
AttnDreamBooth: Towards Text-Aligned Personalized Text-to-Image Generation

Lianyu Pang¹ Jian Yin¹ Baoquan Zhao¹ Feize Wu¹ Fu Lee Wang²
Qing Li³ Xudong Mao^{1*}

¹Sun Yat-sen University ²Hong Kong Metropolitan University

³The Hong Kong Polytechnic University

<https://attdreambooth.github.io>

Abstract

Recent advances in text-to-image models have enabled high-quality personalized image synthesis based on user-provided concepts with flexible textual control. In this work, we analyze the limitations of two primary techniques in text-to-image personalization: Textual Inversion and DreamBooth. When integrating the learned concept into new prompts, Textual Inversion tends to overfit the concept, while DreamBooth often overlooks it. We attribute these issues to the incorrect learning of the embedding alignment for the concept. To address this, we introduce AttnDreamBooth, a novel approach that separately learns the embedding alignment, the attention map, and the subject identity across different training stages. We also introduce a cross-attention map regularization term to enhance the learning of the attention map. Our method demonstrates significant improvements in identity preservation and text alignment compared to the baseline methods.

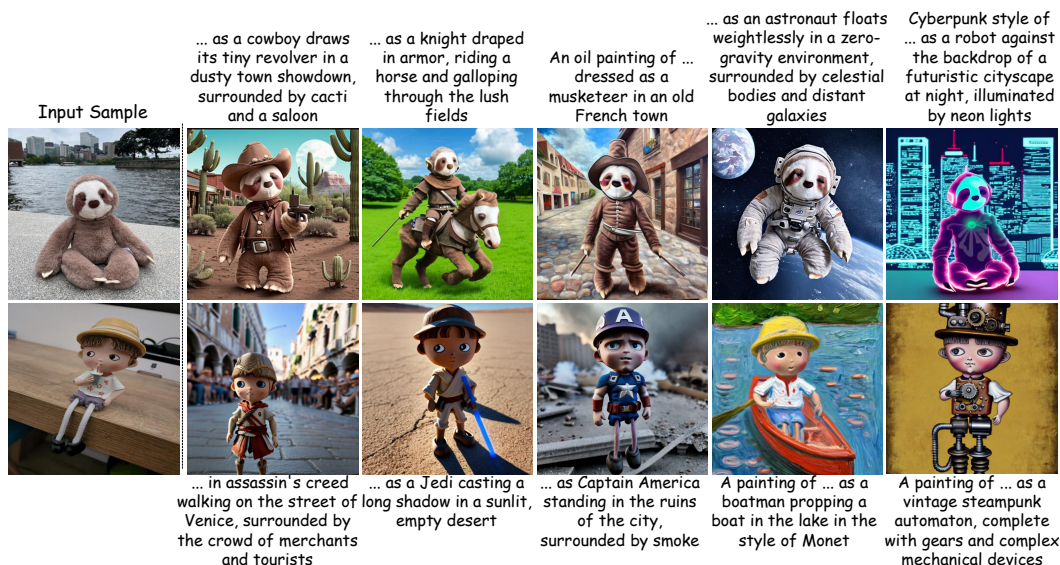


Figure 1: Our method enables text-aligned text-to-image personalization with complex prompts.

*Corresponding author (xudong.xdmao@gmail.com).

1 Introduction

Text-to-image personalization [24, 71, 50] is the task of customizing a pre-trained diffusion model to produce images of user-provided concepts in novel scenes or styles. By providing several examples of a new concept, personalization techniques enable users to employ novel prompts to generate personalized images containing that concept. Current personalization techniques primarily fall into two categories: the first approach involves inverting the new concept into the textual embedding [24]; the second approach involves fine-tuning the diffusion model to learn the new concept [71]. Personalization techniques aim to generate high-quality images of user-provided concepts, achieving high identity preservation and text alignment. However, despite the significant progress in personalization techniques, balancing the trade-off between identity preservation and text alignment remains a challenge for current approaches.

Figure 2 shows the personalization results from Textual Inversion [24] and DreamBooth [71]. Textual Inversion tends to generate images that focus primarily on the learned concept, often neglecting other elements of the prompt. In contrast, DreamBooth appears to overlook the learned concept, producing images that are more influenced by other prompt tokens. These issues can be attributed to the incorrect learning of embedding alignment for the new concept, i.e., the embedding of the new concept is not functionally compatible with the embeddings of existing tokens.

Based on these observations, our approach aims to properly learn not only the subject identity but also the embedding alignment and the attention map for the new concept. Our key insights are as follows: 1) In the early stages of optimization, Textual Inversion effectively learns the embedding alignment but tends to overfit after extensive optimization steps; 2) DreamBooth accurately captures the subject identity but struggles with learning the embedding alignment.

In this paper, we propose a method named AttnDreamBooth, which separates the learning processes of the embedding alignment, the attention map, and the subject identity. Specifically, our approach consists of three main training stages, as illustrated in Figure 3. First, we optimize the textual embedding to learn the embedding alignment while preventing the risk of overfitting, which results in a coarse attention map for the new concept. Next, we fine-tune the cross-attention layers of the U-Net to refine the attention map. Lastly, we fine-tune the entire U-Net to capture the subject identity. Note that the text encoder remains fixed throughout all training stages to preserve its prior knowledge of contextual understanding.

Furthermore, we introduce a cross-attention map regularization term to enhance the learning of the attention map. Throughout the three training stages, we use a consistent training prompt, “a photo of a [V] [super-category]”, where [V] and [super-category] denote the tokens for the new concept and its super-category, respectively. Our attention map regularization term encourages similarity between the attention maps of the new concept and its super-category.

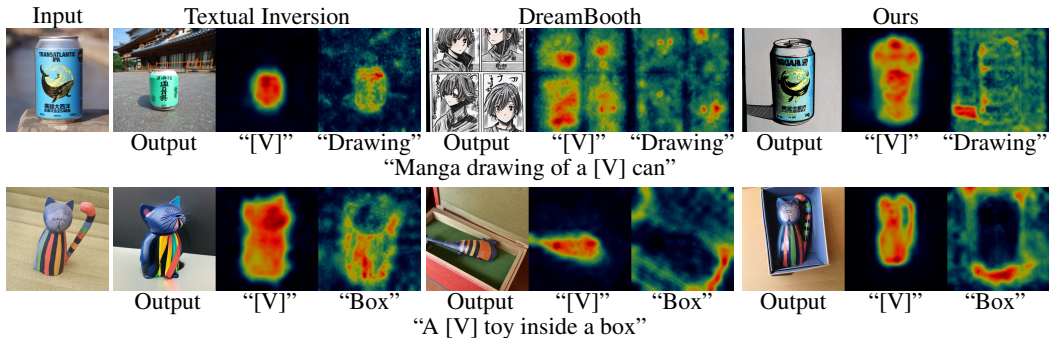


Figure 2: **Analysis of two principal methods.** We visualize the cross-attention maps corresponding to the new concept and other tokens in the prompt. Textual Inversion [24] tends to overfit the textual embedding of the learned concept, resulting in incorrect attention map allocations to other tokens (e.g., “drawing” or “box”). In contrast, DreamBooth [71] appears to overlook the learned concept, producing images primarily based on other tokens.

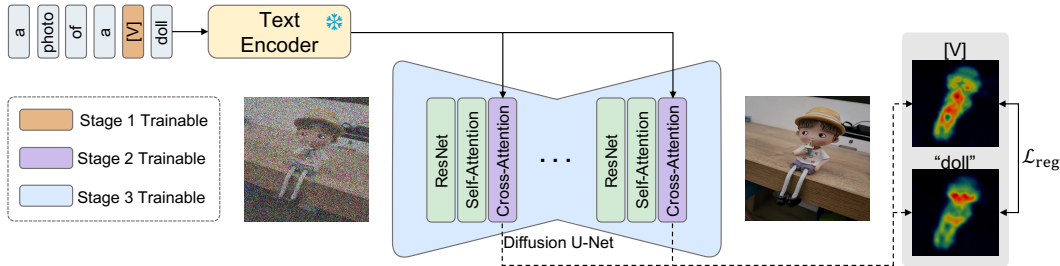


Figure 3: **Overview of AttnDreamBooth.** Our method consists of three training stages. In Stage 1, we optimize the textual embedding of the new concept to align its embedding with existing tokens. In Stage 2, we fine-tune the cross-attention layers to refine the attention map. In Stage 3, we fine-tune the entire U-net to capture the subject identity. Moreover, we introduce a cross-attention map regularization term to guide the learning of the attention map.

To demonstrate the effectiveness of AttnDreamBooth, we compare it with four state-of-the-art baseline methods through both qualitative and quantitative evaluations. Our method achieves superior performance in terms of identity preservation and text alignment compared to the baselines. More importantly, AttnDreamBooth enables a variety of text-aligned personalized generations with complex prompts.

2 Related Work

Text-to-Image Generation. Generative models are designed to create new samples that resemble the patterns observed in their training data. There are various types of generative models, including VAEs [47, 76, 14], GANs [27, 6, 44], auto-regressive models [66, 91], flow-based models [21, 48], and diffusion models [75, 33, 59, 18]. These models can be enhanced by conditioning on text prompts, which are known as text-to-image models [68, 66, 60, 20, 23, 16, 5]. Recent advancements [73, 67, 70] in text-to-image generation, powered by training on extremely large-scale datasets, have demonstrated an impressive ability to generate diverse and generalized outputs.

Text-to-Image Personalization. Leveraging the impressive capabilities of diffusion models, text-to-image personalization involves adapting pre-trained diffusion models to capture new concepts depicted in several given images. Pioneering works approach this by inverting the concept into the textual embedding [24], or by fine-tuning the diffusion model [71]. However, these methods often struggle to balance the trade-off between identity preservation and text alignment, and typically require substantial time for optimization. To overcome these limitations, some studies focus on enhancing the identity preservation of the concept [81, 1, 94, 36, 31, 43, 40], while others aim to improve text alignment [78, 3, 4, 90, 37]. Additionally, there is a growing trend of research attempting to accelerate the personalization process, either by reducing the number of tuning parameters [50, 29, 54, 34, 28, 57], or by pre-training on large datasets [85, 74, 39, 2, 25, 10, 72, 51, 87, 56, 11, 55]. Given the widespread interest in human synthesis, many studies also concentrate on the personalized synthesis of human faces [92, 62, 89, 53, 79, 83, 8, 49, 61, 45, 86, 17, 13, 82, 12].

Cross-Attention Control. The cross-attention layers [70] have been shown to play a crucial role in diffusion models. The control of cross-attention layers has proven effective in a variety of tasks, including image editing [32], compositional synthesis [22, 52, 7, 26, 46], and layout-controlled synthesis [63, 9, 88]. In text-to-image personalization, several studies [50, 87, 4, 78, 30, 42, 58, 93] also have explored the control of cross-attention layers. Custom Diffusion [50] illustrates how incorrect attention maps of the learned concepts can lead to unsuccessful synthesis. FastComposer [87] and Break-A-Scene [4] propose using segmentation masks of the target concepts to guide the learning of the attention maps, thereby enhancing text alignment, especially in scenarios involving multiple concepts. Perfusion [78] identifies the attention overfitting issue and addresses it by fixing the cross-attention key matrices of the target concepts to their super-category tokens.

Multi-Stage Personalization. Several studies [24, 54, 4, 41, 38] have explored combining the strengths of different methods into more efficient models through a multi-stage approach. Inspired by

PTI [69], Textual Inversion [24] investigated a two-stage approach to enhance identity preservation by first optimizing the textual embedding and then fine-tuning the diffusion model to better capture the subject identity. Similarly, MagiCapture [38] first optimizes the textual embedding and then applies LoRA [34] in the U-Net. Break-A-Scene [4] proposes initially optimizing the textual embedding with a high learning rate, followed by fine-tuning both the U-Net and the text encoder using a significantly lower learning rate. Our method differs in several aspects. First, the motivation for our first stage (i.e., optimizing the textual embedding) differs from previous methods. Specifically, we focus on learning the embedding alignment while mitigating the risk of overfitting, thus significantly reducing the optimization steps and lowering the learning rate. Second, we decompose the learning process into three stages: learning the embedding alignment, refining the attention map, and capturing the subject identity. Third, we introduce a cross-attention map regularization to guide the learning of the attention map in a self-supervised manner.

3 Preliminaries

Latent Diffusion Models. Our approach is based on the publicly available Stable Diffusion model, a type of Latent Diffusion Model (LDM) [70] for text-to-image generation. In LDM, an autoencoder is utilized to provide a lower-dimensional representational space, where an encoder \mathcal{E} transforms an image x into a latent representation $z = \mathcal{E}(x)$, and a decoder \mathcal{D} reconstructs the image from this latent code, i.e., $\mathcal{D}(\mathcal{E}(x)) \approx x$. Additionally, a Denoising Diffusion Probabilistic Model (DDPM) [33] is employed to produce latent codes within the latent space of the autoencoder. To generate images from text, the model leverages a conditioning vector $c(y)$, derived from a given text prompt y . The training objective of LDM is given by:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{z \sim \mathcal{E}(x), y, \varepsilon \sim \mathcal{N}(0,1), t} \left[\|\varepsilon - \varepsilon_{\theta}(z_t, t, c(y))\|_2^2 \right], \quad (1)$$

where the denoising network ε_{θ} is tasked with recovering the original latent code z_0 from the noised latent code z_t , given a specific timestep t and the conditioning vector $c(y)$.

Textual Inversion. Textual Inversion (TI) [24] personalizes a pre-trained diffusion model by encoding the target concept into the textual embedding. Given several images of a target concept, TI introduces a new token S_* and its associated textual embedding v_* to represent the concept. The learning process of TI involves initializing v_* with a coarse descriptor and then optimizing it to minimize the diffusion objective (Eq. 1).

DreamBooth. DreamBooth (DB) [71] learns the target concept by fine-tuning the pre-trained diffusion model. Given several images of a target concept, DB labels all the images with the prompt “a [V] [super-category]”, where [V] is a rare token in the vocabulary. The learning process of DB involves fine-tuning the entire U-Net (and possibly the text encoder) using the diffusion objective (Eq. 1) combined with a prior preservation loss [71].

4 Method

In this section, we first analyze the problems associated with Textual Inversion and DreamBooth, as discussed in Section 4.1. To address these issues, we propose a novel method named AttnDreamBooth, as detailed in Section 4.2. To further enhance text alignment, we introduce a cross-attention map regularization term in Section 4.3.

4.1 Analysis of Existing Methods

Problems and Analysis. As illustrated in Figure 2, Textual Inversion and DreamBooth encounter distinct challenges when integrating the learned concept into novel prompts. For Textual Inversion, the generated images often excessively focus on the learned concept, overlooking other prompt tokens. To investigate this issue, we present the attention map visualization using DAAM [77] for different tokens in Figure 2. This visualization reveals an embedding misalignment issue in novel compositions containing the concept, leading to incorrect attention map allocations for other tokens. A typical example is shown where the attention map corresponding to the “drawing” token focuses on incorrect regions. This misalignment occurs because Textual Inversion tends to overfit

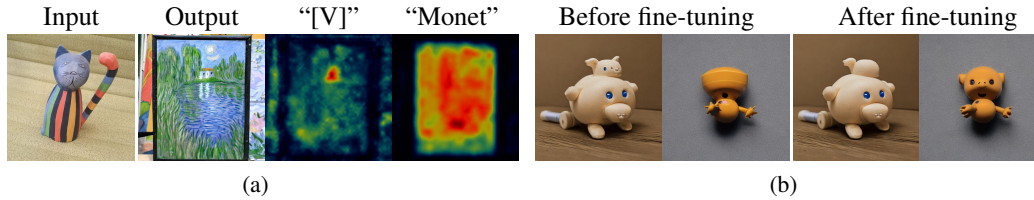


Figure 4: **Analysis of TI+DB.** Column (a) demonstrates that TI+DB neglects the learned concept when integrating it into a new prompt, “A painting of a [V] toy in the style of Monet”. Column (b) shows the generated images based on a single word prompt, “[V]”, both before and after fine-tuning, using the diffusion model without fine-tuning. These images are notably similar to each other, which indicates that the learned textual embedding remains largely unchanged from its initial state.

the input embedding of the text encoder, responsible for managing the contextual understanding of the prompt. Conversely, images generated by DreamBooth sometimes focus solely on other prompt tokens, neglecting the learned concept. This occurs because DreamBooth uses a rare token for the new concept while keeping its textual embedding fixed, thereby leading to insufficient learning of the embedding alignment for the new concept.

A Naive Solution. As analyzed previously, Textual Inversion and DreamBooth exhibit distinct issues related to the embedding alignment: Textual Inversion tends to overfit the embedding alignment for the new concept, while DreamBooth demonstrates insufficient learning of the embedding alignment. A straightforward solution is to combine Textual Inversion with DreamBooth by jointly tuning the textual embedding and the U-Net, a method we denote as TI+DB. We observe that TI+DB enhances performance over Textual Inversion or DreamBooth individually. However, it still tends to neglect the learned concept when integrating it into new prompts. This issue arises from the slow update of the textual embedding relative to the U-Net. As illustrated in Figure 4, the learned textual embedding remains very close to its initial state. Furthermore, we calculate the cosine similarity between the learned and initial embeddings, which averages about 0.9997, indicating that TI+DB still suffers from insufficient learning of the embedding alignment.

4.2 AttnDreamBooth

To address the issues described in Section 4.1, we propose a method named AttnDreamBooth, inspired by two key observations. First, while Textual Inversion often fails to capture the subject identity and tends to overfit the embedding alignment for the new concept, it can effectively learn the embedding alignment in the very early stages of optimization. However, at these early stages, the model only learns a coarse cross-attention map for the new concept. Second, although DreamBooth fails to learn the embedding alignment, it can accurately capture the subject identity. Based on these observations, we propose to decompose the personalization process into three training stages: 1) learning the embedding alignment; 2) refining the attention map; and 3) acquiring the subject identity. An overview of our proposed AttnDreamBooth is illustrated in Figure 3.

Learning the Embedding Alignment. As previously stated, learning the embedding alignment for the new concept is critical for properly allocating the cross-attention maps for novel prompts, which in turn influences the text alignment of the personalized generation results. To achieve this, we optimize the input textual embedding of the text encoder, since the text encoder manages the contextual understanding of the prompt. However, as analyzed in Section 4.1, this approach is prone to overfitting the embedding, leading to an embedding misalignment issue. Therefore, our objective at this stage is to learn the embedding alignment while minimizing the risk of overfitting. To this end, we adapt Textual Inversion [24] with three main modifications. First, we significantly reduce the number of optimization steps (to 60 steps in our experiments) and lower the learning rate (to 10^{-3}). Second, we introduce a cross-attention map regularization (see Section 4.3) to guide the learning of the cross-attention map. Third, to facilitate the incorporation of the cross-attention map regularization, we set the training prompt as “a photo of a [V] [super-category]”. To prevent overfitting, we stop the optimization at very early stages, thereby resulting in a coarse cross-attention map for the new concept, as depicted in Figure 5. A full analysis of attention map allocations for each token is

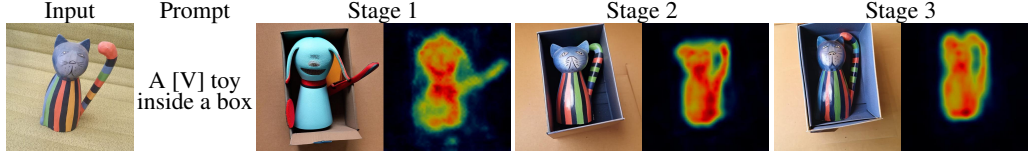


Figure 5: **Results after each training stage.** We present the generations along with the attention maps of “[V]” for each stage. In stage 1, the model properly aligns the embedding of [V] with other tokens, “inside a box”, but learns a very coarse attention map and subject identity. In stage 2, the model refines the attention map and subject identity. In stage 3, the model accurately captures the identity of the concept.

presented in Appendix E. The cross-attention map, as well as the subject identity, are addressed in subsequent steps.

Refining the Cross-Attention Map. To mitigate the embedding misalignment issue, our model initially learns a relatively coarse cross-attention map for the new concept. At this stage, we focus on refining the cross-attention map. Since these attention maps are embedded within the cross-attention layers, inspired by Custom Diffusion [50], we fine-tune all the cross-attention layers in the U-Net. Additionally, we employ the proposed cross-attention map regularization (see Section 4.3) to aid in refining the attention map. Furthermore, we keep the textual embedding and the text encoder fixed to prevent further embedding misalignment.

Capturing the Subject Identity. As illustrated in Figure 5, the previous stage produces images that are similar to the target concept but still exhibit significant distortions. Therefore, in the third stage, following DreamBooth [71], we unfreeze all layers of the U-Net to more accurately capture the subject identity of the target concept. We choose not to adopt the prior preservation loss [71], as we empirically find that it leads to poor identity preservation and requires significantly more training steps. A detail discussion of our models with or without the prior preservation loss could be found in Appendix H. Moreover, similar to the previous stage, we keep the textual embedding and the text encoder fixed to prevent embedding misalignment, and we continue to apply the cross-attention map regularization to guide the learning of the attention map.

4.3 Cross-Attention Map Regularization

We set the training prompt as “a photo of a [V] [super-category]”, where [V] and [super-category] denote the tokens for the new concept and its super-category, respectively. To enhance the learning of the attention map, we introduce a regularization term that encourages similarity between the attention maps of [V] and [super-category]. This regularization term serves two purposes. First, since the new concept and its super-category belong to the same object category, the attention map of the super-category token can serve as a reference for the new concept. Second, since [V] and [super-category] are used together to describe the new concept when integrating it into new prompts, the attention maps of [V] and [super-category] should refer to the same region.

Formally, for the 16 attention maps $\{M_1, M_2, \dots, M_{16}\}$ from 16 different cross-attention layers, we minimize the squared differences in the mean and variance of the attention map values for [V] and [super-category] as follows:

$$\mathcal{L}_{\text{reg}} = \lambda_{\mu} [\mu(M_{1:16}^V) - \mu(M_{1:16}^{\text{category}})]^2 + \lambda_{\sigma} [\sigma^2(M_{1:16}^V) - \sigma^2(M_{1:16}^{\text{category}})]^2, \quad (2)$$

where $\mu(M_{1:16})$ and $\sigma^2(M_{1:16})$ denote the mean and variance of all the values across the 16 attention maps, respectively. This constraint helps ensure that the new concept exhibits a similar level of concentration or dispersion in the attention map as the super-category token. Note that we avoid directly applying the constraint to the attention map values themselves because we empirically find that such a constraint is too restrictive and difficult to optimize.

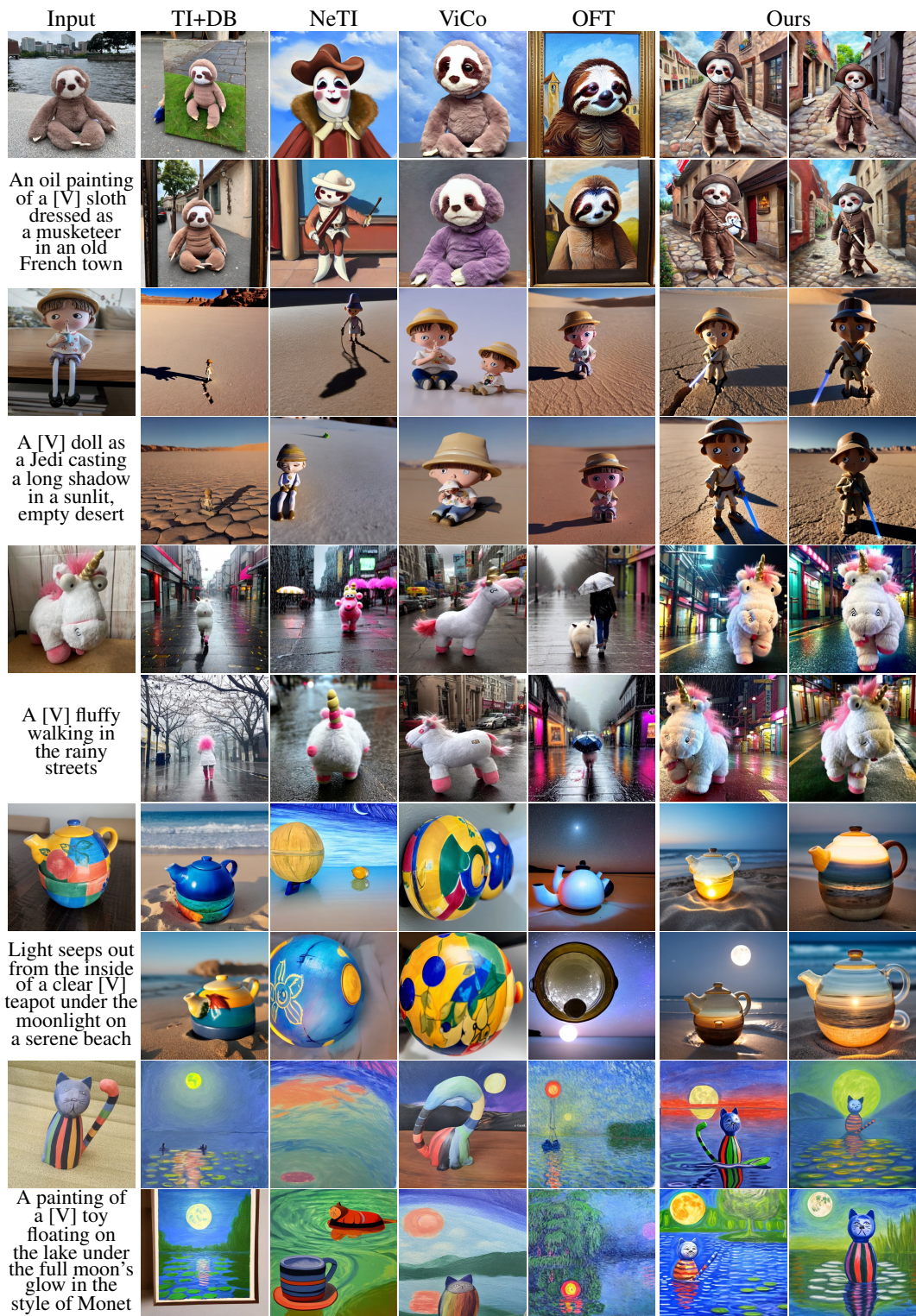


Figure 6: **Qualitative comparison.** We present four images generated by our method and two images from each of the baseline methods, including TI+DB [24, 71], NeTI [1], ViCo [30], and OFT [64]. Our method demonstrates superior performance in text alignment and identity preservation compared to these baselines.

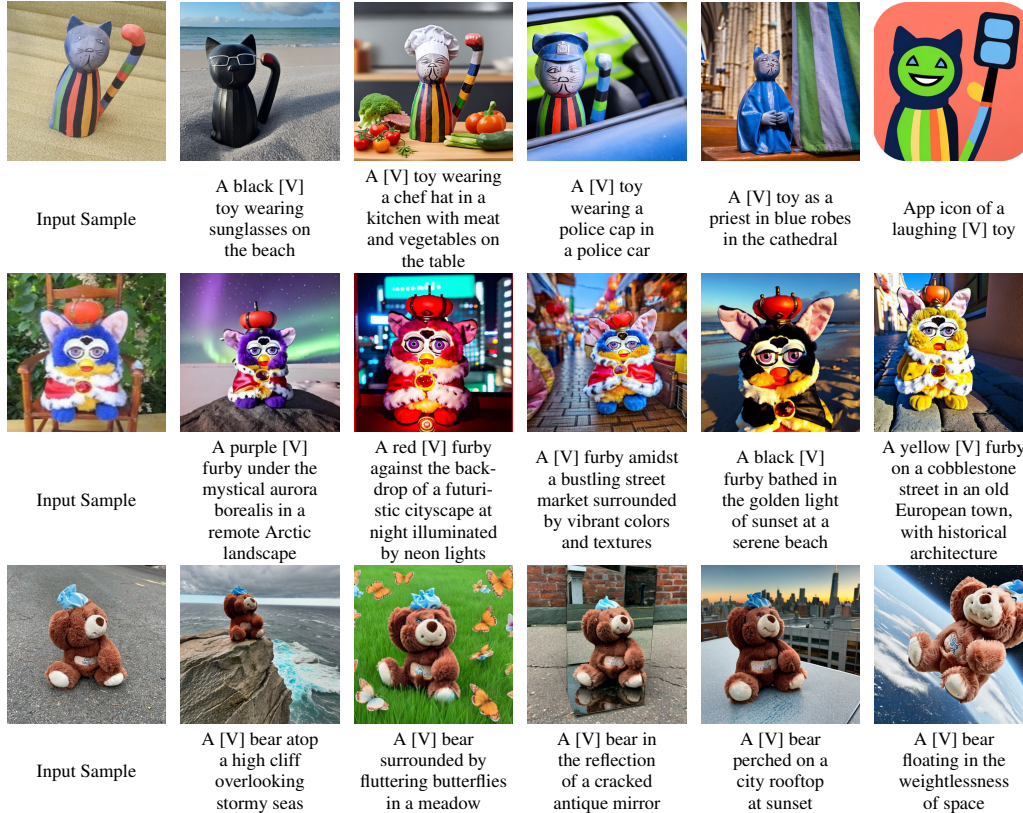


Figure 7: Examples of personalized generations obtained using AttnDreamBooth.

5 Experiments

In this section, we first present the implementation details of our method. Subsequently, we evaluate its performance by conducting a comparative analysis with four state-of-the-art personalization methods. Lastly, we conduct an ablation study to demonstrate the effectiveness of each sub-module.

5.1 Implementation and Evaluation Setup

Implementation Details. Our implementation is based on the publicly available Stable Diffusion V2.1 [70]. The textual embedding of the new concept is initialized using the embedding of the super-category token. We keep a fixed batch size of 8 across all training stages but vary the learning rates and training steps. Specifically, we train with a learning rate of 10^{-3} for 60 steps in stage 1, followed by a learning rate of 2×10^{-5} for 100 steps in stage 2, and conclude with a learning rate of 2×10^{-6} for 500 steps in stage 3. λ_μ and λ_σ are set to 0.1 and 0 in stage 1, respectively, and are adjusted to 2 and 5 in subsequent stages. All experiments are conducted on a single Nvidia A100 GPU. The training process of our method takes about 20 minutes to learn a concept.

Evaluation Setup. We compare our method with four state-of-the-art personalization methods, including TI+DB [24, 71], NeTI [1], ViCo [30], and OFT [64]. The implementation details of the baseline methods are provided in Appendix A. We collect 22 concepts from TI [24] and DB [71]. For the quantitative evaluation, each method is evaluated using a set of 24 text prompts, see Appendix B for a complete list. These prompts cover background change, environment interaction, concept color change, and artistic style.

5.2 Results

Qualitative Evaluation. In Figure 6, we present a visual comparison of personalized generation for various concepts. We employ a set of complex prompts for evaluation, where one prompt simultane-

Table 1: **Quantitative comparisons.** “Identity” denotes the identity preservation, and “Text” denotes the text alignment.

Methods	Identity \uparrow	Text \uparrow
TI+DB [24, 71]	0.7017	0.2578
NeTI [1]	0.6901	0.2522
ViCo [30]	0.7507	0.2106
OFT [64]	0.7257	0.2445
Ours-fast	<u>0.7268</u>	<u>0.2536</u>
Ours	0.7257	0.2532

Table 2: **User study.** We asked the participants to select the image that better preserves the identity and matches the prompt.

Baselines	Prefer Baseline	Prefer Ours
TI+DB [24, 71]	32.0%	68.0%
NeTI [1]	20.6%	79.4%
ViCo [30]	16.6%	83.4%
OFT [64]	22.4%	72.6%

ously incorporates several editing elements such as style change (e.g., “oil painting”), scene change (e.g., “old French town”), and appearance change (e.g., “dressed as a musketeer”). As observed, ViCo tends to overfit the new concept, failing to compose it in novel scenes or styles. Conversely, TI+DB sometimes overlooks the learned concept, producing images that solely reflect other prompt tokens. NeTI and OFT also struggle to achieve text-aligned generations, especially when the prompts are complex. Our method, AttnDreamBooth, is the only method that successfully generates identity-preserved and text-aligned personalized images for these complex prompts. Figures 1 and 7 show more personalized generations using complex prompts from our method. Additional qualitative results can be found in Appendix C.

Quantitative Evaluation. We conduct a quantitative evaluation of each method in terms of identity preservation and text alignment. Identity preservation is measured by the cosine similarity between the CLIP [65] embeddings of generated and real images, while text alignment is measured by the cosine similarity between the CLIP embeddings of generated images and their corresponding prompts. Each method is evaluated using 24 text prompts, generating 32 images per prompt. The results are presented in Table 1. TI+DB excels in text alignment but performs poorly in identity preservation. This is consistent with the qualitative observation that TI+DB often neglects the learned concept and generates images based solely on other prompt tokens. In contrast, ViCo achieves the best identity preservation but ranks lowest in text alignment, indicative of its tendency to overfit the new concept. Besides these two extreme cases, our approach exhibits superior performance in both identity preservation and text alignment.

Fast Version of Our Method. The average training time using our method is about 20 minutes. To reduce the training time, we developed a fast version of our method by increasing the learning rate while simultaneously decreasing both the training steps and the batch size for the third training stage. Originally, this stage involves performing 500 steps with a learning rate of 2×10^{-6} and a batch size of 8. The fast version now completes training in just 200 steps with a learning rate of 1×10^{-5} and a batch size of 4. These adjustments significantly reduce the training time from 20 minutes to an average of 6 minutes. Interestingly, this fast version maintains performance comparable to our original model, likely because the first two stages provide a convenient starting point, allowing for a higher learning rate in the third stage. Notably, the fast version performs similarly to the original model on short prompts, as shown in Table 1, but it shows slight degradation on complex prompts, as depicted in Figure 10.

User Study. We further evaluate our method by conducting a user study. Personalized images are generated using various prompts and concepts for each method. In each question of the study, participants are presented with an input image and a text prompt, along with two generated images: one from our method and another from a baseline method. Participants are asked to select the image that better achieves identity preservation and text alignment. We collected a total of 700 responses from 35 participants, as presented in Table 2. The results demonstrate a clear preference for our method.

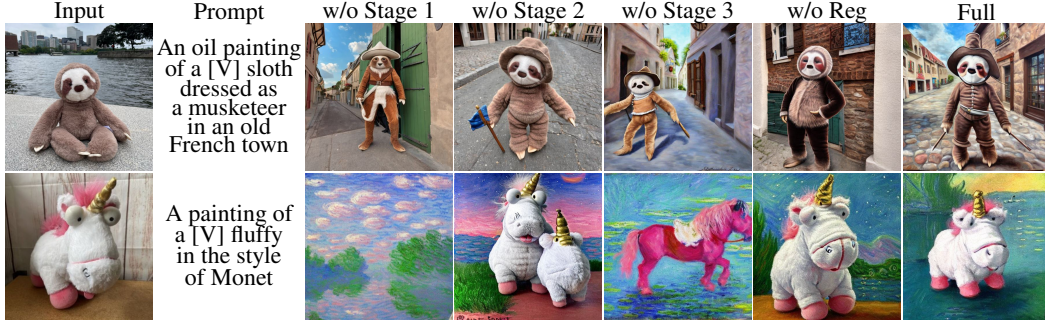


Figure 8: **Ablation study.** We compare models trained without optimizing the textual embedding (w/o Stage 1), without fine-tuning the cross-attention layers (w/o Stage 2), without fine-tuning the U-Net (w/o Stage 3), and without the cross-attention map regularization (w/o Reg). As can be observed, all sub-modules are essential for achieving identity-preserved and text-aligned personalized generations.

5.3 Ablation Study

In this section, we evaluate the effectiveness of each sub-module within our framework. Specifically, we conduct an ablation study by separately removing each training stage or the attention map regularization term. Figure 8 presents a visual comparison of personalized images generated by each variant. The results indicate that all sub-modules are crucial for achieving identity-preserved and text-aligned personalized generations. Specifically, the model without optimizing the textual embedding (w/o Stage 1) tends to neglect the learned concept or generate it with significant distortions due to insufficient learning of the embedding alignment. Models without fine-tuning the cross-attention layers (w/o Stage 2) or without the regularization term (w/o Reg) suffer from degraded text alignment or identity preservation. The model without fine-tuning the U-Net (w/o Stage 3) leads to significant degradation in identity preservation. Additional ablation study results are provided in Appendix F. Similar behavior is observed in the quantitative ablation study, as detailed in Table 4.

6 Conclusions and Limitations

In this paper, we identified and analyzed the embedding misalignment issue encountered by Textual Inversion and DreamBooth. Our proposed method, named AttnDreamBooth, addresses this issue by decomposing the personalization process into three stages: learning the embedding alignment, refining the attention map, and acquiring the subject identity. AttnDreamBooth enables identity-preserved and text-aligned text-to-image personalization, even with complex prompts.

In our experiments, we used consistent training steps across different concepts; however, we observed that performance could be further improved by tuning the training steps for specific concepts. This limitation might be addressed by adopting adaptive training strategies, which we leave for future work. A second limitation is that our three-stage training method requires approximately 20 minutes on average to learn a concept, as it involves fine-tuning all parameters in the U-Net for 500 steps. To mitigate this, we introduced a fast version that reduces the training time to approximately 6 minutes while still maintaining performance comparable to the original model.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No. 62176223 and No. 62302535), Guangdong Basic and Applied Basic Research Foundation (No. 2023A1515012897 and No. 2023A1515011639), and Zhuhai Basic and Applied Basic Research Foundation (No. 2320004002745).

References

- [1] Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization. *arXiv preprint arXiv:2305.15391*, 2023.
- [2] Moab Arar, Rinon Gal, Yuval Atzmon, Gal Chechik, Daniel Cohen-Or, Ariel Shamir, and Amit H. Bermano. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06925*, 2023.
- [3] Moab Arar, Andrey Voynov, Amir Hertz, Omri Avrahami, Shlomi Fruchter, Yael Pritch, Daniel Cohen-Or, and Ariel Shamir. Palp: Prompt aligned personalization of text-to-image models. *arXiv preprint arXiv:2401.06105*, 2024.
- [4] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2305.16311*, 2023.
- [5] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.
- [6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.
- [7] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. In *SIGGRAPH*, 2023.
- [8] Li Chen, Mengyi Zhao, Yiheng Liu, Mingxu Ding, Yangyang Song, Shizun Wang, Xu Wang, Hao Yang, Jing Liu, Kang Du, and Min Zheng. Photoverse: Tuning-free image customization with text-to-image diffusion models. *arXiv preprint arXiv:2309.05793*, 2023.
- [9] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *WACV*, 2024.
- [10] Wenhui Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*, 2023.
- [11] Xi Chen, Lianghai Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023.
- [12] Zhuowei Chen, Shancheng Fang, Wei Liu, Qian He, Mengqi Huang, Yongdong Zhang, and Zhendong Mao. Dreamidentity: Improved editability for efficient face-identity preserved image generation. *arXiv preprint arXiv:2307.00300*, 2023.
- [13] Jiayang Cheng, Pan Xie, Xin Xia, Jiashi Li, Jie Wu, Yuxi Ren, Huixia Li, Xuefeng Xiao, Min Zheng, and Lean Fu. Resadapter: Domain consistent resolution adapter for diffusion models. *arXiv preprint arXiv:2403.02084*, 2024.
- [14] Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. In *ICLR*, 2021.
- [15] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP*, 2023.
- [16] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *ECCV*, 2022.
- [17] Siying Cui, Jiankang Deng, Jia Guo, Xiang An, Yongle Zhao, Xinyu Wei, and Ziyong Feng. Idadapter: Learning mixed features for tuning-free personalization of text-to-image models. *arXiv preprint arXiv:2403.13535*, 2024.

- [18] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.
- [19] 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.
- [20] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. In *NeurIPS*, 2021.
- [21] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.
- [22] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *ICLR*, 2023.
- [23] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *ECCV*, 2022.
- [24] 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.
- [25] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *TOG*, 2023.
- [26] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *ICCV*, 2023.
- [27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.
- [28] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, Yixiao Ge, Ying Shan, and Mike Zheng Shou. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *arXiv preprint arXiv:2305.18292*, 2023.
- [29] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023.
- [30] Shaozhe Hao, Kai Han, Shihao Zhao, and Kwan-Yee K. Wong. Vico: Detail-preserving visual condition for personalized text-to-image generation. *arXiv preprint arXiv:2306.00971*, 2023.
- [31] Xingzhe He, Zhiwen Cao, Nicholas Kolkin, Lantao Yu, Helge Rhodin, and Ratheesh Kalarot. A data perspective on enhanced identity preservation for diffusion personalization. *arXiv preprint arXiv:2311.04315*, 2023.
- [32] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, 2023.
- [33] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.
- [34] 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. In *ICLR*, 2022.
- [35] Hexiang Hu, Kelvin CK Chan, Yu-Chuan Su, Wenhui Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, et al. Instruct-imagen: Image generation with multi-modal instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4754–4763, 2024.

- [36] Miao Hua, Jiawei Liu, Fei Ding, Wei Liu, Jie Wu, and Qian He. Dreamtuner: Single image is enough for subject-driven generation. *arXiv preprint arXiv:2312.13691*, 2023.
- [37] Mengqi Huang, Zhendong Mao, Mingcong Liu, Qian He, and Yongdong Zhang. Realcustom: Narrowing real text word for real-time open-domain text-to-image customization. *arXiv preprint arXiv:2403.00483*, 2024.
- [38] Junha Hyung, Jaeyo Shin, and Jaegul Choo. Magicapture: High-resolution multi-concept portrait customization. *arXiv preprint arXiv:2309.06895*, 2023.
- [39] Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023.
- [40] Jiaxiu Jiang, Yabo Zhang, Kailai Feng, Xiaohe Wu, and Wangmeng Zuo. Mc²: Multi-concept guidance for customized multi-concept generation. *arXiv preprint arXiv:2404.05268*, 2024.
- [41] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *CVPR*, 2024.
- [42] Chen Jin, Ryutaro Tanno, Amrutha Saseendran, Tom Diethe, and Philip Teare. An image is worth multiple words: Learning object level concepts using multi-concept prompt learning. *arXiv preprint arXiv:2310.12274*, 2023.
- [43] Maxwell Jones, Sheng-Yu Wang, Nupur Kumari, David Bau, and Jun-Yan Zhu. Customizing text-to-image models with a single image pair. *arXiv preprint arXiv:2405.01536*, 2024.
- [44] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.
- [45] Chanran Kim, Jeongin Lee, Shichang Joung, Bongmo Kim, and Yeul-Min Baek. Instantfamily: Masked attention for zero-shot multi-id image generation. *arXiv preprint arXiv:2404.19427*, 2024.
- [46] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *ICCV*, 2023.
- [47] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [48] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018.
- [49] Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizhu Wu, Guanying Chen, Wei Liu, and Wenhan Luo. Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. *arXiv preprint arXiv:2403.10983*, 2024.
- [50] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023.
- [51] Dongxu Li, Junnan Li, and Steven C. H. Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023.
- [52] Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. Divide & bind your attention for improved generative semantic nursing. In *BMVC*, 2023.
- [53] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. *arXiv preprint arXiv:2312.04461*, 2023.
- [54] LoRA. Low-rank adaptation for fast text-to-image diffusion fine-tuning. <https://github.com/cloneofsimo/lora>, 2022.

- [55] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion:open domain personalized text-to-image generation without test-time fine-tuning. *arXiv preprint arXiv:2307.11410*, 2023.
- [56] Yiyang Ma, Huan Yang, Wenjing Wang, Jianlong Fu, and Jiaying Liu. Unified multi-modal latent diffusion for joint subject and text conditional image generation. *arXiv preprint arXiv:2303.09319*, 2023.
- [57] Shyam Marjit, Harshit Singh, Nityanand Mathur, Sayak Paul, Chia-Mu Yu, and Pin-Yu Chen. Diffusekrona: A parameter efficient fine-tuning method for personalized diffusion model. *arXiv preprint arXiv:2402.17412*, 2024.
- [58] Jisu Nam, Heesu Kim, DongJae Lee, Siyoon Jin, Seungryong Kim, and Seunggyu Chang. Dreammatcher: Appearance matching self-attention for semantically-consistent text-to-image personalization. *arXiv preprint arXiv:2402.09812*, 2024.
- [59] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021.
- [60] 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.
- [61] Daniil Ostashev, Yuwei Fang, Sergey Tulyakov, Kfir Aberman, et al. Moa: Mixture-of-attention for subject-context disentanglement in personalized image generation. *arXiv preprint arXiv:2404.11565*, 2024.
- [62] Lianyu Pang, Jian Yin, Haoran Xie, Qiping Wang, Qing Li, and Xudong Mao. Cross initialization for personalized text-to-image generation. *arXiv preprint arXiv:2312.15905*, 2023.
- [63] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. *arXiv preprint arXiv:2306.05427*, 2023.
- [64] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. In *NeurIPS*, 2024.
- [65] 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 *ICML*, 2021.
- [66] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021.
- [67] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [68] Scott Reed, Zeynep Akata, Xinchun Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016.
- [69] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *TOG*, 2022.
- [70] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.
- [71] 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 *CVPR*, 2023.
- [72] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023.

- [73] 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. In *NeurIPS*, 2022.
- [74] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023.
- [75] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.
- [76] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *NeurIPS*, 2016.
- [77] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *ACL*, 2023.
- [78] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *SIGGRAPH*, 2023.
- [79] Dani Valevski, Danny Wasserman, Yossi Matias, and Yaniv Leviathan. Face0: Instantaneously conditioning a text-to-image model on a face. *arXiv preprint arXiv:2306.06638*, 2023.
- [80] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers>, 2022.
- [81] 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.
- [82] Qinghe Wang, Xu Jia, Xiaomin Li, Taiqing Li, Liqian Ma, Yunzhi Zhuge, and Huchuan Lu. Stableidentity: Inserting anybody into anywhere at first sight. *arXiv preprint arXiv:2401.15975*, 2024.
- [83] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.
- [84] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *CVPR*, 2020.
- [85] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023.
- [86] Yi Wu, Ziqiang Li, Heliang Zheng, Chaoyue Wang, and Bin Li. Infinite-id: Identity-preserved personalization via id-semantics decoupling paradigm. *arXiv preprint arXiv:2403.11781*, 2024.
- [87] Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023.
- [88] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *ICCV*, 2023.
- [89] Yuxuan Yan, Chi Zhang, Rui Wang, Yichao Zhou, Gege Zhang, Pei Cheng, Gang Yu, and Bin Fu. Facestudio: Put your face everywhere in seconds. *arXiv preprint arXiv:2312.02663*, 2023.
- [90] Hu Ye, Jun Zhang, Sibio 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.
- [91] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022.

- [92] Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. Inserting anybody in diffusion models via celeb basis. In *NeurIPS*, 2023.
- [93] Yanbing Zhang, Mengping Yang, Qin Zhou, and Zhe Wang. Attention calibration for disentangled text-to-image personalization. In *CVPR*, 2024.
- [94] Yufan Zhou, Ruiyi Zhang, Tong Sun, and Jinhui Xu. Enhancing detail preservation for customized text-to-image generation: A regularization-free approach. *arXiv preprint arXiv:2305.13579*, 2023.

A Implementation Details of Baselines

We compare our method with four baseline methods, including TI+DB [24, 71], NeTI [1], ViCo [30], and OFT [64]. For TI+DB, we implement it based on the diffusers library [80] without employing the prior preservation loss. We perform 660 training steps, which matches the total number of steps for our method, with a learning rate of 2×10^{-6} and a batch size of 8. For the other baselines, we use the official implementations and follow the hyper-parameters described in their papers.

In the Appendix, we further compare our method with four other baseline methods, including DreamMatcher [58], FreeCustom [19], SuTI [10], and Instruct-Imagen [35]. For DreamMatcher and FreeCustom, we use their official implementations. Due to the unavailability of open-source models for SuTI and Instruct-Imagen, we rely on the examples provided in their papers for comparison.

B Text Prompts

In Table 3, we list all 24 text prompts used in the quantitative evaluation. These prompts cover a range of modifications, including background change, environment interaction, concept color change, and artistic style.

C Additional Qualitative Results

In Figures 9, 10, and 11, we provide additional qualitative comparisons to the baseline methods across a wide range of prompts. Furthermore, Figure 12 presents a qualitative comparison of our method with Textual Inversion (TI), DreamBooth (DB), and two different configurations of TI+DB: 1) first TI and then DB (TI→DB), and 2) first DB and then TI (DB→TI). In Figure 13, we provide additional qualitative results generated by AttnDreamBooth.

D Single Image Personalization

In this section, we compare AttnDreamBooth with the baseline methods when only a single image is used for training. In Figure 14, we present the generation results of each method under this challenging setting. Our method demonstrates superior text alignment and identity preservation compared to the baselines.

E Attention Maps for Each Token

In the main text, we provide the cross-attention maps of Textual Inversion and DreamBooth in Figure 2 and the cross-attention map of “[V]” for each stage in Figure 5. To vividly demonstrate the efficacy of our method in training the cross-attention layer, we present the generated images along with the cross-attention maps for each token in the prompt in Figure 15. As can be seen, our method accurately assigns the attention maps for each token, demonstrating correct embedding alignment for the new concept.

F Additional Ablation Study

In Figure 16, we provide additional ablation study results for each variant of our method. Table 4 presents the quantitative results of our ablation study. Specifically, the model without Stage 1 achieves better text alignment but significantly poorer identity preservation compared to the full model. This is because, without sufficient training of the textual embedding, the model tends to overlook the learned concept or generate it with significant distortions. Please note that the text alignment score is calculated without considering the new concept; therefore, omitting the new concept can inadvertently boost this score. Similarly, models without Stage 2 or Stage 3 also exhibit higher text alignment scores but lower identity preservation scores, due to insufficient learning of the attention maps and the subject identity, respectively. Additionally, the model without the regularization term shows degraded text alignment.

G Different Version of Stable Diffusion

A visual comparison between our models with SD1.5 or SD2.1 is presented in Figure 17. As shown, the model with SD2.1 achieves superior performance in text alignment and identity preservation. Nevertheless, our method is also effective for SD1.5, and it outperforms the baseline methods.

H Prior Preservation Loss

We present a visual comparison between models with or without the prior preservation loss in Figure 18. The results show that incorporating the prior preservation loss leads to degradation in identity preservation.

I User Study

As described in Section 5.2, we conducted a user study to evaluate our method against the baseline methods. Here, we present the details of this user study. Figure 19 shows an example question from the user study. Given a concept image and a text prompt, along with two generated images (one from our method and another from a baseline method), participants were asked to select the image that better preserves the identity of the concept image and aligns with the text prompt. The results are presented in Table 2.

J Societal Impact

Similar to existing text-to-image personalization techniques, our approach provides broader users access to effectively fine-tuning large-scale pre-trained diffusion models. By enabling users to personalize these models with their own data, our approach can be used for numerous applications, including image editing, artistic creations, and industrial production. However, the use of generative techniques comes with risks, such as the creation of misleading or false information. To mitigate these concerns, it is vital to develop effective methods for identifying fake generations [84, 15].

K Licenses for Pre-trained Models and Datasets

Our implementation is based on the publicly available Stable Diffusion V2.1 [70], which is under the CreativeML Open RAIL++-M License. The datasets used for evaluation are from TI [24] and DB [71]. The data from DB is under the Unsplash license, while the license information for the data from TI is not available online.

Table 3: The prompts used in the quantitative evaluation.

a photo of a [V] [category]
a photo of a [V] [category] in Times Square
a photo of two [V] [category] on a table
a [V] [category] in the jungle
a [V] [category] on a stone wall in the countryside
a [V] [category] on a brick pathway in a garden
a [V] [category] on a pile of fallen leaves in a forest
a [V] [category] at a picnic spot with a checkered blanket
a [V] [category] nestled among rocks
a [V] [category] inside a basket
a [V] [category] inside a metal cage
a [V] [category] drenched in the rainy streets
a [V] [category] in a grassy park with a sunglasses
a [V] [category] floats on the water
a [V] [category] covered by snow
a red [V] [category] wearing bowtie
a purple [V] [category]
a black [V] [category]
a [V] [category] latte art
pencil drawing of a [V] [category]
manga drawing of a [V] [category]
a watercolor painting of a [V] [category]
vector art of a [V] [category]
a painting of a [V] [category] in the style of Monet

Table 4: Quantitative ablation study.

Methods	Identity Preservation \uparrow	Text Alignment \uparrow
W/o Stage 1	0.7031	0.2595
W/o Stage 2	0.7145	0.2541
W/o Stage 3	0.6821	0.2650
W/o Reg	0.7269	0.2502
Full Model	0.7257	0.2532

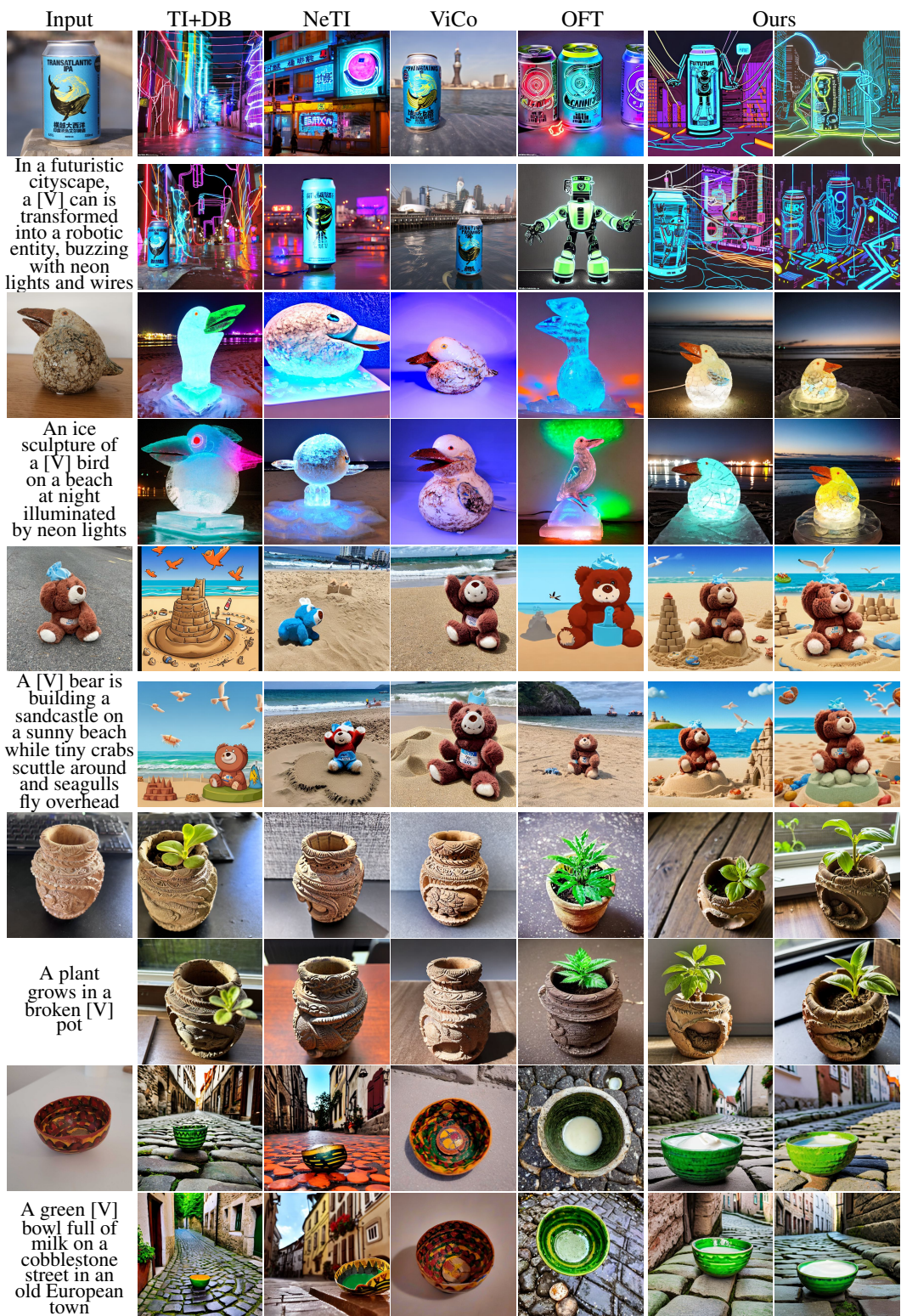


Figure 9: **Additional qualitative comparison.** We present four images generated by our method and two images generated by each of the baseline methods, including TI+DB [24, 71], NeTI [1], ViCo [30], and OFT [64]. Our method demonstrates superior performance in text alignment and identity preservation compared to these baselines.

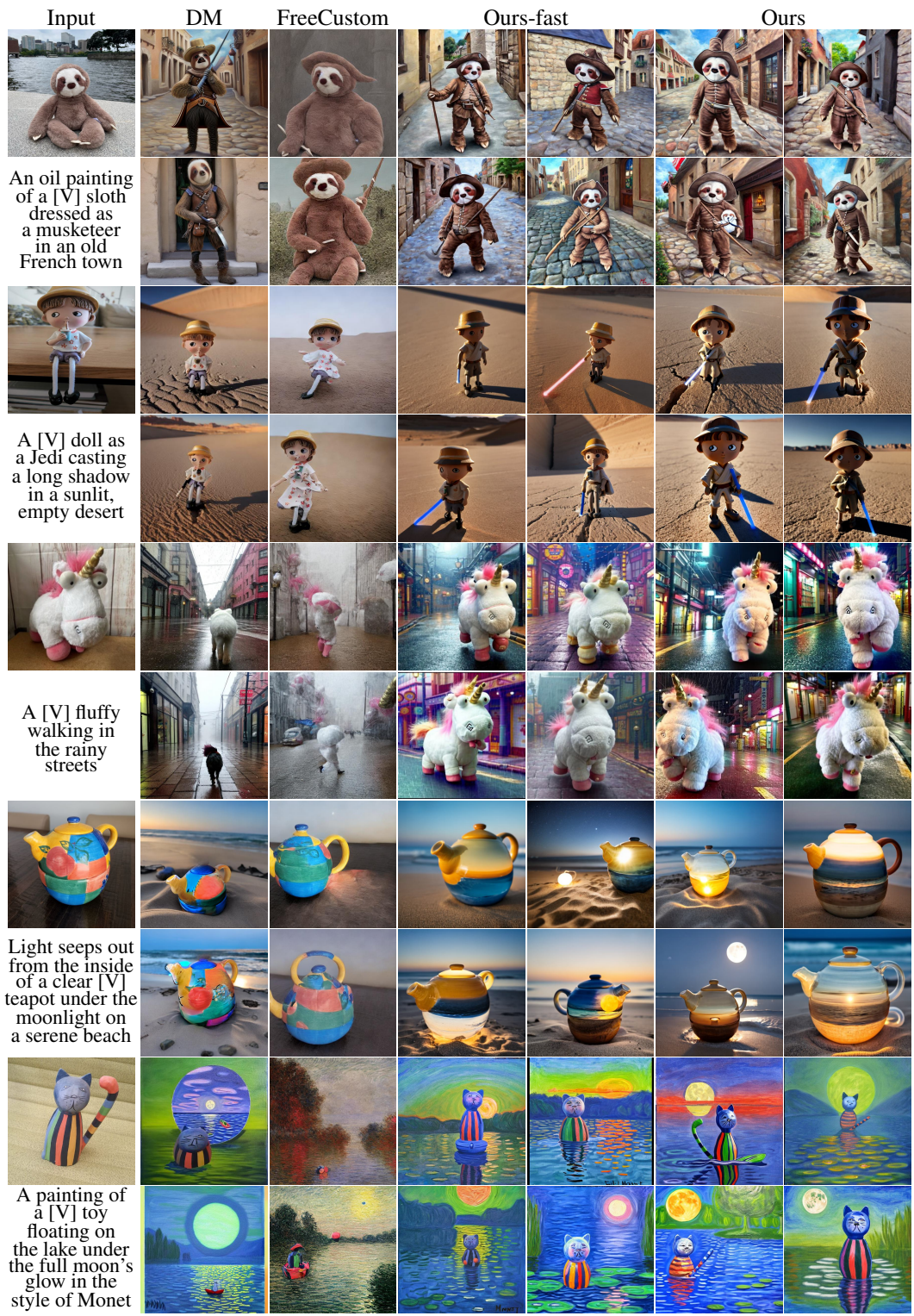


Figure 10: Qualitative results of the fast version of our method, compared with DreamMatcher (DM) [58] and FreeCustom [19].



Figure 11: Qualitative comparison to SuTI [10] and Instruct-Imagen [35] using the examples from their papers.

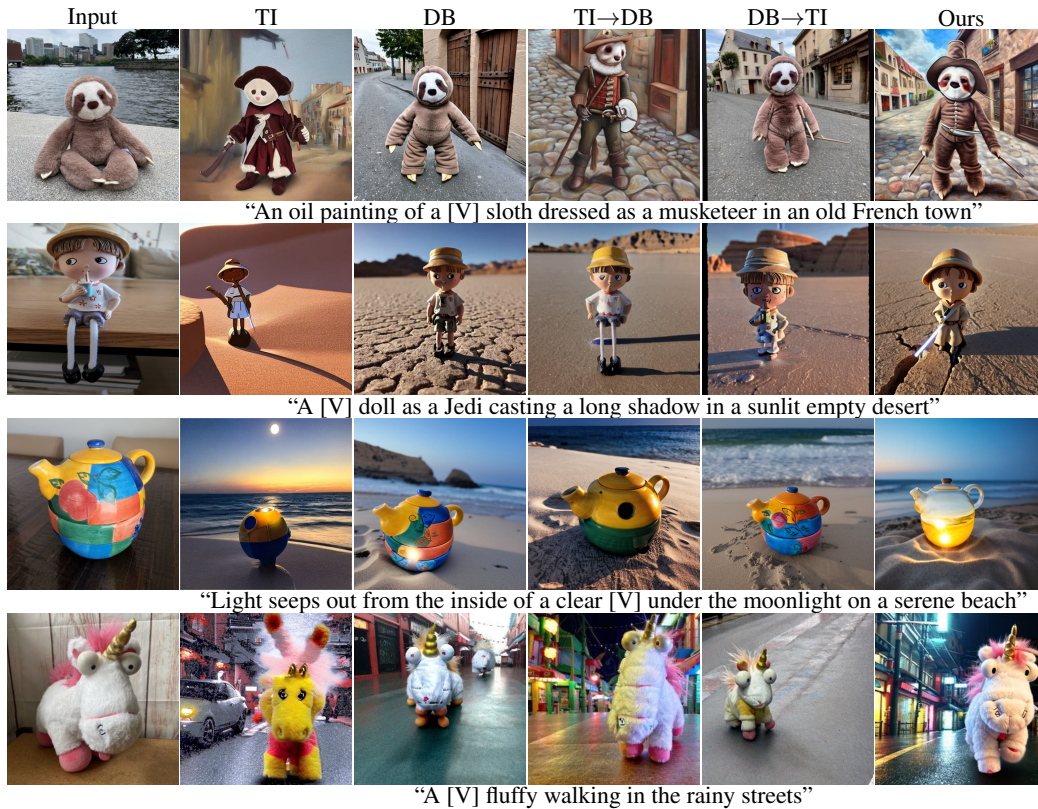


Figure 12: Qualitative comparison of our method with Textual Inversion (TI), DreamBooth (DB), and two different configurations of TI+DB: 1) first TI and then DB (TI→DB), and 2) first DB and then TI (DB→TI).




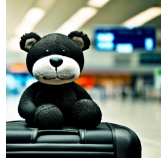

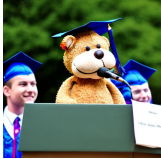












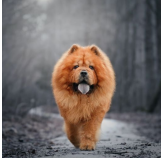



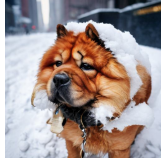



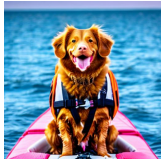
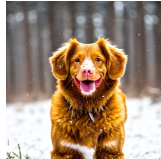
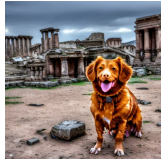
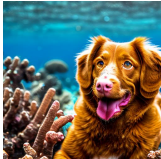
					
Input Sample	A purple [V] bear writing a paper in the conference room	A red [V] bear holding up his accepted paper in the jungle	A black [V] bear sitting on his suitcase at the airport	A [V] bear presenting a poster at a conference with people around	A [V] bear delivering his graduation speech at the podium
					
Input Sample	A cube shaped [V] pot	A [V] pot made out of pure gold with a metallic luster	A clear [V] pot full of milk	A plant grows in a broken [V] pot	Water pouring out of a [V] pot
					
Input Sample	A [V] clock floating on the water with cyberpunk cityscape in the background	A [V] clock in a whimsical, enchanted forest, surrounded by fairies and soft magical light, fantasy illustration	A [V] clock illuminated by the soft glow of a candle, nears a window on a rainy night	A [V] clock embedded in the bark of an ancient oak tree with sunset in the background	A [V] clock in an ancient library with books scattered around
					
Input Sample	A [V] dog surfing on a wave, wearing a floral lei	A [V] dog floating on the water, wearing sunglasses	A wet [V] dog drenched in the rainy streets	A [V] dog covered by snow in New York city	A [V] dog burns in the fire in a burning wood
					
Input Sample	A [V] dog lounging in a hammock on a tropical beach, wearing sunglasses	A [V] dog wearing a life jacket on a boat	A [V] dog caught in a gentle snowfall in a serene winter landscape	A [V] dog amidst the ruins of an ancient, forgotten civilization	A [V] dog nestled within a vibrant coral reef underwater

Figure 13: Additional qualitative results by AttnDreamBooth.



Figure 14: **Single image personalization results.** We present four images generated by our method and two images from each of the baseline methods, including TI+DB [24, 71], NeTI [1], ViCo [30], and OFT [64]. Our method shows better text alignment and identity preservation than the baselines.

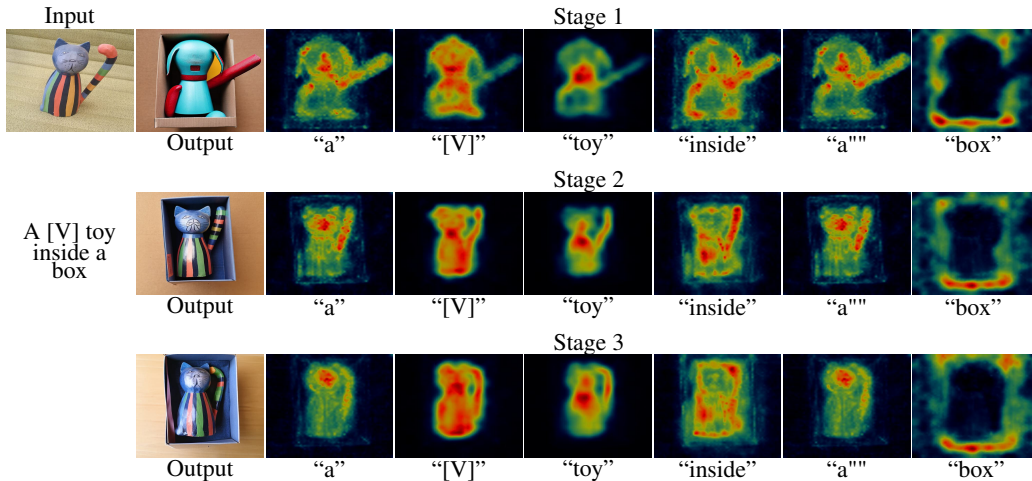


Figure 15: The attention maps for each token after each training stage.



Figure 16: Additional ablation study results.

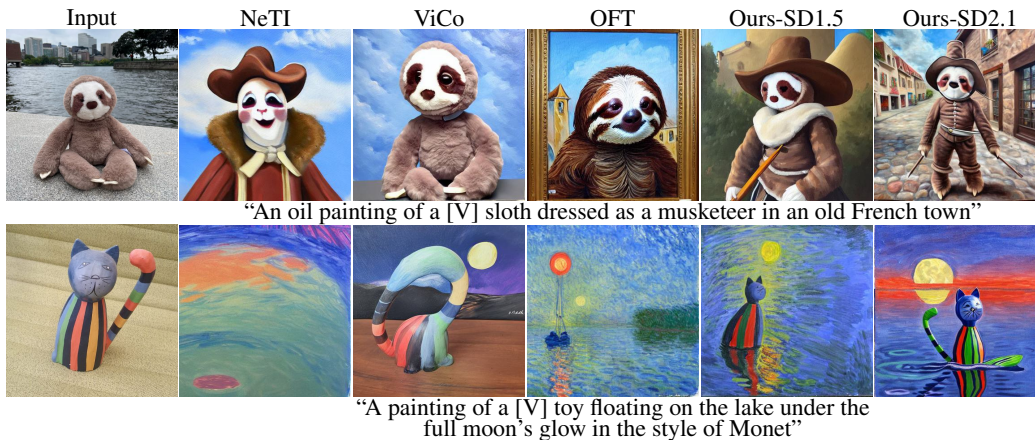


Figure 17: Results of our model using SD1.5.

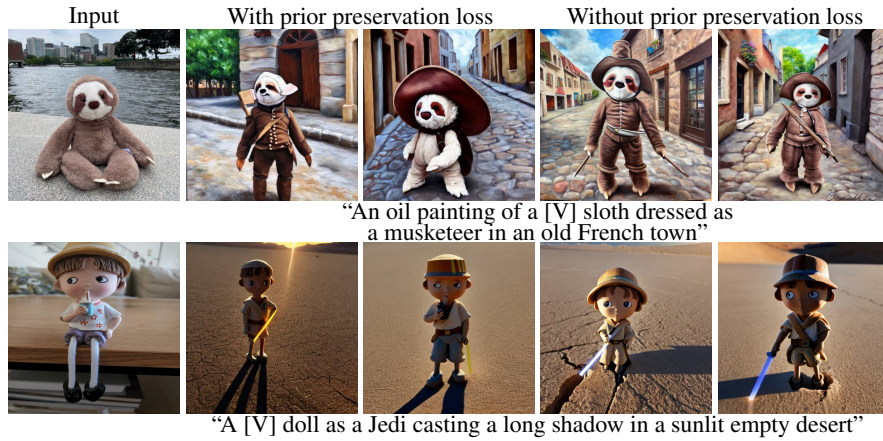


Figure 18: Results of our models with or without the prior preservation loss.

* 1. Concept image (denoted as [V]):



Text prompt: “A [V] burns in the fire”.

Please select the image from the options below that better preserves the identity of the concept image shown above and aligns with the text prompt provided.

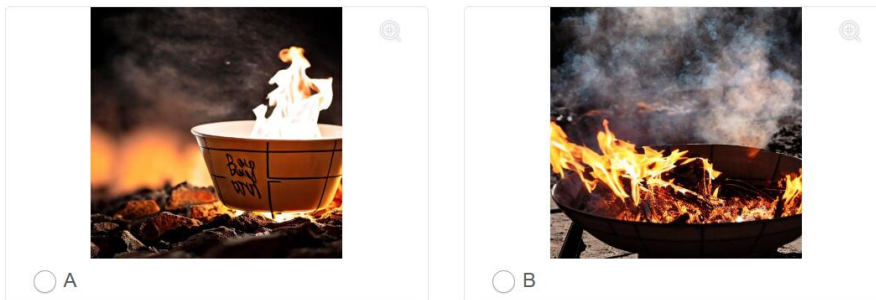


Figure 19: **An example question of the user study.** Given a concept image and a text prompt, along with two generated images, participants are asked to select the image that better preserves the identity of the concept image and aligns with the text prompt.

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: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

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: We have discussed the limitations of our method in Section 6.

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: There is no theory in this 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 provided the implementation details of our method in Section 5.1.

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: We have made the code publicly available with sufficient 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, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have provided the experimental details in Sections 5.1, A, and B.

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: Error bars are not applicable to the experiments conducted in this paper. The qualitative comparison is the most important evaluation for this paper.

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 have provided the information on the computer resources in Section 5.1.

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: Our paper conforms to 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: We have discussed the potential impacts of our work in Section J.

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: [NA]

Justification: We do not release data or pre-trained models.

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: We have discussed the licenses of existing assets in Section K.

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: This 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: [Yes]

Justification: We have provided the full text of instructions given to participants and screenshots in Section I.

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: [No]

Justification: In our country, there is no equivalent organization for 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.