32] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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, 2015.

[38] Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. *arXiv preprint arXiv:2410.10792*, 2024.

[39] 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.

[40] 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.

[41] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. In *European Conference on Computer Vision*, pages 422–438. Springer, 2024.

[42] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[43] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.

[44] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. p+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023.

[45] 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.

[46] Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746*, 2024.

[47] Qian Wang, Weiqi Li, Chong Mou, Xinhua Cheng, and Jian Zhang. 360dvd: Controllable panorama video generation with 360-degree video diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6913–6923, 2024.

[48] Shuyu Wang, Weiqi Li, Qian Wang, Shijie Zhao, and Jian Zhang. Mind-edit: Mllm insight-driven editing via language-vision projection. *arXiv preprint arXiv:2505.19149*, 2025.

[49] 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.

[50] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[51] Wei Zhang, Chen Cao, Shifeng Chen, Jianzhuang Liu, and Xiaoou Tang. Style transfer via image component analysis. *IEEE Transactions on multimedia*, 15(7):1594–1601, 2013.

[52] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–8, 2022.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes. Our main contributions are detailed in Section 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: We did not include theoretical results in our paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have described the details of this method and carefully described the experimental evaluation in the experimental chapter.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Once the blind review period is finished, we will open-source all codes and instructions.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have carried out a detailed narration in the implementation details in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Based on our experimental experience, the reproducibility of the experiments involved in this work is high, with results that are replicable and stable, rather than simply reporting the highest outcomes. Additionally, previous related work [16] has also not reported error bars. We thus do not run the statistical significance test.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We state this detailed information of computer resources in Section 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper have conducted a discussion of broader impacts at Section 1.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: We do not foresee any high risk for misuse of this work.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, we credited them in appropriate ways.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

# A Feature Extraction Pipeline

**Algorithm 1** Algorithm for Caching Q, K, V from a User-Provided Reference Image

**Require:** User-provided reference image $I$, number of inference timesteps $T$
**Ensure:** Cached attention Q, K, V pairs: cached_qkv

Initialize a random noise: noise $\sim \mathcal{N}(0, 1)$
latent $\leftarrow$ vae_encode$(I)$
cached_qkv $\leftarrow \{\}$ ▷ Initialize an empty dictionary
**for** $t \leftarrow T$ **down to** $0$ **do** ▷ Iterate over $T$ timesteps in reverse
    noise_input $\leftarrow \frac{t}{T} \cdot$ noise $+ (1 - \frac{t}{T}) \cdot$ latent ▷ A standard interpolation forward
    $Q_t, K_t, V_t \leftarrow$ DiT(noise_input, $t$, others) ▷ e.g., prompt embeds
    cached_qkv$[t] \leftarrow (Q_t, K_t, V_t)$
**end for**
**return** cached_qkv

# B Additional Compare With Image Editing Methods

Some zero-shot image editing methods can also be applied to style-aligned image generation tasks. We selected two Flux-based editing methods for comparison: RF-Solver [46] and RF-Edit [38]. Figure 12 presents qualitative comparison results between our method and editing methods. Image editing methods often exhibit insufficient editing strength when handling significant image edits such as modifying the main content of the image, leading to persistent retention of original content in the edited results. Table 4 presents the quantitative comparison results between our method and these two methods. As can be observed, our method achieves significantly higher scores on $S_{text}$, $S_{sty}$, and $S_{dino}$ compared to these two baseline methods. It is worth noting that zero-shot editing methods, limited by the instability of editing processes, often require meticulous parameter adjustments to achieve satisfactory results, which limits their practical applicability. In contrast, our method exhibits strong robustness and can typically achieve favorable outcomes in most scenarios without necessitating fine-tuning.

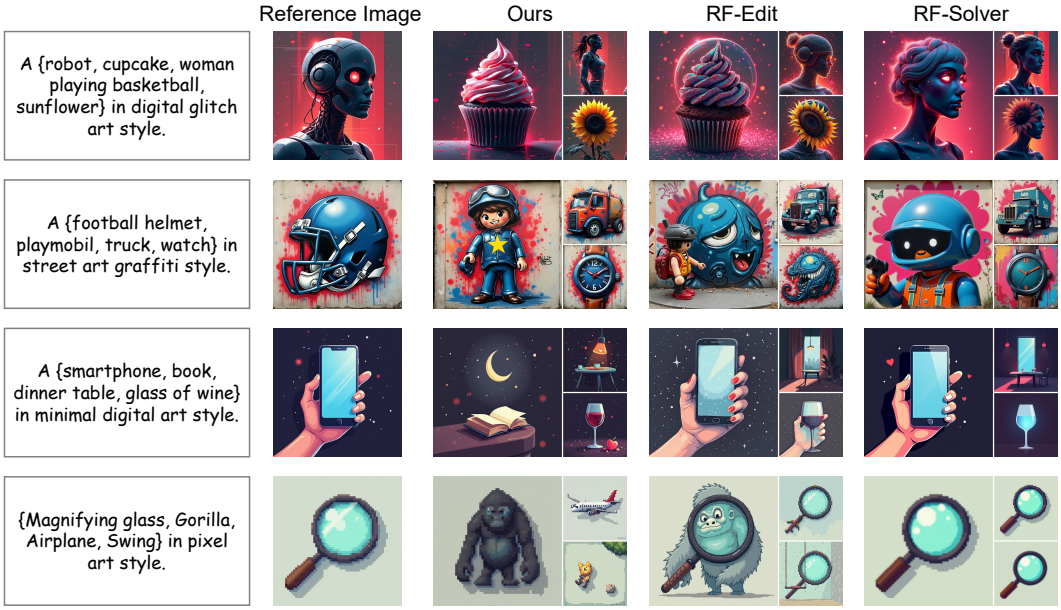



Figure 12: Qualitative comparison between our method and other zero-shot image editing methods.

| Metric | Ours | RF-Solver [46] | RF-Edit [38] |
|---|---|---|---|
| $S_{text} \uparrow$ | **0.282** | 0.278 | 0.280 |
| $S_{sty} \uparrow$ | **0.740** | 0.703 | 0.701 |
| $S_{dino} \uparrow$ | **0.554** | 0.487 | 0.456 |

Table 4: **Quantitative comparison with zero-shot image editing methods.** Best result is marked in **bold**.

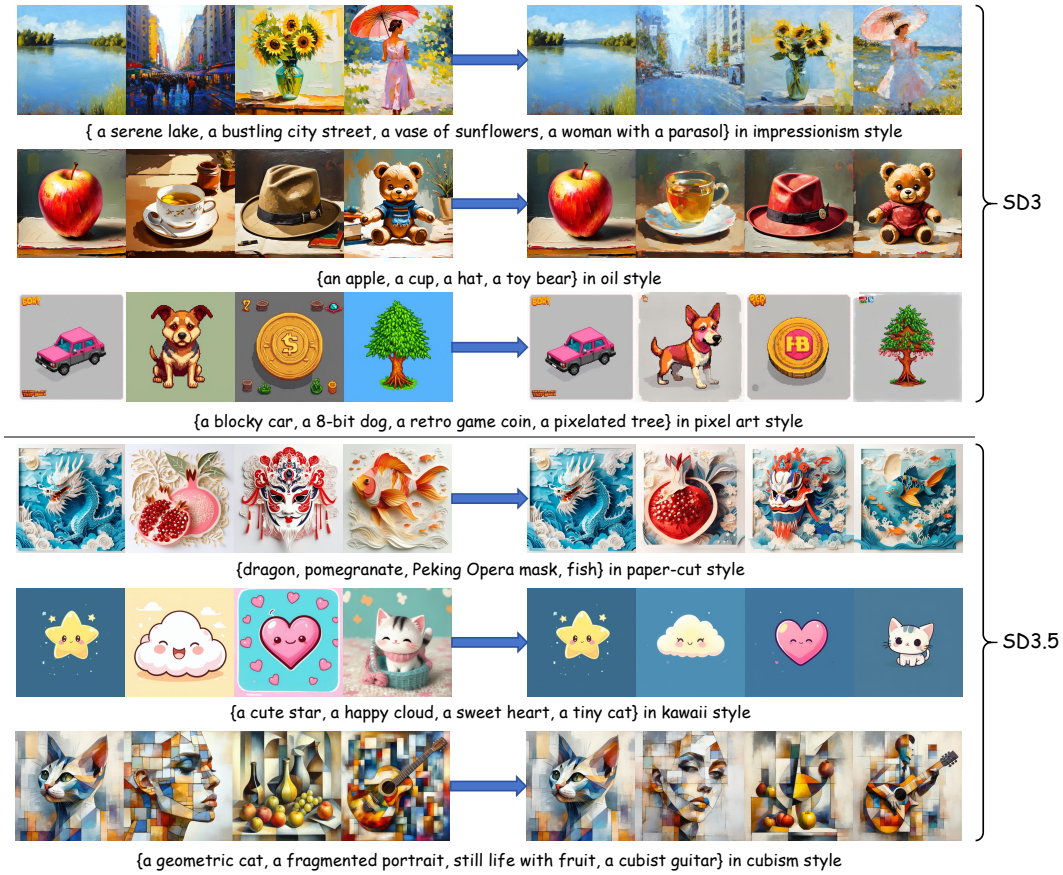



Figure 13: Generation results of applying AlignedGen to other MM-DiT models. The left side presents images generated by the original model, while the right side displays images generated after applying our method.

## C More Visual Results

Figure 13 presents the generation results of applying AlignedGen to other MM-DiT architecture diffusion models. Figure 14 demonstrates the generation outcomes from combining AlignedGen with Dreambooth. Figure 15 illustrates the generation results achieved by integrating depth control with AlignedGen.

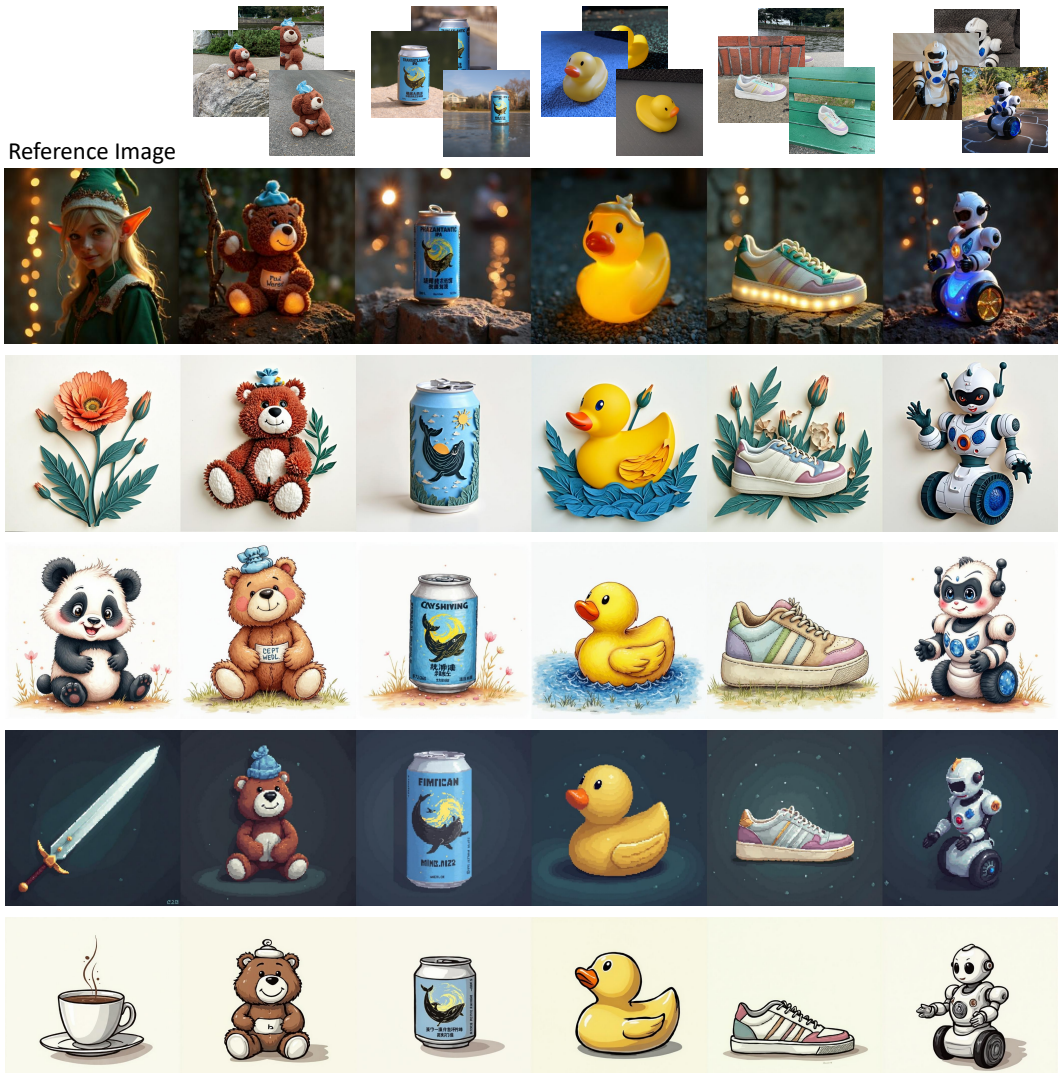



Figure 14: Subject-driven image generation with AlignedGen. Each row shows style aligned image set using the reference image on the left, applied on different personalized diffusion models, fine-tuned over the personalized content on top.

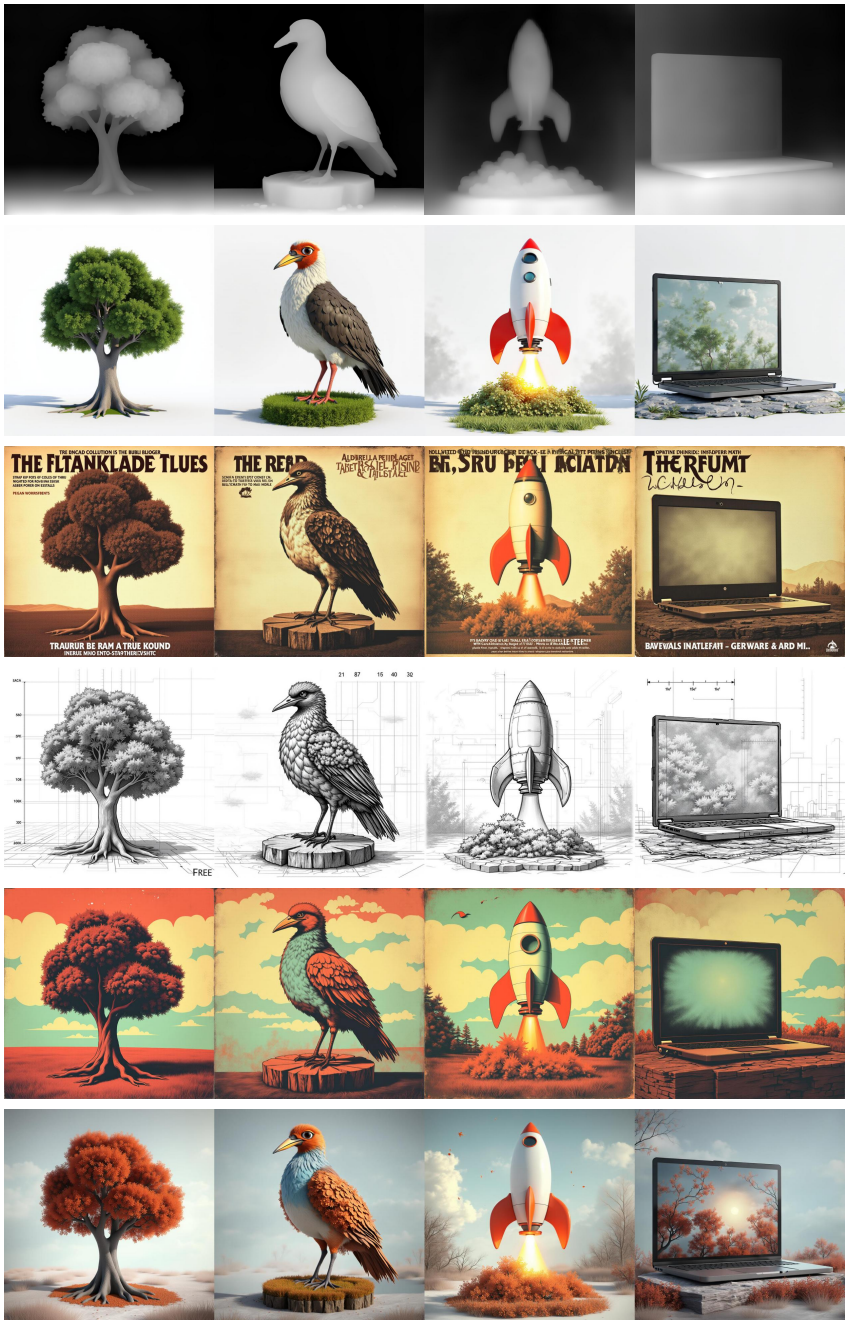


Figure 15: Depth Control with AlignedGen.