ANCHOR TO EXPAND: SEMANTIC ANCHORING FOR PERSONALIZED TEXT-TO-IMAGE DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalizing text-to-image diffusion models extends pretrained models to represent novel user-specific concepts from only a few reference images. However, learning a new concept while building on the prior knowledge of the pretrained model remains a key challenge. When personalization focuses on learning the target concept, the model tends to overfit the reference examples and degrade its general capability. In contrast, emphasizing prior preservation can hinder capturing distinctive personalized attributes. In this paper, we address this challenge by viewing a personalized concept as an underrepresented concept whose semantic counterpart is well represented in the pretrained model. Rather than treating the learning of a new concept and prior preservation as separate objectives, we reformulate them as a single anchored learning problem. We therefore introduce *Semantic Anchoring Personalization* (SAP), which keeps concept learning grounded in the pretrained semantic structure while capturing subject-specific attributes. The proposed objective offers a simple yet effective formulation that can be applied across different model backbones without architectural modifications or auxiliary networks. Extensive experiments across various settings demonstrate that SAP achieves a better balance between subject fidelity and text-image alignment than baseline methods. Further ablation studies validate the contribution of semantic anchoring to personalization.

1 INTRODUCTION

Recent advancements in text-to-image diffusion models (Ramesh et al., 2021; 2022; Rombach et al., 2022; Schuhmann et al., 2022; Podell et al., 2023; Esser et al., 2024) have established powerful foundation models for generic content generation. Despite their strong performance, these models still struggle to depict user-preferred details that are not captured by textual conditioning. To address this gap, personalized generation has emerged as a major application of text-to-image generative models (Gal et al., 2023; Ruiz et al., 2023). Rather than relying on samples drawn from the learned distribution, personalization aims to introduce user-specific attributes into the generation process.

However, a key challenge in personalization is capturing user-desired attributes from only a handful of examples while leveraging the prior knowledge of the pretrained model. This few-shot personalization often leads the model to overfit to limited data and lose the knowledge encoded in the pretrained prior. To mitigate this issue, existing methods constrain the update components of the pretrained model to prevent such drift. These approaches include selective layer updates (Kumari et al., 2023; Tewel et al., 2023), additional normalization (Han et al., 2023; Qiu et al., 2023), and parameter-efficient modules (Hu et al., 2022; Yeh et al., 2024; Chen et al., 2025). While restricting parameter updates enhances training stability, it also limits sufficient learning of target attributes.

Alongside efforts to preserve the prior knowledge of the pretrained model, recent studies have explored ways to better capture fine-grained personalized attributes. Building on Classifier-Free Guidance (Dhariwal & Nichol, 2021; Ho & Salimans, 2022), some methods leverage guidance from the pretrained model (Park et al., 2025b; Phunyaphibarn et al., 2026). Other approaches disentangle the target concept from irrelevant attributes using image-specific descriptions obtained from additional vision-language models (Kim et al., 2026; Liu et al., 2026). These approaches improve the representation of personalized attributes using additional semantic information. Prior

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

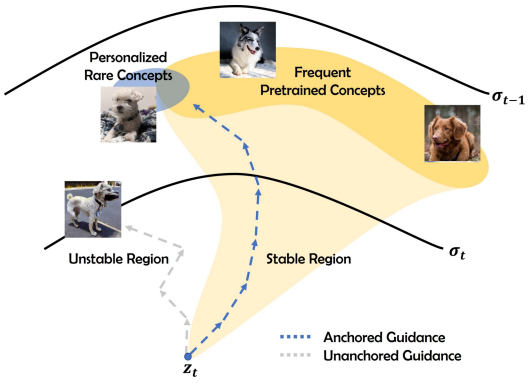

Figure 1: Conceptual illustration of our method, Semantic Anchoring Personalization (SAP). Pretrained semantics provide stable guidance, whereas guidance in newly introduced regions remains unstable. Our approach anchors personalization to pretrained semantics, enabling stable guidance as the model expands toward novel concepts.

studies have explored these complementary directions independently, treating prior preservation and personalized attribute learning as separate problems.

In this paper, we view learning personalized attributes and preserving pretrained semantics through the lens of rare concept generation (Samuel et al., 2023; 2024; Um & Ye, 2025). Rare concept generation aims to improve the generation of specific concepts that are underrepresented in pretrained models. This setting resembles personalization, where the model learns fine-grained user-specific attributes from only a few examples. In particular, Rare-to-Frequent (Park et al., 2025a) provides a theoretical basis by showing that a concept underrepresented in the pretrained model can benefit from guidance provided by its more common semantic counterpart. We extend this insight to personalization by treating a personalized concept represented by only a few reference images as a rare concept and anchoring its learning to its more common semantic counterpart.

Therefore, we propose *Semantic Anchoring Personalization* (SAP), a reformulated personalization objective that models the relation between the target concept and its pretrained semantic counterpart. As illustrated in Fig. 1, SAP learns the personalized concept as an extension of the pretrained semantic structure with the general semantic counterpart serving as an anchor. Our analysis further shows that the relation between a personalized concept and its general semantic counterpart weakens as personalization proceeds. This divergence indicates that the personalized concept loses alignment with the pretrained semantic structure. This observation suggests that SAP can mitigate such divergence by grounding subject-specific learning in pretrained semantics. Quantitative and qualitative results demonstrate consistent improvements across different model backbones, while ablation studies validate the role of semantic anchoring in balancing subject fidelity and text-image alignment. In summary, our main contributions are as follows:

- We introduce a perspective on concept learning and pretrained semantic preservation in personalization based on the relationship between underrepresented and well-represented concepts.
- We propose *Semantic Anchoring Personalization* (SAP), a reformulated personalization objective that incorporates the relation between the target concept and its pretrained semantic counterpart.
- Our experiments and analyses demonstrate that our method achieves consistent improvements in fidelity and generalization across diverse backbones and outperforms existing baselines.

2 Related Work

Personalization methods have explored which model components should be adapted for a new concept. DreamBooth (Ruiz et al., 2023) trains all model parameters, which can make personalization sensitive

to overfitting on limited reference images. In contrast, Textual Inversion (Gal et al., 2023) optimizes a token-level representation, which often suffers from limited expressiveness in capturing the target concept. These challenges have motivated subsequent studies to identify which parameters should be adapted for personalization. Custom Diffusion (Kumari et al., 2023) trains the attention layers, while other studies modify specific components such as singular values (Han et al., 2023) or angular features (Qiu et al., 2023) of weight matrices. Low-rank adaptation strategies (Hu et al., 2022) are also widely studied for personalization. They introduce a small set of trainable parameters while keeping the original weights fixed. Recent work investigates how to improve the initialization (Yeh et al., 2024) or training efficiency (Chen et al., 2025). While restricting parameter updates can stabilize personalization, selecting which components to adapt may need to be reconsidered for different model architectures. Moreover, these approaches preserve pretrained semantics through frozen parameters rather than incorporating them into the learning of the target concept.

Recent studies have shifted their focus from parameter adaptation toward better capturing fine-grained personalized attributes. Park et al. (2025b) and Phunyaphibarn et al. (2026) leverage pretrained unconditional guidance as an inference-time control signal. However, these methods do not address the learning of the personalized representation itself (Soboleva et al., 2025). Liu et al. (2026) and Kim et al. (2026) use image-specific captions to disentangle target attributes from irrelevant factors in the reference images. However, these approaches rely on additional vision-language models to obtain such descriptions during personalization. Some methods instead use superclass semantics available within the personalization setup (Ruiz et al., 2023; Tewel et al., 2023). Huang et al. (2025) aligns personalized token representations with the superclass, while Qiao et al. (2024) regularizes personalized predictions to remain within the pretrained semantic space using unconditional and class-level outputs. These approaches use general semantics as preservation constraints during adaptation. In contrast, we view the personalized concept and its general counterpart as a rare-to-frequent pair and incorporate their relationship into the learning objective.

3 METHOD

In this section, we first introduce the notation and the baseline formulation used in personalization. Next, we show that the learned personalized representation diverges from the pretrained semantics throughout personalization. Motivated by this observation, we reformulate the personalization objective through semantic anchoring to retain its connection to the pretrained semantic structure.

3.1 BACKGROUND

Text-to-image (T2I) diffusion models learn to generate an image latent z_0 from a textual condition c by reversing a gradual noising process (Song et al., 2020a;b; Nichol & Dhariwal, 2021; Luo, 2022). At each diffusion timestep t, a noisy latent z_t is obtained as $z_t = \alpha_t z_0 + \sigma_t \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$, and α_t and σ_t control the signal and noise scales. A denoising network $\epsilon_\theta(z_t, c, t)$ is trained to predict the added noise ϵ through the noise-prediction objective (Ho et al., 2020):

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{z_t, c, \epsilon, t}\left[\|\epsilon - \epsilon_\theta(z_t, c, t)\|_2^2\right]. \tag{1}$$

Personalization extends a pretrained T2I diffusion model parameterized by θ' into a personalized model θ that captures a specific subject. Let c^{sbj} denote the subject prompt (e.g., "`a photo of a [V*] dog`"), where `[V*]` denotes a special token representing the subject attribute (Ruiz et al., 2023). For a subject image, let z_t denote its noisy latent at timestep t. The personalized model learns the new subject from (z_t, c^{sbj}) following Eq. 1:

$$\mathcal{L}_{\text{sbj}} = \mathbb{E}_{z_t, c^{\text{sbj}}, \epsilon, t}\left[\|\epsilon - \epsilon_\theta(z_t, c^{\text{sbj}}, t)\|_2^2\right]. \tag{2}$$

Prior Preservation Loss (PPL) introduces an additional objective using class-level samples generated by the pretrained model (Ruiz et al., 2023). The pretrained model generates z^{cls} from a class prompt c^{cls} (e.g., "`a photo of a dog`"), encouraging the personalized model to preserve the pretrained behavior on these samples during personalization. Combining Eq. 2 with PPL has become a standard training scheme for diffusion-based personalization. However, the two objectives treat subject learning and prior preservation separately without considering their semantic relation.

3.2 Semantic Divergence in Personalization

In this paper, we treat pretrained semantics not as an independent preservation target, but as guidance for learning the personalized concept. Our motivation comes from rare concept generation, which aims to improve the generation of underrepresented concepts. Such concepts are less reliably represented because they receive limited support in the training distribution. In particular, Rare-to-Frequent (Park et al., 2025a) studies whether a related frequent concept can provide guidance for such underrepresented concepts. It formalizes the relation between a rare concept c^{rar} and its related frequent concept c^{frq} through score interpolation:

Theorem 1. *Under a Gaussian model where the frequent concept is well represented and the rare concept has estimation uncertainty characterized by σ, consider the interpolated score*

$$s_\alpha(z_t) = \alpha \nabla_{z_t} \log p_\theta(z_t|c^{\text{rar}}) + (1-\alpha)\nabla_{z_t} \log p_\theta(z_t|c^{\text{frq}}) \tag{3}$$

Let $p_{\text{lerp}}(x|c^{\text{rar}};\alpha)$ denote the distribution corresponding to this interpolated score, and let μ_{rar} and μ_{frq} denote the means of the rare and frequent concept distributions. If $\sigma \geq 1 + \sqrt{\|\mu_{\text{frq}} - \mu_{\text{rar}}\|_2^2 + 0.2}$, then

$$\min_\alpha W_2(p_{\text{lerp}}(x|c^{\text{rar}};\alpha), p_{\text{data}}(x|c^{\text{rar}})) < W_2(p_\theta(x|c^{\text{rar}}), p_{\text{data}}(x|c^{\text{rar}}))\,, \tag{4}$$

where W_2 denotes the 2-Wasserstein distance.

Eq. 4 shows that an underrepresented concept can be better approximated by leveraging a related well-represented concept rather than using the rare concept alone. This highlights the importance of considering an underrepresented concept in relation to its well-represented semantic counterpart.

A similar relation arises in personalization, where a personalized concept learned from only a few reference images can be regarded as an underrepresented concept, while its general semantic counterpart is already well represented in the pretrained model. Under this correspondence, we assume that the personalized concept should maintain its relation to its pretrained semantic counterpart. To examine how this relation evolves, we measure $\|\epsilon_\theta(z_t, c^{\text{sbj}}, t) - \epsilon_{\theta'}(z_t^{\text{cls}}, c^{\text{cls}}, t)\|_2$ between the personalized subject prediction and its pretrained class-level counterpart throughout personalization. As shown by the green curve in Fig. 2, the distance increases when optimizing the objective in Eq. 2. Even with PPL (blue curve), the increasing trend persists as personalization proceeds. This observation indicates that prior preservation does not ensure that the personalized concept remains aligned with its pretrained semantic counterpart. This divergence can isolate the personalized concept from the pretrained semantic structure as illustrated in Fig. 1. From the perspective of Theorem 1, preserving this relation is essential for leveraging guidance from the well-represented semantic counterpart.

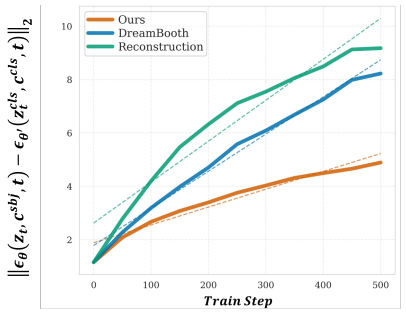

Figure 2: L_2 distance between the personalized subject prediction and its pretrained class-level counterpart throughout the personalization process.

3.3 Semantic Anchoring Personalization

Motivated by the analysis above, we consider a personalization objective that learns the target concept while accounting for its relation to the pretrained semantic counterpart. To this end, we reformulate the rare-to-frequent guidance in Eq. 3 as a training objective for personalization. We first express the score-based guidance under the noise-prediction parameterization. By Tweedie's formula (Efron, 2011; Song et al., 2020b), the conditional score at timestep t can be written as $\nabla_{z_t} \log p(z_t|c) = -\epsilon(z_t, c, t)/\sigma_t$. Substituting this relation into Eq. 3, we obtain

$$s_\alpha(z_t) \propto -\left[\lambda \epsilon^{\text{rar}} + (1-\lambda)\epsilon^{\text{frq}}\right]. \tag{5}$$

We then set the interpolated noise prediction in Eq. 5 as the training target itself. This formulation incorporates the pretrained counterpart into the training target rather than adding it as a separate preservation term. The following proposition shows that optimizing this target can be reformulated as rare-concept reconstruction anchored to its frequent semantic counterpart:

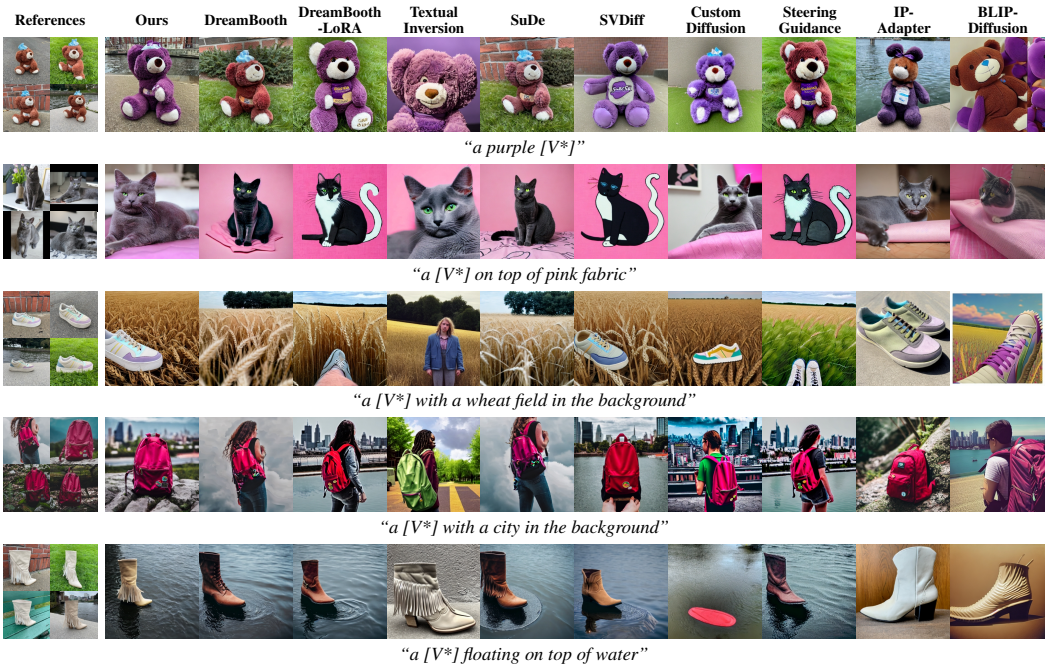

Figure 3: Visual comparison of few-shot personalization and encoder-based methods on SD1.5.

Proposition 1. *Let* $\lambda \in (0, 1]$ *and define* $\epsilon^* = \lambda\epsilon^{\text{rar}} + (1-\lambda)\epsilon^{\text{frq}}$. *Using* ϵ^* *as the training target yields the same optimal parameters as*

$$\theta^* = \arg\min_\theta \mathbb{E}\left[\|\epsilon^{\text{rar}} - \epsilon_\theta\|_2^2 + \frac{1-\lambda}{\lambda}\left\|\epsilon_\theta - \epsilon^{\text{frq}}\right\|_2^2\right]. \tag{6}$$

The proof is provided in the appendix E.

Eq. 6 shows that learning a rare concept requires not only reconstructing its target representation but also remaining anchored to its frequent semantic counterpart.

Following the correspondence established above, we instantiate the derived formulation for personalization by treating the personalized subject as the rare concept and its class-level semantics as the frequent counterpart. Accordingly, ϵ^{rar} corresponds to the ground-truth noise ϵ used as the subject target in Eq. 2, while ϵ^{frq} is given by the prediction of the pretrained model $\epsilon_{\theta'}(z_t, c^{\text{cls}}, t)$. Substituting these terms into Eq. 6 yields:

$$\begin{aligned}\mathcal{L}_{\text{SAP}} = \mathbb{E}_{z_t, c^{\text{sbj}}, c^{\text{cls}}, \epsilon, t}\Big[&\|\epsilon - \epsilon_\theta(z_t, c^{\text{sbj}}, t)\|_2^2 \\ &+ w\|\epsilon_\theta(z_t, c^{\text{sbj}}, t) - \epsilon_{\theta'}(z_t, c^{\text{cls}}, t)\|_2^2\Big],\end{aligned} \tag{7}$$

where $w = (1-\lambda)/\lambda$ and θ' remains frozen during personalization.

We introduce the objective in Eq. 7 as *Semantic Anchoring Personalization* (SAP). The first term learns the subject-specific attributes, and the second term anchors the personalized prediction to its pretrained class-level counterpart. Anchoring the personalized prediction to its pretrained class-level counterpart preserves its connection to the pretrained semantic structure as illustrated in Fig. 1. Unlike prior approaches that separate concept learning from the preservation of pretrained semantics, our objective reframes personalization as learning the target concept in relation to its pretrained semantic counterpart. As a result, it reduces the semantic divergence observed in Fig. 2 (orange line). Moreover, this derivation yields an objective that can be applied to the standard personalization setting without additional components. In practice, SAP leads to consistent performance improvements across the settings in Sec. 4, suggesting that preserving this semantic relation contributes to effective personalization.

Table 1: Quantitative comparison with baseline methods on SD1.5. The Rank column is computed by averaging the per-metric ranks.

Method	CLIP-I ↑	CLIP-T ↑	DINO ↑	Rank
Few-shot personalization methods				
Ours	0.7833	0.3064	0.6051	**1**
DreamBooth	0.7761	0.3039	0.5895	4
DreamBooth-LoRA	0.7314	0.3195	0.4866	6
SuDe	0.7800	0.3047	0.5953	3
SVDiff	0.7470	0.3167	0.5171	4
Textual Inversion	0.7599	0.2578	0.5330	9
Custom Diffusion	0.6479	0.3154	0.5568	6
Steering Guidance	0.7381	0.3133	0.5064	8
Encoder-based methods				
IP-Adapter	0.8311	0.2646	0.6290	2
BLIP-Diffusion	0.7016	0.2949	0.4544	10

Table 2: Quantitative comparison on SD1.5, SDXL, and SD3. The Rank column is computed by averaging the per-metric ranks.

Model	Method	CLIP-I ↑	CLIP-T ↑	DINO ↑	Rank
SD1.5	DreamBooth-LoRA	0.7314	0.3195	0.4866	2
	Beyond-Finetuning	0.7294	**0.3212**	0.4786	3
	Ours	**0.7833**	0.3064	**0.6051**	**1**
SDXL	DreamBooth-LoRA	0.7939	0.3107	0.6586	2
	Beyond-Finetuning	0.7872	0.3141	0.6429	3
	Ours	**0.8005**	**0.3154**	**0.6661**	**1**
SD3	DreamBooth-LoRA	0.7449	0.3185	0.5109	2
	Beyond-Finetuning	0.7443	**0.3226**	0.5037	2
	Ours	**0.7579**	0.3187	**0.5679**	**1**

4 EXPERIMENTS

In this section, we present experiments and analyses to validate the effectiveness of our method across various subjects and diffusion backbones. For further analysis beyond metrics, we also conduct a user study to assess perceptual quality and preference. Additional qualitative results and detailed analyses are provided in Appendix G.

4.1 EXPERIMENTAL SETUP

Datasets. We conduct experiments on the widely used personalization benchmark (Ruiz et al., 2023). The benchmark contains 30 subjects, including both objects and live subjects. Each subject provides 4–6 reference images, and each model is evaluated using 25 distinct textual templates with four generated samples per prompt to measure subject and text consistency. We use c^{sbj}: "`a photo of [V*] [Subject]`" as the personalized prompt and c^{cls}: "`a photo of a [Subject]`" as the class-level anchor prompt. This setup follows the standard personalization setting, and we introduce no additional prompt modifications.

Evaluation Metrics. We evaluate the generated results along two dimensions. Subject fidelity, quantified by CLIP-I (Radford et al., 2021) and DINO (Caron et al., 2021), measures the cosine similarity between image embeddings of generated images and reference samples. Text fidelity, measured by CLIP-T, evaluates text-image alignment through the cosine similarity between CLIP text and image embeddings. Since subject fidelity and text fidelity often exhibit a trade-off, evaluating either dimension alone may not reflect the overall personalization performance. We therefore report an aggregate rank by averaging the per-metric ranks across CLIP-I, CLIP-T, and DINO, where a lower rank indicates a better balance between subject fidelity and text alignment.

Baselines. We compare our method against several personalization approaches. The baselines include DreamBooth (Ruiz et al., 2023) with both full fine-tuning and LoRA-based variants, SuDe (Qiao et al., 2024), SVDiff (Han et al., 2023), Textual Inversion (Gal et al., 2023), Custom Diffusion (Kumari et al., 2023), and Steering Guidance (Park et al., 2025b). We also include encoder-based methods, including IP-Adapter (Ye et al., 2023) and BLIP-Diffusion (Li et al., 2023a). For comparisons across personalization methods, we adopt SD1.5, where official implementations and training recipes are available, facilitating consistent comparison across methods. Furthermore, we evaluate our method across SD1.5 (Rombach et al., 2022), SDXL (Podell et al., 2023), and SD3 (Esser et al., 2024) to examine its applicability across different model backbones. For this cross-backbone evaluation, we compare with Beyond-Finetuning (Soboleva et al., 2025), which applies class-level guidance at inference time, allowing comparison with our training-time formulation. For all experiments, our method builds upon LoRA-based adaptation.

4.2 COMPARISON WITH BASELINE METHODS

Table 1 summarizes quantitative comparisons with baseline methods across subject fidelity and text alignment metrics. Our method obtains the top rank under this evaluation, indicating the best

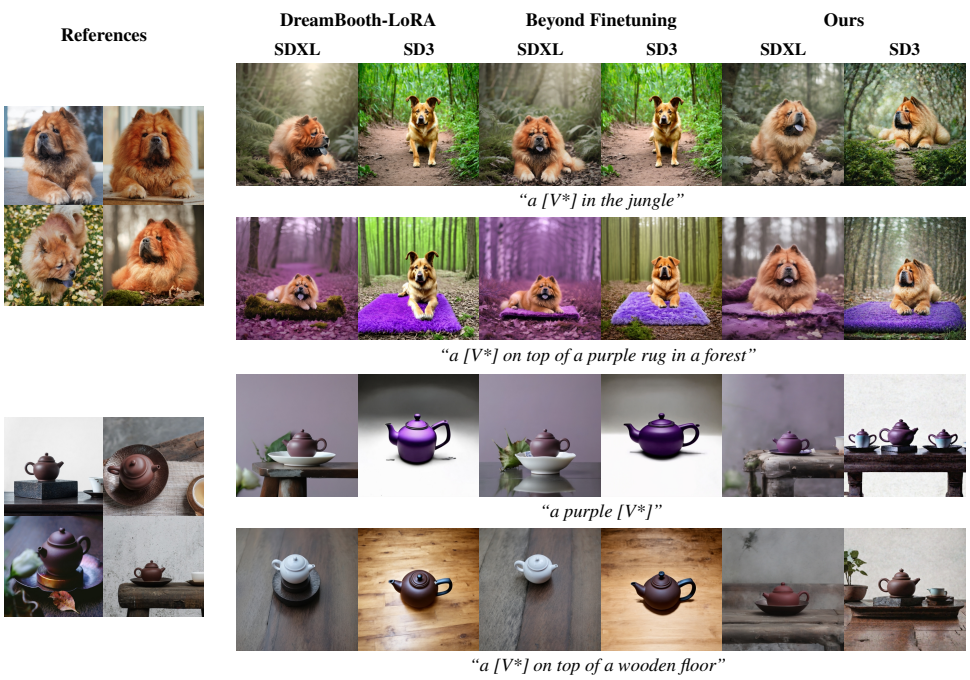

Figure 4: Qualitative comparison on SDXL and SD3 backbones among DreamBooth-LoRA, Beyond-Finetuning, and our method.

balance between subject fidelity and text alignment among the compared methods. In particular, SAP improves upon DreamBooth and SuDe across all three metrics. Compared with DreamBooth-LoRA, SAP yields substantial gains in CLIP-I and DINO under the same LoRA adaptation setting, highlighting the effect of the proposed training objective on learning subject-specific attributes. Although DreamBooth-LoRA achieves a higher CLIP-T score, further qualitative analysis suggests that this gain is associated with underfitting to the personalized subject, leading to weaker subject fidelity despite stronger text alignment.

The importance of balancing subject fidelity and text alignment becomes more evident in the qualitative results as illustrated in Fig. 3. In the third and fifth rows, SVDiff, Custom Diffusion, and Steering Guidance follow the target prompts but fail to preserve distinctive subject details. In contrast, DreamBooth and SuDe preserve the subject appearance, but the first and fourth rows show cases where the generated images remain close to the reference images and do not reflect the target prompt. These examples illustrate the trade-off between subject fidelity and prompt alignment and motivate evaluating both aspects jointly. SAP maintains a better balance across these cases. Although subject–prompt mixing appears in the second row, the generated results preserve subject identity while following the target prompt. We further support these findings with a user study and additional analyses provided in the Appendices G and I.

4.3 Comparison across Backbone Models

We further evaluate SAP across SD1.5, SDXL, and SD3 backbones to examine whether its effectiveness extends beyond a single model architecture, as shown in Table 2. SAP attains the top rank across all three backbones, based on the same ranking procedure used in Table 1. On SDXL, SAP achieves the highest scores across all metrics. On SD1.5 and SD3, SAP yields higher subject fidelity while maintaining comparable text alignment with DreamBooth-LoRA and Beyond-Finetuning. The gains over DreamBooth-LoRA demonstrate the benefit of the proposed training objective across different backbones. The comparison with Beyond-Finetuning further suggests that incorporating class-level guidance into the training objective provides benefits beyond applying it only at inference time.

Qualitative results further support these findings as illustrated in Fig. 4. For the dog subject, DreamBooth-LoRA and Beyond-Finetuning produce faithful personalized results on SDXL, but fail to preserve subject identity on SD3. For the teapot subject, both methods show the opposite trend,

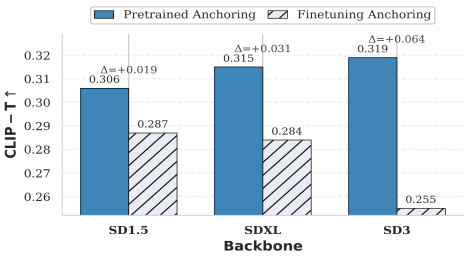

Figure 5: Comparison of semantic alignment (CLIP-T) using two different anchoring strategies. Pretrained anchoring outperforms finetuning anchoring.

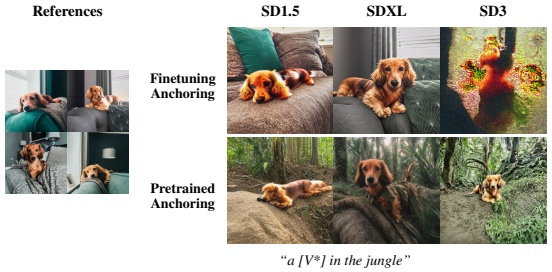

Figure 6: Comparison of qualitative results using two different anchoring strategies. Both preserve subject fidelity, and the pretrained anchoring shows better consistency to the text description.

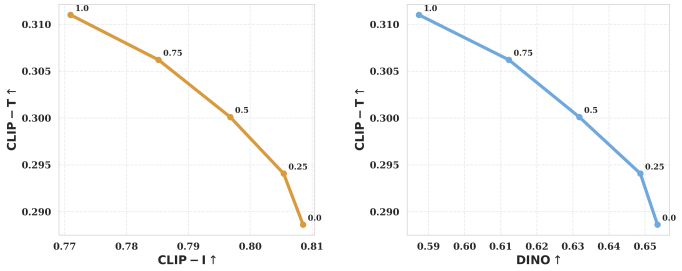

Figure 7: Analysis of subject fidelity and text alignment under different weighting parameters w. Left: relationship between CLIP-I and CLIP-T. Right: relationship between DINO and CLIP-T.

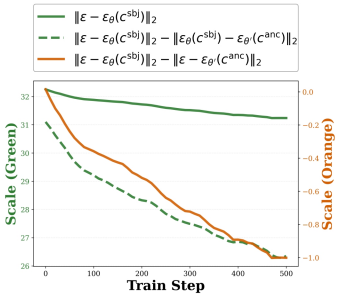

Figure 8: Analysis of distance relationships among reference, subject, and anchor predictions.

producing appropriate results on SD3 but either failing to reflect the target prompt or losing subject identity on SDXL. In contrast, SAP preserves distinctive subject details such as fur color, snout shape, and posture, resulting in the highest CLIP-I and DINO scores. It also captures contextual cues such as "`on top of a purple rug`" and "`in a forest`". A similar pattern is observed in the teapot examples (third and fourth rows). These results indicate that semantic anchoring yields more consistent personalization across diffusion backbones.

5 Ablation Studies

In this section, we conduct ablation studies to analyze the contribution of each component in semantic anchoring and its impact on the overall personalization process. We also provide additional ablation studies in Appendix H.

5.1 Anchoring Effect on Personalization

We investigate how different balances between learning the personalized subject and maintaining its relation to the pretrained semantic counterpart affect the final personalization results. In Eq. 7, we vary the anchoring weight $w \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$ to control the relative strength of subject learning and semantic anchoring. As shown in Fig. 7, increasing w improves CLIP-T but decreases CLIP-I and DINO, whereas reducing w produces the opposite trend. At $w = 0$, CLIP-I exceeds 0.8 whereas CLIP-T falls below 0.3. This setting favors subject fidelity over text alignment. At $w = 1.0$, CLIP-T exceeds 0.3 whereas DINO falls below 0.6, showing the opposite trade-off. An intermediate value of $w = 0.5$ maintains both visual fidelity and semantic consistency. This behavior indicates that the proposed objective balances personalized learning with pretrained semantics across different weighting regimes, rather than relying on a single operating point.

5.2 Reference Model Selection for Semantic Anchoring

To examine reference model selection for semantic anchoring, we vary the source of the anchor prediction between the pretrained weights θ' and the current personalized weights θ. This comparison evaluates which reference better serves as a semantic anchor for personalization. Fig. 5 shows that the pretrained anchor setting achieves higher CLIP-T scores across all backbones. CLIP-T remains above 0.3 with the pretrained anchor, whereas it falls below 0.3 with the jointly finetuned anchor. Since text coherence reflects the pretrained semantic knowledge, these results indicate that preserving the pretrained reference is essential for maintaining the original contextual semantics. When the anchor reference is also finetuned, both the subject and anchor predictions evolve simultaneously, causing the semantic reference point to drift. Consequently, the model loses text alignment capability. This observation is consistent with our formulation in Eq. 7, encouraging the personalized prediction to remain connected to the pretrained semantic structure during adaptation.

5.3 Ablation on the Semantic Anchor Choice

To verify that SAP benefits from the semantic relation to the anchor rather than from arbitrary anchoring, we vary the anchor semantics. Compared to using the corresponding class concept as the anchor (e.g., dog→dog), we vary the anchor to a related category (e.g., dog→cat), a broader superclass (e.g., dog→animal), or an unrelated concept (e.g., dog→shark). In our experiments, for a live subject (dog), we test a set of anchors, including *dog*, *cat*, *animal*, and *shark*. We further conduct the same analysis for an object concept (teapot). The results are summarized in Table 3. Using an unrelated semantic anchor (teapot→clock) leads to a clear performance drop. Similarly, overly broad anchors (teapot→container) also degrade performance. These results indicate that our method leverages the pretrained semantic prior through the anchor. They also show that the corresponding class concept provides an effective and reliable reference for semantic anchoring.

Table 3: Ablation study on anchor semantics.

Group	Setting	CLIP-I ↑	CLIP-T ↑	DINO ↑
dog	**dog→dog (Ours)**	**0.8108**	**0.3059**	**0.6230**
	dog→cat	0.8076	0.3031	0.6170
	dog→animal	0.8027	0.3039	0.6130
	dog→shark	0.7965	0.3026	0.5918
teapot	**teapot→teapot (Ours)**	**0.8329**	**0.3400**	**0.5002**
	teapot→kettle	0.8024	0.3318	0.4488
	teapot→container	0.8184	0.3287	0.4488
	teapot→clock	0.8026	0.3270	0.4331

5.4 Semantic Space Dynamics

To better understand the semantic dynamics throughout personalization, we analyze the distances among the components of our objective. As shown in Fig. 8, the distance $\|\epsilon - \epsilon_\theta(z_t, c^{\text{sbj}}, t)\|_2$ decreases as training progresses. This indicates that the personalized representation learns subject-specific attributes. The dashed green line shows that the personalized prediction moves away from the pretrained class-level anchor as the subject is learned. At the same time, the orange line shows that the personalized prediction becomes closer to the subject target than the pretrained class-level prediction. However, as shown in Fig. 2, SAP exhibits the smallest divergence from the pretrained counterpart among the compared objectives. Together, these observations are consistent with the intended behavior of SAP. Rather than constraining the personalized concept within pretrained semantics (Qiao et al., 2024), SAP allows it to expand from the pretrained semantic reference toward subject-specific semantics while maintaining its connection to the anchor as illustrated in Fig. 1.

6 Conclusion

In this paper, we took a joint view of concept learning and prior preservation, which are often treated as separate objectives in personalization. Inspired by the relation between underrepresented and well-represented concepts, we reformulated the learning target to incorporate the corresponding semantic counterpart of the target concept. The proposed method anchors the personalized concept to its pretrained semantic counterpart throughout personalization, allowing subject-specific semantics to develop from the pretrained semantic structure rather than in isolation. Experiments show that the proposed method achieves a better balance between subject fidelity and text alignment than existing baselines. Results across multiple diffusion backbones further demonstrate the effectiveness of the proposed objective across different model architectures. Our ablation studies validate the role of semantic anchoring and provide further analysis of its key design choices. In conclusion, we present Semantic Anchoring Personalization, a simple yet effective personalization framework that learns novel user-specific concepts as extensions of pretrained semantics.

REFERENCES

Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Minhao Cheng, Boqing Gong, and Cho-Jui Hsieh. Understanding the impact of negative prompts: When and how do they take effect? In *european conference on computer vision*, pp. 190–206. Springer, 2024.

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*, pp. 9650–9660, 2021.

Shangyu Chen, Zizheng Pan, Jianfei Cai, and Dinh Phung. Para: Personalizing text-to-image diffusion via parameter rank reduction. In *ICLR*, 2025.

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36:30286–30305, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

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 *International Conference on Learning Representations*, 2023.

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*, pp. 7323–7334, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Jiannan Huang, Jun Hao Liew, Hanshu Yan, Yuyang Yin, Yao Zhao, Humphrey Shi, and Yunchao Wei. Classdiffusion: More aligned personalization tuning with explicit class guidance. In *International Conference on Learning Representations*, 2025.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024.

Minseo Kim, Minchan Kwon, Dongyeun Lee, Yunho Jeon, and Junmo Kim. Conceptprism: Concept disentanglement in personalized diffusion models via residual token optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2381–2390, June 2026.

Felix Koulischer, Johannes Deleu, Gabriel Raya, Thomas Demeester, and Luca Ambrogioni. Dynamic negative guidance of diffusion models. *arXiv preprint arXiv:2410.14398*, 2024.

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*, pp. 1931–1941, 2023.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, 37:122458–122483, 2024.

Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 2023a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023b.

Shi Zhan Liu, Hao Zheng, Hang Yu, and Jianguo Li. Accord: Alleviating concept coupling through dependence regularization for text-to-image diffusion personalization. In *International Conference on Learning Representations*, 2026.

Calvin Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022.

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*, pp. 1–12, 2024.

Amin Karimi Monsefi, Mridul Khurana, Rajiv Ramnath, Anuj Karpatne, Wei-Lun Chao, and Cheng Zhang. Taxadiffusion: Progressively trained diffusion model for fine-grained species generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8579–8589, 2025.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Dongmin Park, Sebin Kim, Taehong Moon, Minkyu Kim, Kangwook Lee, and Jaewoong Cho. Rare-to-frequent: Unlocking compositional generation power of diffusion models on rare concepts with llm guidance. In *International Conference on Learning Representations*, 2025a.

Sunghyun Park, Seokeon Choi, Hyoungwoo Park, and Sungrack Yun. Steering guidance for personalized text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15907–15916, 2025b.

Prin Phunyaphibarn, Phillip Y. Lee, Jaihoon Kim, and Minhyuk Sung. Unconditional priors matter! improving conditional generation of fine-tuned diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3982–3992, 2026.

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.

Pengchong Qiao, Lei Shang, Chang Liu, Baigui Sun, Xiangyang Ji, and Jie Chen. Facechain-sude: Building derived class to inherit category attributes for one-shot subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7215–7224, 2024.

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.

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*, pp. 8748–8763. PMLR, 2021.

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*, pp. 8821–8831. Pmlr, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

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*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

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*, pp. 22500–22510, 2023.

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.

Dvir Samuel, Rami Ben-Ari, Nir Darshan, Haggai Maron, and Gal Chechik. Norm-guided latent space exploration for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 57863–57875, 2023.

Dvir Samuel, Rami Ben-Ari, Simon Raviv, Nir Darshan, and Gal Chechik. Generating images of rare concepts using pre-trained diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4695–4703, 2024.

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.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8543–8552, 2024.

Vera Soboleva, Maksim Nakhodnov, and Aibek Alanov. Beyond fine-tuning: A systematic study of sampling techniques in personalized image generation. *arXiv preprint arXiv:2502.05895*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

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*, pp. 1–11, 2023.

Soobin Um and Jong Chul Ye. Minority-focused text-to-image generation via prompt optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20926–20936, 2025.

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.

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*, pp. 15943–15953, 2023.

Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, 133(3):1175–1194, 2025.

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

Shih-Ying Yeh, Yu-Guan Hsieh, Zhidong Gao, Bernard BW Yang, Giyeong Oh, and Yanmin Gong. Navigating text-to-image customization: From lycoris fine-tuning to model evaluation. In *The Twelfth International Conference on Learning Representations*, 2024.

Yu Zeng, Vishal M Patel, Haochen Wang, Xun Huang, Ting-Chun Wang, Ming-Yu Liu, and Yogesh Balaji. Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6786–6795, 2024.

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*, pp. 3836–3847, 2023.

A Discussion and Future Work

Our findings demonstrate that introducing a semantic anchor offers a principled way to structure the interaction between pretrained and subject-specific representations. While our formulation anchors each corresponding semantics, the appropriate granularity and adaptiveness of this anchoring remain open questions. Exploring richer semantic structures and understanding how different semantic factors contribute to each subject's adaptation dynamics presents a promising direction for future work.

B Societal Impact

This work improves personalization in text-to-image diffusion models, enabling more controllable and expressive image generation. However, such capabilities may increase the risk of misuse, particularly in the creation of deepfakes or misleading visual content. To address these concerns, responsible use is essential, including transparency, detection tools, and safeguards such as watermarking. We emphasize the importance of considering these societal risks when deploying such systems.

C AI Use Statement

We used generative AI tools to assist with translation, improve the readability of the manuscript, and format LaTeX tables and equations. We did not use these tools for other research tasks requiring disclosure under the ICLR AI policy. All AI-assisted edits were reviewed by the authors to verify that they accurately reflected the intended meaning and technical content. The authors take full responsibility for the final content of this paper.

D Reproducibility Statement

Section 3.3 presents the formulation and training objective of SAP, and Sec. 4.1 describes the experimental setup and evaluation protocol. Appendix E provides the proof of Proposition 1, while Appendix F details the implementation settings, hyperparameters, baseline configurations, and computational resources. Additional qualitative results, ablation studies, and user-study details are provided in Appendices G and I.

E Proof of Proposition 1

Proposition (Restatement of Proposition 1)**.** *Let $\lambda \in (0, 1]$ and define $\epsilon^* = \lambda\epsilon^{\mathrm{rar}} + (1-\lambda)\epsilon^{\mathrm{frq}}$. Using ϵ^* as the training target yields the same optimal parameters as*

$$\theta^* = \arg\min_{\theta} \mathbb{E}\left[\|\epsilon^{\mathrm{rar}} - \epsilon_\theta\|_2^2 + \frac{1-\lambda}{\lambda}\left\|\epsilon_\theta - \epsilon^{\mathrm{frq}}\right\|_2^2\right].$$

Proof. For clarity, let

$$A = \epsilon_\theta - \epsilon^{\mathrm{rar}}, \qquad B = \epsilon_\theta - \epsilon^{\mathrm{frq}},$$

where we omit the input arguments of ϵ_θ for brevity. Since $\epsilon^* = \lambda\epsilon^{\mathrm{rar}} + (1-\lambda)\epsilon^{\mathrm{frq}}$, the prediction error can be written as

$$\begin{aligned}\epsilon_\theta - \epsilon^* &= \epsilon_\theta - \left(\lambda\epsilon^{\mathrm{rar}} + (1-\lambda)\epsilon^{\mathrm{frq}}\right) \\ &= \lambda\left(\epsilon_\theta - \epsilon^{\mathrm{rar}}\right) + (1-\lambda)\left(\epsilon_\theta - \epsilon^{\mathrm{frq}}\right) \\ &= \lambda A + (1-\lambda)B.\end{aligned}$$

Therefore,

$$\begin{aligned}\|\epsilon_\theta - \epsilon^*\|_2^2 &= \|\lambda A + (1-\lambda)B\|_2^2 \\ &= \lambda^2\|A\|_2^2 + (1-\lambda)^2\|B\|_2^2 + 2\lambda(1-\lambda)\langle A, B\rangle. \end{aligned} \tag{8}$$

Using

$$\langle A, B\rangle = \frac{1}{2}\left(\|A\|_2^2 + \|B\|_2^2 - \|A - B\|_2^2\right),$$

Eq. 8 becomes

$$\begin{aligned}\|\epsilon_\theta - \epsilon^*\|_2^2 = \lambda\|A\|_2^2 + (1-\lambda)\|B\|_2^2 \\ - \lambda(1-\lambda)\|A - B\|_2^2.\end{aligned} \tag{9}$$

Substituting A and B into Eq. 9 gives

$$\begin{aligned}\|\epsilon_\theta - \epsilon^*\|_2^2 = \lambda\,\|\epsilon_\theta - \epsilon^{\mathrm{rar}}\|_2^2 + (1-\lambda)\,\left\|\epsilon_\theta - \epsilon^{\mathrm{frq}}\right\|_2^2 \\ - \lambda(1-\lambda)\,\left\|\epsilon^{\mathrm{rar}} - \epsilon^{\mathrm{frq}}\right\|_2^2.\end{aligned} \tag{10}$$

Since ϵ^{rar} and ϵ^{frq} do not depend on the trainable parameters θ, the last term in Eq. 10 is constant with respect to θ. Taking the expectation gives

$$\mathbb{E}\left[\|\epsilon_\theta - \epsilon^*\|_2^2\right] = \lambda\mathbb{E}\left[\|\epsilon_\theta - \epsilon^{\mathrm{rar}}\|_2^2 + \frac{1-\lambda}{\lambda}\left\|\epsilon_\theta - \epsilon^{\mathrm{frq}}\right\|_2^2\right] + C,$$

where C is independent of θ.

Since $\lambda > 0$, the additive constant and positive scaling factor do not change the minimizer. Therefore,

$$\arg\min_\theta \mathbb{E}\left[\|\epsilon^* - \epsilon_\theta\|_2^2\right] = \arg\min_\theta \mathbb{E}\left[\|\epsilon^{\mathrm{rar}} - \epsilon_\theta\|_2^2 + \frac{1-\lambda}{\lambda}\left\|\epsilon_\theta - \epsilon^{\mathrm{frq}}\right\|_2^2\right] = \theta^*.$$

□

F Additional Implementation Details

This section provides additional implementation details for both the proposed method and the baseline setups in Sec. 4. All experiments were conducted on a single RTX 3090 or a single A6000 GPU.

Ours. We apply the proposed method to multiple backbones, and the corresponding hyperparameters are as follows. For SD 1.5, we use a learning rate of 5×10^{-4}, a semantic anchoring weight of $w = 0.8$, and a guidance scale of 7.5 during inference. For SDXL, the learning rate is also set to 5×10^{-4}, with $w = 1.0$ and a guidance scale of 7.0. For SD3, we use a learning rate of 2.5×10^{-4}, a semantic anchoring weight of $w = 0.05$, and an inference guidance scale of 10.0. The special token for all training prompts is set to "sks", and the number of inference steps is fixed to 50. Our method is implemented on top of the Diffusers library[1], and all settings follow the default configurations unless otherwise specified.

Baselines. We follow the configurations specified in the original papers or the official libraries when conducting the baseline experiments. For Textual Inversion (Gal et al., 2023)[2], Custom Diffusion (Kumari et al., 2023)[3], SVDiff (Han et al., 2023)[4], DreamBooth full fine-tuning, and DreamBooth (Ruiz et al., 2023) with LoRA (Hu et al., 2022)[5], we use the publicly available implementations provided in the corresponding repositories and follow their default training settings. For BLIP-Diffusion (Li et al., 2023a)[6] and IP-Adapter (Ye et al., 2023)[7], we use the publicly released pretrained models and evaluate them through inference. For SuDe (Qiao et al., 2024) and Steering Guidance (Park et al., 2025b), we re-implemented the methods following the descriptions in the papers based on the Diffusers library. For SuDe, we observed training instability, and therefore removed the denominator term σ_t in the objective. The modified version was used for training and subsequent evaluation.

[1] https://github.com/huggingface/diffusers

[2] https://github.com/huggingface/diffusers/tree/main/examples/textual_inversion

[3] https://github.com/huggingface/diffusers/tree/main/examples/custom_diffusion

[4] https://github.com/mkshing/svdiff-pytorch

[5] https://github.com/huggingface/diffusers/tree/main/examples/dreambooth

[6] https://github.com/salesforce/LAVIS/tree/main/projects/blip-diffusion

[7] https://github.com/tencent-ailab/IP-Adapter

G Additional Visual Results in Main Experiments

In this section, we present additional visual results and offer extended interpretation. We further include a detailed examination of failure cases.

G.1 Additional Comparison with Baseline Methods

Fig. 9 presents additional visual comparisons for Sec. 4.2. Our method demonstrates a balanced trade-off between subject fidelity and prompt coherence across a wide range of subjects. In the first, second, and third rows, the model trained with our approach preserves the subject identity while faithfully reflecting the given prompts. In contrast, DreamBooth introduces artifacts in the first row or fails to retain identity in cases such as the second row, and similar issues are also observed in SuDe. These observations are consistent with the improved image-level metrics reported in Table 1. SVDiff and Custom Diffusion generally follow the text prompts well; however, SVDiff tends to lose subject identity, as seen in the second row, while Custom Diffusion misses fine-grained details such as color in the first and third rows. Steering Guidance similarly reflects the requested attributes and contexts, but does not consistently preserve the distinctive appearance of the reference subjects. This aligns with their relatively high text-alignment scores but lower image-fidelity scores in Table 1. These results indicate that our model achieves a more balanced optimization across both dimensions, which is reflected in its highest position in the comprehensive ranking.

Although our method achieves meaningful improvements, several cases remain challenging. We examine these examples to analyze the limitations of the approach. While our method provides improved image fidelity and preserves subject identity more effectively than other baselines, the first row in Fig. 10 shows that the prompt is not fully reflected. In the second row, the generated results primarily capture the subject's texture and the prompt, but fail to represent identity-specific characteristics. The last row further illustrates that the model struggles to capture detailed attributes for rare concepts such as cartoon-style subjects. These observations indicate that, despite the overall improvements and balanced performance, the effectiveness of our method varies across subjects. As discussed in the Future Work section, this highlights the need for subject-dependent anchoring strategies and more adaptive balancing during the personalization process.

G.2 Additional Comparison across Backbone Models

Fig. 11 demonstrates additional visual comparisons for Sec. 4.3. The proposed method produces the most balanced results across different backbones. For the first subject, DreamBooth-LoRA and Beyond-Finetuning fail to preserve subject-specific characteristics in SD1.5 and SD3, while our approach generates coherent results across all backbones. This observation aligns with the substantial gains in the image fidelity scores (CLIP-I and DINO) reported in Table 2. A similar trend can be seen for the second subject in Fig. 11.

For the third subject, our method achieves the strongest subject fidelity in SDXL and SD3. Notably, SDXL yields strong results across all methods, which is consistent with its superior performance in Table 2. However, the SD1.5–based result of our method exhibits insufficient prompt adherence, corresponding to its lower text-alignment score. In contrast, the results of other methods fail to preserve subject characteristics, again reflecting the improvements in image fidelity enabled by our method. Taken together with the observations discussed in Sec. G.1, this analysis illustrates that personalization remains challenging and highly subject-dependent, highlighting the need for more advanced techniques that adapt to the characteristics of each subject.

H Extended Analyses in Ablation Studies

In this section, we provide extended analyses of the ablation studies. We expand the experiments in Sec. 5 with additional visual comparisons and further evaluations of alternative inference-time strategies.

H.1 Extended Analysis of Anchoring Effects on Personalization

Fig. 12 presents additional visual results that correspond to the analysis in Sec. 5.1. As shown in Fig. 12, varying the weighting parameter w produces a transition between subject fidelity and semantic alignment. For many subjects, the intermediate region around $w = 0.5$ yields the most coherent balance between retaining the reference identity and incorporating the desired attributes. Lower values of w retain more subject-specific appearance, while increasing w strengthens adherence to the textual attributes introduced by the pretrained anchor. The smooth transition observed across different w values empirically supports the interpretation in Fig. 7 that the proposed objective behaves as an interpolation between subject-specific semantics and the pretrained anchor prior. This also justifies our design choice that the training objective should guide the model toward an ideal blended target direction, reflecting the intended balance of identity and prompt attributes.

However, the bottom row in Fig. 12 shows that the optimal choice of w is not universal. In this example, higher values of w capture the prompt more faithfully without fully compromising identity, suggesting that different subjects may require different interpolation strengths. These subject-dependent variations indicate that no single fixed w is optimal across all cases. This observation motivates future work on adaptive anchoring mechanisms, where the effective anchor strength is dynamically adjusted based on the characteristics of each reference subject.

Table 4: Comparison on SD1.5, SDXL, SD3 using two different anchoring strategies.

Model	Method	CLIP-I ↑	CLIP-T ↑	DINO ↑
SD1.5	Finetuning Anchoring	0.8112	0.2870	0.6563
	Pretrained Anchoring (Ours)	0.7833	0.3064	0.6051
SDXL	Finetuning Anchoring	0.8407	0.2838	0.7109
	Pretrained Anchoring (Ours)	0.8005	0.3154	0.6661
SD3	Finetuning Anchoring	0.6182	0.2548	0.2685
	Pretrained Anchoring (Ours)	0.7579	0.3187	0.5679

H.2 Additional Comparison of Anchoring Strategy

Table 4 and Fig. 13 provide an extended comparison of the two anchoring strategies discussed in Sec. 5.2 to more comprehensively assess the differences between the two strategies. The qualitative results in Fig. 13 reveal a clear difference between the two anchoring strategies. Finetuning anchoring preserves subject fidelity well, but it fails to reflect the text information. In contrast, our semantic anchoring personalization captures the text prompt much more faithfully while still maintaining the subject characteristics provided in the references. When viewed together with Table 4, the results reveal an inherent trade-off that finetuned anchoring achieves slightly higher subject similarity, but pretrained anchoring consistently improves prompt alignment while maintaining reasonable identity preservation.

However, finetuned anchoring introduces an inherent limitation. Because the anchor branch is updated during personalization, it becomes entangled with the subject and is more susceptible to optimization instability. This effect is particularly evident in the SD3 results, where finetuned anchoring underperforms across all metrics. As further illustrated in Fig. 13, finetuned anchoring often causes mode collapse and fails to represent both the subject and the text prompt. For these reasons, we adopt the pretrained anchoring strategy, which provides a more stable reference signal and achieves a balance between subject fidelity and prompt adherence.

Table 5: Ablation of inference-time guidance strategies on SD1.5 under the model trained with our semantic anchoring objective.

Method	CLIP-I ↑	CLIP-T ↑	DINO ↑
Ours	0.7833	0.3064	0.6051
Pretrained Unconditional Guidance	0.7869	0.2791	0.6050
Finetuned Anchored Guidance	0.7250	0.2497	0.4935
Pretrained Anchored Guidance	0.7888	0.2506	0.6049

H.3 Additional Ablation on Inference-Time Guidance Strategies

To better understand how inference-time guidance interacts with the proposed training objective, we evaluate four variants of Classifier-Free Guidance (CFG) that differ only in how the unconditional branch is computed. Let z_t denote the noised latent at timestep t, and $\epsilon_\theta(z_t, c)$ be the noise prediction of the personalized model θ under condition c. We denote the target prompt as c^{sbj} (e.g., "a photo of a [V*] dog") and its anchor prompt as c^{anc} (e.g., "a photo of a dog"). The pretrained model is represented by θ'.

Using the same guidance scale $s = 7.5$, we evaluate four choices of the guidance:
(1) **Unconditional Guidance**: empty prompt using the personalized model

$$\epsilon_{\text{guide}} = \epsilon_\theta(z_t, c^{\text{sbj}}) + s\big(\epsilon_\theta(z_t, c^{\text{sbj}}) - \epsilon_\theta(z_t, c^{\text{uncond}})\big),$$

(2) **Pretrained Unconditional Guidance**: empty prompt using the pretrained model

$$\epsilon_{\text{guide}} = \epsilon_\theta(z_t, c^{\text{sbj}}) + s\big(\epsilon_\theta(z_t, c^{\text{sbj}}) - \epsilon_{\theta'}(z_t, c^{\text{uncond}})\big),$$

(3) **Finetuned Anchored Guidance**: prediction of anchor prompt using the personalized model

$$\epsilon_{\text{guide}} = \epsilon_\theta(z_t, c^{\text{sbj}}) + s\big(\epsilon_\theta(z_t, c^{\text{sbj}}) - \epsilon_\theta(z_t, c^{\text{anc}})\big),$$

(4) **Pretrained Anchored Guidance**: prediction of anchor prompt using the pretrained model

$$\epsilon_{\text{guide}} = \epsilon_\theta(z_t, c^{\text{sbj}}) + s\big(\epsilon_\theta(z_t, c^{\text{sbj}}) - \epsilon_{\theta'}(z_t, c^{\text{anc}})\big).$$

Table 5 demonstrates the quantitative results under different inference strategies. Our method achieves the most balanced performance across both image fidelity and text alignment metrics. In contrast, using pretrained information at inference time leads to a noticeable loss in text alignment, and leveraging finetuned class information results in inferior performance across all metrics.

Fig. 14 further illustrates how each inference strategy interacts with the learned personalized guidance, providing a perspective that is not captured in Table 5. When pretrained information is applied, the generated images contain visible artifacts. This suggests that the fixed pretrained features are not fully compatible with the personalized guidance and therefore introduce inaccurate guidance signals. In comparison, the proposed strategy mitigates these artifacts, suggesting that it provides the most reliable and coherent guidance during inference.

Taken together, inference-time modification does not universally provide benefits with the proposed method. As our training objective already enforces a semantically aligned relation between the subject and its class semantics, the standard unconditional CFG is sufficient and, in practice, produces the most reliable balance of subject fidelity and text alignment. This means that our method does not require any additional guidance strategy at inference time, and replacing the unconditional branch to other guidance can even deteriorate the generated results.

Table 6: User study results across four criteria (all values in %).

Method	Identity Preserv.	Prompt Align.	Visual Quality	Overall
Ours	**64.00**	**74.50**	**65.17**	**78.50**
DreamBooth	22.67	12.50	18.00	14.33
DreamBooth-LoRA	1.33	10.50	6.83	4.50
IP-Adapter	12.00	2.50	10.00	2.67

I User Study

To assess perceptual quality beyond quantitative metrics, we conduct a user study with 20 participants along four criteria. For each criterion, evaluators are presented with outputs in a shuffled order and select the best image according to the corresponding metric. The results are summarized in Table 6. Low-rank adaptation exhibits the lowest identity preservation, suggesting difficulty in incorporating new concepts, consistent with the qualitative examples in Fig. 3. Encoder-based methods achieve relatively low prompt alignment, in line with the quantitative results in Table 1, which also leads to poor overall scores. In contrast, our method attains the best performance across all criteria, indicating a more balanced trade-off between learning new concepts and preserving the pretrained semantics. This suggests that our semantic anchoring enables concept learning with minimal deviation from the pretrained model.

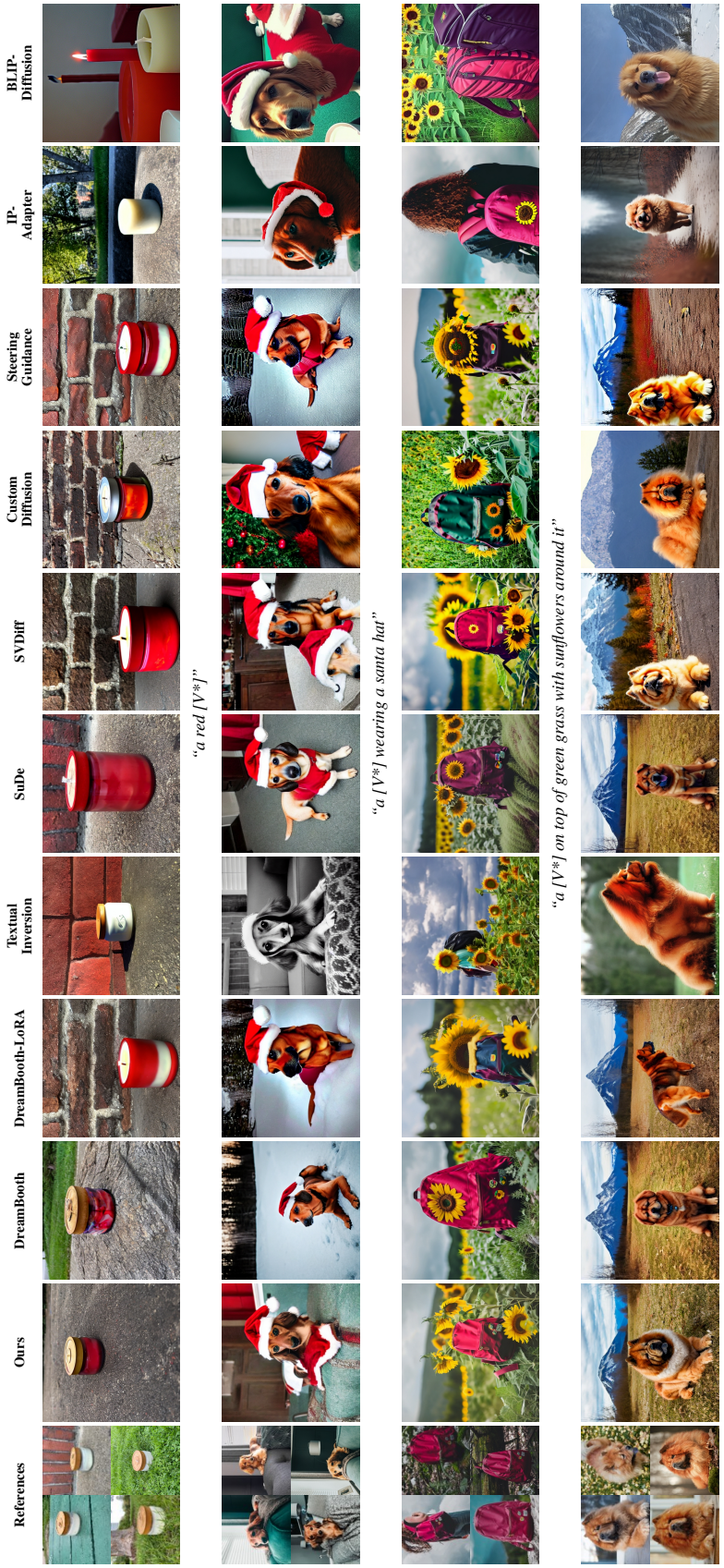

Figure 9: Additional visual comparison of few-shot personalization and encoder-based methods on the SD1.5 backbone.

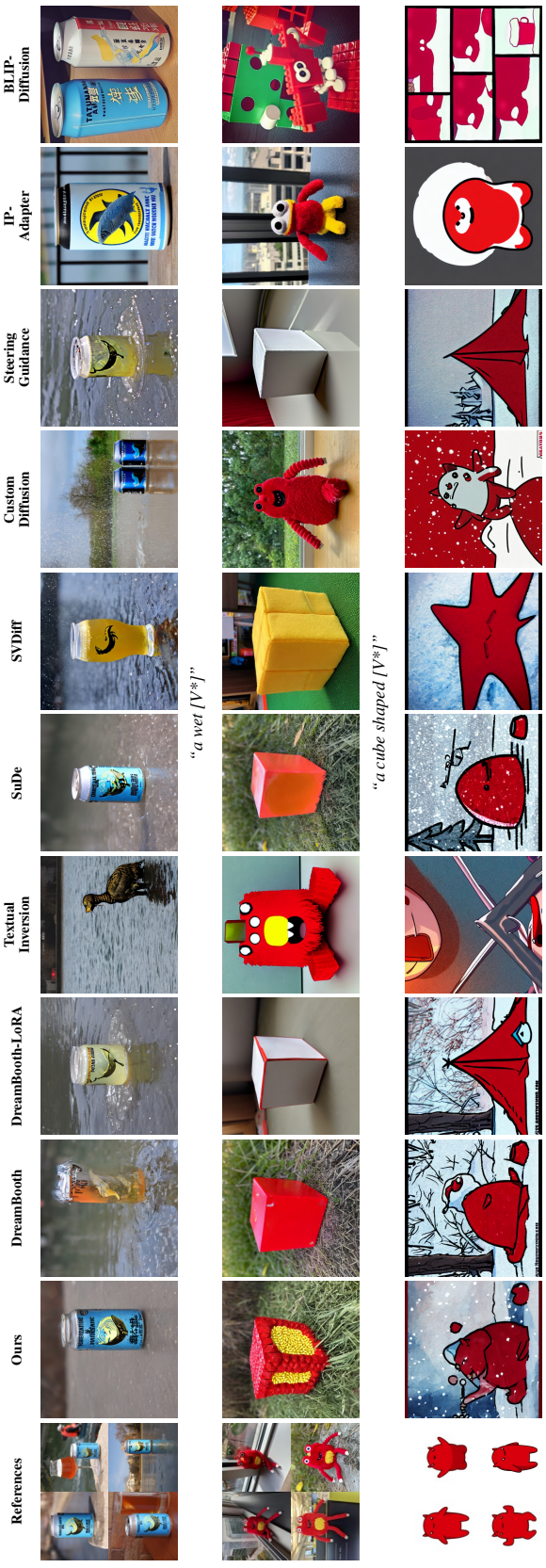

Figure 10: Additional visual comparison of few-shot personalization and encoder-based methods on the SD1.5 backbone, including representative failure cases.



Figure 11: Additional qualitative comparison on SD1.5, SDXL, and SD3 backbones among DreamBooth-LoRA, Beyond-Finetuning, and our method.



Figure 12: Comparison of qualitative results under different weighting parameters w.

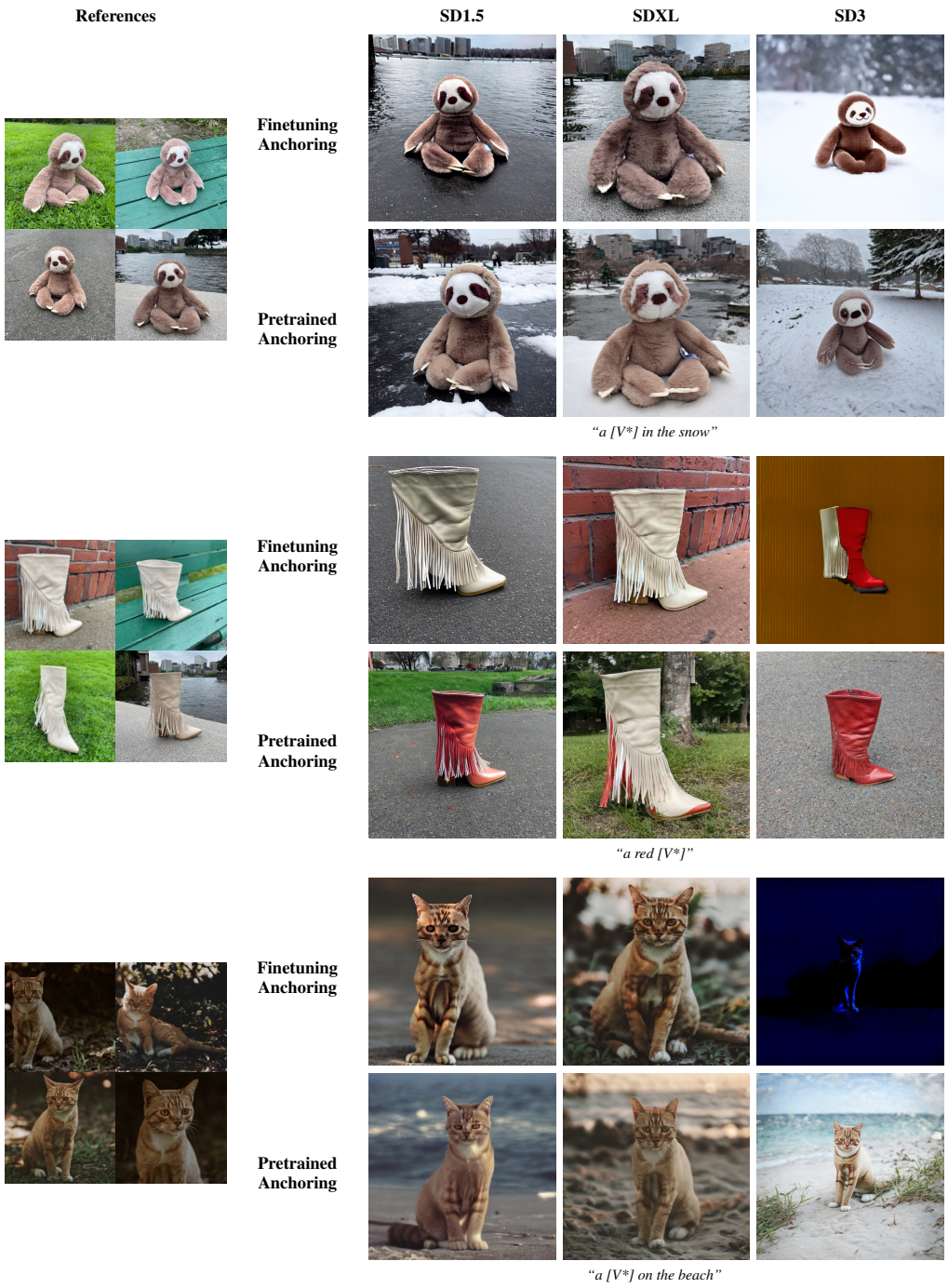

Figure 13: Comparison of qualitative results using two different anchoring strategies.

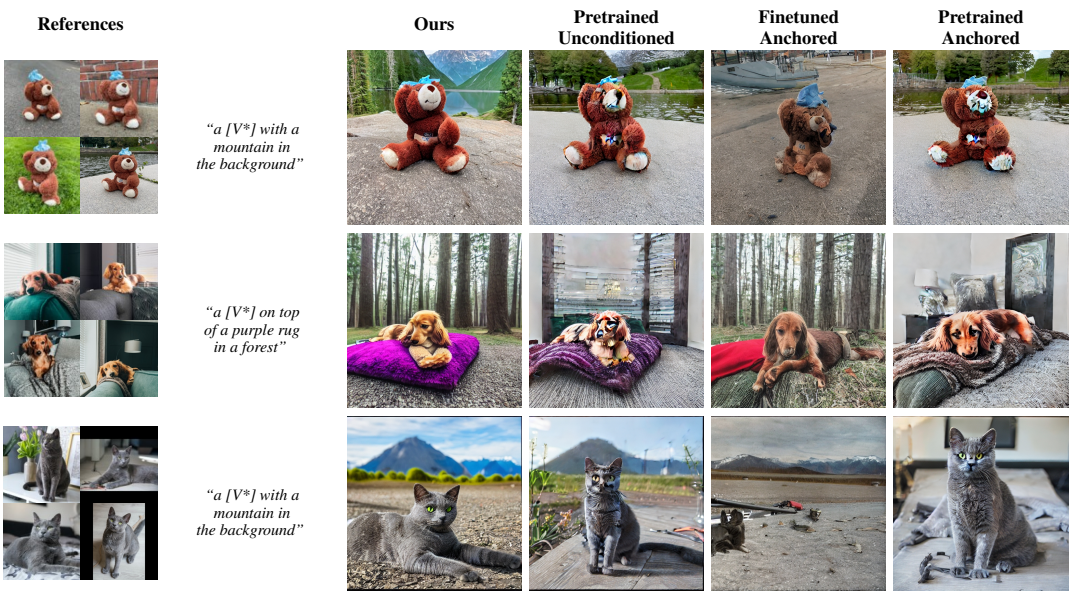

Figure 14: Visual comparison of different guidance strategies during inference.