

# Noise Consistency Regularization for Improved Subject-Driven Image Synthesis

Ni, Yao; Wen, Song; Koniusz, Piotr; Cherian, Anoop

TR2025-073 June 07, 2025

## Abstract

Fine-tuning Stable Diffusion enables subject-driven image synthesis by adapting the model to generate images containing specific subjects. However, existing fine-tuning methods suffer from two key issues: underfitting, where the model fails to reliably capture subject identity, and overfitting, where it memorizes the subject image and reduces background diversity. To address these challenges, we propose two auxiliary consistency losses for diffusion fine-tuning. First, a prior consistency regularization loss ensures that the predicted diffusion noise for prior (non- subject) images remains consistent with that of the pretrained model, improving fidelity. Second, a subject consistency regularization loss enhances the fine-tuned model's robustness to multiplicative noise modulated latent code, helping to preserve subject identity while improving diversity. Our experimental results demonstrate that incorporating these losses into fine-tuning not only preserves subject identity but also enhances image diversity, outperforming DreamBooth in terms of CLIP scores, background variation, and overall visual quality.

*IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR) 2025*



# Noise Consistency Regularization for Improved Subject-Driven Image Synthesis

Yao Ni<sup>1,†</sup> Song Wen<sup>2</sup> Piotr Koniusz<sup>1,3</sup> Anoop Cherian<sup>4,\*</sup>

<sup>1</sup>The Australian National University <sup>2</sup>Rutgers University <sup>3</sup>Data61♥CSIRO

<sup>4</sup>Mitsubishi Electric Research Laboratories (MERL)

<sup>1</sup>yao.ni@anu.edu.au <sup>2</sup>song.wen@rutgers.edu <sup>3</sup>piotr.koniusz@data61.csiro.au <sup>4</sup>cherian@merl.com

## Abstract

*Fine-tuning Stable Diffusion enables subject-driven image synthesis by adapting the model to generate images containing specific subjects. However, existing fine-tuning methods suffer from two key issues: underfitting, where the model fails to reliably capture subject identity, and overfitting, where it memorizes the subject image and reduces background diversity. To address these challenges, we propose two auxiliary consistency losses for diffusion fine-tuning. First, a prior consistency regularization loss ensures that the predicted diffusion noise for prior (non-subject) images remains consistent with that of the pre-trained model, improving fidelity. Second, a subject consistency regularization loss enhances the fine-tuned model’s robustness to multiplicative noise modulated latent code, helping to preserve subject identity while improving diversity. Our experimental results demonstrate that incorporating these losses into fine-tuning not only preserves subject identity but also enhances image diversity, outperforming DreamBooth in terms of CLIP scores, background variation, and overall visual quality.*

## 1. Introduction

Recent advancements in deep learning have significantly enhanced generative modeling capabilities, driven by the availability of large-scale datasets such as LAION-5B [51] and ImageNet [10]. Among various generative architectures, including VAEs [23] and GANs [17, 33], diffusion models [42, 43, 49] have emerged as state-of-the-art, demonstrating remarkable performance in image synthesis [2, 6, 30, 36], video generation [3, 4], and 3D content creation [37, 41, 58]. Notably, Stable Diffusion [45] has gained widespread adoption as a foundational model for solving various downstream generative tasks through fine-tuning. Applications include subject-driven image personalization [25, 26, 47] and controlled image synthesis, as ex-

emplified by ControlNet [21, 64, 67].

Subject-driven personalization [25, 26, 47] aims to adapt pre-trained generative models to synthesize images of specific subjects while preserving both subject identity and the model’s generalization ability for diverse image generation. A common approach is to fine-tune diffusion models on a limited set of subject images using LoRA [20], while leveraging linguistic prompts to encode subject concepts [39]. For example, fine-tuning with the prompt “A photo of a [V] dog” allows Stable Diffusion to associate the subject with its broader class (“dog”), facilitating contextual understanding for more diverse synthesis and direct text control. An example is shown in Figure 1.

However, fine-tuning poses a major challenge: catastrophic forgetting [25, 47], where the model loses its pre-trained knowledge. To counter this, DreamBooth [47] introduces a prior preservation loss, generating a set of (e.g., 100–200 images) class-specific images (e.g., “A photo of a dog”) using the original model and enforcing consistency during fine-tuning. While this helps retain prior knowledge, empirical results show that it often fails to fully preserve subject identity and reduces image diversity, limiting its effectiveness [7, 52, 61].

To tackle these issues, we introduce novel auxiliary consistency losses for fine-tuning Stable Diffusion, aiming to preserve subject identity while improving image fidelity and diversity. We argue that LoRA, which adapts Stable Diffusion for subject images, does not need to re-learn class prior images, as the pre-trained model already captures the class concept. We also observe that fine-tuning LoRA on ground-truth noise for prior images lowers fidelity by causing the model to forget the realism learned by pre-trained.

To address this, we propose feeding prior images to both the fine-tuned and pre-trained models, enforcing consistency between the noise predicted by both models. For subject images, we maintain the standard approach where the fine-tuned model predicts the ground-truth noise added to its latent code. This approach improves the fidelity of generated subject images by retaining the realism developed in the pre-trained model while allowing LoRA to focus its ca-

\*The corresponding author. <sup>†</sup>Work done during internship at MERL.

capacity on fitting the subject images.

While this prior consistency regularization preserves identity and improves fidelity, we find that it reduces diversity in the generated images due to the limited number of subject images. To enhance diversity, we introduce multiplicative Gaussian noise modulation [32, 35] to the latent vectors, which diversifies the pattern of subject images without altering their semantics. We then enforce consistency between the noised and clean latent codes. Combining these approaches improves background diversity while ensuring high fidelity and preserving subject identity.

To validate the effectiveness of our proposed regularizations, we conduct experiments on the benchmark dataset from [41], which includes 30 subject classes spanning both live subjects and objects. Our method achieves state-of-the-art performance in subject-driven text-to-image synthesis, as measured by CLIP [39] and DINO [5] scores. Additionally, We demonstrate the generality of our approach by integrating it into both DreamBooth [47] and DCO [26], yielding clear improvements. Qualitative results further highlight the diversity of generated images while preserving subject identity, reinforcing the effectiveness of our method.

Before we proceed with describing our technical details, we summarize the key contributions of this paper below:

- We address the important issue of preserving subject identity while avoiding overfitting in subject-driven Stable Diffusion-based image synthesis.
- To preserve identity, we propose enforcing a LoRA-based fine-tuning model with an auxiliary noise consistency regularization using prior synthesized images.
- To improve the diversity of synthesized images, we propose adding multiplicative noise to the subject images and enforcing the fine-tuned Stable Diffusion model to be consistent with both the noised and clean latent codes.
- We present experiments demonstrating that our method can improve the fidelity of the subject image while also enhancing diversity compared to previous methods.

## 2. Related work

**Subject-driven personalization.** These approaches [25, 26, 28, 47, 48, 54] can be broadly categorized into three broad directions: full model fine-tuning, partial model fine-tuning, and embedding optimization. In full model fine-tuning, DreamBooth [47] pioneered the idea of fine-tuning all model weights to encode subjects as unique identifiers, enabling the generation of the learned subjects in various contexts and styles. HyperDreambooth [48] uses a hyper-network that can efficiently produce a small set of personalized weights from a single image of a person. However, this comprehensive fine-tuning approach often faces challenges in computational efficiency and overfitting. To address these limitations, several works have explored partial model fine-tuning. CustomDiffusion [25] demonstrated

that fine-tuning only the cross-attention layers is sufficient for concept learning while achieving superior performance in multi-concept compositional generation. StyleDrop [54] enables to synthesize an image following a specific style while only fine-tuning 1% of total model parameters. Similarly, DCO [26] proposed a fine-tuning method that preserves the pre-trained model’s compositional capabilities through implicit reward models. An alternative approach focuses on embedding optimization. Textual Inversion [14] proposed optimizing word embeddings while keeping the pre-trained diffusion model frozen, effectively capturing unique concepts with minimal parameter updates. Building upon this, hybrid approaches [1, 16, 59] have emerged that combine both word embedding optimization and selective model weight updates. Another notable work in this direction [13] further explored gradient-based embedding optimization techniques, while [31] focused on zero-shot image-conditioned text-to-image synthesis. In contrast to these approaches, we consider consistency losses with a focus in preserving the subject identity and prior knowledge of diffusion models during fine-tuning, making our approach applicable to any prior fine-tuning technique.

**Parameter-efficient fine-tuning.** The emergence of large-scale foundation models has introduced significant challenges for fine-tuning on limited datasets due to their massive parameter counts. To address the efficiency and overfitting concerns inherent in traditional fine-tuning approaches, Parameter Efficient Fine-Tuning (PEFT) methods have emerged as a promising solution. PEFT techniques can be categorized into several major approaches. Adapter tuning [8, 22, 44] introduces trainable layers within the pre-trained model architecture, creating efficient bottlenecks for adaptation. Soft prompt tuning [15, 55, 57, 59] is an efficient technique for adapting large foundation models to specific tasks by learning task-relevant prompt embeddings, which guide model outputs without modifying the core model weights. Low-Rank Adaptation (LoRA) [11, 19, 20, 24, 65, 68] has gained particular prominence by decomposing weight updates into products of low-rank matrices, dramatically reducing the parameter count while maintaining performance. Recent innovations have explored alternative decomposition strategies. SVDiff [18] employs singular value decomposition of pre-trained weights, focusing adaptation on singular values alone. OFT [27, 38] preserves model structure by incorporating trainable orthogonal matrices while maintaining hyperspherical energy. KronA [12, 29] introduces a novel approach using Kronecker products of smaller matrices to construct weight updates efficiently. SODA [66] decomposes the pretrained weight matrix and fine-tunes its individual components. Our proposed consistency losses attempt to balance PEFT weight adaption while minimizing its impact on the prior knowledge already learned in the pre-trained model.

**Consistency regularization.** The concept of consistency regularization has been explored periodically in various contexts. FixMatch [53] applies it to augmented images for semi-supervised learning, while OpenMatch [50] leverages it for outlier predictions in open-set semi-supervised learning. R-Drop [60] introduces consistency regularization in transformers [56] with dropout for NLP tasks. In GAN training, CR [63] enforces consistency between augmented real and fake images, and CAGAN [34] applies it to discriminators with dropout. Sem-GAN [9] enforces semantic consistency for image-to-image translation. Despite the empirical success of consistency regularization, theoretical analysis has been limited. NICE [32] addresses this gap by showing that consistency regularization reduces latent feature gradients, stabilizing GAN training. PACE [35] further connects consistency regularization to improved generalization in deep learning. Additionally, consistency regularization has been applied to fine-tuned and pre-trained models [35, 40, 46] to retain pre-trained knowledge. However, its potential for preserving subject identity while enhancing background diversity remains unexplored. Our method addresses this gap, offering a novel approach that improves both subject consistency and background diversity, providing a compelling advancement over previous methods.

### 3. Background

Before detailing our approach, we introduce key background concepts in deep generative modeling that underpin our contributions. This section covers the fundamental principles of Stable Diffusion and subject-driven image synthesis while establishing our notation.

#### 3.1. Stable Diffusion

A breakthrough in deep generative modeling happened with the introduction of the Stable Diffusion model, a text-to-image model pretrained on extensive text-image pairs  $\{(p, x)\}$ , with each pair linking a text prompt  $p$  to its corresponding image  $x$ . Such a paired dataset forms the foundation for training the model’s architecture, which includes an auto-encoder with encoder  $\mathcal{E}(\cdot)$  and decoder  $\mathcal{D}(\cdot)$ , a CLIP text encoder  $\tau(\cdot)$ , and a UNet-based conditional diffusion model  $f(\cdot)$ . Within this system, the encoder transforms images  $x$  into latent representations  $z = \mathcal{E}(x)$ , creating compact codes that enable the decoder to effectively reconstruct images as needed. Operating within this latent space rather than raw pixels, the diffusion process introduces noise at time step  $t$  with  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ , producing noisy latent code  $z_t = \alpha_t z + \beta_t \epsilon$  controlled by coefficients  $\alpha_t$  and  $\beta_t$  that manage the denoising schedule. These noisy representations drive the training of the conditional diffusion model  $f(\cdot)$  through the objective:

$$\min_{\mathbf{p}, \mathbf{z}, \epsilon, t} [\|\epsilon - f(z_t, t, \tau(\mathbf{p}))\|_2^2]. \quad (1)$$

The noise prediction objective, guided by text embedding  $\tau(\mathbf{p})$  and conditioned on time step  $t$ , enables diffusion models to iteratively refine noisy latent codes  $z_t$ . This refinement process, enhanced by an expressive transformer architecture and knowledge derived from diverse training data, endows Stable Diffusion model with robust semantic priors. Leveraging these priors, the model generates high-quality and diverse images directly from natural language prompts while ensuring precise alignment between textual input and visual output, delivering reliable text-to-image generation.

#### 3.2. Subject-driven Synthesis

While Stable Diffusion models excel at general text-to-image synthesis, they struggle to faithfully replicate specific subjects from a reference set or synthesize these subjects in new contexts. To overcome this limitation, subject-driven fine-tuning extends the model’s vision-language dictionary by binding a new token  $[V]$  to user-defined subjects.

Let  $s$  represent a subject represented by a small set of subject images  $\mathcal{S} = \{x_i\}_{i=1}^K$ , where the number of available images  $K$  is very small. To expand the Stable Diffusion vision-language dictionary, subject-driven synthesis fine-tunes the model using a subject-specific prompt  $p_s$ , e.g., “a photo of  $[V]$  dog” and optimize the objective:

$$\min_{\Delta\theta} \mathcal{L}_s := \mathbb{E}_{x \sim \mathcal{S}, z = \mathcal{E}(x), \epsilon, t} [\|\epsilon - f_{\Delta\theta}(z_t, t, \tau(p_s))\|_2^2], \quad (2)$$

where  $\Delta\theta$  denotes the trainable LoRA weights added to the otherwise frozen pre-trained model, resulting in the fine-tuned model  $f_{\Delta\theta}$ .

The naïve fine-tuning scheme in Eq. (2) aligns the generated image space of Stable Diffusion with subject images. However, this alignment, when applied to a limited number of subject images, risks overfitting to these images, discarding pre-trained knowledge in the model, thereby reducing output diversity. To preserve diversity while retaining pre-trained knowledge, DreamBooth [47] introduces prior preservation, generating a set of class images  $\mathcal{P} = \{x_i\}_{i=1}^{K_p}$  (where  $K_p$ , e.g., 100–200 is the number of prior images) using a generic prompt  $p_p$  such as “a photo of dog”. The model is then fine-tuned using a combined objective:

$$\begin{aligned} \min_{\Delta\theta} \mathcal{L}_{\text{prior}} &:= \mathbb{E}_{x \sim \mathcal{P}, z = \mathcal{E}(x), \epsilon, t} [\|\epsilon - f_{\Delta\theta}(z_t, t, \tau(p_p))\|_2^2], \\ \min_{\Delta\theta} \mathcal{L}_{\text{dreambooth}} &:= \mathcal{L}_s + \lambda_{\text{prior}} \mathcal{L}_{\text{prior}}, \end{aligned} \quad (3)$$

where  $\lambda_{\text{prior}}$  regulates the balance between subject fidelity and pre-trained knowledge retention.

### 4. Proposed Approach

While DreamBooth effectively integrates new concepts into Stable Diffusion, its emphasis on optimizing performance for a limited set of subject and class prior images introduces

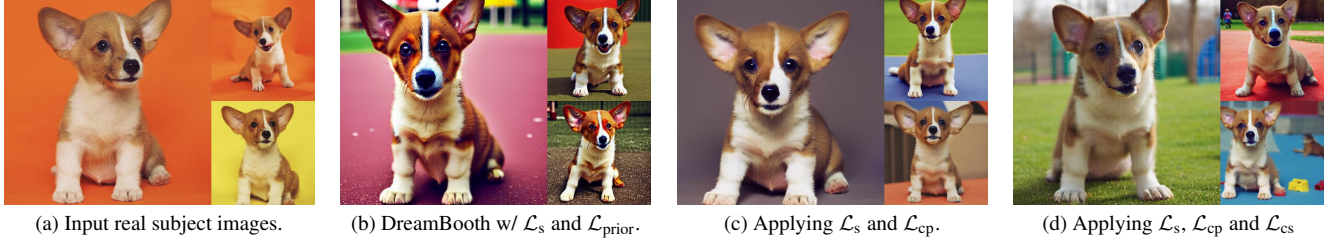


Figure 1. Generated images using different methods with the prompt “A photo of a [V] dog on the playground”. DreamBooth reduces fidelity and loses subject identity. Applying  $\mathcal{L}_{cp}$  improves fidelity but reduces background diversity, while  $\mathcal{L}_{cs}$  enhances it.

critical flaws. The combination of scarce training data and the model’s vast parameter space makes it highly prone to overfitting, distorting its original representation of natural image distributions. This problem is further exacerbated by a fundamental contradiction in the training process, where the ground-truth noise added to class images differs from the noise the pretrained model naturally predicts for those inputs. Simultaneously, forcing the LoRA weights  $\Delta\theta$  to adapt to both subject and prior images introduces redundancy, as the pretrained model already encodes prior semantics. This mismatch skews the latent diffusion space toward the prior images, corrupting the model’s learned data manifold during pre-training. Consequently, the susceptibility to overfitting and the noise-prediction mismatch degrades its ability to generate diverse and realistic outputs that preserve subject identity and lose background details. This problem is illustrated in Figure 1, where for the subject images in Figure 1a, DreamBooth is seen to lose realism and subject identity as shown in Figure 1b.

#### 4.1. Consistency Regularization for Fidelity

To resolve the noise-prediction mismatch and eliminate redundant adaptation to prior images, we propose enforcing consistency between the fine-tuned and pretrained models for prior images  $\mathcal{P}$ . Specifically, we input diffusion-noised prior images into both the fine-tuned and pre-trained models and optimize the LoRA weights  $\Delta\theta$  in the fine-tuned model to ensure that the predicted noise remains consistent with the prediction from the pre-trained model. Our objective can be mathematically stated as:

$$\min_{\Delta\theta} \mathcal{L}_{cp} := \mathbb{E}_{\mathbf{x} \sim \mathcal{P}, \mathbf{z}^p = \mathcal{E}(\mathbf{x}), \epsilon, t} [\|f_{\Delta\theta}(\mathbf{z}_t^p, t, \tau(\mathbf{p}_p)) - f_{\text{ref}}(\mathbf{z}_t^p, t, \tau(\mathbf{p}_p))\|_2^2], \quad (4)$$

where  $f_{\text{ref}}$  denotes the pretrained conditional latent Stable Diffusion model. By enforcing consistency between the fine-tuned and pretrained models on prior images, the fine-tuned model inherently retains the pre-trained knowledge, eliminating the need to relearn prior images. This allows the fine-tuned model to fully allocate the LoRA weights  $\Delta\theta$  to learning the subject identity, ensuring that the fine-tuned

model synthesizes the subject faithfully while preserving the natural and realistic outputs inherited from  $f_{\text{ref}}$ .

#### 4.2. Consistency Regularization for Diversity

While the consistency regularization on prior images effectively preserves subject identity and fidelity, synthesizing diverse outputs remains challenging due to the limited number of subject images. As illustrated in Figure 1c, combining  $\mathcal{L}_s$  and  $\mathcal{L}_{cp}$  improves the realism and preserves the subject identity compared to DreamBooth. However, this comes at the cost of reduced background diversity.

To address this, we propose multiplicative noise modulation, a technique inspired by its success in stabilizing GAN training [32] and improving generalization in parameter-efficient fine-tuning [35], to expand the latent space of subject images without altering the core image semantics. For each subject image, we perturb its latent code as:

$$\mathbf{z}' = \mathbf{z} \odot \epsilon_m, \quad \epsilon_m \sim \mathcal{N}(\mathbf{1}, \sigma^2 \mathbf{I}), \quad (5)$$

where  $\sigma^2$  is the variance of the multiplicative noise. Notably, while multiplicative noise has been applied in the prior works [32, 35], its use to enhance diversity in diffusion models has never been explored. This semantics-preserving perturbation introduces controlled variations into the latent space, diversifying outputs without distorting the subject’s identity.

To ensure noise-augmented latents retain subject identity while improving diversity, we enforce consistency between predictions for clean latent code  $\mathbf{z}$  and noise-modulated latent codes  $\mathbf{z}'$  of the subject images via:

$$\min_{\Delta\theta} \mathcal{L}_{cs} := \mathbb{E}_{\mathbf{x} \sim \mathcal{S}, \mathbf{z} = \mathcal{E}(\mathbf{x}), \mathbf{z}', \epsilon, t} [\|f_{\Delta\theta}(\mathbf{z}_t, t, \tau(\mathbf{p}_s)) - f_{\Delta\theta}(\mathbf{z}'_t, t, \tau(\mathbf{p}_s))\|_2^2]. \quad (6)$$

Note that  $\mathbf{z}_t = \alpha_t \mathbf{z} + \beta_t \epsilon$  and  $\mathbf{z}'_t = \alpha_t (\mathbf{z} \odot \epsilon_m) + \beta_t \epsilon$  share the same diffusion noise  $\epsilon$  and time step  $t$ .

#### 4.3. Overall Objective

With consistency regularization applied to both prior and subject images, our final objective for robust subject-driven



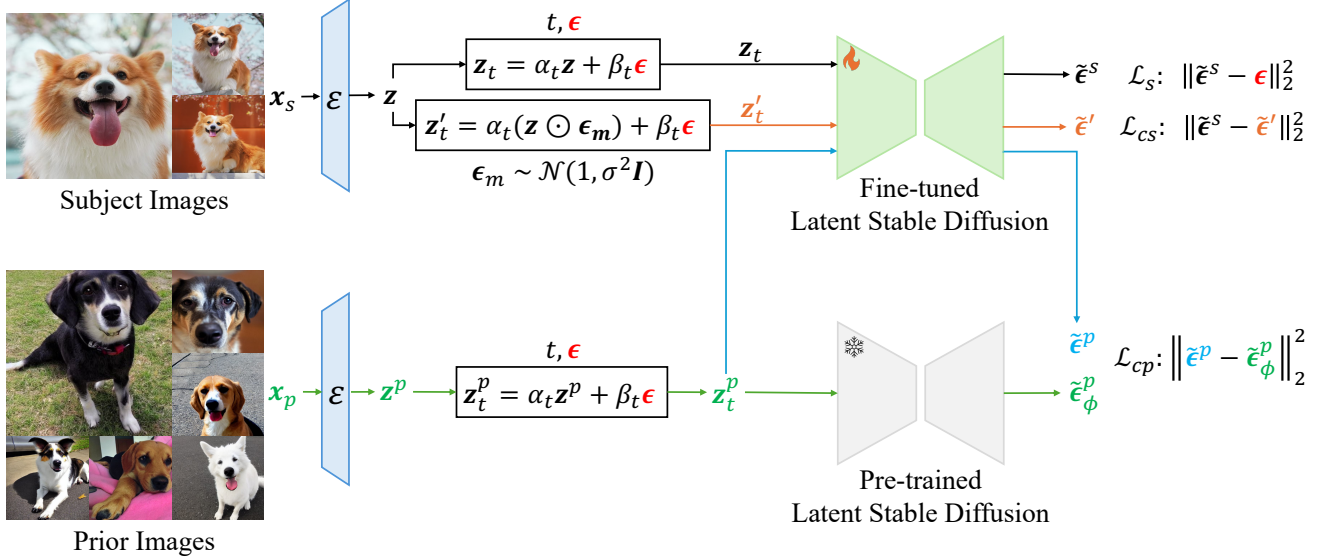


Figure 2. Our pipeline applies two forms of consistency regularization: one over subject images and one over prior images. For the subject images, we modulate the latent code  $z$  with multiplicative noise  $\epsilon_m \sim \mathcal{N}(1, \sigma^2 I)$ , then add identical diffusion noise  $\epsilon$  to both the clean and noise-modulated latent codes, yielding  $z_t$  and  $z'_t$ . These are fed into the fine-tuned Stable Diffusion model. This model has two objectives for subject images: (1)  $\mathcal{L}_s$ , where the predicted noise  $\tilde{\epsilon}^s$  approximates the ground truth noise  $\epsilon$ , and (2)  $\mathcal{L}_{cs}$ , where the prediction  $\tilde{\epsilon}'$  should be consistent with  $\tilde{\epsilon}^s$ . For prior images, we input the diffused latent code  $z_t^p$  into both the fine-tuned and pre-trained latent diffusion models, enforcing consistency between the fine-tuned model’s prediction  $\tilde{\epsilon}^p$  and the pre-trained model’s prediction  $\tilde{\epsilon}^p_\phi$ .

fine-tuning is formulated as:

$$\min_{\Delta\theta} \mathcal{L} := \mathcal{L}_s + \lambda_{cp} \mathcal{L}_{cp} + \lambda_{cs} \mathcal{L}_{cs}, \quad (7)$$

where  $\lambda_{cp}$  and  $\lambda_{cs}$  control the regularization strengths for prior and subject images, respectively.

Our pipeline for implementing these two consistency regularizations is illustrated in Figure 2. For subject images, we apply multiplicative noise modulation to their latent codes and process both the clean and noised codes through the diffusion process. The resulting codes are then fed into the fine-tuned diffusion model, where  $\mathcal{L}_s$  predicts the ground-truth noise and  $\mathcal{L}_{cs}$  enforces consistency between predictions of both clean and noised codes. For prior images, we apply the diffusion process to their latent codes, which are then fed into both the fine-tuned and pre-trained models, followed by enforcing consistency between their outputs. These two consistency losses strengthen the fine-tuning process, enhancing model robustness and enabling the fine-tuned diffusion model to generate diverse and realistic subject images, as shown in Figure 1d.

## 5. Experiments

In this section, we provide experiments demonstrating the benefits of our proposed regularization schemes. To this end, we fine-tune Stable Diffusion V1.5 [45] on subject images for subject-driven text-to-image synthesis. We qualitatively

and quantitatively evaluate our methods and conduct ablation studies on our proposed regularization designs.

### 5.1. Dataset

We use the publicly available dataset from [47]. This dataset consists of 30 subjects and comprising a diverse range of 9 unique live subjects, including animals such as dogs and cats, along with 21 distinct objects. These objects span various categories, such as backpacks, stuffed animals, sunglasses, cartoons, toys, sneakers, teapots, vases, and more. This selection of subjects and objects provides a broad and varied basis for our study.

### 5.2. Evaluation

For each subject, we generate images guided by 10 distinct prompts using the fine-tuned Stable Diffusion model. For each prompt, we synthesize 4 images, resulting in a total of 40 images per subject. We then calculate the CLIP-I, CLIP-T, and DINO scores as follows:

**CLIP-I Score:** We use the CLIP image encoder [39] to obtain embeddings for both the generated images and the real images. We calculate the mean pairwise cosine similarity between embeddings of generated and real images.

**CLIP-T Score:** We input each prompt into the CLIP text encoder to obtain prompt embeddings, then calculate the mean cosine similarity between the prompt embedding and the embeddings of the generated images.

Table 1. Results for subject-driven text-to-image synthesis. We compare our method with DreamBooth [47] and DCO [26], evaluating the performance using three metrics: CLIP-I, CLIP-T, and DINO scores. The reported results represent the average performance across 30 subjects, encompassing both live subjects and object categories. Our method demonstrates superior performance.

Method	DINO $\uparrow$	CLIP-I $\uparrow$	CLIP-T $\uparrow$
Real Images	0.703	0.864	—
DreamBooth	0.602	0.778	<b>0.329</b>
Ours	<b>0.634</b>	<b>0.792</b>	0.324
DreamBooth + DCO	0.638	0.795	0.309
Ours + DCO	<b>0.652</b>	<b>0.811</b>	<b>0.310</b>

**DINO Score:** Using the DINO model [5], we encode the generated and real images and calculate the mean pairwise cosine similarity between their embeddings.

For each method, we report the average of these scores across all 30 subjects, providing a comprehensive assessment of performance across different subjects and prompts.

### 5.3. Hyperparameter settings

We compare our method against DreamBooth [47] and DCO [26] under consistent experimental settings to ensure a fair evaluation. Specifically, we fix the learning rate at  $5e-4$  and use LoRA with a rank of 4 to fine-tune the UNet of the latent Stable Diffusion model across all methods.

For DreamBooth, we set  $\lambda_{\text{prior}} = 1$  and generate 100 prior images to preserve prior knowledge, as recommended in the original paper. We also follow the guidelines provided by DCO and set its parameter  $\beta$  to 1000, as recommended.

In our method, we maintain  $\lambda_{\text{cp}} = \lambda_{\text{cs}} = 0.5$  for consistency regularization and fix the noise modulation parameter  $\sigma = 0.2$ . These consistent configurations ensure a robust and meaningful comparison across all methods.

### 5.4. Quantitative Results

In Table 1, we present a comparison of our method with DreamBooth [47] and DCO [26]. Our approach achieves higher CLIP-I and DINO scores than both DreamBooth and DCO, while maintaining a competitive CLIP-T score. These results highlight the effectiveness of our method in producing subject-driven text-to-image generations that capture both visual and semantic alignment, demonstrating a clear advantage over existing techniques.

### 5.5. Qualitative Results

We provide a qualitative comparison of our method with DreamBooth, as shown in Figure 3. From the figure, it is evident that our method excels in preserving the subject’s identity, maintaining consistent and recognizable features across the generated images. In contrast, the baseline

method, DreamBooth, struggles to retain the subject’s identity, leading to noticeable distortions and inconsistencies. These results clearly demonstrate the superior performance of our approach in subject-driven text-to-image synthesis, underscoring its ability to achieve higher fidelity and visual coherence compared to the baseline.

### 5.6. Ablation studies

Below, we present experiments to justifying  $\mathcal{L}_{\text{cp}}$  and  $\mathcal{L}_{\text{cs}}$  improve fidelity and background diversity.

**Impact of  $\lambda_{\text{cp}}$ .** To investigate the impact of  $\lambda_{\text{cp}}$  on performance, we fixed other hyperparameters and varied  $\lambda_{\text{cp}}$  from 0.0 to 1.0. The generated images for different values of  $\lambda_{\text{cp}}$  using the prompt “A photo of a [V] dog on the playground” are shown in Figure 4. Comparing images generated with  $\lambda_{\text{cp}} = 0.0$  and  $\lambda_{\text{cp}} = 0.2$ , we observe that introducing consistency regularization for class prior images through  $\mathcal{L}_{\text{cp}}$  effectively improves both fidelity and background diversity, confirming its importance and effectiveness. The best fidelity and diversity are achieved at  $\lambda_{\text{cp}} = 0.5$ , with larger values showing minimal visual differences.

**Impact of  $\lambda_{\text{cs}}$ .** We examine the effect of  $\lambda_{\text{cs}}$  on performance. Keeping other hyperparameters fixed, we varied  $\lambda_{\text{cs}}$  from 0.0 to 0.8. The generated images for different values of  $\lambda_{\text{cs}}$  using the prompt “A photo of a [V] dog on the playground” are presented in Figure 5. Comparing the images generated with  $\lambda_{\text{cs}} = 0.0$  and  $\lambda_{\text{cs}} = 0.2$ , we find that introducing consistency regularization for the subject via  $\mathcal{L}_{\text{cs}}$  effectively enhances background diversity, validating its effectiveness. The best balance between fidelity and diversity is achieved at  $\lambda_{\text{cs}} = 0.5$ , while higher values increase background diversity at the cost of subject identity preservation.

**How does  $\mathcal{L}_{\text{cs}}$  improve background diversity?** To clarify why the proposed consistency regularization through multiplicative noise modulation enhances the background diversity, we analyze the latent code of these images using the encoder  $\mathcal{E}(\cdot)$ . Specifically, we generate images by applying different loss combinations, including DreamBooth losses,  $\mathcal{L}_{\text{s}}/\mathcal{L}_{\text{cp}}$ ,  $\mathcal{L}_{\text{s}}/\mathcal{L}_{\text{cs}}$ , and  $\mathcal{L}_{\text{s}}/\mathcal{L}_{\text{cp}}/\mathcal{L}_{\text{cs}}$ , for the subject dataset shown in Figure 1a. We then compute and plot the histogram of the latent code values. Additionally, we calculate the KL divergence between the latent code of generated images and the prior images, as shown in Figure 6.

Among these configurations, applying  $\mathcal{L}_{\text{cs}}$  yields the lowest KL divergence with the prior images, indicating that multiplicative noise modulation effectively diversifies the latent code distribution and enhances the background diversity of the generated images. Moreover, applying  $\mathcal{L}_{\text{cp}}$  alone improves fidelity but reduces diversity, resulting in the highest KL divergence. In contrast, combining  $\mathcal{L}_{\text{cs}}$  with  $\mathcal{L}_{\text{cp}}$  effectively mitigates this loss of diversity while maintaining fidelity, achieving a lower KL divergence than Dream-



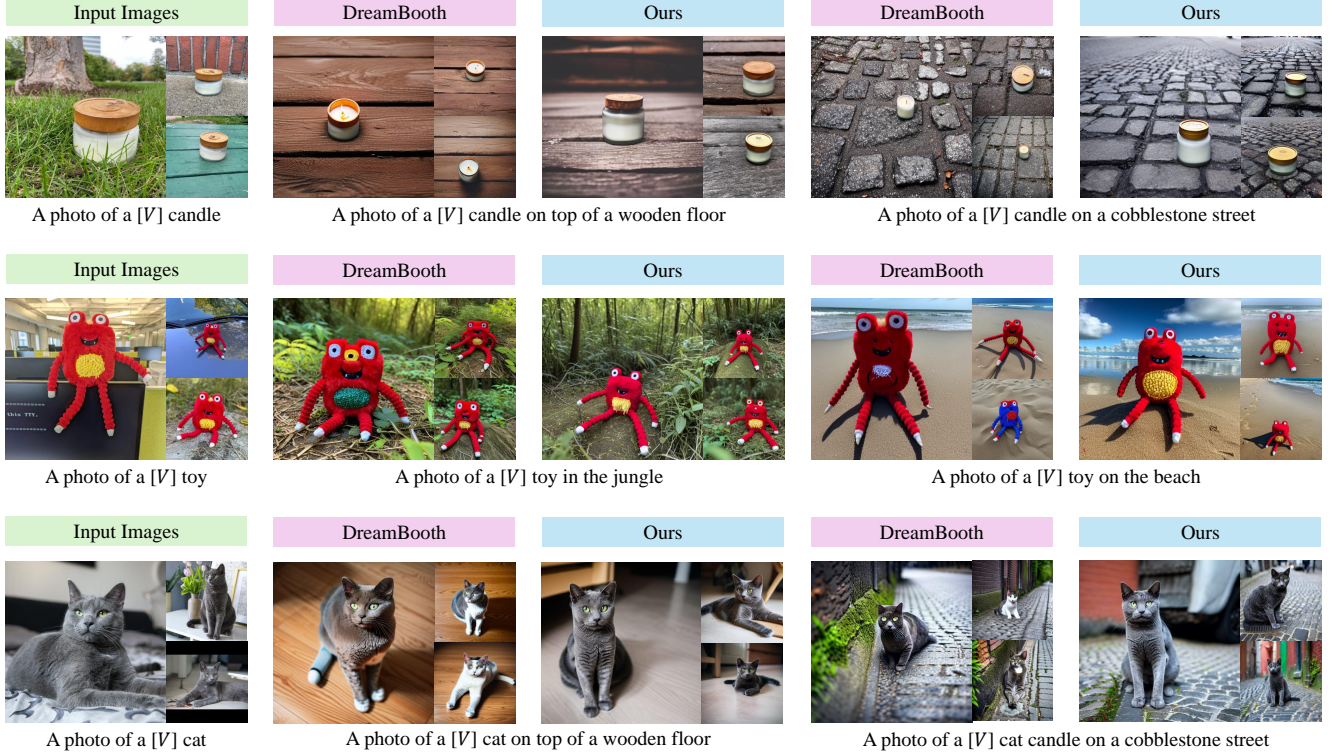


Figure 3. Generated images and their corresponding prompts for both DreamBooth and our method are presented. As shown, our method demonstrates a significantly better ability to preserve the subject’s identity, maintaining consistent and recognizable features throughout the synthesized images. This results in higher fidelity and more accurate representations of the subject compared to DreamBooth, which often fails to retain essential details and features. The superior performance of our approach highlights its effectiveness in subject-driven text-to-image synthesis, ensuring both identity preservation and visual coherence.

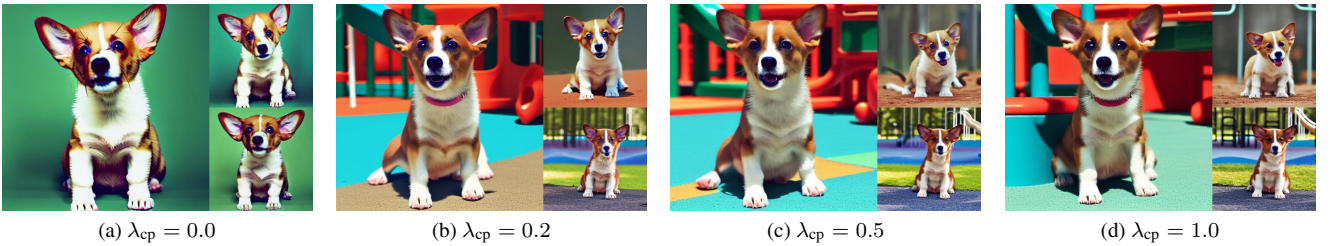


Figure 4. Ablation study for  $\lambda_{cp}$ . Images generated by different  $\lambda_{cp}$  with prompt “A photo of a [V] dog on the playground”.

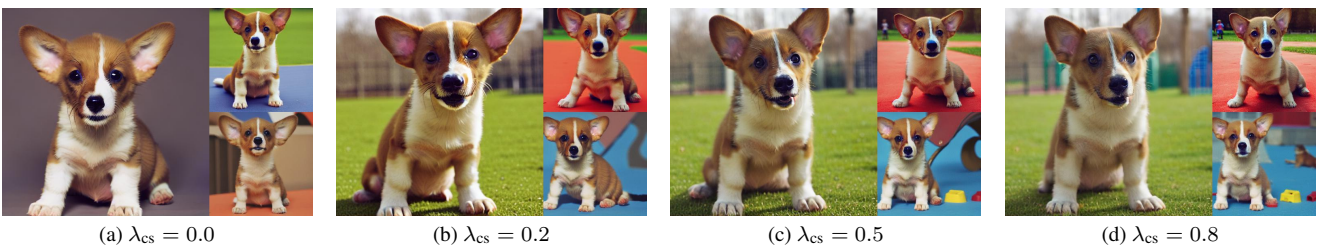


Figure 5. Ablation study for  $\lambda_{cs}$ . Images generated by different  $\lambda_{cs}$  with prompt “A photo of a [V] dog on the playground”.

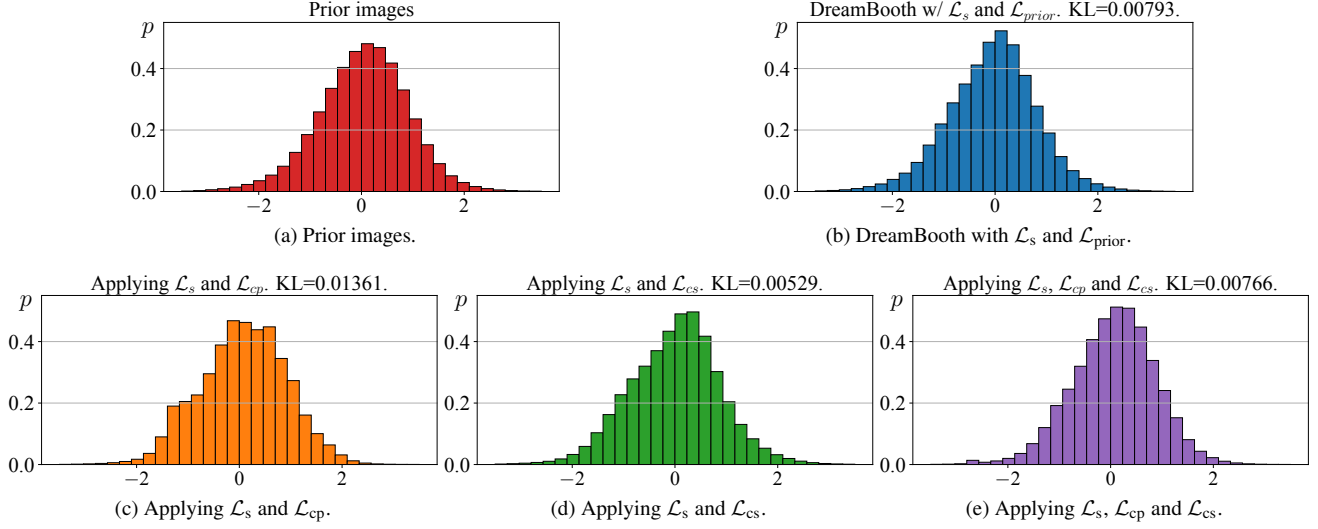


Figure 6. Histogram of the latent codes for generated and prior images under different consistency losses. (a) shows the distribution of prior images, while (c) demonstrates that applying only prior consistency regularization ( $\mathcal{L}_{cp}$ ) shifts the distribution. Our subject consistency loss with multiplicative noise ( $\mathcal{L}_{cs}$ ) in (d) better aligns with the prior distribution. (e) illustrates that combining both regularizations achieves lower KL divergence than DreamBooth in (b), highlighting the effectiveness of our method. The  $x$ -axis represents latent code values and the  $y$ -axis represents density  $p$ .

Booth. This demonstrates that  $\mathcal{L}_{cs}$  helps preserve diversity, leading to overall better performance.

**Additive noise vs. multiplicative noise.** To justify our use of multiplicative noise over additive noise, we replace multiplicative noise with additive noise and analyze the generated images (Figure 7). The results show that additive noise reduces fidelity and fails to preserve subject identity, whereas multiplicative noise maintains fidelity and preserves subject identity. This occurs because additive noise conflicts with the diffusion process, which also relies on additive noise, making it difficult for the fine-tuned diffusion model to learn subject identity. In contrast, multiplicative noise does not interfere with the diffusion process. Instead, it diversifies the latent code patterns while preserving semantics, effectively maintaining subject identity and enhancing background diversity.

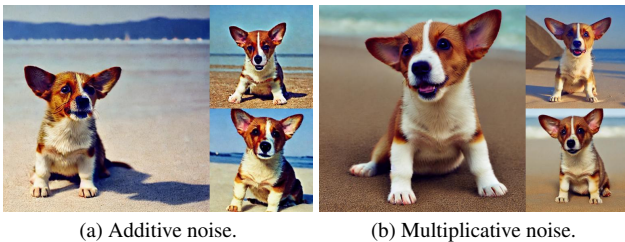


Figure 7. Consistency regularization applied to (a) additive noise modulation and (b) multiplicative noise modulation using the prompt “A photo of a [V] dog on the beach”.

## 6. Limitations and Future work

While the proposed consistency regularization improves the diversity and fidelity of subject-driven synthesis, it approximately doubles the computational cost in terms of FLOPs and GPU memory during training, as both subject and prior images are processed by the model twice. Inference, however, remains unaffected. In future work, we will explore strategies to reduce training overhead, such as applying consistency regularization intermittently or at intermediate layers. We also plan to evaluate these approaches on other Stable Diffusion fine-tuning tasks, including IP-Adapter [62], ControlNet [64] and Textual Inversion [14].

## 7. Conclusions

To enhance both subject identity preservation and image diversity, our approach introduces two distinct consistency regularizations. First, we enforce consistency between the fine-tuned model’s predictions for clean latent codes and those for noise-modulated latent codes of the subject images. Second, we require consistency between the fine-tuned model and the pre-trained model on prior images. These regularizations not only allow the fine-tuned model to better capture and preserve the subject’s unique identity, but they also significantly enhance the diversity of generated images. Experiments on subject-driven text-to-image synthesis demonstrate the effectiveness of our method, as it outperforms two strong baselines, yielding superior scores and producing high-fidelity images that more accurately reflect the subject’s identity.

## References

- [1] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023. 2
- [2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2:3, 2023. 1
- [3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1
- [4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 1
- [5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 6
- [6] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 1
- [7] Hong Chen, Yipeng Zhang, Simin Wu, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2024. 1
- [8] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022. 2
- [9] Anoop Cherian and Alan Sullivan. Sem-gan: Semantically-consistent image-to-image translation. In *2019 IEEE winter conference on applications of computer vision (wacv)*, pages 1797–1806. IEEE, 2019. 3
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1
- [11] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [12] Ali Edalati, Marzieh Tahaei, Ivan Kobyzev, Vahid Partovi Nia, James J Clark, and Mehdi Rezagholizadeh. Krona: Parameter efficient tuning with kronecker adapter. *arXiv preprint arXiv:2212.10650*, 2022. 2
- [13] Zhengcong Fei, Mingyuan Fan, and Junshi Huang. Gradient-free textual inversion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1364–1373, 2023. 2
- [14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 8
- [15] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*, 2023. 2
- [16] 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. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023. 2
- [17] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1
- [18] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7323–7334, 2023. 2
- [19] Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Lora+: Efficient low rank adaptation of large models. *arXiv preprint arXiv:2402.12354*, 2024. 2
- [20] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 1, 2
- [21] Zhihao Hu and Dong Xu. Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *arXiv preprint arXiv:2307.14073*, 2023. 1
- [22] Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Ka-Wei Lee. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 2
- [23] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013. 1
- [24] Dawid Jan Kopiczko, Tijmen Blankevoort, and Yuki Markus Asano. Vera: Vector-based random matrix adaptation. *arXiv preprint arXiv:2310.11454*, 2023. 2
- [25] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 1, 2
- [26] Kyungmin Lee, Sangkyung Kwak, Kihyuk Sohn, and Jinwoo Shin. Direct consistency optimization for compositional text-to-image personalization. *arXiv preprint arXiv:2402.12004*, 2024. 1, 2, 6



- [27] Weiyang Liu, Zeju Qiu, Yao Feng, Yuliang Xiu, Yuxuan Xue, Longhui Yu, Haiwen Feng, Zhen Liu, Juyeon Heo, Songyou Peng, et al. Parameter-efficient orthogonal finetuning via butterfly factorization. In *The Twelfth International Conference on Learning Representations*, 2023. 2
- [28] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 2
- [29] 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. 2
- [30] Nithin Gopalakrishnan Nair, Anoop Cherian, Suhas Lohit, Ye Wang, Toshiaki Koike-Akino, Vishal M Patel, and Tim K Marks. Steered diffusion: A generalized framework for plug-and-play conditional image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20850–20860, 2023. 1
- [31] Haomiao Ni, Bernhard Egger, Suhas Lohit, Anoop Cherian, Ye Wang, Toshiaki Koike-Akino, Sharon X Huang, and Tim K Marks. T2v-zero: Zero-shot image conditioning for text-to-video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9015–9025, 2024. 2
- [32] Yao Ni and Piotr Koniusz. Nice: Noise-modulated consistency regularization for data-efficient gans. *Advances in Neural Information Processing Systems*, 36:13773–13801, 2023. 2, 3, 4
- [33] Yao Ni and Piotr Koniusz. Chain: Enhancing generalization in data-efficient gans via lipschitz continuity constrained normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6763–6774, 2024. 1
- [34] Yao Ni, Dandan Song, Xi Zhang, Hao Wu, and Lejian Liao. Cagan: Consistent adversarial training enhanced gans. In *IJCAI*, pages 2588–2594, 2018. 3
- [35] Yao Ni, Shan Zhang, and Piotr Koniusz. Pace: Marrying generalization in parameter-efficient fine-tuning with consistency regularization. *Advances in Neural Information Processing Systems*, 37:61238–61266, 2024. 2, 3, 4
- [36] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1
- [37] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1
- [38] 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. *Advances in Neural Information Processing Systems*, 36:79320–79362, 2023. 2
- [39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 5
- [40] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023. 3
- [41] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. Dreambooth3d: Subject-driven text-to-3d generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2349–2359, 2023. 1, 2
- [42] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1
- [43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1
- [44] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30, 2017. 2
- [45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 5
- [46] Shuvendu Roy and Ali Etemad. Consistency-guided prompt learning for vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024. 3
- [47] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 1, 2, 3, 5, 6
- [48] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6527–6536, 2024. 2
- [49] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1
- [50] Kuniaki Saito, Donghyun Kim, and Kate Saenko. Openmatch: Open-set semi-supervised learning with open-set consistency regularization. *Advances in Neural Information Processing Systems*, 34:25956–25967, 2021. 3

- [51] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1
- [52] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instant-booth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8543–8552, 2024. 1
- [53] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 3
- [54] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023. 2
- [55] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2
- [56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 3
- [57] 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. 2
- [58] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024. 1
- [59] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023. 2
- [60] Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905, 2021. 3
- [61] Yi Wu, Ziqiang Li, Heliang Zheng, Chaoyue Wang, and Bin Li. Infinite-id: Identity-preserved personalization via id-semantics decoupling paradigm. In *European Conference on Computer Vision*, pages 279–296. Springer, 2024. 1
- [62] 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. 8
- [63] Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. Consistency regularization for generative adversarial networks. *arXiv preprint arXiv:1910.12027*, 2019. 3
- [64] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 8
- [65] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient finetuning. In *The Eleventh International Conference on Learning Representations*, 2023. 2
- [66] Xinxi Zhang, Song Wen, Ligong Han, Felix Juefei-Xu, Akash Srivastava, Junzhou Huang, Hao Wang, Molei Tao, and Dimitris N Metaxas. Spectrum-aware parameter efficient fine-tuning for diffusion models. *arXiv preprint arXiv:2405.21050*, 2024. 2
- [67] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 1
- [68] Jiacheng Zhu, Kristjan Greenewald, Kimia Nadjahi, Haitz Sáez de Ocáriz Borde, Rickard Brüel Gabriellsson, Leshem Choshen, Marzyeh Ghassemi, Mikhail Yurochkin, and Justin Solomon. Asymmetry in low-rank adapters of foundation models. *arXiv preprint arXiv:2402.16842*, 2024. 2