applied sciences

MDPI

Article

Dual-Direction Fine-Tuning of Pre-Trained Text-to-Image Models with Attentive Attribution Alignment

Xinyu Wang, Junsheng Luan, Wei Xing, Huaizhong Lin and Lei Zhao *

The College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China; cswxy@zju.edu.cn (X.W.); l.junsheng121@zju.edu.cn (J.L.); wxing@zju.edu.cn (W.X.); linhz@zju.edu.cn (H.L.)
* Correspondence: cszhl@zju.edu.cn

Abstract

Recent large pre-trained text-to-image models possess massive prior knowledge, which can generate high-quality images with a simple prompt. However, when we need to generate a personalized object with a specific attribution description, namely, an identity V^*, it cannot generate images that match the desired identity. Existing personalization methods for text-to-image models have the following three problems: (1) low identity fidelity, (2) low prompt fidelity and (3) low quality. To address the above issues, we propose a dual-direction fine-tuning method that guides the large pre-trained text-to-image model to generate images corresponding to the input prompt. Specifically, we input a class label into a large pre-trained text-to-image model to generate class-generalized images and caption these images using a Bootstrapping Language–Image Pre-training (BLIP) image-captioning model, obtaining class-generalized prompts. The two fine-tuning directions are ID-strengthen and ID-weaken directions. For the ID-strengthen direction, we append up-weighted V^* at the beginning of a class-generalized prompt to obtain an ID-strengthen prompt, then we use the reference images of V^* and the ID-strengthen prompt to fine-tune the pre-trained model. For the ID-weaken direction, we append down-weighted V^*, obtain an ID-weaken prompt, and fine-tune the pre-trained model. In addition, we propose an attentive attribution alignment strategy to align the semantic information of the weighted prompts to the class-generalized prompts. Qualitative and quantitative experiments show that our method improves identity fidelity and prompt fidelity while maintaining high visual quality.

Keywords: image synthesis; subject-driven generation; diffusion models

Check for updates

Academic Editor: Christos Bouras

Received: 10 May 2026
Revised: 22 June 2026
Accepted: 24 June 2026
Published: 28 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Large pre-trained text-to-image models [1–5] can generate various highly realistic scenes or objects with a simple text prompt, as they are trained on massive text–image pairs and acquire extensive prior knowledge. Specifically, GLIDE [6] proposes classifier-free textual conditions on diffusion models with a pre-trained text encoder. DALL-E 2 [3] proposes a prior that generates an image embedding from a text prompt using Contrastive Language–Image Pre-training (CLIP) [7] and a decoder that generates an image conditioned on the image embedding. CLIP is a standard pre-trained vision–language framework that maps images and text into a shared semantic embedding space; it is not a framework proposed in this work. Imagen [4] proposes a cascade of text conditional diffusion models with increasing resolutions. Stable Diffusion [2] conducts the diffusion process on the

latent domain instead of the pixel domain, thus lowering computational consumption and allowing a larger model scale.

However, the above approaches cannot meet the requirement of "personalization". Following established subject-driven personalization methods [8–13], we use V^* to denote a learnable textual identifier representing a target personalized identity. A small set of reference images provides visual supervision for learning the identity represented by V^*, while a context prompt containing V^* specifies the desired scene or attributes during generation. A personalized prompt containing V^* represents a personalized demand, like "A Siamese cat", as shown in Figure 1. We observe that large pre-trained models do not have sufficient prior knowledge corresponding to V^* due to insufficient text–image pairs used for training, so they generate images that are inconsistent with the input prompt containing V^*. Thus, several personalization methods [8,9,14] propose to embed the identity V^* into the semantic space of large pre-trained text-to-image models. Although these methods make significant progress in solving the above issue, there are still three visual challenges (see Figure 2): (1)low identity fidelity , where the generated subject fails to faithfully preserve the distinctive visual attributes of the reference identity; (2) low prompt fidelity, where the generated content does not accurately follow the semantic requirements specified by the contextual prompt; and (3) low quality, characterized by three representative perceptual defects: unrealistic colors, geometric distortions, or implausible poses. Specifically, Textual Inversion [8] uses pseudo-words to represent the identity and optimizes new embeddings for V^*; then, V^* can be injected into new scenes using simple descriptions. InST [14] inverts reference style images into the corresponding textual embeddings, then obtains personalized style images with prompts. Textual Inversion and InST show relatively low identity fidelity because their generated images preserve fewer fine-grained identity features. DreamBooth [9] fine-tunes the whole pre-trained Stable Diffusion (SD) model with identity-class pairs. Custom Diffusion [10] fine-tunes the pre-trained model on a smaller parameter set, meanwhile coherently composing multiple new identities in the same scene. DreamBooth and Custom Diffusion show strong identity preservation but limited prompt fidelity.

To overcome the three challenges, we present a novel fine-tuning method for large pre-trained text-to-image models. In general, we guide new identities to better fit the existing prior knowledge in large pre-trained models. Specifically, we input a class label into a large pre-trained SD model to generate a group of images, namely, the class-generalized images, and caption these images using the Bootstrapping Language–Image Pre-training (BLIP) image-captioning model, obtaining class-generalized prompts. We append up-weighted V^* at the beginning of a class-generalized prompt to obtain an ID-strengthen prompt; we append down-weighted V^* to obtain an ID-weaken prompt. For the ID-strengthen fine-tuning direction, ID-strengthen prompts and reference images fine-tune the pre-trained SD model; for the ID-weaken direction, ID-weaken prompts and class-generalized images fine-tune it.

The ID-strengthen prompt emphasizes the preservation and amplification of identity features, enhancing the model's ability to generate images that closely align with the specific characteristics of the target identity. By appending the up-weighted V^* to the class-generalized prompt, the fine-tuning process steers the model to prioritize the unique identity features while also maintaining the semantic coherence of the given prompt. This results in the generation of images that exhibit strong, identifiable traits, improving the model's fidelity in reproducing distinct visual characteristics of the target identity.

On the other hand, the ID-weaken prompt focuses on reducing the influence of identity-specific features by appending the down-weighted V^* to the class-generalized prompt. This direction encourages the model to emphasize broader, more generalizable features, which mitigates overfitting to any particular identity. By incorporating class-

generalized images during fine-tuning, the model demonstrates the capability to generate diverse outputs, thus enhancing its ability to create a wider range of images that are still semantically consistent with the prompt. Together, ID-strengthen and ID-weaken prompts offer complementary benefits, allowing for fine-tuning that balances specificity with versatility. In this way, critical features of identity V^* and a strong semantic prior of the prompt words are combined, encouraging the fine-tuned text-to-image SD model to generate diverse, high-quality images matching the given prompt.

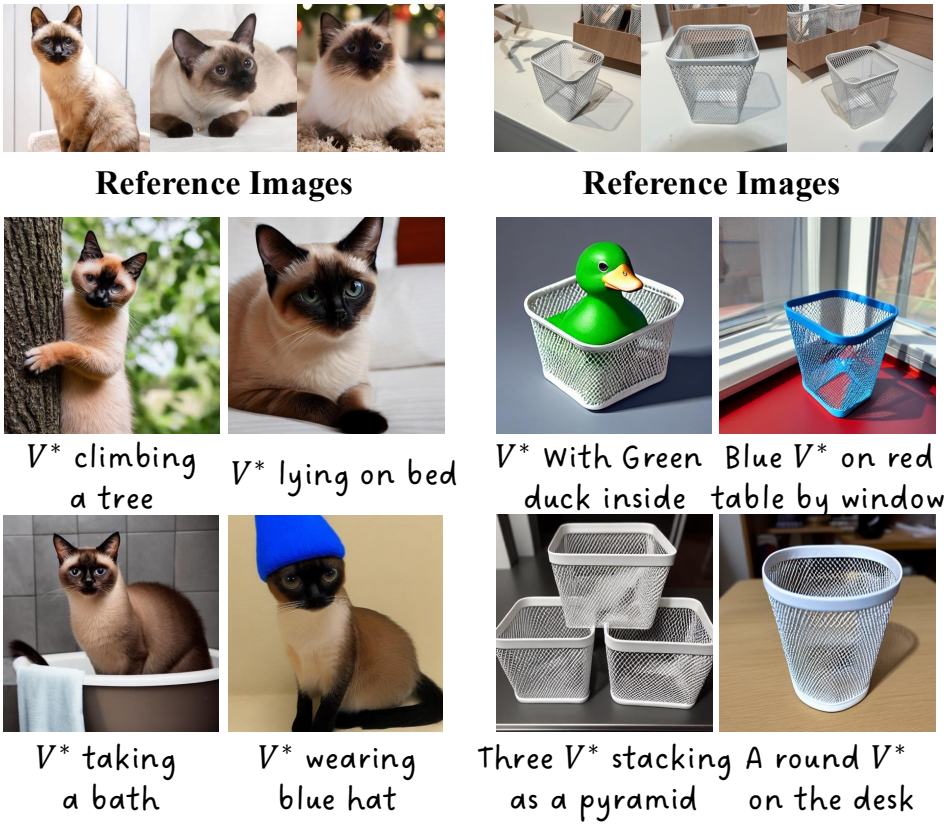


Figure 1. By fine-tuning a pre-trained text-to-image model with 3∼5 reference images and a corresponding identity V^*, our proposed method generates images that match the desired prompt and reference images.

We further propose an attentive attribution alignment strategy for the weighted prompts. In the inference process, cross-attention layers map noisy data to queries and textual condition vectors to key-value pairs [2]. The resulting cross-attention maps describe semantic relationships between prompt tokens and image pixels [15]. We notice that up-/down-weighted V^* influences the cross-attention maps and raises deviated semantic information, thus generating inconsistent images with text prompts. We also notice that the class-generalized prompts generate attention maps with correct semantic information due to the large prior knowledge of the pre-trained SD. Therefore, we augment the fine-tuning process to correct the deviated semantic information by proposing attentive attribution alignment loss. Specifically, a pre-trained SD is copied and frozen at the beginning of the training process. The class-generalized prompts use the frozen pre-trained SD to generate benchmark attention maps A_{ben}, representing the correct semantic information; the ID-strengthen prompt and ID-weaken prompt in the fine-tuning SD generate deviated attention maps A_{dev1} and A_{dev2}, representing the deviated semantic information. We propose mean square error loss to align their common context-token attention-map columns, excluding the additional identity token V^*. In this way, the identity V^* can be properly combined with the other prompt words, improving prompt fidelity in the generated images.

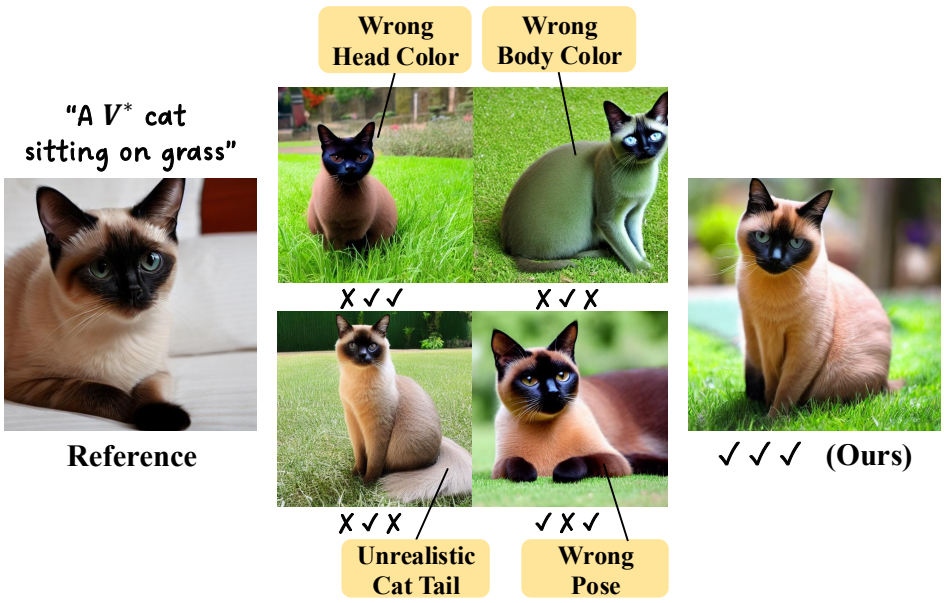


Figure 2. The three visual challenges of personalization methods: low identity fidelity, low prompt fidelity, and low quality. A tick denotes overcoming the corresponding challenge, and a cross denotes a bad performance.

In summary, our contributions are as follows:

- We propose a dual-direction fine-tuning strategy for large pre-trained text-to-image models, encouraging the models to generate diverse, high-quality images that better match the reference images of the identity V^* and the prompts.
- We introduce an attentive attribution alignment strategy to align deviated attention maps to benchmark attention maps, better combining the identity V^* with the other prompt words.
- Qualitative and quantitative evaluations show that our approach achieves stronger prompt fidelity and a better balance between identity preservation and prompt alignment compared with existing personalization methods.

2. Related Works

2.1. Text-to-Image Generation

Text-to-image generation models [1–5,16–32] recently achieve remarkable success, largely due to training on large-scale datasets. Recent journal studies further investigate text-guided latent-space generation, attention-based cross-modal interaction, and the broader development of cross-modal visual content generation [30–32]. Among these, diffusion-based models such as Stable Diffusion, DALL-E 2, and Imagen gain significant attention for their exceptional ability to generate high-quality images guided by text descriptions. However, despite their strong performance, these models face challenges in subject-driven text-to-image generation. Specifically, given an input prompt with a specific identity V^* (like an anime character), they cannot generate the appearance of V^* due to a lack of prior knowledge about V^* in the pre-trained SD models. This challenge sparks growing interest in the emerging field of personalized text-to-image generation.

2.2. Personalized Generation

Personalization needs involve personalized identities that are difficult to describe accurately in text prompts, such as anime characters. When we need to generate an image

with a personalized identity, large pre-trained text-to-image models generate images that are inconsistent with the reference images of the identity. This is because the pre-trained models lack prior knowledge of the identity. To this end, several personalization methods based on large pre-trained text-to-image models are proposed to improve personalized generation. Specifically, Textual Inversion [8] freezes the diffusion U-Net part, appends extra tokens and embeddings for a new identity, and fine-tunes the embeddings with identity features. However, it is difficult to capture the critical features of the identities. DreamBooth [9] fine-tunes a pre-trained text-to-image model such that it learns to bind a unique identity with that specific subject. Once the subject is embedded in the output domain of the model, the unique identity can be used to synthesize novel photorealistic images of the subject contextualized in different scenes. Custom Diffusion [10] fine-tunes the pre-trained model on a smaller parameter set with multiple identities. Recent journal studies further improve personalization through prior-preserved learning without image regularization, sample-aware disentangled tuning, and encoder-based domain tuning [11–13]. Although these methods make great progress in personalization and feature learning, personalized generation still faces challenges in identity fidelity, prompt fidelity, and generation quality. Our proposed method enables high-quality generation given V^* and corresponding reference images, thus solving the above issues.

2.3. Attention Regularization

Attention-based mechanisms are widely explored to control and interpret text-to-image diffusion models. Recent works show that cross-attention layers in diffusion models implicitly capture the correspondence between textual tokens and spatial regions in generated images, making them a practical handle for guiding generation behavior. For example, ref. [15] leverages diffusion attentive attribution maps to better interpret and control stable diffusion, while [33] proposes a simplified, yet more stable and efficient, tuning-free procedure that modifies only the self-attention maps of specified attention layers during the denoising process, for better image editing.

3. Method

3.1. Preliminaries

Given a text prompt P, text-to-image models generate an image z_0 by progressively denoising a noisy vector z_t. The denoising process uses a neural network ϵ_θ, which follows the objective:

$$\min_{\theta} \mathbb{E}_{z_0,\epsilon\sim N(0,I),t\sim U(1,T)}||\epsilon - \epsilon_\theta(z_t, t, \varphi(P))||_2^2 \quad (1)$$

For the Stable Diffusion (SD) backbone, z_0 is encoded from reference images I using a pre-trained autoencoder, and $\varphi(P)$ is the text embedding encoded by a pre-trained text encoder. z_t is generated from z_0 by adding random noise $z_t = \alpha_t z_0 + \beta_t \epsilon$, where α_t and β_t are the diffusion process scales. During inference, the model iteratively denoises random noise to generate an image corresponding to a prompt.

3.2. Dual-Direction Fine-Tuning Strategy with Weighted Prompt

The architecture of our dual-direction fine-tuning strategy is presented in Figure 3. The fine-tuning process inputs a class word (the general class of the specific identity V^*) and 3∼5 reference images of V^*. We first use the class label to generate a group of images in the pre-trained SD, namely, the class-generalized images. We then employ the off-the-shelf Bootstrapping Language–Image Pre-training (BLIP) model [34] to automatically convert these images into semantically descriptive captions, namely, the class-generalized prompts. These prompts provide the textual conditions required by both fine-tuning directions and the attentive attribution alignment strategy, avoiding additional manual annotation.

For example, given the class label "cat", the pre-trained SD generates a batch of class-generalized cat images. BLIP converts these images into class-generalized prompts, such as "A patterned cat lying on the ground". Then, we append up-weighted and down-weighted V^* at the beginning of a class-generalized prompt, obtaining the ID-strengthen prompt and the ID-weaken prompt. An up arrow strengthens the weight of identity V^*, and a down arrow weakens it. A weight strengthening or weakening of V^* is implemented by increasing or decreasing the numerical scale of corresponding text embeddings.

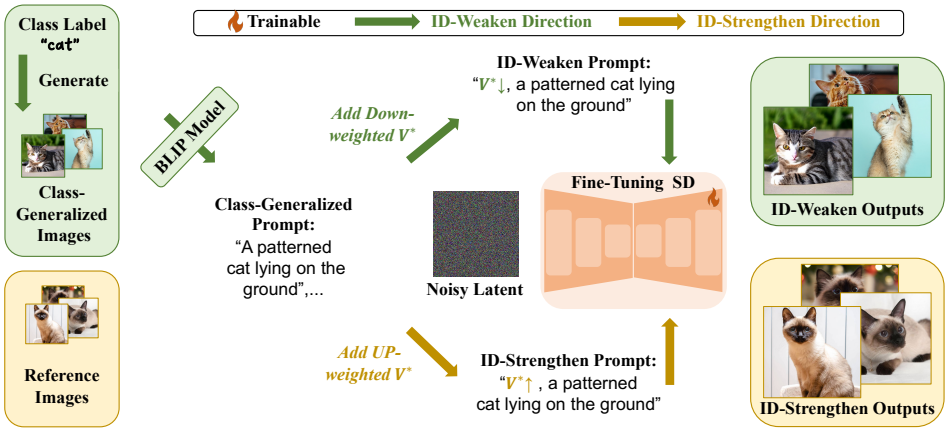


Figure 3. An overview of our proposed method. Our method fine-tunes the pre-trained text-to-image model in two directions. For the ID-strengthen direction, ID-strengthen prompts and reference images form text–image pairs to fine-tune the pre-trained text-to-image model; for the ID-weaken direction, ID-weaken prompts and class-generalized images form the text-image pair. The green and yellow arrows indicate the ID-weaken and ID-strengthen directions, respectively.

The ID-strengthen prompt emphasizes the preservation and amplification of identity features, enhancing the model's ability to generate images that closely align with the specific characteristics of the target identity. By appending the up-weighted V^* to the class-generalized prompt, the fine-tuning process steers the model to focus more on the specific identity while ensuring the semantic coherence of the prompt remains intact. As a result, the model is better equipped to produce images with distinctive and recognizable identity traits, thereby increasing its ability to accurately reproduce the visual characteristics of the target identity.

In contrast, the ID-weaken prompt seeks to reduce the influence of identity-specific features by appending the down-weighted V^* to the class-generalized prompt. This strategy encourages the model to focus on broader, more generalizable features, preventing it from overfitting to any particular identity. When class-generalized images are used during fine-tuning, the model gains the flexibility to generate a wider array of outputs while maintaining semantic alignment with the given prompt. By combining both ID-strengthen and ID-weaken prompts, we provide a balanced fine-tuning process that blends the specificity of identity features with the versatility of generalization. In this way, critical features of identity V^* and a strong semantic prior of the prompt words are combined, encouraging the fine-tuned text-to-image SD model to generate diverse, high-quality images matching the given prompt.

There are two fine-tuning directions in our proposed method. For the ID-strengthen fine-tuning direction, ID-strengthen prompts and reference images form text–image pairs that are used to fine-tune the large pre-trained text-to-image model; for the ID-weaken direction, ID-weaken prompts and class-generalized images form the pair. We use reconstruction loss to fine-tune the SD U-Net using LoRA [35].

3.3. Attentive Attribution Alignment Loss

We further propose an attentive attribution alignment strategy for the ID-strengthen and ID-weaken prompts to align the common context-token columns of the deviated attention maps A_{dev1} and A_{dev2} to those of the benchmark attention maps A_{ben}. The attention map and its query, key, and value projections are defined in Equation (2) and Equation (3), respectively.

$$A(Q,K) = softmax(\frac{QK^{\mathsf{T}}}{\sqrt{d}}) \tag{2}$$

$$\begin{aligned} Q(x_t) = W_Q \cdot \phi_\theta(x_t), K(c) = W_K \cdot \gamma_\delta(c), \\ V(c) = W_V \cdot \gamma_\delta(c) \end{aligned} \tag{3}$$

where x_t denotes noisy data at timestep t, c denotes text embedding, ϕ_θ denotes a function expressed by U-Net, γ_δ is a CLIP text encoder, and W_Q, W_K and W_V represent a trainable linear network layer, respectively. We strengthen and weaken the weight of the identity V^* in prompts, resulting in scaled key–value pairs for V^*. We notice that a weight change on identity V^* will indirectly influence the attention maps and raise deviated semantic information, thus generating inconsistent images with text prompts.

To handle this issue, we propose attentive attribution alignment loss, denoted by $\mathcal{L}_{A^3}$. The attentive attribution alignment process is shown in Figure 4. We first copy and freeze the pre-trained SD. We feed the noisy data and class-generalized prompts into the copied SD U-Net to generate the benchmark attention maps A_{ben} by calculating $A(Q,K) = softmax(\frac{QK^{\mathsf{T}}}{\sqrt{d}})$. Q denotes noisy-latent queries, and K denotes encoded prompt keys in SD U-Net layers. The benchmark attention maps serve as the benchmark of correct semantic information. During training, we feed the noisy data and weighted prompts into the fine-tuning SD U-Net to generate the deviated attention maps A_{dev1} and A_{dev2}, representing the deviated semantic information. Since the class-generalized prompt and the weighted prompts have different token sequences, because the weighted prompts include the additional identity token V^*, we do not align the full token dimension directly. Instead, we define $\mathcal{R}_c$ and $\mathcal{R}_{c_o}$ as token-selection operators that select and order the attention-map columns corresponding to the common context tokens shared by the weighted prompt c and the class-generalized prompt c_o, excluding the additional V^* token. Afterward, we align the selected common-token parts of A_{dev1} and A_{dev2} to A_{ben}. The semantic information of the identity V^* can be more effectively combined with the prompt words, thereby improving prompt fidelity. $\mathcal{L}_{A^3}$ is shown as follows:

$$\begin{aligned} \mathcal{L}_{A^3} = E_{c\sim\{c_u,c_d\},c_o,t,x_t}||\mathcal{R}_c(S(\frac{Q(x_t)K^{\mathsf{T}}(c)}{\sqrt{d}}))- \\ \mathcal{R}_{c_o}(S(\frac{Q(x_t)K^{\mathsf{T}}(c_o)}{\sqrt{d}}))||_2^2 \end{aligned} \tag{4}$$

where S represents the softmax function, c_d represents text embeddings with the down-weighted identity, and c_u represents the up-weighted ones. c_o represents the text embeddings of class-generalized prompts.

The dual-direction reconstruction loss combines critical features of identity V^* and a broad semantically related prior, encouraging the text-to-image model to generate diverse, high-quality images matching the given V^*. $\mathcal{L}_{Dual}$ is shown as follows:

$$\begin{aligned} \mathcal{L}_{Dual} = \lambda_1 E_{\epsilon\sim\mathcal{N}(0,I),c_u,t}||\epsilon_{ref} - \epsilon_\theta(x_t,t,c_u)||_2^2 + \\ \lambda_2 E_{\epsilon\sim\mathcal{N}(0,I),c_d,t}||\epsilon_{cls} - \epsilon_\theta(x_t,t,c_d)||_2^2 \end{aligned} \tag{5}$$

where ϵ_{ref} denotes an added noise corresponding to the reference images in a specific timestep, and ϵ_{cls} denotes an added noise corresponding to the class-generalized images. The overall loss function is shown as follows:

$$\mathcal{L}_{Total} = \alpha_1 \mathcal{L}_{A^3} + \alpha_2 \mathcal{L}_{Dual} \quad (6)$$

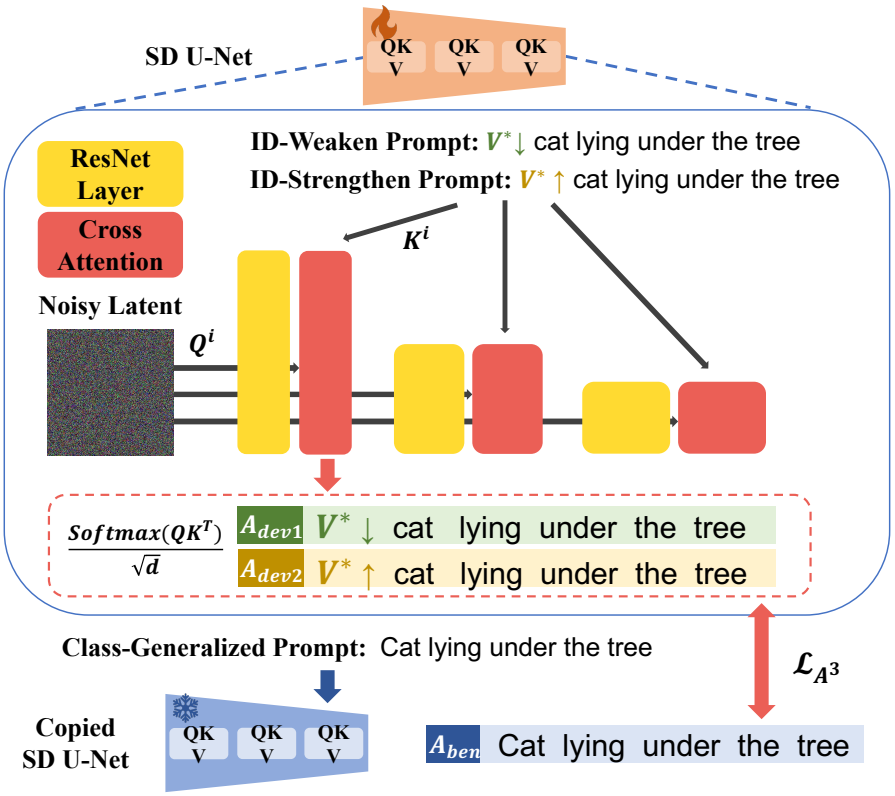


Figure 4. The process of our proposed attentive attribution alignment method. The ID-weaken and ID-strengthen prompts generate the deviated attention maps A_{dev1} and A_{dev2} in the trainable SD U-Net, while the class-generalized prompt generates the benchmark attention map A_{ben} in the copied SD U-Net. We then propose $\mathcal{L}_{A^3}$ to align their common context-token attention maps. The fire symbol denotes the trainable part and the snowflake symbol denotes the frozen part.

4. Experiments

4.1. Implementation Details

We collect 50 groups of images online from Unsplash and Danbooru, including daily objects, animals, and anime characters. Each of them consists of 3∼5 reference images representing a specific identity V^*. There are a total of 200 images in our dataset, including 20 daily-object identities with 95 images, 20 animal identities with 65 images, and 10 anime-character identities with 40 images, as summarized in Table 1. We select images in which the target identity is clearly visible, is not severely occluded, and has sufficient resolution for its distinctive visual attributes to be recognized. Duplicate images within each identity set are excluded. The real-world photographs collected from Unsplash retain natural variations in background, illumination, viewpoint, and object pose, while the anime images collected from Danbooru retain variations in stylized character appearance and pose. These collection conditions reflect realistic personalization scenarios while ensuring that each reference image provides usable identity supervision. Unlike a conventional classification dataset, each personalized identity is fine-tuned independently from its own reference set; samples from one identity are not used to optimize another identity. Therefore, category-level imbalance does not directly bias the fine-tuning loss toward a majority category. We

also report quantitative results separately for daily objects, animals, and anime characters to prevent a category containing more identities from dominating a pooled result.

Table 1. Statistics of the personalized-image dataset.

Category	Identities	Images	Source
Daily objects	20	95	Unsplash
Animals	20	65	Unsplash
Anime characters	10	40	Danbooru
Total	50	200	Unsplash and Danbooru

We use Stable Diffusion v1.5 as the pre-trained text-to-image model and fine-tune it using a single NVIDIA RTX 3090 GPU with 24 GB memory for 1000 iterations and a learning rate of 5×10^{-5}. We use LoRA [35] fine-tuning with a rank of 4. For the anime identity, we fine-tune the pre-trained PastelMix model instead. We use fp16 precision and a batch size of 1 to accelerate the fine-tuning process. The benchmarked identity takes 212.07 s end-to-end to fine-tune. The LoRA module contains 797,184 trainable parameters and the saved LoRA weights occupy 3.1 MB. For single-image 512×512 inference with 50 denoising steps using the prompt "a photo of sks backpack", image generation takes 2.79 s, and the total script time, including model and LoRA loading, is 3.92 s. All model computation is performed on the server side; the user side only submits reference images and prompts and receives generated images, without requiring a dedicated GPU or local model execution. Because client-side CPU, memory, and network usage depend on the deployment interface and user device rather than the proposed model, we report the user-side model-computation requirement as negligible instead of assigning a fixed numerical benchmark. The fine-tuning and inference resource benchmarks are reported in Table 2 and Table 3, respectively.

Table 2. Fine-tuning hardware and computational-resource benchmark of the proposed method.

Metric	User Side	Server Side
Hardware requirement	No dedicated GPU	NVIDIA RTX 3090 (24 GB)
Trainable parameters	Not applicable	797,184
LoRA weight size	Not applicable	3.1 MB
Fine-tuning time	Not applicable	212.07 s
`nvidia-smi` peak GPU memory	Not applicable	5695 MiB
Peak GPU power	Negligible client workload	347.39 W

Table 3. Inference hardware and computational-resource benchmark of the proposed method.

Metric	User Side	Server Side
Hardware requirement	No dedicated GPU	NVIDIA RTX 3090 (24 GB)
Image generation time	Not applicable	2.79 s
Total inference script time	Not applicable	3.92 s
`nvidia-smi` peak GPU memory	Not applicable	3585 MiB
Peak GPU power	Negligible client workload	349.04 W

We set $\lambda_1 = 0.2, \lambda_2 = 1, \alpha_1 = 1$, and $\alpha_2 = 5e-3$ in the fine-tuning process.

Evaluation Metrics. We evaluate subject-driven text-to-image methods using four criteria: (1) CLIP-I , (2) DINO, (3) CLIP-T, and (4) Human-Preference. CLIP-I, DINO, and CLIP-T are jointly adopted in multiple subject-driven generation studies, including DreamBooth, BLIP-Diffusion, JeDi, ELITE, SuTI, and Subject-Diffusion [9,36–40]; DreamTuner also adopts CLIP-I and CLIP-T [41]. CLIP-I and DINO are sensitive to changes in visual subject similarity and are used to estimate identity fidelity, with DINO emphasizing

fine-grained appearance and structural correspondence. CLIP-T measures semantic image–text correspondence and is used to estimate prompt fidelity. Because these embedding-based proxy metrics do not fully capture fine-grained identity details or human perception, we complement them with a randomized human-preference study. Following the DreamBooth evaluation protocol [9], we use 25 context prompts covering recontextualization, accessorization, and property modification for each identity, and generate four images for each identity–prompt pair. For our 50 identities, this produces a total of 5000 generated images for quantitative evaluation. For methods that cannot be reproducibly evaluated under the same protocol, we do not include them in the numerical tables. For all methods reported in the numerical evaluation tables, the same identities, prompts, number of generated images, and metric computation protocol are applied. No generated samples are manually removed, and the reported CLIP-I, DINO, and CLIP-T scores are averaged over all generated images and context prompts in each identity category. CLIP is employed here only as a standard pre-trained evaluation model rather than a component proposed in this work. Let I denote a generated image, $\{I_0^k\}_{k=1}^K$ denote the K reference images of the target identity, and P denote the context prompt. For CLIP-I, we use the image encoder f^I_{CLIP} of a pre-trained CLIP ViT-L/16 model to calculate the average cosine similarity between the generated image and the reference images:

$$\text{CLIP-I}(I, \{I_0^k\}) = \frac{1}{K}\sum_{k=1}^{K} \frac{f^I_{\text{CLIP}}(I)^\top f^I_{\text{CLIP}}(I_0^k)}{\|f^I_{\text{CLIP}}(I)\|_2 \|f^I_{\text{CLIP}}(I_0^k)\|_2}. \tag{7}$$

For DINO [42], we use a pre-trained ViT-S/16 DINO model, where $f_{\text{DINO}}(\cdot)$ denotes its image encoding function, and calculate the average pairwise cosine similarity:

$$\text{DINO}(I, \{I_0^k\}) = \frac{1}{K}\sum_{k=1}^{K} \frac{f_{\text{DINO}}(I)^\top f_{\text{DINO}}(I_0^k)}{\|f_{\text{DINO}}(I)\|_2 \|f_{\text{DINO}}(I_0^k)\|_2}. \tag{8}$$

DINO emphasizes fine-grained visual correspondence and is commonly used to evaluate subject fidelity [9]. For CLIP-T, we use the CLIP image encoder f^I_{CLIP} and text encoder f^T_{CLIP} to calculate the cosine similarity between the generated image and the context prompt:

$$\text{CLIP-T}(I, P) = \frac{f^I_{\text{CLIP}}(I)^\top f^T_{\text{CLIP}}(P)}{\|f^I_{\text{CLIP}}(I)\|_2 \|f^T_{\text{CLIP}}(P)\|_2}. \tag{9}$$

Higher CLIP-I and DINO scores indicate stronger identity fidelity, while a higher CLIP-T score indicates stronger context fidelity. Image quality is primarily assessed through subjective human perception because visual plausibility and aesthetic quality cannot be fully characterized by a single objective metric. For Human-Preference, we collected 500 online questionnaires. Each questionnaire contains three groups of generated images, and users answer three questions concerning identity fidelity, prompt fidelity, and overall generation quality. In total, we collected 4500 answers and report the percentage of users selecting each method; a higher percentage indicates a higher user preference.

4.2. Qualitative Comparison with Personalization Approaches

We first conduct a qualitative comparison with personalization approaches, as shown in Figure 5. The qualitative examples are selected from the same test identities and context prompts used in the quantitative evaluation. All compared methods use identical reference images and identical prompts. The displayed cases are chosen to cover the three evaluated identity categories, namely, daily objects, animals, and anime characters, and to illustrate representative successes and common failure modes observed in the full test set. These displayed examples are not used as the sole evidence for performance claims; the compara-

tive conclusions are supported by the complete quantitative evaluation over all generated samples and by the randomized human-preference study. We carry out evaluations on the three visual challenges: low identity fidelity, low prompt fidelity, and low quality. In the qualitative comparison, we mark a low-quality failure when the generated image contains visually observable defects, including unrealistic color distribution, distorted object geometry or broken structure, and implausible object pose. These marks are used as qualitative indicators, while overall perceptual quality is further evaluated through the user study. A tick denotes a good performance on the corresponding challenge, and a cross denotes a bad performance. Compared personalization methods include Textual Inversion [8], InST [14], DreamBooth [9], Custom Diffusion [10], OminiControl [43], Nano Banana 2 [44], and OmniGen [45]. The first column corresponds to prompts. Textual Inversion only optimizes the corresponding embeddings while preserving the original weights of the pre-trained model, resulting in limited identity fidelity. InST improves feature learning but still misses detailed identity features. DreamBooth and Custom Diffusion preserve identity features well, but some generated images exhibit limited prompt fidelity. OminiControl sometimes fails to preserve or generate the target subject, whereas Nano Banana 2 often produces outputs visually close to the reference image in our qualitative examples, which may limit generation diversity. OmniGen also tends to closely fit the reference image, leading to weaker text–image consistency. Although the DiT-based methods OminiControl and OmniGen benefit from stronger generative priors and generally produce high-quality images, our SD1.5-based method achieves comparable visual quality with stronger prompt fidelity.



Figure 5. The comparison with state-of-the-art methods. The first row is a detailed example, where colored boxes and the corresponding annotations indicate representative observations on identity fidelity, prompt fidelity, and visual quality. Our proposed method achieves strong identity fidelity and prompt fidelity while maintaining competitive visual quality.

4.3. Quantitative Comparison with Personalization Approaches

We compare three categories of identity: daily objects, animals, and anime characters. The critical identity features should be faithfully preserved, and the generated content should also align with the given context prompts. As formally defined in Equations (7)–(9), CLIP-I and DINO evaluate identity fidelity, while CLIP-T evaluates prompt fidelity. As shown in Table 4, OminiControl obtains relatively low CLIP-I and DINO scores, consistent with its occasional failure to preserve or generate the target subject. OmniGen obtains high CLIP-I and DINO scores because it tends to closely reproduce the visual appearance of the reference images. However, its lower CLIP-T scores indicate limited alignment with the context prompts. Our method achieves the highest DINO score for animals, remains competitive with OmniGen across all identity categories, and obtains the highest CLIP-T scores, demonstrating a better balance between identity preservation and prompt fidelity. We further perform paired Wilcoxon signed-rank tests between our method and each baseline using per-sample CLIP-I, DINO, and CLIP-T scores under the same identity-prompt-seed settings. All comparisons are significant at the uncorrected $p < 0.05$ level, supporting the reliability of the observed metric differences and quantitative trends. Nano Banana 2 is included only in the qualitative comparison because its closed-source interface does not allow us to guarantee identical batch-generation settings, random seeds, or reproducible large-scale evaluation. As an additional practical consideration, generating the complete test set through its credit-based service would incur substantial monetary cost. Therefore, we generate a limited number of samples for qualitative comparison and exclude Nano Banana 2 from quantitative metrics and user-study ranking to avoid unfair or non-reproducible comparisons.

Table 4. Quantitative comparison with state-of-the-art methods and statistical significance analysis of automatic quantitative metrics. CLIP-I and DINO represent identity fidelity, and CLIP-T represents prompt fidelity. We make the comparison on three categories of identity and conduct paired Wilcoxon signed-rank tests between our method and each baseline using per-sample scores under the same identity-prompt-seed settings. The **bold** number denotes the best one.

(a) Average Metric Scores on Three Identity Categories.

	Daily Objects			Animals			Anime Characters		
	CLIP-I	**DINO**	**CLIP-T**	**CLIP-I**	**DINO**	**CLIP-T**	**CLIP-I**	**DINO**	**CLIP-T**
Textual Inversion	0.773	0.571	0.287	0.584	0.492	0.290	0.745	0.569	0.245
InST	0.798	0.569	0.290	0.633	0.461	0.284	0.763	0.495	0.271
DreamBooth	0.903	0.683	0.301	**0.892**	0.652	0.298	0.911	0.589	0.283
Custom Diffusion	0.895	0.681	0.304	0.884	0.639	0.301	0.900	0.565	0.289
OminiControl	0.632	0.497	0.253	0.597	0.372	0.236	0.662	0.383	0.259
OmniGen	**0.916**	**0.715**	0.275	0.882	0.679	0.281	**0.919**	**0.675**	0.286
Ours	0.911	0.693	**0.307**	0.889	**0.681**	**0.310**	0.913	0.662	**0.306**

(b) Paired Wilcoxon signed-rank tests.

Comparison	CLIP-I	DINO	CLIP-T
Ours vs. Textual Inversion	$p = 0.0007$	$p = 0.0008$	$p = 0.0036$
Ours vs. InST	$p = 0.0045$	$p = 0.0036$	$p = 0.0091$
Ours vs. DreamBooth	$p = 0.0071$	$p = 0.0062$	$p = 0.0019$
Ours vs. Custom Diffusion	$p = 0.0004$	$p = 0.0070$	$p = 0.0065$
Ours vs. OminiControl	$p = 0.0070$	$p = 0.0081$	$p = 0.0249$
Ours vs. OmniGen	$p = 0.0392$	$p = 0.0417$	$p = 0.0484$

4.4. User Study

We further conducted a randomized human-preference study to evaluate subjective perceptual quality. We collected 500 online questionnaires from users. Each questionnaire

consists of three randomly selected groups of generated images from different subjects. Each image group is evaluated by five different participants to enable inter-rater reliability analysis. Each group contains seven anonymized generated images from the compared methods and has three forced-choice questions: "Which of the following seven images preserves the most details of the reference image?", "Which of the following images best corresponds to the prompt?", and "Which image do you like best?". The questionnaire therefore adopts a forced-choice nominal scale. The first two questions evaluate identity fidelity and prompt fidelity, respectively, while the final question evaluates users' overall preference regarding generation quality.

To minimize positional, method-order, and recognition bias, the seven generated images are displayed in a randomized order for each questionnaire group, and the corresponding method names are hidden from participants. The seven compared methods exclude Nano Banana 2 because its closed-source interface does not allow identical and reproducible large-scale user-study generation settings. In total, 4500 answers were collected, including 1500 answers for each question.

As shown in Table 5, the DiT-based OmniGen and OminiControl benefit from stronger generative priors than SD1.5-based methods. Ours ranks first in identity fidelity and prompt fidelity, while OmniGen ranks first in overall quality. The subject-generation failures of OminiControl limit its user preference, whereas our SD1.5-based method achieves comparable overall-quality preference.

We further conducted statistical validation of the user study. Since the questionnaire uses a forced-choice nominal scale rather than a continuous rating scale, we used the Chi-square test instead of ANOVA to evaluate whether the observed preference distribution differs significantly across methods. The Chi-square tests produce $p < 0.001$ for all three questions, indicating that the user preferences are not uniformly distributed and that the questionnaire is sensitive to method-level differences. We also report Cramer's V to measure the effect size. In addition, we computed Fleiss' κ to evaluate inter-rater reliability among the five participants assigned to each image group. The Fleiss' κ values are 0.70, 0.50, and 0.30 for identity fidelity, prompt fidelity, and overall quality, respectively, indicating substantial, moderate, and fair agreement under the forced-choice perceptual setting. We use the user study as complementary evidence rather than the sole evidence for method superiority, together with the automatic metrics and qualitative comparisons.

Table 5. Results of the user study. The questionnaire adopts a forced-choice nominal scale. The three questions correspond to identity fidelity, prompt fidelity, and overall generation quality. A higher percentage represents a higher user preference. Each question contains 1500 answers. The **bold** number denotes the best one.

Method	Identity Fidelity	Prompt Fidelity	Overall Quality
Textual Inversion	1%	6%	1%
InST	10%	14%	5%
DreamBooth	19%	15%	21%
Custom Diffusion	21%	16%	18%
OminiControl	1%	3%	12%
OmniGen	23%	18%	**25%**
Ours	**25%**	**28%**	18%
No. of answers	1500	1500	1500
Chi-square test	$p < 0.001$	$p < 0.001$	$p < 0.001$
Cramer's V	0.271	0.216	0.231
Fleiss' κ	0.70	0.50	0.30

4.5. Ablation Studies

We conducted ablation studies on the dual-direction fine-tuning and the attentive attribution alignment loss.

4.5.1. Fine-Tuning Direction Ablation

We fine-tune the large pre-trained text-to-image model with varied V^* and corresponding reference image sets with weighted fine-tuning directions. We set $\lambda_1 = 0$ for the absence of the ID-strengthen direction and $\lambda_2 = 0$ for the absence of the ID-weaken direction. The qualitative result is shown in Figure 6, with prompt "V^* on a grey sofa". Without the ID-strengthen direction, the generated object lacks the target identity-specific features, producing an ordinary glass bottle rather than the demanded one. The absence of the ID-weaken direction corresponds to direct fine-tuning on SD, resulting in overfitting. We further evaluate reduced ID-strengthen and ID-weaken directions by decreasing λ_1 and λ_2, respectively. As can be seen, a decreased λ_1 induces feature loss, and a decreased λ_2 induces content inconsistency with the given prompt "grey sofa". The quantitative results are shown in Table 6. Although $\lambda_1 = 0.2, \lambda_2 = 1$ does not achieve the highest single CLIP-I or CLIP-T score, it provides a balanced trade-off between identity fidelity and prompt fidelity, with relatively high values on both metrics. A suitable combination of the ID-strengthen and ID-weaken directions can generate diverse, high-quality images matching the given V^*.

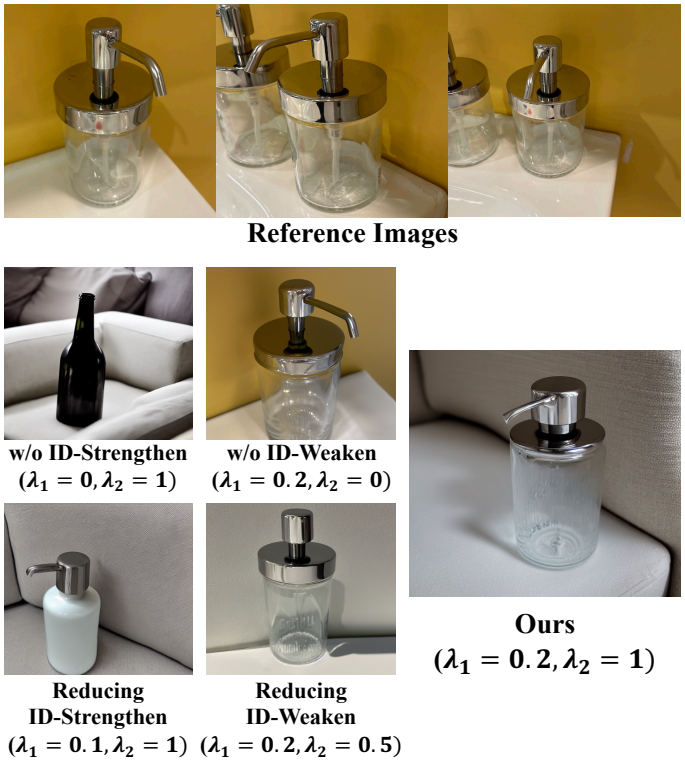


Figure 6. Qualitative ablation study on fine-tuning direction. The prompt is "V^* on a grey sofa".

Table 6. Quantitative ablation study on fine-tuning direction. A higher CLIP-T represents higher prompt fidelity, and a higher CLIP-I represents higher identity fidelity. Bold italic numbers indicate relatively high levels of CLIP-I and CLIP-T.

Combination		CLIP-I	CLIP-T
$\lambda_1 = 0$	$\lambda_2 = 0.5$	0.365	0.263
$\lambda_1 = 0$	$\lambda_2 = 1$	0.301	**0.312**
$\lambda_1 = 0.1$	$\lambda_2 = 0$	0.793	0.193
$\lambda_1 = 0.1$	$\lambda_2 = 0.5$	0.778	0.247
$\lambda_1 = 0.1$	$\lambda_2 = 1$	0.754	0.298
$\lambda_1 = 0.2$	$\lambda_2 = 0$	**0.925**	0.163
$\lambda_1 = 0.2$	$\lambda_2 = 0.5$	0.917	0.221
$\lambda_1 = 0.2$ **(Ours)**	$\lambda_2 = 1$ **(Ours)**	***0.909***	***0.295***

4.5.2. Attentive Attribution Alignment Ablation

We further conduct the ablation on different weights of attentive attribution alignment loss. The results are shown in Figure 7. The horizontal axis is CLIP-I, representing identity fidelity, and the vertical axis is CLIP-T, representing prompt fidelity. We set $\alpha_1 = 1$ in Equation (6) and modify α_2 to give the results. A smaller α_2 leads to better prompt fidelity but lower identity fidelity, while a larger α_2 increases identity fidelity but decreases prompt fidelity. Therefore, $\alpha_2 = 5 \times 10^{-3}$ provides a suitable trade-off between identity fidelity and prompt fidelity. With a suitable attentive attribution alignment loss, the new identity can be properly combined with prompts, leading to stronger prompt fidelity.

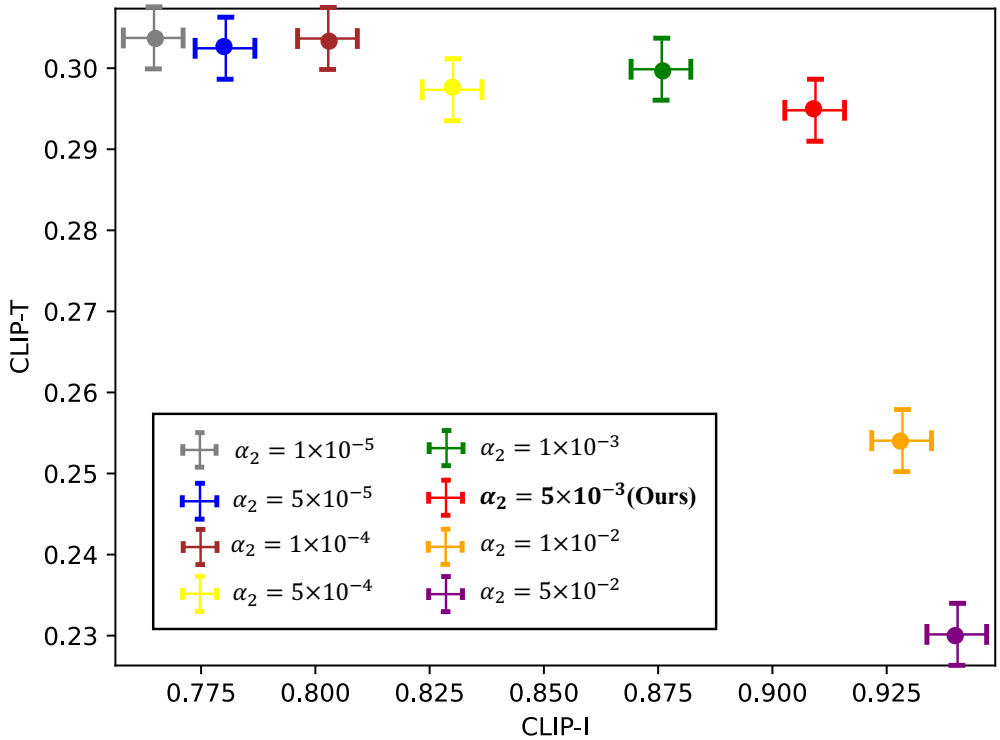


Figure 7. Ablation study on attentive attribution alignment loss. A higher CLIP-T represents a higher prompt fidelity, and a higher CLIP-I represents a higher identity fidelity. The **bold** number denotes the best one.

5. Discussion

Our method is designed for subject-driven personalized image generation with a small number of reference images. It is suitable for everyday creative use, such as generating personalized objects, animals, or anime characters under user-specified prompts,

and for lightweight commercial workflows such as design prototyping, advertising material preparation, customized visual assets, and content creation services. Since the method is built on the SD1.5 backbone and uses LoRA fine-tuning, it can be integrated into existing Stable Diffusion-based software pipelines or deployed as a personalization module in image-generation services with moderate hardware requirements.

Although our method improves identity fidelity and prompt fidelity preferences while maintaining comparable overall-quality preference, several limitations remain. First, the large pre-trained text-to-image model requires greater robustness in the fine-tuning process, as the BLIP model may generate a mistaken caption. Moreover, the pre-trained text-to-image model may generate low-quality class-generalized images. In addition, the weights of the two directions in the fine-tuning process may vary with the prior capacity of different data distributions, such as real-world and anime character image distributions. This may reduce robustness and increase the difficulty of fine-tuning. Second, although our method is more lightweight than large unified content-generation models, fine-tuning still requires GPU resources and may be costly for large-scale commercial deployment. Third, the method is sensitive to the quality of source images; low-resolution images, occlusion, or background noise may weaken identity learning. Fourth, complex geometry, body structures, and object compositions may still cause distortions or imperfect spatial layouts. Fifth, generated images may show limited diversity when the reference set contains similar poses or visual patterns. Dataset imbalance can also affect the breadth and reliability of the evaluation: a category containing fewer identities provides a less representative performance estimate, while identities containing fewer, lower-quality, or less diverse reference images may show weaker identity preservation, reduced generation diversity, or greater sensitivity to pose and background. Finally, ethical and legal concerns should be considered in practical use, especially when using human images, copyrighted identities, or private visual content. In future work, we will focus on improving robustness during the fine-tuning process, reducing computational cost, expanding the dataset with a more balanced number of identities and reference images, evaluating additional collection conditions, enhancing diversity, and incorporating safeguards for responsible personalization.

6. Conclusions

Existing personalization methods for text-to-image models have the following three problems: (1) low identity fidelity, (2) low prompt fidelity, and (3) low quality. In this paper, we propose a novel dual-direction fine-tuning strategy for personalizing text-to-image models with attentive attribution alignment. Our model can produce high-quality images based on several reference images of a specific identity. Experimental results, including automated metrics and statistically validated user-study analysis, demonstrate that our method achieves strong identity fidelity and prompt fidelity while maintaining competitive visual quality.

Author Contributions: Conceptualization, X.W.; methodology, X.W.; software, J.L.; validation, X.W.; formal analysis, J.L.; investigation, J.L.; resources, L.Z.; data curation, X.W.; writing—original draft preparation, X.W.; writing—review and editing, X.W.; visualization, J.L.; supervision, W.X., H.L., and L.Z.; project administration, W.X., H.L., and L.Z.; funding acquisition, L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by Zhejiang Province Program (2024C03263, 2025C01068, 2024C01110, 2025C02023, LZ25F020006), and Ningbo Science and Technology Plan Project (2025Z052, 2025Z062, 2022Z167, 2023Z137).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets presented in this article are not readily available because the data are part of an ongoing study.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT-5.5 for the purposes of language polishing. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Abdal, R.; Zhu, P.; Femiani, J.; Mitra, N.; Wonka, P. Clip2stylegan: Unsupervised extraction of stylegan edit directions. In Proceedings of the ACM SIGGRAPH 2022 Conference Proceedings, Vancouver, BC, Canada, 7–11 August 2022; pp. 1–9.
2. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
3. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv* **2022**, arXiv:2204.06125.
4. Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. Photorealistic text-to-image diffusion models with deep language understanding. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 36479–36494. [CrossRef]
5. Ho, J.; Saharia, C.; Chan, W.; Fleet, D.J.; Norouzi, M.; Salimans, T. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.* **2022**, *23*, 2249–2281.
6. Nichol, A.Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; Chen, M. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proceedings of the 39th International Conference on Machine Learning*; Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S., Eds.; Proceedings of Machine Learning Research; PMLR: New York, NY, USA, 2022; Volume 162, pp. 16784–16804.
7. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021; PMLR: New York, NY, USA, 2021; pp. 8748–8763.
8. Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A.H.; Chechik, G.; Cohen-or, D. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In Proceedings of the The Eleventh International Conference on Learning Representations, Virtual, 25–29 April 2022.
9. Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; Aberman, K. 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, Vancouver, BC, Canada, 18–24 June 2023; pp. 22500–22510.
10. Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; Zhu, J.Y. Multi-concept customization of text-to-image diffusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–24 June 2023; pp. 1931–1941.
11. Wang, Z.; Li, O.; Wang, T.; Wei, L.; Hao, Y.; Wang, X.; Tian, Q. Prior Preserved Text-to-Image Personalization without Image Regularization. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *35*, 1318–1330. [CrossRef]
12. Chen, H.; Zhang, Y.; Wang, X.; Duan, X.; Zhou, Y.; Zhu, W. DisenDreamer: Subject-Driven Text-to-Image Generation with Sample-aware Disentangled Tuning. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 6860–6873. [CrossRef]
13. Gal, R.; Arar, M.; Atzmon, Y.; Bermano, A.H.; Chechik, G.; Cohen-Or, D. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Trans. Graph.* **2023**, *42*, 150. [CrossRef]
14. Zhang, Y.; Huang, N.; Tang, F.; Huang, H.; Ma, C.; Dong, W.; Xu, C. Inversion-based style transfer with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–24 June 2023; pp. 10146–10156.
15. Tang, R.; Liu, L.; Pandey, A.; Jiang, Z.; Yang, G.; Kumar, K.; Stenetorp, P.; Lin, J.; Türe, F. What the daam: Interpreting stable diffusion using cross attention. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, ON, Canada, 9–14 July 2023; pp. 5644–5659.
16. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.
17. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
18. Yang, C.; Shen, Y.; Zhou, B. Semantic hierarchy emerges in deep generative representations for scene synthesis. *Int. J. Comput. Vis.* **2021**, *129*, 1451–1466. [CrossRef]

19. Esser, P.; Rombach, R.; Ommer, B. Taming transformers for high-resolution image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12873–12883.
20. Bond-Taylor, S.; Leach, A.; Long, Y.; Willcocks, C.G. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7327–7347. [CrossRef]
21. Yang, M.; Niu, S.; Wang, Z.; Li, D.; Du, W. DFSGAN: Introducing editable and representative attributes for few-shot image generation. *Eng. Appl. Artif. Intell.* **2023**, *117*, 105519. [CrossRef]
22. Liu, C.; Shao, M.; Gong, Z.; Lv, X.; Meng, L. MirrorDiff: Prompt redescription for zero-shot grounded text-to-image generation with attention modulation. *Eng. Appl. Artif. Intell.* **2025**, *153*, 110741. [CrossRef]
23. Chen, Y.; Xia, R.; Yang, K.; Zou, K. Image inpainting algorithm based on inference attention module and two-stage network. *Eng. Appl. Artif. Intell.* **2024**, *137*, 109181. [CrossRef]
24. Luan, J.; Zhang, Z.; Xing, W.; Zhao, L. Personalized text-to-image generation with Large Language and Vision Assistant enhanced training. *Eng. Appl. Artif. Intell.* **2025**, *161*, 112116. [CrossRef]
25. Luan, J.; Li, G.; Zhang, Z.; Zhao, L.; Xing, W. IP-Controller: Decomposition and Optimization of Cross-Attention Maps for Accurate Subject-Driven Text-to-Image Diffusion Generation. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *36*, 5274–5287. [CrossRef]
26. Luan, J.; Li, G.; Zhao, L.; Xing, W. Mc-vton: Minimal control virtual try-on diffusion transformer. *arXiv* **2025**, arXiv:2501.03630.
27. Zhang, Z.; Zhang, Q.; Xing, W.; Li, G.; Zhao, L.; Sun, J.; Lan, Z.; Luan, J.; Huang, Y.; Lin, H. Artbank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 7396–7404.
28. Zhang, Z.; Zhang, Q.; Luan, J.; Yang, M.; Wang, Y.; Zhao, L. SPAST: Arbitrary style transfer with style priors via pre-trained large-scale model. *Neural Netw.* **2025**, *189*, 107556. [CrossRef] [PubMed]
29. Chen, H.; Wang, Z.; Zhao, L.; Li, J.; Yang, J. TRTST: Arbitrary High-Quality Text-Guided Style Transfer With Transformers. *IEEE Trans. Image Process.* **2025**, *34*, 759–771. [CrossRef] [PubMed]
30. Wang, Y.; Zhou, W.; Bao, J.; Wang, W.; Li, L.; Li, H. CLIP2GAN: Toward Bridging Text With the Latent Space of GANs. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 6847–6859. [CrossRef]
31. Tan, H.; Yin, B.; Xu, K.; Wang, H.; Liu, X.; Li, X. Attention-Bridged Modal Interaction for Text-to-Image Generation. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 5400–5413. [CrossRef]
32. Nazarieh, F.; Feng, Z.; Awais, M.; Wang, W.; Kittler, J. A Survey of Cross-Modal Visual Content Generation. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 6814–6832. [CrossRef]
33. Liu, B.; Wang, C.; Cao, T.; Jia, K.; Huang, J. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 7817–7826.
34. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning, Baltimore, MY, USA, 17–23 July 2022; PMLR: New York, NY, USA, 2022; pp. 12888–12900.
35. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
36. Li, D.; Li, J.; Hoi, S. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 30146–30166. [CrossRef]
37. Zeng, Y.; Patel, V.M.; Wang, H.; Huang, X.; Wang, T.C.; Liu, M.Y.; Balaji, Y. JeDi: Joint-Image Diffusion Models for Finetuning-Free Personalized Text-to-Image Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–24 June 2024; pp. 6786–6795.
38. Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; Zuo, W. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 4–6 October 2023; pp. 15943–15953.
39. Chen, W.; Hu, H.; Li, Y.; Rui, N.; Jia, X.; Chang, M.W.; Cohen, W.W. Subject-driven text-to-image generation via apprenticeship learning. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; Volume 36.
40. Ma, J.; Liang, J.; Chen, C.; Lu, H. Subject-Diffusion: Open Domain Personalized Text-to-Image Generation without Test-time Fine-tuning. In Proceedings of the ACM SIGGRAPH 2024 Conference Papers, Denver, CO, USA, 27 July–1 August 2024; ACM: New York, NY, USA, 2024. [CrossRef]
41. Hua, M.; Liu, J.; Ding, F.; Liu, W.; Wu, J.; He, Q. Dreamtuner: Single image is enough for subject-driven generation. *arXiv* **2023**, arXiv:2312.13691.

42. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 9650–9660.
43. Tan, Z.; Liu, S.; Yang, X.; Xue, Q.; Wang, X. Ominicontrol: Minimal and universal control for diffusion transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–20 October 2025; pp. 14940–14950.
44. Raisinghani, N. Nano Banana 2: Combining Pro Capabilities with Lightning-Fast Speed, Google DeepMind. 2026. Available online: https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/ (accessed on 6 June 2026).
45. Xiao, S.; Wang, Y.; Zhou, J.; Yuan, H.; Xing, X.; Yan, R.; Li, C.; Wang, S.; Huang, T.; Liu, Z. Omnigen: Unified image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 11–15 June 2025; pp. 13294–13304.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.