TextBoost: Towards One-Shot Personalization of Text-to-Image Models via Fine-tuning Text Encoder

NaHyeon Park[*]   **Kunhee Kim**[*]   **Hyunjung Shim**

KAIST AI

{julia19, kunhee.kim, kateshim}@kaist.ac.kr

Abstract

Recent breakthroughs in text-to-image models have opened up promising research avenues in personalized image generation, enabling users to create diverse images of a specific subject using natural language prompts. However, existing methods often suffer from performance degradation when given only a single reference image. They tend to overfit the input, producing highly similar outputs regardless of the text prompt. This paper addresses the challenge of one-shot personalization by mitigating overfitting, enabling the creation of controllable images through text prompts. Specifically, we propose a selective fine-tuning strategy that focuses on the text encoder. Furthermore, we introduce three key techniques to enhance personalization performance: (1) augmentation tokens to encourage feature disentanglement and alleviate overfitting, (2) a knowledge-preservation loss to reduce language drift and promote generalizability across diverse prompts, and (3) SNR-weighted sampling for efficient training. Extensive experiments demonstrate that our approach efficiently generates high-quality, diverse images using only a single reference image while significantly reducing memory and storage requirements.

Project page — https://textboost.github.io

Introduction

Recent breakthroughs in large-scale text-to-image models (Nichol et al. 2023; Podell et al. 2024; Ramesh et al. 2022; Rombach et al. 2022; Saharia et al. 2022; Sauer et al. 2024) have opened up the new era of image generation, enabling the creation of diverse and imaginative visuals with natural language prompts. Building on this success, a burgeoning area of research focuses on *personalization* (Ruiz et al. 2023; Gal et al. 2023), which allows users to customize these models to generate images featuring specific concepts.

However, existing personalization approaches (Gal et al. 2023; Gu et al. 2023; He et al. 2023; Kumari et al. 2023; Ruiz et al. 2023; Wang et al. 2023) still require at least 3 to 5 reference images to produce high-quality outputs. When only a single reference image is provided, these methods often struggle to reflect the user's text prompt effectively, reproducing almost identical images regardless of the text

input (see Figure 4). This limitation undermines their practicality in real-world scenarios where users frequently want to customize images using just one reference photo. For example, users may wish to modify a cherished photograph – such as a picture of a grandmother, a childhood memory, or a unique drawing – by applying creative text prompts to both preserve and enrich their memories.

In this paper, we aim to achieve high-quality personalization using only a *single* reference image. We propose a novel approach focusing exclusively on fine-tuning the text encoder, different from existing approaches that directly fine-tune the image module (Ruiz et al. 2023; Kumari et al. 2023; Chen et al. 2024a; Zhang et al. 2024c). Our idea is motivated by the observation that the text encoder' weights exhibit the most significant changes during fine-tuning (see Figure 1). Furthermore, we introduce three innovative techniques tailored for image customization: (1) introduction of augmentation token to reduce overfitting and promote disentanglement of subject-relevant and irrelevant features, (2) knowledge preservation loss to prevent text encoder from language drift, thus preserving its general capability across diverse text prompts, and (3) SNR-weighted timestep sampling to further facilitate efficient training. Since our method significantly enhances one-shot personalization performance through the exclusive fine-tuning of the text encoder, we name our approach **TextBoost**. Practical experiments demonstrate that our approach effectively achieves high-quality personalization and creative control through text prompts, utilizing solely a single reference image across diverse real-world applications.

To summarize, our contributions are as follows:

- To the best of our knowledge, we are the first to exclusively fine-tune the text encoder for one-shot text-to-image personalization.

- We propose three novel techniques – paired data augmentation, knowledge-preservation loss, and SNR-weighted sampling – to enhance personalization capability.

- We demonstrate our method's effectiveness in generating high-quality, diverse outputs across various prompts and datasets, using only a single reference image as input.

- Our approach is memory and storage-efficient, requiring only 0.7M parameters and 5.1MB of storage, making it applicable to a broader range of real-world applications.

---

[*]These authors contributed equally.

Background

Text-to-Image Diffusion Models

Diffusion models (Sohl-Dickstein et al. 2015) have recently become the most widely adopted generative models for text-to-image generation (Saharia et al. 2022; Ramesh et al. 2022; Rombach et al. 2022). These models aim to closely approximate the original data distribution $q(\mathbf{x}_0)$ with $p_\theta(\mathbf{x}_0)$. Here, $p_\theta(\mathbf{x}_0) := \int p_\theta(\mathbf{x}_{0:T})$, where $p_\theta(\mathbf{x}_{0:T})$ is termed as the reverse process being Markov chain with learned Gaussian transitions. The approximate posterior $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ termed a forward process, of which the noise is gradually added to the original data point $\mathbf{x}_0$ as $\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1 - \alpha_t}\epsilon$, can be expressed as $q(\mathbf{x}_{1:T}|\mathbf{x}_0) := \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$. The training objective is formulated with variational bound on negative log-likelihood:

$$\mathbb{E}[-\log p_\theta(\mathbf{x}_0)] \leq \mathbb{E}_q[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}] := L, \quad (1)$$

which is further simplified into the following objective:

$$L_{simple}(\theta) := \mathbb{E}_{t,\mathbf{x}_0,\epsilon}[||\epsilon - \epsilon_\theta(\mathbf{x}_t, t)||_2^2]. \quad (2)$$

Text-to-image (T2I) diffusion models (Rombach et al. 2022) take text condition $\mathbf{c}$ as additional input, where the text prompt y is encoded with text encoder $\mathcal{E}_T$ as $\mathbf{c} = \mathcal{E}_T(y)$. These models usually work in latent space by using encoder $\mathcal{E}_I$ to encode image x to latent $z = \mathcal{E}_I(x)$. The training objective hence takes an additional $\mathbf{c}$ and minimizes the following loss:

$$\mathbb{E}_{z,t,y,\epsilon}[||\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})||_2^2]. \quad (3)$$

Although the same concept can be applied to various text-to-image diffusion models, this paper utilizes Stable Diffusion (Rombach et al. 2022) primarily due to its public accessibility. Note that the U-Net (Ronneberger, Fischer, and Brox 2015) architecture is commonly employed for the diffusion module, with the text condition incorporated through the cross-attention layers.

Personalized Text-to-Image Generation

Personalization aims to create images that align with user-provided text descriptions, using a small set of reference images supplied by the user (Zhang et al. 2024b). The most well-known methods are Textual Inversion (Gal et al. 2023) and DreamBooth (Ruiz et al. 2023). In both works, a unique identifier V^* is used to represent the subject to customize and optimize the reconstruction objective (Eq. 3). However, what to fine-tune is different. Textual Inversion optimizes the text embedding of V^* to represent the subject in the latent space, while DreamBooth fine-tunes the diffusion model's weights using reconstruction loss and class-specific prior-preservation loss. These two approaches each have their advantages and disadvantages. Textual Inversion is lightweight, as it only optimizes the text embeddings; however, the generated outputs often suffer from low subject fidelity. On the other hand, DreamBooth produces high-quality images by fine-tuning the entire image generation module but requires significant storage space to save the U-Net parameters for each subject of interest. Various studies have been conducted related to these two methods. XTI (Voynov et al. 2023) introduces an extended textual conditioning space, and NeTI (Alaluf et al. 2023) further extends this by incorporating timesteps. Custom Diffusion (Kumari et al. 2023) has gained significant attention as it fine-tunes only the key and value projection matrices in the U-Net's cross-attention layers, thereby requiring much fewer parameters to fine-tune while still achieving high-quality generation. Therefore, in later sections, we compare our approach with Custom Diffusion. Additionally, SVD-iff (Han et al. 2023) fine-tunes the singular values of model weights, and Perfusion (Tewel et al. 2023) introduces a key-locking strategy along with rank-1 editing. StyleDrop (Sohn et al. 2023) focuses on style generation using adapters with iterative feedback. Other works (Liu et al. 2023; Chen et al. 2024a; Zhang et al. 2024a,c) address the challenges of compositional generation. Recently, another line of approaches has explored learning-based strategies, where additional networks are pre-trained for fast inference (Chen et al. 2024b; Hua et al. 2023; Li, Li, and Hoi 2023; Wei et al. 2023; Xiao et al. 2023; Sauer et al. 2024; Shi et al. 2024). However, these techniques often require a massive amount of data and extensive pre-training of the encoder. In this work, we take an optimization-based approach, focusing on fine-tuning the text encoder with simple yet novel techniques to enhance performance for one-shot personalization.

Need for Fine-tuning Text Encoder

Limitation of Existing Methods

Existing approaches typically require at least 3 to 5 images to produce high-quality output. However, our work aims to achieve high-quality personalization using only a single reference image. Notably, Ruiz et al. (2023) has reported a rapid decline in quality when fewer than 3 images are provided. Similarly, Gal et al. (2023) also demonstrates that personalization requires 3 to 5 samples. This is because training a U-Net on a single data point is akin to fitting a diffusion model to a single data point, inevitably leading to overfitting. As a result, the customized model forgets prior distribution, fails to effectively interpret the user's text prompt, and consistently generates images that closely resemble the reference image with little variation.

Instead of direct fine-tuning, recent studies, such as Custom Diffusion (Kumari et al. 2023), have explored fine-tuning specific portions of the U-Net. However, even partial fine-tuning leads to overfitting due to the inherent difficulty of the task, especially when only a single reference image is available (see Figure 4).

Another approach to address overfitting incorporates regularization dataset during training (Ruiz et al. 2023). However, this method requires generating or retrieving samples of 100 to 200, which is computationally intensive. Additionally, this regularization effectively doubles the batch size, significantly increasing GPU memory requirements.
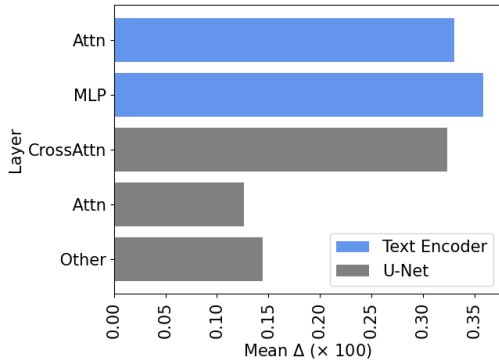
Figure 1: **Change in weights of different layers during fine-tuning.** The mean weight change of text encoder layers is relatively greater than that of U-Net parameters.

Figure 2: **Method overview.** We selectively fine-tune text encoder for one-shot personalization. We utilze three novel techniques to further boost the personalization performance.

Personalizing the Text Encoder

In the previous section, we discussed that direct fine-tuning of the image module (i.e., U-Net parameters) leads to overfitting. Then, what module should we fine-tune to both mitigate overfitting and achieve high-quality personalization at the same time? We begin by assuming that a foundation model (e.g., Stable Diffusion) already encompasses the distribution of the samples we aim to generate, as these models are trained to cover a diverse range of image distributions. Therefore, identifying the proper conditioning input to the foundation model should produce personalized samples, since personalization requires the model to capture the unique characteristics of the subject by associating a unique identifier, V^*, with the text prompt. Unfortunately, crafting a precise prompt to describe a specific image is challenging. Textual Inversion attempted to find a single token embedding to represent this, but it is time-consuming and lacks expressiveness.

Given this, we must reconsider which part of the text-to-image model should be fine-tuned. Kumari et al. (2023) have justified their focus on U-Net fine-tuning based on their observations that the parameters of the cross-attention layers within the U-Net change the most during fine-tuning (Li et al. 2020). However, their analysis is limited to the U-Net, without considering potential changes in the text encoder's weights. In response, we expand the investigation to include the text encoder layers, re-examining the fine-tuning process. Precisely, we measure the change in parameters for each module using the following equation: $\Delta = \frac{||\hat{\theta} - \theta||}{||\theta||}$, where θ and $\hat{\theta}$ represent the original and fine-tuned weights, respectively, with the reconstruction loss (Eq. 3). Consistent with Custom Diffusion's findings, we observe that the cross-attention layer parameters of the U-Net exhibit relatively larger changes compared to other layers.

More interestingly, as illustrated in Figure 1, the text encoder's weights undergo the most significant changes, even surpassing those observed in the U-Net. This finding strongly suggests that fine-tuning the text encoder is crucial, as its parameters play a pivotal role in personalized model
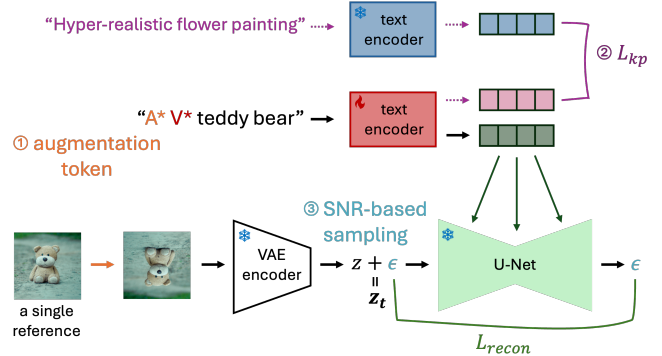
adaptation. To the best of our knowledge, our work is the first to focus exclusively on fine-tuning the text encoder for customized text-to-image generation.

Furthermore, to enhance parameter efficiency, we adopt the Low-Rank Adaptation (LoRA) technique (Hu et al. 2022), appending the adapter to the attention and fully-connected layers of the text encoder. This approach offers significant advantages as it requires much fewer parameters than existing methods. Specifically, our method reduces the number of trainable parameters and the storage space needed for each customized model, making it more practical. Our approach significantly reduces the number of parameters to 0.7 M, which is substantially lower than DreamBooth (865.9 M) and Custom Diffusion (19.2 M).

Method

In this section, we propose three novel techniques to fine-tune the text encoder for one-shot personalization. First, we use paired data augmentation by introducing an augmentation token to encourage the disentanglement of subject-relevant and subject-irrelevant features. Second, we introduce a knowledge-preservation loss to mitigate language drift and maintain the generalization capability of the text encoder across diverse prompts. Lastly, we implement SNR-weighted sampling to enhance training efficiency.

Augmentation Token

Fine-tuning the text encoder alleviates direct memorization of the input at a pixel level. However, to disentangle the subject from its background and induce the text encoder to learn subject-relevant features, we incorporate appropriate data augmentation techniques to further enhance generalization capability.

The most naive approach is to apply data augmentation to the image itself. During training, a data-augmented input image is fed with the text prompt, 'A photo of V^*' Despite being simple, we find that the resulting images suffer from distorted output due to *augmentation leaking*, which is known to be a common problem when training generative models with data augmentation (Karras et al. 2020). We speculate

that this is due to the binding of augmentation into the subject token. For example, when vertical-flip augmentation is applied during training, the model is likely to learn to map both the subject and vertically flipped subject into the subject token V^*. Examples of generated images are available in Appendix A.1.

To this end, we craft precise text prompts describing the desired augmentations. This approach is also partially employed in Kumari et al. (2023), where they involve 'resize' augmentation: appending phrases like 'zoomed in' or 'close up' for 1.2-1.4× scaling and 'far away' or 'very small' for 0.4-0.6× scaling. This approach has a noticeable effect in successfully mitigating augmentation leaking. However, the augmentation prompt needs to be manually selected with care for each augmentation, and has to be ensured that the vocabularies in the selected phrase do not interfere with the model's original embedding.

To allow for more versatility, we suggest a novel method, which is to automatically map the applied transformation to the augmentation token A^*. For instance, a sample image of a dog rotated by 90 degrees would be accompanied by the prompt 'A^* photo of V^*'. Through this approach, we learn to associate A^* for corresponding type of image transformation applied to the input, inducing the model to disentangle V^* from subject-irrelevant information. Therefore, during training, we jointly optimize V^* and A^*, while using token V^* for customized generation at inference time. We interpret this effect as being similar to that of DistAug (Jun et al. 2020). Please refer to Appendix A.2 for further details on the implementation and analysis.

Knowledge Preservation Loss

Fine-tuning a pre-trained model for a specific task is known to cause a problem called *language drift* (Lee, Cho, and Kiela 2019), a phenomenon where the fine-tuned model gradually loses the syntactic and semantic properties of the language. To mitigate this issue, DreamBooth (Ruiz et al. 2023) introduces a class-specific prior preservation loss by adding a reconstruction loss term on the predicted noise from diffusion models, using 100 to 200 images generated from a pre-trained model.

The significant difference in our work is that we fine-tune the text encoder. Accordingly, we devise a novel knowledge preservation loss to ensure that the online text encoder retains its prior knowledge. We achieve this by using the cosine similarity between the text embeddings of the pre-trained and fine-tuned text encoders, similar to studies that used feature distance to mitigate language drift (Rolnick et al. 2019; Kang et al. 2023).

Given a set of text prompts $\Gamma = \{\gamma_1, \gamma_2, ..., \gamma_n\}$ and text embeddings from the pre-trained text encoder $\mathcal{E}_T(\gamma_i)$ and the online text encoder $\hat{\mathcal{E}}_T(\gamma_i)$,

$$\mathcal{L}_{kp} = \text{sim}(\mathcal{E}_T(\gamma_i), \hat{\mathcal{E}}_T(\gamma_i)) = \frac{\mathcal{E}_T(\gamma_i) \cdot \hat{\mathcal{E}}_T(\gamma_i)}{|\mathcal{E}_T(\gamma_i)| \, |\hat{\mathcal{E}}_T(\gamma_i)|} \quad (4)$$

where Γ can be any set of text prompts. For our experiments, we used captions from InstructPix2Pix (Brooks, Holynski, and Efros 2023), as it is human-written text.

Our knowledge preservation loss offers two practical advantages. First, it eliminates the need to generate prior images during fine-tuning, reducing overall training time compared to class prior preservation loss. Second, it avoids the requirement for additional GPU memory to store batches of class-prior images, allowing for more efficient memory utilization during training.

SNR-Weighted Timestep Sampling

To further enhance training, we propose a timestep sampling method based on the signal-to-noise ratio (SNR) of noisy input $\mathbf{x}_t$. During the fine-tuning of text-to-image diffusion models, the model predicts the added noise for a given noisy image $\mathbf{x}_t$, where the timestep t is traditionally sampled from a uniform distribution (Ho, Jain, and Abbeel 2020). However, we hypothesize that the impact of the text embedding on the model's prediction varies depending on the specific timestep and conducted a small experiment to test this hypothesis. As expected, given two distinct text prompts (e.g., 'A photo of a dog' and 'A photo of a cat'), the network predictions (ϵ_t) are noticeably different at high noise levels, where the model is primarily constructing the content of the image (Choi et al. 2022). Conversely, at low noise levels, where only minimal noise is added to the input $\mathbf{x}_t$, the model primarily focuses on denoising the remaining noise, resulting in the text prompt having a reduced impact. To summarize, we find out that the impact of text prompts on the network output is proportional to the input's noise level.

Based on this observation, during the fine-tuning of the text encoder, we sample the timestep from the categorical distribution, with normalized probability $p \propto -\log(\text{SNR})$, where $\text{SNR}(t) = \alpha_t^2/\sigma_t^2$ for any $t \in [0, 1]$ with distribution of noisy input being $q(\mathbf{x}_t|\mathbf{x}) = \mathcal{N}(\alpha_t \mathbf{x}, \sigma_t^2 \mathbf{I})$ (Kingma et al. 2021).

We find that this timestep sampling strategy noticeably enhances the output image quality. More details of this section can be found in Appendix B.

Overall Training

To sum up, we train our text encoder with the following objective:

$$\mathcal{L} = \mathbb{E}_{z,t,y,\epsilon}[||\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \hat{\mathcal{E}}_T(y))||_2^2] + \lambda \mathcal{L}_{kp}. \quad (5)$$

Here, text prompt y is sampled from a pre-defined set such as 'A^* photo of V^* with background', and γ_i is a text sequence from regularization dataset Γ. It is important to note that the subject token V^* and augmentation token A^* are jointly optimized together in order to minimize the objective above. Following relevant studies that use LoRA for text-to-image model fine-tuning (Fan et al. 2023; Black et al. 2024; Chen et al. 2024a), we use rank 4, and set λ to 0.1.

Experiments

Experimental Setup

Dataset and evaluation metrics. We employed the benchmark introduced by Ruiz et al. (2023), which comprises 30 subjects and 25 text prompts, with each subject

Methods	CLIP-T ↑	CLIP-I ↑ (seen)	CLIP-I ↑ (unseen)	# Params ↓	GPU M ↓ (batch size=8)	Storage ↓ (per concept)
BLIP Diffusion	0.608	0.787	0.743	-	-	-
Textual Inversion	0.460	0.726	0.678	**0.0 M**	**21.6 GB**	**3.2 KB**
DreamBooth	0.572	0.827	0.764	865.9 M	48.9 GB	3.3 GB
DreamBooth-LoRA	0.573	0.771	0.721	0.8 M	33.7 GB	3.3 MB
Custom Diffusion	0.538	**0.866**	**0.795**	19.2 M	34.4 GB	74 MB
TextBoost (ours)	**0.628**	0.767	0.728	0.7 M	22.6 GB	5.1 MB

Table 1: **Quantitative comparison on Stable Diffusion v1.5.** We measure CLIP-T scores for image-text fidelity. CLIP-I scores are evaluated with the reference images that are used for training (seen) and remaining reference images (unseen). For practicality, we compare the number of parameters, required memory for training, and storage to save customized model.

Methods	Image & Text Fidelity
Textual Inversion	2.8 %
DreamBooth	23.95 %
Custom Diffusion	20.6 %
TextBoost (ours)	**52.65 %**

Table 2: **User study.** To assess real-world user preferences, we ask 100 participants to choose the image that best resembles the given subject and matches the provided text prompt. Each participant answers 20 questions through Amazon Mechanical Turk, resulting in a total of 2,000 responses.

Figure 3: **Qualitative comparison on Stable Diffusion v1.5.** We compare images generated by each method using various types of text prompts on different subjects. All models are trained using a *single* reference image.

Methods	CLIP-T ↑	CLIP-I ↑ (seen)	CLIP-I ↑ (unseen)	# Params ↓	GPU M ↓ (batch size=2)	Storage ↓ (per concept)
DreamBooth	0.656	0.777	0.732	865.9 M	35.7 GB	3.3 GB
Custom Diffusion	0.609	**0.845**	**0.780**	19.2 M	23.7 GB	98 MB
TextBoost (ours)	**0.675**	0.783	0.739	**1.7 M**	**17.4 GB**	**6.6 MB**

Table 3: **Quantitative comparison on Stable Diffusion v2.1.** We evaluate our method on different version of Stable Diffusion.

having 4 to 6 associated images. To test applicability in diverse datasets, we also incorporate reference images from Custom Diffusion (Kumari et al. 2023) for qualitative results. We trained all of the models using a single reference image. Following the DreamBooth, we generated 4 images per prompt (3,000 images in total) for evaluation.

We assessed subject fidelity by reporting the similarity between the generated image and reference image(s) in the CLIP image embedding space (CLIP-I). Our evaluation includes results for both 'seen' images (a reference image used for training) and 'unseen' images (subject images not used as training data). To assess text-image alignment, we cal-culated the similarity between each generated image and its corresponding prompt (CLIP-T). Since the CLIP-I score does not evaluate the extent of overfitting, memorization of reference images can result in inflated scores. Considering this, we assess the inter-similarity between generated images using the DINOv2 (Oquab et al. 2024) score. Additionally, to assess the practical applicability of our method, we conducted a large-scale user study involving 100 participants on Amazon Mechanical Turk to evaluate user preferences.

Baselines. We compare our approach with three widely-used personalization models: DreamBooth (Ruiz et al. 2023), Textual Inversion (Gal et al. 2023), and Custom Diffusion (Kumari et al. 2023). For reference, we also report the result from the zero-shot personalization method (Li, Li, and Hoi 2023), which requires two-stage pre-training beforehand.

Implementation details. We utilize the Stable Diffusion v1.5[1] text-to-image model, integrating the text encoder from the CLIP ViT-L/14 model (Radford et al. 2021). All of our
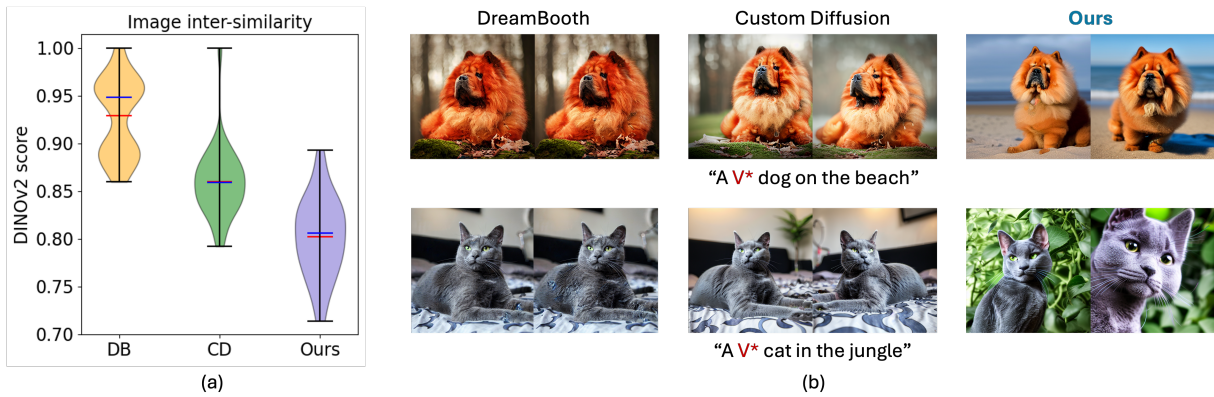
---
[1]https://github.com/runwayml/stable-diffusion

Figure 4: **Diversity comparison.** (a) We calculate the inter-similarity of 100 generated images using the DINOv2 score and plot the distribution, given the same reference image and identical prompts. Blue and red horizontal lines indicate the median and mean of each distribution, respectively. (b) Qualitative examples of each method, with two subjects, each with two images per prompt. Note that for a fair comparison, the random seeds are fixed.

experiments were conducted in a one-shot setting. Our text encoder is trained using the AdamW optimizer (Loshchilov and Hutter 2019) with a learning rate of 5e-3 for text encoder and 1e-2 for V^* token for 250 steps. We also employed a linear decay learning rate for the V^* token. For augmentations, we employ a number of color and geometric augmentations such as grayscale conversion, brightness adjustment, vertical flip, crop, and cutout. All experiments were conducted with batch size 8 on a single NVIDIA A6000 GPU.

Results

We compare our method with several other methods, both quantitatively and qualitatively. For quantitative evaluation, we measure CLIP-I and T scores and conduct a user study to assess real-world preferences. Additionally, we evaluate the diversity of generated images to determine whether each method can produce a variety of outputs, potentially satisfying user needs. We also conduct an ablation study to confirm the effectiveness of the three novel techniques used in training our text encoder. Moreover, we demonstrate our method in broader application scenarios to showcase its practicality.

Quantitative comparison. We measured various quantitative metrics, and the results are presented in Table 1. Our method achieves performance comparable to existing approaches, although Custom Diffusion exhibits higher CLIP-I scores than the others. We speculate that this may be due to the high diversity of our generated outputs, which we evaluate in later sections. Another important aspect is text fidelity, which we evaluated using the CLIP-T score. As shown in the table, our method noticeably outperforms existing approaches, indicating superior text-based editing performances. We also test our method on different versions of Stable Diffusion (v2.1), where results show similar tendency (Table 3)

Computational efficiency. We emphasize the practical advantages of our method, which demands significantly fewer trainable parameters and GPU memory compared to

other approaches, as detailed in Table 1. Moreover, personalization presents storage challenges, as the fine-tuned model should be saved for each concept. While Textual Inversion's storage efficiency is notable due to its focus on text embedding fine-tuning, its limited performance hinders practical application, evidenced by scores of the table and qualitative results in later sections. On the other hand, DreamBooth requires the storage of all U-Net parameters, consuming a substantial 3.3 GB. Custom Diffusion reduces this to 74 MB, but our method further optimizes storage by requiring only 5.1 MB — a mere 0.15 % and 6.89 % of DreamBooth and Custom Diffusion, respectively. This compact size allows our model to be seamlessly stored in constrained environments like the cloud or portable storage devices.

User study. In real-world scenarios, users need to select the output image that both satisfies (1) subject fidelity and (2) alignment between the text prompt and the generated image. To evaluate existing methods in this context and understand user preferences, we conducted a large-scale user study. Participants were tasked with selecting the best image from multiple options generated by different methods. We employed diverse subjects and text prompts sourced from DreamBench (Ruiz et al. 2023) (templates in Appendix C). To ensure fairness, random seeds were fixed, and image order was randomized for each question. Through Amazon Mechanical Turk, 100 participants completed 20 questions each, yielding a total of 2,000 responses. As indicated in Table 2, our method was favored by 52.65% of users, demonstrating its superior ability to meet user demands for subject fidelity and text-image alignment in practical settings.

Qualitative results. We evaluated our method's performance across diverse text prompts and subjects, visualizing the generated images for each method in Figure 3. Our method effectively modifies object properties, accurately applying colors to specified subject (e.g., 'colorful V^* dog', 'red V^* bowl'), as illustrated in the images of the first row. It also adeptly generates images across various scenes (e.g.,

input text: "A V* dog in the jungle"
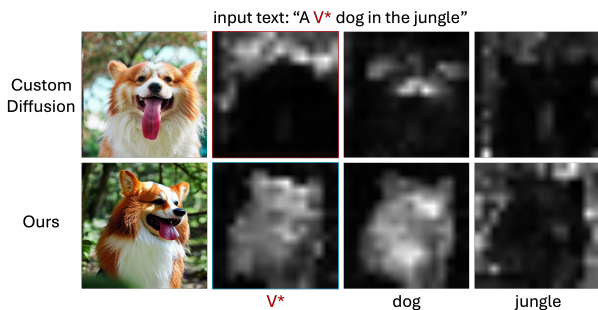
Custom Diffusion

Ours

V*   dog   jungle

Figure 5: **Comparison of generated attention maps.** We compare cross-attention maps of Custom Diffusion and our method. Our approach successfully disentangles subject-relevant information from irrelevant details.

Aug.	KPL	WS	CLIP-T	CLIP-I
✗	✓	✓	0.625	0.703
✓	✗	✓	**0.632**	0.723
✓	✓	✗	0.623	0.727
✓	✓	✓	0.628	**0.728**

Table 4: **Ablation study.** Aug., KPL, and WS denote our paired data augmentation, knowledge preservation loss, and weighted sampling, respectively.

'night sky', 'ice in the ocean', 'in the city'), as demonstrated in the left image of the second row and the images in the last row. Furthermore, our method produces images in diverse styles (e.g, 'watercolor painting' and 'drawing'). Qualitative comparisons reveal that TextBoost excels at simultaneously preserving subject fidelity and adhering to text prompts, outperforming other methods.

Diversity. To evaluate the degree of overfitting that leads to reduced diversity in outputs, we measure output diversity using the inter-similarity between generated images with the DINOv2 score (Figure 4 (a)). We also examine qualitative outputs across different subjects and prompts in (b). Compared to other two methods, the results confirm that our approach produces highly diverse images with varied poses, effectively incorporating the specified text prompts.

Disentanglement. To verify the effectiveness of our method in disentangling subject-irrelevant information (e.g., background) from subject-relevant features, we analyze the attention maps of corresponding tokens. As depicted in Figure 5, the attention maps for the identifier token V^* reveal a clear contrast between our method and Custom Diffusion. While Custom Diffusion's V^* token exhibits a strong focus on irrelevant background information, our method demonstrates that V^* accurately attends to subject-relevant features.

Ablation study

To verify the effectiveness of the components in our method, we conducted an ablation study, with the results presented in Table 4. Additionally, to assess whether the augmentation

V* dog

<v-flip> V* dog

Figure 6: **Ablation on augmentation token.** To test whether the augmentation token has learned the corresponding augmentation, we generate images with and without the augmentation token as the input prompt. We showcase a vertical flip as an example of intuitive visualization.

tokens were learned effectively, we performed an ablation specifically on the augmentation token. In this analysis, we used a vertical flip to visually and intuitively evaluate its effectiveness. Examples are provided in Figure 6, where the dog, prompted with the augmentation token, generates vertically flipped dog images, confirming the token's intended effect.

Stylization

To test whether our method is applicable to style personalization, we conducted the following experiment. We employed a single comprehensive caption, following the approach of Sohn et al. (2023) using Stable Diffusion v2.1. As illustrated in Figure 7, our method successfully learns the style from just a single image.

"A cat in V* style"

Figure 7: **Stylization.** We use a single style image (bottom left) as a reference to generate customized images.

Conclusion

In this paper, we aimed to develop high-quality personalized text-to-image generation method that enables creative control through text prompts, with a single reference image. Our TextBoost, which focuses on fine-tuning the text encoder with innovative training methods, effectively mitigates overfitting and delivers superior quality, particularly in text control. We believe our approach paves the way for tailored text-to-image generation, making one-shot personalization a practical reality in various real-world applications.

Acknowledgments

References

Alaluf, Y.; Richardson, E.; Metzer, G.; and Cohen-Or, D. 2023. A Neural Space-Time Representation for Text-to-Image Personalization. *ACM Transactions on Graphics (TOG)*, 42.

Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; Catanzaro, B.; Karras, T.; and Liu, M.-Y. 2022. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. ArXiv:2211.01324 [cs].

Black, K.; Janner, M.; Du, Y.; Kostrikov, I.; and Levine, S. 2024. Training Diffusion Models with Reinforcement Learning.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instruct-Pix2Pix: Learning to Follow Image Editing Instructions. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Chen, H.; Zhang, Y.; Wu, S.; Wang, X.; Duan, X.; Zhou, Y.; and Zhu, W. 2024a. DisenBooth: Identity-Preserving Disentangled Tuning for Subject-Driven Text-to-Image Generation. In *International Conference on Learning Representations*.

Chen, W.; Hu, H.; Li, Y.; Ruiz, N.; Jia, X.; Chang, M.-W.; and Cohen, W. W. 2024b. Subject-driven Text-to-Image Generation via Apprenticeship Learning. In *Advances in Neural Information Processing Systems*. ArXiv:2304.00186 [cs].

Choi, J.; Lee, J.; Shin, C.; Kim, S.; Kim, H.; and Yoon, S. 2022. Perception Prioritized Training of Diffusion Models. In *CVPR*. ArXiv:2204.00227 [cs].

Fan, Y.; Watkins, O.; Du, Y.; Liu, H.; Ryu, M.; Boutilier, C.; Abbeel, P.; Ghavamzadeh, M.; Lee, K.; and Lee, K. 2023. DPOK: Reinforcement Learning for Fine-tuning Text-to-Image Diffusion Models.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *International Conference on Learning Representations*.

Gu, Y.; Wang, X.; Wu, J. Z.; Shi, Y.; Chen, Y.; Fan, Z.; Xiao, W.; Zhao, R.; Chang, S.; Wu, W.; Ge, Y.; Shan, Y.; and Shou, M. Z. 2023. Mix-of-Show: Decentralized Low-Rank Adaptation for Multi-Concept Customization of Diffusion Models. In *Advances in Neural Information Processing Systems*.

Han, L.; Li, Y.; Zhang, H.; Milanfar, P.; Metaxas, D.; and Yang, F. 2023. SVDiff: Compact Parameter Space for Diffusion Fine-Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, X.; Cao, Z.; Kolkin, N.; Yu, L.; Wan, K.; Rhodin, H.; and Kalarot, R. 2023. A Data Perspective on Enhanced Identity Preservation for Diffusion Personalization. ArXiv:2311.04315 [cs].

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*. ArXiv: 2006.11239.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models.

Hua, M.; Liu, J.; Ding, F.; Liu, W.; Wu, J.; and He, Q. 2023. DreamTuner: Single Image is Enough for Subject-Driven Generation. ArXiv:2312.13691 [cs].

Jun, H.; Child, R.; Chen, M.; Schulman, J.; Ramesh, A.; Radford, A.; and Sutskever, I. 2020. Distribution Augmentation for Generative Modeling. In *Proceedings of the 37th International Conference on Machine Learning*, 5006–5019. PMLR.

Kang, M.; Zhang, J.; Zhang, J.; Wang, X.; Chen, Y.; Ma, Z.; and Huang, X. 2023. Alleviating Catastrophic Forgetting of Incremental Object Detection via Within-Class and Between-Class Knowledge Distillation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 18848–18858. Paris, France: IEEE. ISBN 9798350307184.

Karras, T.; Aittala, M.; Hellsten, J.; Laine, S.; Lehtinen, J.; and Aila, T. 2020. Training Generative Adversarial Networks with Limited Data. In *Advances in Neural Information Processing Systems*.

Kingma, D. P.; Salimans, T.; Poole, B.; and Ho, J. 2021. Variational Diffusion Models. In *Advances in Neural Information Processing Systems*. ArXiv:2107.00630 [cs, stat].

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-Concept Customization of Text-to-Image Diffusion. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Lee, J.; Cho, K.; and Kiela, D. 2019. Countering Language Drift via Visual Grounding. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4385–4395. Hong Kong, China: Association for Computational Linguistics.

Li, D.; Li, J.; and Hoi, S. C. H. 2023. BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing. In *Advances in Neural Information Processing Systems*.

Li, Y.; Zhang, R.; Lu, J. C.; and Shechtman, E. 2020. Few-shot Image Generation with Elastic Weight Consolidation. In *Advances in Neural Information Processing Systems*.

Liu, Z.; Feng, R.; Zhu, K.; Zhang, Y.; Zheng, K.; Liu, Y.; Zhao, D.; Zhou, J.; and Cao, Y. 2023. Cones: Concept Neurons in Diffusion Models for Customized Generation. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2023. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*.

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

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents.

Rolnick, D.; Ahuja, A.; Schwarz, J.; Lillicrap, T. P.; and Wayne, G. 2019. Experience Replay for Continual Learning. In *Advances in Neural Information Processing Systems*. arXiv. ArXiv:1811.11682 [cs, stat].

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. ArXiv:2208.12242 [cs].

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems*.

Sauer, A.; Boesel, F.; Dockhorn, T.; Blattmann, A.; Esser, P.; and Rombach, R. 2024. Fast High-Resolution Image Synthesis with Latent Adversarial Diffusion Distillation. ArXiv:2403.12015 [cs].

Shi, J.; Xiong, W.; Lin, Z.; and Jung, H. J. 2024. InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sohl-Dickstein, J.; Weiss, E. A.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, 2256–2265. PMLR.

Sohn, K.; Ruiz, N.; Lee, K.; Chin, D. C.; Blok, I.; Chang, H.; Barber, J.; Jiang, L.; Entis, G.; Li, Y.; Hao, Y.; Essa, I.; Rubinstein, M.; and Krishnan, D. 2023. StyleDrop: Text-to-Image Generation in Any Style. In *Advances in Neural Information Processing Systems*.

Tewel, Y.; Gal, R.; Chechik, G.; and Atzmon, Y. 2023. Key-Locked Rank One Editing for Text-to-Image Personalization. *ACM SIGGRAPH 2023 Conference Proceedings*.

Voynov, A.; Chu, Q.; Cohen-Or, D.; and Aberman, K. 2023. P+: Extended Textual Conditioning in Text-to-Image Generation. ArXiv:2303.09522 [cs].

Wang, Z.; Wei, W.; Zhao, Y.; Xiao, Z.; Hasegawa-Johnson, M.; Shi, H.; and Hou, T. 2023. HiFi Tuner: High-Fidelity Subject-Driven Fine-Tuning for Diffusion Models. ArXiv:2312.00079 [cs].

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiao, G.; Yin, T.; Freeman, W. T.; Durand, F.; and Han, S. 2023. FastComposer: Tuning-Free Multi-Subject Image Generation with Localized Attention. ArXiv:2305.10431 [cs].

Zhang, X.; Wei, X.-Y.; Wu, J.; Zhang, T.; Zhang, Z.; Lei, Z.; and Li, Q. 2024a. Compositional Inversion for Stable Diffusion Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhang, X.; Wei, X.-Y.; Zhang, W.; Wu, J.; Zhang, Z.; Lei, Z.; and Li, Q. 2024b. A Survey on Personalized Content Synthesis with Diffusion Models. ArXiv:2405.05538 [cs].

Zhang, Y.; Yang, M.; Zhou, Q.; and Wang, Z. 2024c. Attention Calibration for Disentangled Text-to-Image Personalization. In *IEEE Conference on Computer Vision and Pattern Recognition*.

A. Data Augmentation

Naive Augmentation

$$s = \sigma(\log \text{SNR}) \qquad (6)$$

When input images are augmented without corresponding modifications to the text prompt (i.e., naive augmentation), the resulting images often suffer from augmentation leakage, as described in Karras et al. (2020) (refer to Figure 8). This issue notably degrades image alignment scores, as demonstrated in Table 5. Consequently, it is essential that the text prompt is modified in tandem with image augmentations to preserve accurate alignment and maintain the integrity of the generated results (Figure 9).
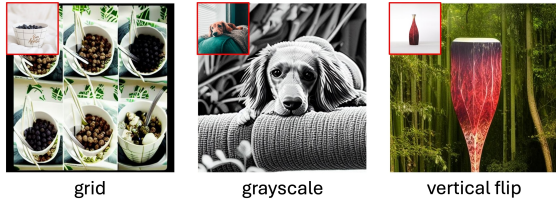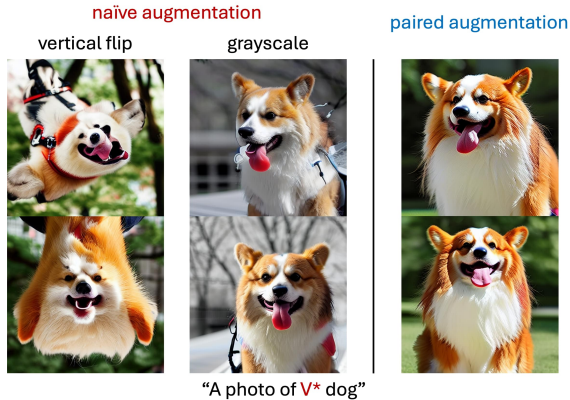

grid   grayscale   vertical flip

Figure 8: **Examples of Augmentation leaking.**

Figure 9: **Comparison on naive augmentation and paired augmentation.**

Prompt augmentation	CLIP-T	CLIP-I
✓	**0.629**	0.702
✗	0.628	**0.728**

Table 5: **Effect of prompt augmentation.**

Relationship with DistAug

Custom Diffusion (Kumari et al. 2023) adopted a similar augmentation strategy to ours, albeit focusing on relatively weak transformations like scaling. In contrast, we employed a broader range of stronger augmentations, such as cutout or

Augmentation	Initialization token	Num. tokens
horizontal flip	flip	1
vertical flip	flip	1
scale	zoom in / zoom out	2
translate	on the left / on the right	3
random crop & resize	crop	1
grayscale	gray	1
brightness	bright / dark	1
cutout	hole	1
grid	grid of	2

Table 6: **Augmentation and initialization tokens.**

grid, without augmentation leaking. We argue that our approach aligns with the princples of Distribution Augmentation (DistAug) (Jun et al. 2020). DistAug states that training a generative model on various augmented distributions (including identity transformation) can prevent overfitting and enable sampling from the original distribution by conditioning on the identity transformation. Our method aligns with this principle, as consistent token-augmentation parings would potentially achieve similar results without requiring augmentation inversion. However, we observed potential coupling between specific tokens and augmentations (e.g. "right" and horizontal shift), necessitating the introduction of augmentation tokens.

Implementation detail

We employed eight distinct augmentations during the training. With a probability of 0.8, we randomly selected one of these augmentations to apply to the image and prompt. Otherwise, no augmentation was performed. When an augmentation was applied, we inserted between 1 to 3 augmentation tokens (A^*) before or after the prompt. Each set of augmentation tokens was initialized with unique words. For more details on specific augmentations, please refer to Table 6. Visualization of the augmentations can be found in Figure 10.

B. Timestep Sampling

Effect of text embedding per timestep. Previous studies (Choi et al. 2022; Balaji et al. 2022) have shown that diffusion models are can be formulated as a mixture-f-experts, with each timestep conditioned U-Net playing a different role. Notably, Balaji et al. (2022) demonstrated that diffusion models become less reliant on text input as noise level decreases. Inspired by these findings, we investigated the influence of text conditioning on diffusion model outputs. Specifically, we computed the difference in model outputs, $\epsilon(\mathbf{x}_t, \mathbf{y})$, for identical latent inputs, $\mathbf{x}_t$, but varying text prompts, $\mathbf{y}$. For instance, we compared the outputs for the base prompt "photo of a dog" to alternative prompts like "photo of a cat".

Considering the reverse process of the denoising diffusion probabilistic model (DDPM) (Ho, Jain, and Abbeel 2020), expressed as:

$$\mathbf{x}_t = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_{t-1} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)) + \sigma_t \mathbf{z}. \qquad (7)$$
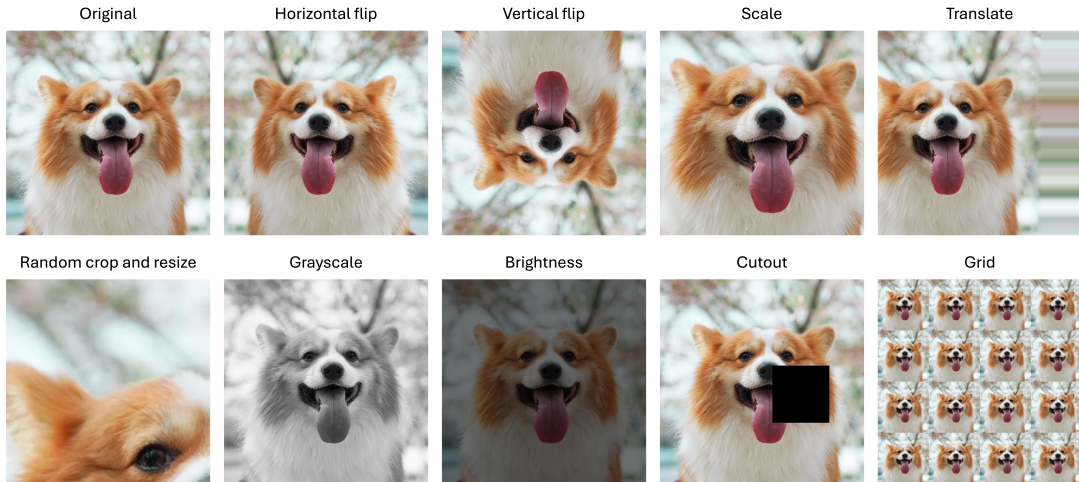
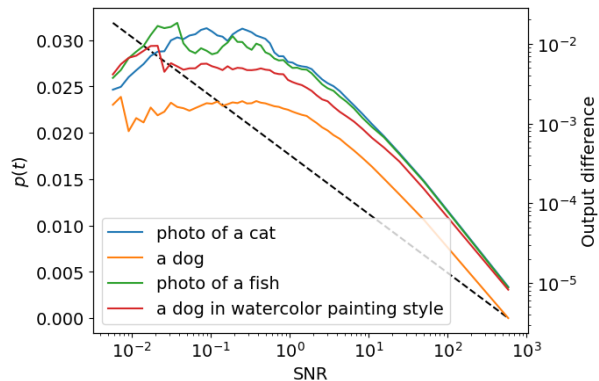Figure 10: **Augmentations used during training TextBoost.**

Figure 11: **Effect of text input.** We measured the effect of the text conditioning by $\beta_t|\epsilon(\mathbf{x}_t, y_{base}) - \epsilon(\mathbf{x}_t, y_{other})|$, where a base prompt y_{base} is 'photo of a dog'.

Here, β_t, α_t, and $\bar{\alpha}_t$ are predefined time-dependant scaling factors, we can see that the model output's impact is scaled by $\beta_t/\sqrt{1 - \bar{\alpha}_t}$. Consequently, we scaled the difference by β_t.

As illustrated in Figure 11, the effect of text input intensifies as the signal-to-noise ratio (SNR) decreases. To emphasize this region during training, we biased our timestep sampling toward lower SNR values.

Implementation detail. In practice, we sampled timestep t from a categorical distribution. The probability of each timestep, $p(t)$, was defined as follows:

$$p(t = t_i) = \frac{-\log \text{SNR}(t_i) + C}{\sum_{i=0}^{T}(-\log \text{SNR}(t_i) + C)}, \qquad (8)$$

where $C = \max \log \text{SNR}(t_i)$ is a constant shift ensuring $p(t)$ is positive. The probability distribution of timesteps is visualized as a dashed line in Figure 11.

C. User Study Template

We ask users to choose one image from the four outputs generated by each of Textual Inversion, DreamBooth, Custom Diffusion, and TextBoost (ours). Users are instructed to select the image that best (1) resembles the given subject and (2) adheres to the given prompt simultaneously. Example questions are shown in Figure 12. Note that output images are sampled with a fixed random seed across methods for fair comparison and are presented in shuffled order.

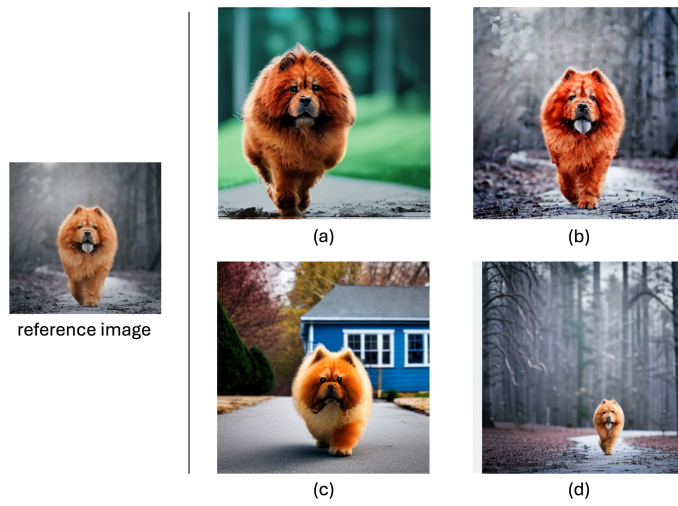## D. More Results

Comparison with Custom Diffusion

Here, we present more qualitative results compared to Custom Diffusion in Figure 13. Our generated images closely resemble the subject matter of the reference image. Moreover, our method significantly outperforms Custom Diffusion in capturing the essence of the text prompt.

More qualitative results of our method

Figures 14, 15, and 16 offer additional examples showcasing the diverse subjects our TextBoost can handle across a wide range of text prompts. TextBoost consistently generates high-quality outputs, enabling creative control through imaginative prompts.

reference image

(a)

(b)

(c)

(d)

(a) Text prompt: "A boot with a tree and autumn leaves in the background"

reference image

(a)

(b)

(c)

(d)

(b) Text prompt: "A dog with a blue house in the background"

Figure 12: **User study template.** We ask 100 participants to answer 20 questions each on Amazon Mechanical Turk.
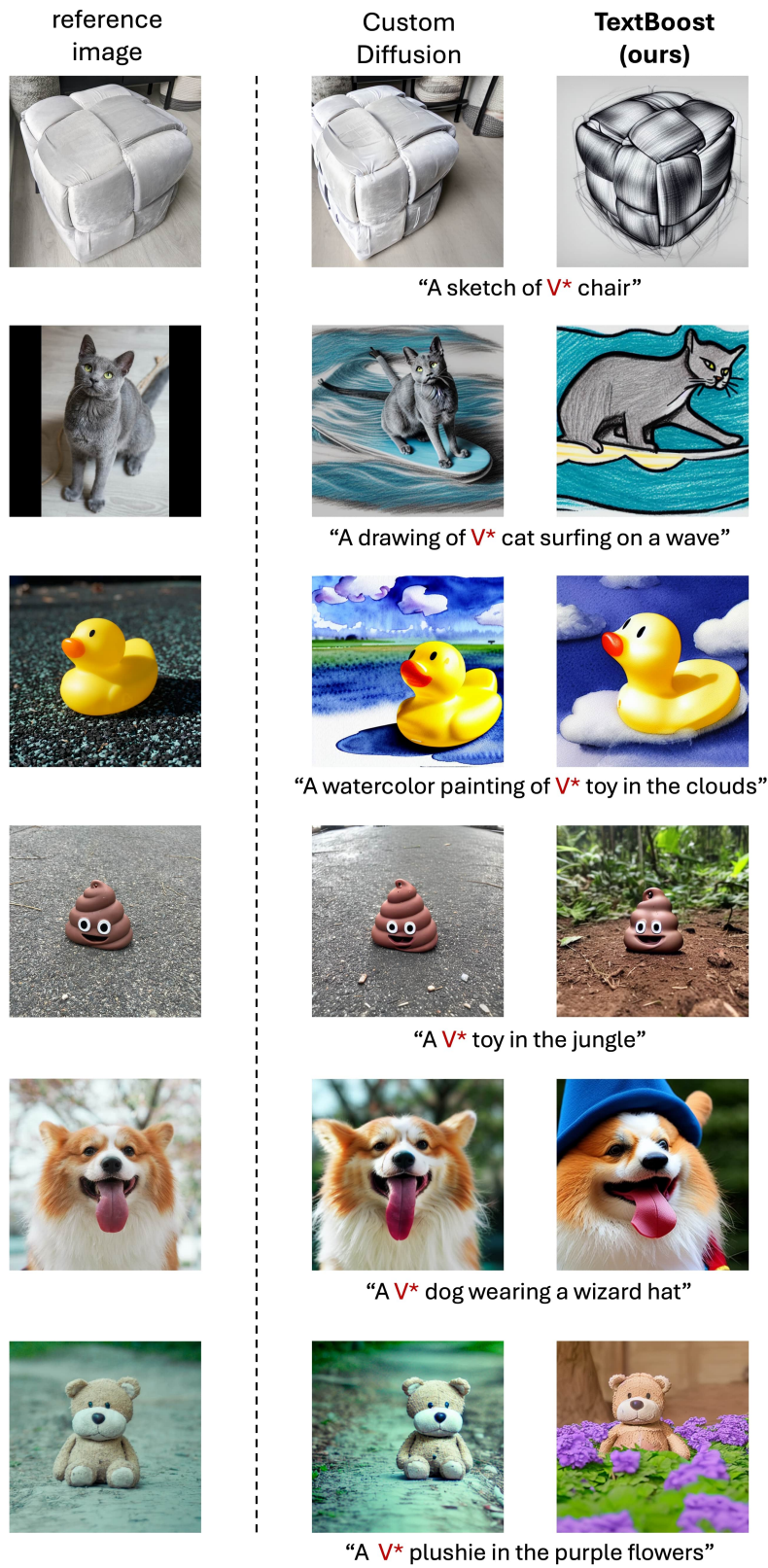
Figure 13: **Comparison with Custom Diffusion.** More qualitative results on comparison with Custom Diffusion. Random seeds are fixed for fair comparison.

TextBoost (ours)

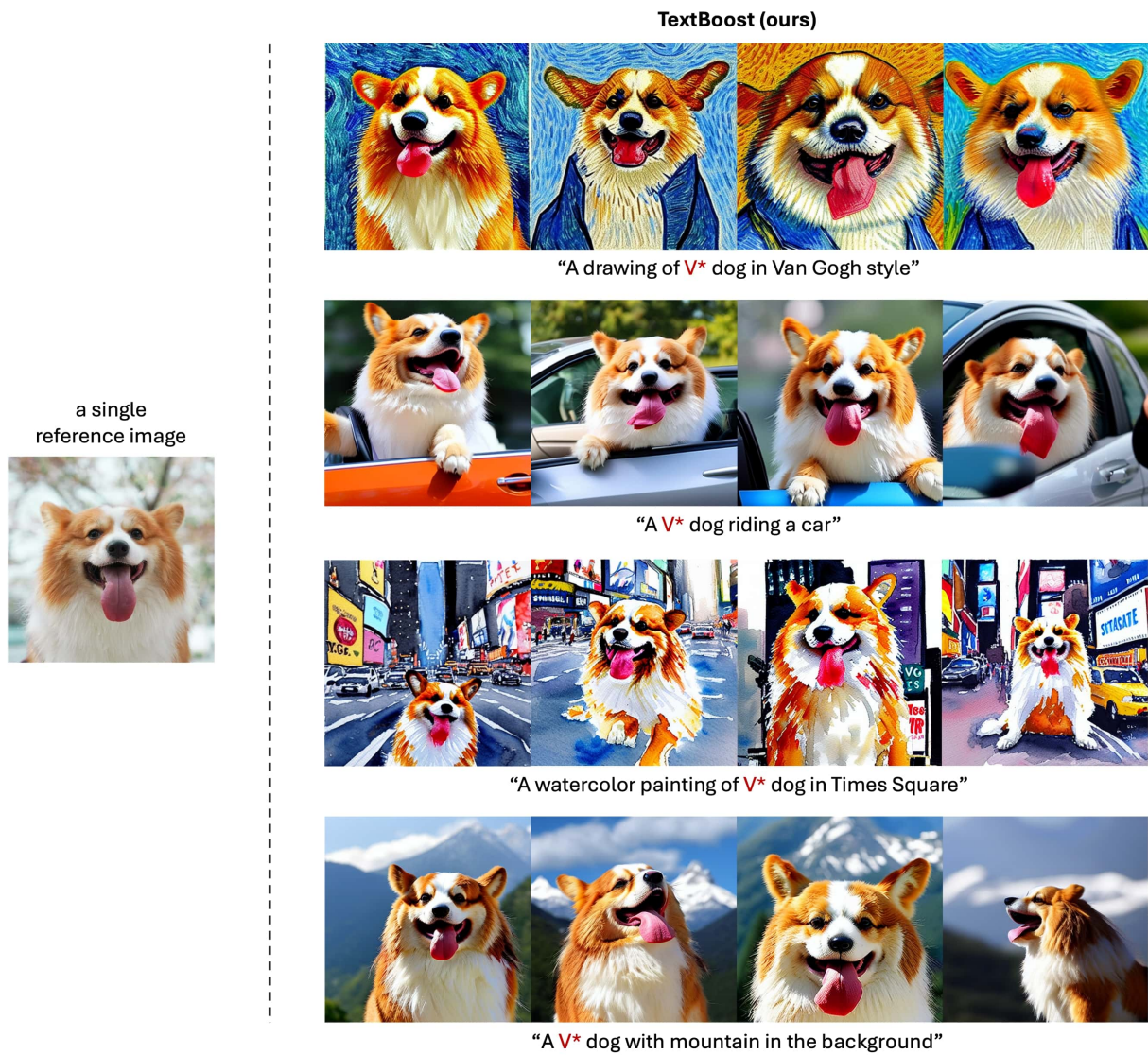"A drawing of V* dog in Van Gogh style"

"A V* dog riding a car"

"A watercolor painting of V* dog in Times Square"

"A V* dog with mountain in the background"

a single
reference image

Figure 14: **More qualitative results of our TextBoost (dog).**

a single
reference image

"Two V* cats"

"A V* cat in the style of watercolor painting"

"A V* cat on the sofa"

"A sketch of V* cat wearing a tie"

Figure 15: **More qualitative results of our TextBoost (cat).**

TextBoost (ours)

a single
reference image

"A drawing of V* plushie in the sky"

"A V* plushie in the desert"

a single
reference image

"A V* cartoon in the night sky"

a single
reference image

"A painting of V* sneaker in the jungle"

Figure 16: **More qualitative results of our TextBoost (several subjects).**