
Preserving Compositionality for Robust Multi-Subject Personalization in Text-to-Image Generation

Sangho Lee^{*1}

Eugene Baek^{*1}

Suho Ryu¹

Dongsoo Shin¹

Joonseok Lee^{†1}

¹Seoul National University, Seoul, Korea

Abstract

Personalized text-to-image generation faces a fundamental trade-off between learning new concepts and preserving existing knowledge, and balancing these two goals gets particularly more challenging with multiple subjects to be personalized. Existing methods have relieved this by providing additional layout constraints, but it substantially undermines flexibility of the generation process. To address this issue, we propose a teacher–student architecture equipped with explicit regularizers to mitigate this trade-off. Specifically, we regulate the internal representations and cross-attention maps not to significantly deviate from the original foundation model, balanced with the reconstruction objective to internalize new concepts. Through extensive experiments, we demonstrate that our method enables reliable layout-free multi-subject generation, achieving the state-of-the-art performance on both single- and multi-concept personalization.

1 INTRODUCTION

Text-to-image diffusion models [Ramesh et al., 2021, Saharia et al., 2022, Rombach et al., 2022] have recently improved their ability to synthesize high-quality images in resolution, photorealism, style controllability, and conditional guidance; *e.g.*, pose, layout, color via external signals [Zhang et al., 2023, Mou et al., 2024]). Sampling speed and stability also have been improved through distillation-based accelerations [Salimans and Ho, 2022, Song and Dhariwal, 2024, Luo et al., 2023]. However, personalizing these models to faithfully render user-specific concepts, *e.g.*, a particular pet or a toy, remains challenging due to its fundamental identity-composition trade-off: pushing for high

subject fidelity often overfits the model to those new concepts, degrading the model’s compositional ability, whereas excessive conservation of the general knowledge underfits the model on the new concepts, failing to capture their visual details.

Existing personalization approaches have adopted two strategies. First, Textual Inversion and subsequent methods [Gal et al., 2023, Wei et al., 2023, Kumari et al., 2023] add a new identity token into the text manifold of an existing foundation model, keeping the original embedding space itself unchanged. This approach preserves the foundation model’s compositional machinery, but is prone to underfitting the identity of the new concepts, especially when the new token needs to compete with strongly pretrained categories. In contrast, DreamBooth and LoRA [Ruiz et al., 2023, Kumari et al., 2023] allow the model parameters to be updated, effectively remapping how text embeddings bind to visual features. This approach amplifies subject fidelity, but risks the latent space drift that perturbs the pretrained semantic and spatial priors, eventually degrading composition ability.

The resulting tension becomes particularly pronounced with multiple personalized concepts, leading to chronic issues such as attribute mixing, concept omission, and weak relational binding, illustrated in Fig. 1. A prevalent workaround is to add explicit spatial constraints, *e.g.*, layouts or poses [Li et al., 2023b, Gu et al., 2023, Kwon et al., 2024, Xiao et al., 2024, Kwon and Ye, 2025, Jang et al., 2024], which are well suited for applications requiring precise spatial control. When such explicit control is unnecessary, however, these guidance signals restrict the open-ended adaptability inherent in layout-free text-to-image generation by rigidly pre-determining object scale, placement, and interactions, as illustrated in Fig. 2.

Putting these together, an ideal personalized text-to-image model should be able to learn new concepts while retaining its layout-free compositional ability. Importantly, our objective is not to replace layout-guided generation, but to preserve robust compositional behavior in scenarios where

^{*}Equal Contribution

[†]Corresponding author



Figure 1: **Common failure modes in multi-concept generation.** (top) *Attribute mixing* (e.g., v1 dog entangled with v2 robot); (middle) *Concept omission* (e.g., v2 can missing); (bottom) *Relation/binding error* (e.g., implausible pairing of v1 backpack with v2 dog).

explicit spatial guidance is unnecessary, while remaining complementary to layout-guided approaches when such guidance is available. In other words, the model parameters need to be updated to sufficiently grasp the identity of new concepts, while minimizing collateral distortion to the underlying latent space. Existing methods only partially address this trade-off and often rely on prompt-specific or layout-dependent strategies. In this paper, we aim to mitigate the latent space drift and improve robustness of the generation process by explicitly anchoring the target model (student) to a frozen foundation model (teacher)’s class-level structure. Unlike conventional teacher–student schemes [Hinton et al., 2015] adopted for efficiency, however, our approach leverages distillation as an explicit *regularizer* to balance between the two conflicting goals of learning new concepts and preserving compositional structure. Specifically, we introduce two regularizers to preserve the teacher’s prior: (i) feature distillation of intermediate activations, and (ii) cross-attention alignment that matches the student’s attention on the new-concept-specific token (e.g., v1 dog) to the teacher’s attention on its general class token (e.g., dog) at the same denoising step. We empirically verify that this explicit guidance ensures the student to absorb the new identity without significantly distorting the latent structure required for composition.

Our main contributions are summarized as follows:

- We identify that the compositional failures in layout-free, multi-subject personalization are primarily rooted in *identity–composition trade-off*.
- We propose explicit *regularization* via intermediate



Figure 2: **Layout-guided vs. layout-free generation.** (left) Layout guidance fixes boxes and scales, yielding rigid, template-like results. (right) Our layout-free approach produces diverse, natural multi-subject compositions without manual spatial constraints.

feature distillation and cross-attention alignment to mitigate the latent-space drift and resolve the identity–composition trade-off.

- We demonstrate consistent improvements in both fidelity and composition across single- and multi-concept benchmarks, achieving *robust layout-free generation* where prior methods often struggle.

2 RELATED WORK

Single-Concept Personalization. Textual Inversion [Gal et al., 2023] and its successors add new tokens for novel concepts into the fixed text-embedding manifold; e.g., stronger token priors [Voynov et al., 2023, Tewel et al., 2023, Wei et al., 2023, Zhang et al., 2024b] and pre-trained subject encoders [Li et al., 2023a, Jia et al., 2023, Chen et al., 2023, Alaluf et al., 2023, Hao et al., 2023]. These methods generally preserve compositional priors but often underfit new identities. In contrast, parameter fine-tuning approaches remap text–visual alignment by updating parts of the denoiser and attention; e.g., DreamBooth [Ruiz et al., 2023], LoRA-based methods [Hu et al., 2022, Han et al., 2023, Kumari et al., 2023], and implicit adaptation Ruiz et al. [2024]. While effective for single subjects, many struggle with layout-free multi-subject composition, often causing identity blending.

Multi-Concept Personalization. Recent advances in personalization have extended single-subject synthesis to the more challenging task of composing multiple identities, typically attempting to mitigate concept interference through various architectural or optimization strategies. Some approaches focus on structural separation, utilizing decentralized adapter fusion or isolation [Gu et al., 2023], selective fine-tuning or decoupling of cross-attention [Kumari et al., 2023, Kwon et al., 2024], or attention masking and gating [Jang et al., 2024]. Others emphasize feature-level alignment via appearance matching [Nam et al., 2024], learned

adapters or localized subject conditioning [Kwon and Ye, 2025, Xiao et al., 2024], or decoupling of textual embeddings [Wei et al., 2024]. Tuning-free methods, such as FastComposer [Xiao et al., 2024], FreeCustom [Ding et al., 2024], and MS-Diffusion [Wang et al., 2025], utilize pre-trained subject encoders to inject multiple identities without additional optimization. Another line of research focuses on composing multiple personalized concepts from independently trained adapters or LoRA modules at inference time. Representative examples include LoRAshop [Dalva et al., 2026], FreeFuse [Liu et al., 2025], and CLoRA [Meral et al., 2025], which combine pre-trained concept-specific adapters while attempting to mitigate interference among concepts. These approaches assume that a separate personalization model has already been trained for each concept and primarily address adapter composition at inference. In contrast, our work focuses on the earlier personalization stage itself, aiming to learn new identities from scratch while preserving the compositional structure of the foundation model during fine-tuning. Despite this progress, existing methods rely on explicit spatial guidance such as layouts, masks, or poses to prevent concept blending [Li et al., 2023b, Gu et al., 2023, Kwon et al., 2024, Xiao et al., 2024, Kwon and Ye, 2025, Jang et al., 2024]. While effective when precise spatial control is desired, such guidance often sacrifices flexibility of generation. HATIE [Ryu et al., 2025] proposed fully-automated evaluation of image personalization.

Compositional Generation. To ensure precise object presence and attribute binding, various strategies have been integrated into text-to-image pipelines and large-scale T2I models [Ramesh et al., 2021, 2022, Saharia et al., 2022, Rombach et al., 2022, Podell et al., 2024]. These range from cross-modal alignment [Zhang et al., 2021, Brooks et al., 2023] and attention-based controls [Rassin et al., 2023, Zhuang et al., 2024] to extended textual conditioning [Vovnov et al., 2023]. While broader conditioning frameworks like ControlNet [Zhang et al., 2023], T2I-Adapter [Mou et al., 2024], and GLIGEN [Li et al., 2023b] achieve robust binding by injecting external structures (*e.g.*, bounding boxes, edges, or poses), rely on external structural inputs, making them less suitable when diverse layout-free generation is desired.

Distillation as Regularization. Knowledge distillation [Hinton et al., 2015, Romero et al., 2015] in diffusion has primarily focused on inference acceleration and training stability [Salimans and Ho, 2022, Song and Dhariwal, 2024, Luo et al., 2023]. However, distillation can also serve as a powerful regularization tool to mitigate catastrophic forgetting and interference—long-standing issues in model adaptation [Li and Hoiem, 2016, Ruiz et al., 2023, Kumari et al., 2023]. While existing personalization methods attempt to curb latent-space drift and cross-concept interference through orthogonality constraints through LoRA [Hu et al., 2022] or parameter isolation [Gu et al., 2023, Kwon

et al., 2024], they often fall short in complex multi-concept scenarios. Arar et al. [2024] enhances prompt alignment at personalization, but does not explicitly address the preservation of layout-free compositional structure. In contrast, our method employs a frozen teacher as an explicit anchor to regularize the personalization process. By aligning mid-level features and cross-attention maps, our framework aims to better preserve the foundation model’s semantic and spatial priors during personalization, reducing latent-space drift without requiring external spatial templates.

3 BACKGROUND

Preliminaries on Diffusion Models. Text-to-image diffusion models (*e.g.*, Rombach et al. [2022]) generate images by an iterative denoising process, predicting the noise added to the latent representation at each diffusion timestep t . The model $F(\cdot)$ is trained to minimize the reconstruction loss $\mathcal{L}_{\text{recon}}(F; x, y, t) = \mathbb{E}_{\epsilon, t} [\|\epsilon - F(x, y, t)\|_2^2]$, where x is the latent at timestep t , y is the text conditioning, and ϵ is the ground-truth noise. This process relies on the cross-attention mechanism, where attention maps for textual tokens dictate the spatial manifestation of concepts. Our method leverages these internal structures—intermediate features and cross-attention maps—to maintain model integrity during personalization.

Problem Formulation. Let F_θ denote a frozen foundation text-to-image diffusion (teacher) model with fixed parameters θ , and let F_ϕ be a personalized student model with learnable parameters ϕ . We denote a set of user-defined concepts by $\mathcal{C} = \{c_1, \dots, c_M\}$. Each concept $c \in \mathcal{C}$ is represented by a few-shot reference image set $\mathcal{X}_c = \{x_i\}_{i=1}^{N_c}$ and is mapped to a unique identifier token τ_c within the text-embedding space. A textual prompt y may incorporate one or more identifier tokens (*e.g.*, “a photo of τ_{c_a} and τ_{c_b} at the beach”) without any auxiliary layout guidance.

Our goal is twofold: (i) to render each concept c with high identity fidelity, and (ii) to seamlessly compose multiple concepts, maintaining accurate spatial and semantic relationships among subjects. We expect this approach would address the compositional degradations prevalent in multi-subject personalization, such as the attribute leakage and concept omission cases evidenced in Figs. 1 and 3.

Trade-off in Personalization Mechanisms. As introduced in Sec. 2, personalization methods can be categorized by their approach to the latent space: fixed *vs.* adaptive. The former such as Textual Inversion and its extensions [Gal et al., 2023, Wei et al., 2023, Zhang et al., 2024b] keeps the model weights fixed ($\phi = \theta$) and learns only the embeddings $e(\tau_c)$ of the new concepts through reconstruction, *e.g.*,

$$\min_{e(\tau_c)} \mathbb{E}_{x \in \mathcal{X}_c, t, y} [\mathcal{L}_{\text{recon}}(F_\theta, e(\tau_c); x, y, t)]. \quad (1)$$

Since θ remains unchanged, the foundation model’s compo-



Figure 3: **Qualitative comparison of multi-subject personalization.** For a multi-concept scenario “v1 and v2 on the cobble stone”, our method preserves both identities without identity leakage. See Fig. III–VII in Appendix F for more examples.

sitional machinery is almost completely preserved. However, the identity signal for the new concept c often *underfits*, failing to effectively compete against strong native tokens in the cross-attention layers. This eventually weakens the identity of the new concepts in the downstream composition.

On the other hand, fine-tuning approaches including Dream-Booth [Ruiz et al., 2023] and LoRA-style methods [Hu et al., 2022, Han et al., 2023, Kumari et al., 2023] update the model weights ($\phi = \theta + \Delta\theta$). Typically anchoring the new concept to a class-level prompt, they minimize

$$\min_{\Delta\theta} \mathbb{E}_{x, t, y} [\mathcal{L}_{\text{recon}}(F_{\theta+\Delta\theta}; x, y, t)], \quad (2)$$

which injects a strong identity signal, in turn potentially distorts the pre-trained text-visual binding excessively. This latent space drift leads to interference or even catastrophic forgetting of semantic priors, thereby degrading overall composition performance [Ruiz et al., 2023, Kumari et al., 2023, Li and Hoiem, 2016]. In essence, these two paradigms represent the opposite ends of the *identity-composition trade-off*, where gaining fidelity in one often sacrifices the structural integrity of the other, and vice versa.

Challenges in Multi-Subject Scenarios. The tension becomes particularly exacerbated when adapting multiple concepts simultaneously; for instance, in a layout-free prompt like “a photo of τ_{c_a} and τ_{c_b} at the beach”, which mention more than one personalized identifiers.

The challenge basically lies in the increased competition between the new concepts and pre-trained knowledge. When the θ is fixed, it is hard for the new tokens to locate themselves in the rigid latent space, which has been optimized without considering them at all. Obviously, this difficulty gets significantly harder with more such new tokens. As a result, their relatively weak identity cues struggle to compete against pre-trained categories, especially the related ones to the new concepts.

On the other hand, this issue is not simply resolved by allowing the parameters θ to be updated. The fine-tuning approaches face another difficulty, when it squeezes multiple new concepts into the previously optimized latent space. This significantly intensifies latent-space drift, often causing multiple concepts to conflict and compete. Such cross-concept interference often yields attribute mixing, concept omission, or incorrect relational bindings.

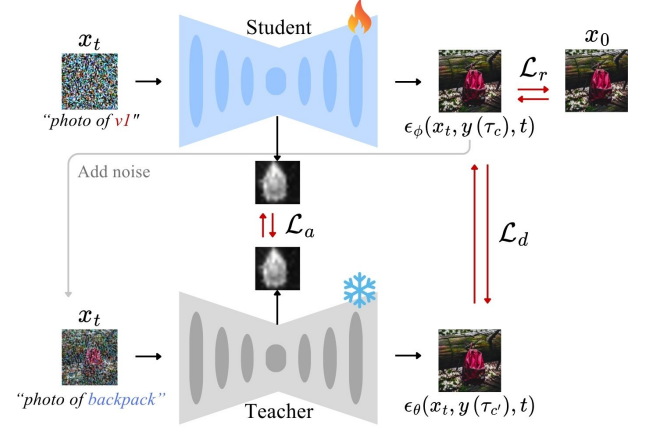


Figure 4: **Overview of our proposed method.** The student F_ϕ learns a new concept c by reconstructing its few-shot reference images with an identifier token τ_c . To preserve general compositional knowledge, a pre-trained teacher F_θ processes the same inputs with the identifiers replaced to general classes (e.g., $\tau_c \rightarrow$ “backpack”). We explicitly align intermediate activations (*feature distillation*) and match the student’s attention on τ_c to the teacher’s attention on the corresponding class tokens (*cross-attention alignment*).

Overall, a naive multi-subject composition easily pushes the identity-composition trade-off to its extreme, making it hard to balance avoiding both weak identities and distorted priors. We therefore propose a systematic approach to learn new concepts while being explicitly regularized to remain aligned with the original compositional structure in Sec. 4.

4 PROPOSED APPROACH

To seamlessly integrate multiple novel concepts into a pre-trained embedding space without requiring external spatial prior such as layout, mask, or pose, we propose a training framework in Fig. 4 with explicit regularizers to effectively avoid both underfitting and overfitting of the new concepts.

Teacher-Student Architecture. We employ a teacher-student scheme where a frozen foundation model F_θ serves as the source of undistorted semantic and spatial priors (teacher). The student F_ϕ is a trainable copy of F_θ that learns novel concepts $c \in \mathcal{C}$ using identifier tokens τ_c and

few-shot reference images $\mathcal{X}_c$. During training, the student is conditioned on the concept identifier tokens τ_c , while the teacher is conditioned on the corresponding general class tokens for c (e.g., $\tau_c \rightarrow$ “backpack”), allowing the teacher to provide a “clean” reference of how a generic object of the class should be composed.

To achieve this, we formulate the personalization objective by combining a reconstruction term for identity learning with an explicit regularizer that prevents the student from deviating from the teacher’s compositional priors:

$$\min_{\phi} \mathbb{E}_{c \sim \mathcal{C}, x \in \mathcal{X}_c, y, t} \left[\underbrace{\mathcal{L}_{\text{recon}}(F_{\phi}; x, y, t)}_{\text{to learn identity}} + \lambda \cdot \underbrace{\mathcal{L}_{\text{reg}}(F_{\phi}, F_{\theta}; x, y, t)}_{\text{to preserve composition}} \right],$$

where $\lambda > 0$ balances identity fidelity and compositional integrity, and both models are evaluated at the same diffusion timestep t . Crucially, the teacher parameters θ remain frozen throughout training to ensure that the identifier’s specific semantics do not leak into the teacher, preserving its role as a stable anchor for semantic and spatial priors. Below, we further instantiate $\mathcal{L}_{\text{reg}}$ via intermediate-layer feature distillation and cross-attention alignment.

Learning New Concepts. To ensure the model to capture the specific visual characteristics of each concept, The student model F_{ϕ} is trained on few-shot reference images $\mathcal{X}_c$ using prompts that include the concept-specific identifier τ_c through the standard diffusion reconstruction objective:

$$\mathcal{L}_r = \mathbb{E}_{x \in \mathcal{X}_c, y, t} \left[\mathcal{L}_{\text{recon}}(F_{\phi}; x, y(\tau_c), t) \right]. \quad (3)$$

It tightly couples the identifier token τ_c to the visual identity in $\mathcal{X}_c$, yielding a strong concept signal that can effectively compete at multi-subject composition.

Preserving Priors by Regularization. While the reconstruction loss ensures fidelity, it poses a risk of distorting the pre-trained latent space, which can erode the semantic and spatial priors essential for layout-free composition. To mitigate this, we introduce an explicit regularizer consisting of two complementary terms: one for preserving the semantic features and another for maintaining spatial priors.

First, we align the intermediate activations between the student and teacher to preserve semantic priors:

$$\mathcal{L}_d = \mathbb{E}_{x, y, t} \left[\sum_{\ell \in \mathcal{H}} w_{\ell} \left\| \mathbf{h}_{\phi}^{(\ell)}(x_t, t, y(\tau_c)) - \text{sg}[\mathbf{h}_{\theta}^{(\ell)}(x_t, t, y(\tau_{c'}))] \right\|_2^2 \right],$$

where c' is the generic class of the new concept c , $\mathcal{H}$ is a specific set of feature blocks, $w_{\ell} \geq 0$ with $\sum_{\ell} w_{\ell} = 1$, x_t is the noisy input at diffusion timestep t , and $\text{sg}[\cdot]$ denotes stop-gradient. Intuitively, the student’s internal logic for generating c is nudged to behave like the teacher’s logic for the general class, thereby preventing latent-space drift.

Moreover, to preserve spatial priors and prevent attribute entanglement, we regularize the cross-attention maps for the new concept to match those for its general class:

$$\mathcal{L}_a = \mathbb{E}_{x, y, t} \left[\sum_{l \in A} \lambda_l \left\| \mathbf{A}_{\phi}^{(l)}(x_t, t, \tau_c) - \text{sg}[\mathbf{A}_{\theta}^{(l)}(x_t, t, \tau_{c'})] \right\|_1 \right],$$

where $\mathbf{A}^{(l)}$ is the token-specific cross-attention map at layer l (again, identifier token slice for the student; class slice for the teacher), and $\lambda_l \geq 0$ controls relative weights among layers. This loss discourages alignment remapping that might otherwise misplace concepts or entangle attributes (e.g., Fig. 1) in layout-free multi-subject scenes.

Multi-Subject Extension. For prompts involving multiple identifiers (e.g., “a photo of τ_{c_a} and τ_{c_b} at the beach”), we aggregate the regularizers across all identifier-class pairs by summing the respective losses for each $c_k \in \mathcal{C}$ mentioned in the prompt:

$$\mathcal{L}_d^m = \mathbb{E} \left[\sum_k \sum_{\ell \in \mathcal{H}} w_{\ell} \left\| \mathbf{h}_{\phi}^{(\ell)}(x_t, t, y(\tau_{c_k})) - \text{sg}[\mathbf{h}_{\theta}^{(\ell)}(x_t, t, y(\tau_{c'_k}))] \right\|_2^2 \right],$$

$$\mathcal{L}_a^m = \mathbb{E} \left[\sum_k \sum_{l \in A} \lambda_l \left\| \mathbf{A}_{\phi}^{(l)}(x_t, t, \tau_{c_k}) - \text{sg}[\mathbf{A}_{\theta}^{(l)}(x_t, t, \tau_{c'_k})] \right\|_1 \right].$$

Overall Objective. We minimize a weighted sum of the reconstruction loss and the multi-subject regularizers:

$$\mathcal{L} = \mathcal{L}_r + \alpha \cdot \mathcal{L}_d^m + \beta \cdot \mathcal{L}_a^m, \quad (4)$$

where α and β control the strength of semantic and spatial regularizations, respectively. We apply $\mathcal{L}_d$ to the subset of the model layers, and $\mathcal{L}_a$ to the corresponding cross-attention layers.

Training Protocol. To ensure a stable and consistent transfer of knowledge, at each iteration, we obtain x_t by noising each reference image and feed the *same* x_t to both the student and the teacher, with the new concept identifier to the student and a general class token to the teacher. The teacher remains frozen to serve as a stable anchor, preventing the specific semantics of the identifier from leaking into the teacher’s priors. For further stability, we optionally apply EMA smoothing to the teacher’s feature and attention targets. More details are provided in Algorithm 1.

5 EXPERIMENTS

We conduct experiments to compare our method with state-of-the-art baselines, focusing on its capacity for layout-free multi-concept composition and identity fidelity.

5.1 EXPERIMENTAL SETTINGS

Datasets. We use a dataset collected from DreamBooth [Ruiz et al., 2023] and Textual Inversion [Gal et al., 2023], consisting of 43 distinct concepts with an average of 5.4 reference images per subject. For multi-concept evaluations, we randomly select 15 combinations of two-class pairs. The complete collection of images and prompts across all evaluation scenarios is provided in Appendix C.

Baselines. We compare with multi-concept personalization baselines, including Custom Diffusion [Kumari et al.,

Algorithm 1: Our Proposed Fine-Tuning Framework

Require: Frozen teacher F_θ ; VAE encoder E ; Student F_ϕ (init from F_θ);
Reference set $\mathcal{D}_{\text{ref}} = \{(c, X_c)\}$; Layer sets $\mathcal{H}, A$; Weights
 $\alpha, \beta, \{w_\ell\}, \{\lambda_l\}$; Learning rate $\eta > 0$.

Output: Updated student parameters ϕ .

while not converged **do**

```

// 1. Sample and encode
Sample concept  $c \in \mathcal{C}$  and  $x \in \mathcal{X}_c$ ;  $z_0 \leftarrow E(x)$ 
Sample timestep  $t$  and noise  $\epsilon \sim \mathcal{N}(0, I)$ ;  $z_t \leftarrow \alpha_t z_0 + \sigma_t \epsilon$ 
Build identifier prompt  $y_{\tau_c}$  and class prompt  $y_{\text{class}(c)}$ 
Student:
Run  $F_\phi(z_t, t, y_{\tau_c})$  to obtain:
noise pred  $\hat{\epsilon}_\phi$ , mid-layer features  $\{\mathbf{h}_\phi^{(\ell)}\}_{\ell \in \mathcal{H}}$ , and token-slice
cross-attention  $\{\mathbf{A}_\phi^{(l)}[\tau_c]\}_{l \in A}$ 
Teacher (stop-grad):
Run  $F_\theta(z_t, t, y_{\text{class}(c)})$  to obtain:
mid-layer features  $\{\mathbf{h}_\theta^{(\ell)}\}_{\ell \in \mathcal{H}}$  and token-slice cross-attention
 $\{\mathbf{A}_\theta^{(l)}[\text{class}(c)]\}_{l \in A}$ 
// 2. Loss Calculation
 $\mathcal{L}_r \leftarrow \|\hat{\epsilon}_\phi - \epsilon\|_2^2$ 
 $\mathcal{L}_d \leftarrow \sum_{\ell \in \mathcal{H}} w_\ell \|\mathbf{h}_\phi^{(\ell)} - \text{sg}[\mathbf{h}_\theta^{(\ell)}]\|_2^2$ 
 $\mathcal{L}_a \leftarrow \sum_{l \in A} \lambda_l \|\mathbf{A}_\phi^{(l)}[\tau_c] - \text{sg}[\mathbf{A}_\theta^{(l)}[\text{class}(c)]]\|_1$ 
// 3. Update
 $\mathcal{L} \leftarrow \mathcal{L}_r + \alpha \mathcal{L}_d + \beta \mathcal{L}_a$ 
 $\phi \leftarrow \phi - \eta \nabla_\phi \mathcal{L}$ 

```

end

2023], MuDI [Jang et al., 2024], Tweedie [Kwon and Ye, 2025], FastComposer [Xiao et al., 2024], and MS-Diffusion [Wang et al., 2025]. We also include established single-concept methods, *e.g.*, DreamBooth, Textual Inversion, DE-TEX [Wei et al., 2024], and DreamMatcher [Nam et al., 2024].

Evaluation Metrics. We evaluate two core capabilities for high-quality personalization: object and global fidelity.

First, the *object fidelity* measures the degree of visual preservation of the target concept(s). We report the ℓ_2 -normalized cosine similarity between the generated output $\hat{x}$ and the reference image x : $I(\hat{x}, x) = \cos(f(\hat{x}), f(x))$, where we use $f \in \{\text{CLIP}, \text{DINOv2}\}$. For multi-concept scenarios, we independently assess each concept’s appearance using object-level crops, following Jang et al. [2024].

On the other hand, global fidelity measures the preservation of scene structure and relational binding. More specifically, we report the *prompt fidelity* (T_{CLIP}), measured by the CLIP image-to-text similarity between the generated image $\hat{x}$ and a class-level prompt y : $T_{\text{CLIP}}(\hat{x}, y) = \cos(f_{\text{CLIP}}^{\text{text}}(y), f_{\text{CLIP}}^{\text{img}}(\hat{x}))$, and the *composition fidelity*, measured by I_{CLIP} and I_{DINO} between $\hat{x}$ and the golden reference x^* to evaluate adherence to foundational layout and structural priors. Here, we take the output from the frozen foundation model with the highest T_{CLIP} for a class-level prompt as the golden reference x^* .

Implementation Details. We use pre-trained Stable Diffusion v2.1 as the default teacher F_θ . To verify the generality of our approach, we also conduct experiments using SDXL [Podell et al., 2024]. The student model F_ϕ is initialized from F_θ and fine-tuned for 800 epochs with a batch

Table 1: **Single-concept evaluation.** Our method shows strong object and global fidelity. Note that I_{CLIP} for composition fidelity is omitted as I_{CLIP} in object fidelity already measures the global fidelity.

Method	Object Fidelity		Global Fidelity	
	$I_{\text{CLIP}} \uparrow$	$I_{\text{DINO}} \uparrow$	Prompt ($T_{\text{CLIP}} \uparrow$)	Composition ($I_{\text{DINO}} \uparrow$)
DreamBooth	0.7412	0.6128	0.2462	0.5781
DeTex	0.7265	0.5941	0.2587	0.5593
Dreammatcher	0.7689	0.6273	0.2672	0.5478
Ours	0.7604	0.6347	0.3447	0.6150

size of 4 using the AdamW optimizer (learning rate 10^{-5}). The resolution is 512×512 for SDv2.1 and 1024×1024 for SDXL. We regularize the intermediate features and cross-attention layers within the U-Net’s middle block. Loss weights are set to $\alpha = 0.7$ and $\beta = 1.0$. New identifier tokens are first initialized from the corresponding general class token embeddings and optimized following Textual Inversion. Afterward, the learned identifier embeddings are frozen, and only the student U-Net denoiser parameters are updated. The VAE and text encoder remain frozen throughout training. We apply both feature distillation and cross-attention alignment to the U-Net middle block, which provides a favorable efficiency-performance trade-off due to its lowest spatial resolution. More details on hyperparameters and baseline configurations are provided in Appendix A.

5.2 QUANTITATIVE RESULTS

We test our hypothesis that the identity-composition trade-off stems from the tension between identity underfitting and latent-space drift, a phenomenon that is significantly amplified in layout-free, multi-subject compositions.

Single-concept Generation. Tab. 1 compares the personalization performance when integrating only a single new concept. It is noteworthy that we assess the full text-image alignment and scene composition beyond “v1-only” evaluation, by prompting the model to co-generate the new concept with another general instance (*e.g.*, “v1 and dog”) through prompt fidelity (T_{CLIP}) and composition fidelity (I_{DINO}).

As shown in Tab. 1, fine-tuning baselines (*e.g.*, DreamBooth, DeTex) deliver competitive concept fidelity, while they often suffer from lower prompt (T_{CLIP}) and composition (I_{DINO}) fidelity. This degradation is prominently evidenced when a personalized concept co-exists with another native object (*e.g.*, “v1 and dog”); as reported in the ‘Composition’ column in Tab. 1, baselines incur substantial performance drops in composition fidelity (I_{DINO}) in such scenarios. We attribute this to the distorted latent space, which inadvertently arises when they struggle to reconcile the newly learned concept with pre-existing native categories.

In contrast, our method achieves the state-of-the-art per-

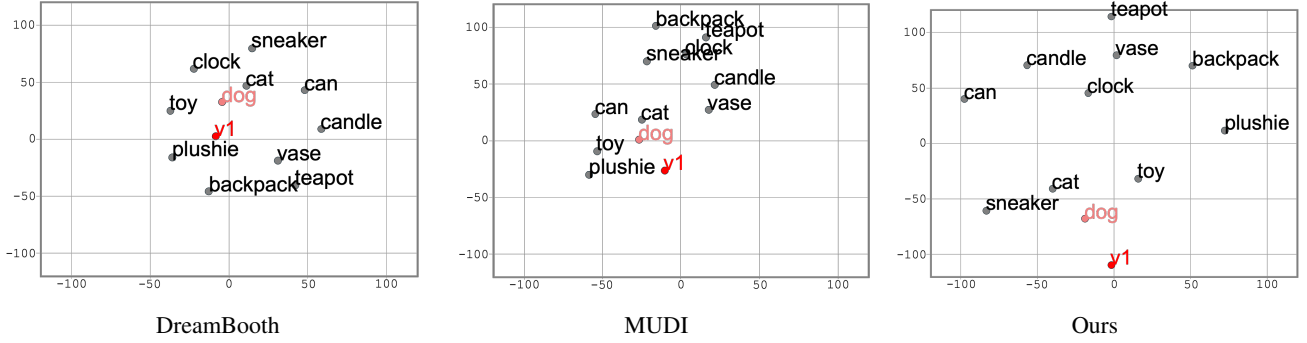


Figure 5: ***t*-SNE visualization of token embeddings.** Baseline methods produce entangled embeddings with ‘v1’ near unrelated concepts, while our method places ‘v1’ in a well-isolated region, indicating better concept disentanglement.

formance in both object and composition fidelity metrics, significantly surpassing all baselines. These results indicate that our approach effectively mitigates semantic distortion even in the single-concept regime.

Multi-concept Generation. The interference gets severer with a larger number of new concepts. As reported in Tab. 2, the performance of baselines degrades sharply in multi-subject scenarios. Custom Diffusion and MuDI struggle to balance multiple identities within a shared latent space, leading to severe cross-concept interference. This is evidenced by the substantial decrease in global fidelity, where attribute mixing and concept omission become prevalent.

Our approach, in contrast, consistently maintains superior performance across all metrics over all baselines, with a larger gap compared to the single object personalization cases. These results suggest that anchoring the student to the teacher’s semantic and spatial priors helps preserve the foundation model’s compositional capability for layout-free generation. By preventing the latent space from collapsing or drifting excessively, our method successfully renders multiple subjects with high identity fidelity ($I_{\text{CLIP}} = 0.56$), while ensuring the generated scene remains aligned with the intended global structure.

5.3 QUALITATIVE RESULTS

Fig. 6 presents several qualitative results to visually assess the performance and flexibility of our method. In Fig. 6(a) demonstrates that our method reliably generates the learned concept with another general object (e.g., “v1 and a dog”), without suffering from concept blending or omission, common with other models. Fig. 6(b) highlights that a single learned concept (‘v1’) can be flexibly adapted to diverse attributes and styles (e.g., “firefighter outfit” or “Van Gogh style”), behaving as naturally as a native token within the foundation model. Most importantly, Fig. 6(c) underscores our core contribution: the ability to coherently generate multiple personalized concepts simultaneously across various scene context (e.g., “on beach” or “in snow”) without any

Table 2: **Multi-concept evaluation.** Our method excels at both object-level and scene-level fidelity.

Method	Object Fidelity (MuDI)		Global Fidelity		
	CLIP-I $\uparrow$	DINOv2 $\uparrow$	Prompt $T_{\text{CLIP}} \uparrow$	Composition $I_{\text{CLIP}} \uparrow$	$I_{\text{DINO}} \uparrow$
FastComposer	0.22	0.11	0.22	0.59	0.15
MS-Diffusion	0.24	0.17	0.25	0.62	0.19
Custom Diffusion	0.22	0.18	0.21	0.48	0.17
MuDI	0.33	0.27	0.28	0.61	0.37
Tweedie	0.31	0.27	0.25	0.64	0.31
Ours	0.56	0.52	0.31	0.67	0.55

layout constraints.

We further demonstrate that the learned identities remain disentangled for composing even when composing three personalized concepts simultaneously. As shown in Fig. 7, our method successfully preserves the appearance of all subjects while maintaining plausible spatial arrangements and prompt-consistent interactions, highlighting the scalability of the proposed regularization beyond pairwise composition. This generalizability also extends to advanced foundation models, as evidenced by the strong out-of-the-box compatibility with SDXL in Fig. 8. Additional examples are provided in Appendix F.

5.4 ABLATION STUDY

We verify the contributions of our proposed regularizers in mitigating the two primary failure modes: identity underfitting and latent-space drift. Beyond the regularization strengths, we additionally analyze several design choices of our framework, including the selection of supervised U-Net layers, the sensitivity to layer weighting schemes, and computational efficiency. Detailed results are provided in Appendix G.

Semantic-prior Regularization ($\mathcal{L}_d$). First, we vary the weights α on the semantic-prior regularizer $\mathcal{L}_d$. Tab. 3 indicates that reconstruction-only training without feature distil-



Figure 6: **Qualitative illustration.** (a) Successful co-generation of learned concepts with native objects. (b) High flexibility in adapting personalized concepts to diverse styles and attributes. (c) Robust multi-concept composition without explicit spatial or layout guidance.

Table 3: **Ablation on distillation weight α .** The results show an optimal range of 0.7 - 1.1.

Metric	α	0.0	0.3	0.7	0.9	1.1	1.5	1.9
Prompt Fidelity	T_{CLIP}	0.2351	0.2543	0.2972	0.2943	0.2925	0.2729	0.2516
Composition	I_{CLIP}	0.5912	0.6178	0.6574	0.6527	0.6541	0.6410	0.6220
Fidelity	I_{DINO}	0.4123	0.4418	0.4815	0.4723	0.4671	0.4487	0.4293

lation ($\alpha=0$) significantly drops prompt/composition fidelity and increases omissions. Moderately higher α restores composition fidelity by pulling mid-level features toward the teacher’s manifold, confirming the drift-suppression role of $\mathcal{L}_{\text{distill}}$. However, an overly large α slightly dulls identity-specific details. Fig. 9 (and Fig. VIII in Appendix G) qualitatively illustrates progressive mitigation of concept omission as α increases.

Spatial-prior Regularization ($\mathcal{L}_a$). Tab. 4 implies that raising β consistently improves composition fidelity and object presence, indicating a reduction in spatial misbinding and subject displacement. This supports that our cross-attention alignment helps to preserve spatial priors critical for layout-free multi-subject composition.

5.5 ANALYSIS AND DISCUSSION

Mitigation of Latent-Space Drift. We further investigate if our performance gains stem from the mitigated latent-space

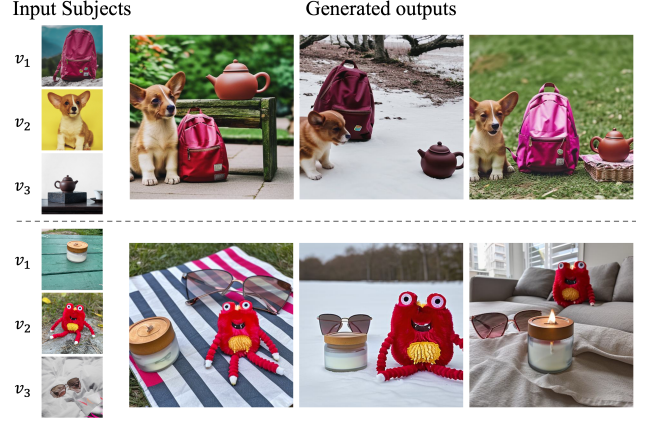


Figure 7: **Three-subject composition.** Our method successfully composes three personalized concepts without explicit layout guidance while preserving the identity of each subject across diverse scene contexts.



Figure 8: **Generated examples with our method on SDXL** for the prompt “v1 dog and v2 backpack”.

drift achieved by our regularizers. We first visualize the token embeddings using t -SNE in Fig. 5, which shows the distribution of new identifier token embeddings in relation to the teacher model’s pre-existing manifold. As observed, parameter fine-tuning baselines (e.g., DreamBooth, Custom Diffusion) tend to entangle the new identifier token (τ_c) with semantically related but distinct base tokens (e.g., “dog”, “backpack”). This overlapping distribution indicates a significant deformation of the pre-trained manifold, where the unique identity of the new concept is poorly resolved from its base category, leading to attribute mixing. In stark contrast, our method places the identifier τ_c in a distinct, well-isolated region that remains adjacent to, yet clearly separated from, its base class. This suggests that our approach successfully learns the new identity while minimizing the erosion of the teacher’s semantic structure.

We further assess concept separability quantitatively in Tab. 5. We define a *Disentanglement Gap* by contrasting the cosine similarity between embeddings of the same class (intra-class) against that with different classes (inter-class):

$$\text{Gap}(c) = \underbrace{\cos(e(\tau_c), e(c))}_{\text{Intra-class similarity}} - \underbrace{\frac{1}{|\mathcal{C}'|} \sum_{c' \in \mathcal{C}'} \cos(e(\tau_c), e(c'))}_{\text{Inter-class similarity}},$$

where $e(\cdot)$ denotes the text embedding and $\mathcal{C}'$ represents the set of non-target concepts. A larger gap score indicates

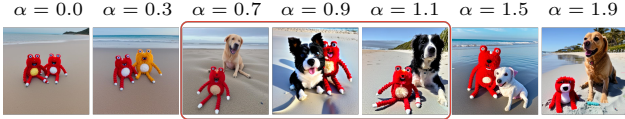


Figure 9: **Qualitative results for various α on “photo of v^* and dog”.** Low α causes concept omission (missing dog), while high α degrades the identity of v^* .

Table 4: **Ablation on attention weight β .** Performance improves with higher β .

Metric	β	0.0	0.1	0.3	0.5	0.7	0.9	1.0
Prompt Fidelity	T_{CLIP}	0.2881	0.2964	0.2909	0.2903	0.3027	0.3064	0.3102
Composition	I_{CLIP}	0.6272	0.6186	0.6357	0.6443	0.6529	0.6614	0.6730
Fidelity	I_{DINO}	0.5012	0.5127	0.5231	0.5316	0.5398	0.5426	0.5452

superior concept separability.

As reported in Tab. 5, our method yields a markedly larger gap compared to all baselines. For instance, while baselines often exhibit higher inter-class similarity due to collapsed semantic boundaries, our approach maintains a clearer margin. This verifies the improved separability observed in the t -SNE plots and explains the superior composition fidelity gains reported in Tabs. 1 and 2.

User Study. Since automated metrics may not fully capture human preferences and perceptual quality [Ryu et al., 2025], we additionally conduct a user preference study with 40 participants. Each user was given a set of blind questions to choose the best generated outputs from Custom Diffusion, MuDI, Tweedie, and ours, given a reference image of the target concept and a text prompt. As reported in Tab. 6, participants voted to the images generated by our method for a majority of test examples (72.5%). This confirms that our method generates visually appealing and contextually more accurate images.

Discussion. A potential counter-argument to our findings would be that the superior compositional reasoning of larger-scale models might render our explicit regularization redundant. However, we emphasize that the identity-composition trade-off is a fundamental consequence of the fine-tuning process itself. Regardless of the teacher’s strength, unconstrained optimization inevitably drifts the latent-space, leading to a degradation of the model’s reasoning capabilities.

In this regard, our framework serves as an architecture-agnostic safeguard. As demonstrated in our experiments with SD2.1 and SDXL, it provides a robust foundation that inherits and amplifies the strengths of the teacher while ensuring that high identity fidelity does not compromise the model’s inherent structural integrity and compositional knowledge. We anticipate that our approach will remain as an essential component in maintaining the delicate balance between learning new concepts and preserving foundational

Table 5: **Latent space disentanglement.** Higher $\text{Sim}_{\text{intra}}$ and lower $\text{Sim}_{\text{inter}}$ between the learned ‘ v_1 ’ token and other class tokens indicate superior disentanglement.

Method	$\text{Sim}_{\text{intra}} \uparrow$	$\text{Sim}_{\text{inter}} \downarrow$	Gap $\uparrow$
DreamBooth	0.4157	0.2569	0.1588
DeTex	0.4067	0.2589	0.1478
Dreammatcher	0.3865	0.2722	0.1143
MuDI	0.5309	0.3631	0.1678
Tweedie	0.7955	0.5295	0.2660
Ours	0.6232	0.2695	0.3537

Table 6: **Summary of user preference scores.**

	Custom Diffusion	MuDI	Tweedie	Ours	Total
Total Votes	19	148	53	580	800
(Ratio)	(2.4%)	(18.5%)	(6.6%)	(72.5%)	(100%)

expertise in the evolving landscape of generative models.

Limitations. Although our distillation-based approach greatly improves multi-concept generation, it still inherits some intrinsic limitations of the foundation model. Fig. II in Appendix E illustrates occasional failures in fine-grained attribute binding can still occur in exceptionally complex prompts, reflecting known issues in the teacher’s prior knowledge [Rassin et al., 2023, Zhang et al., 2024a, Zhuang et al., 2024]. Also, while our method effectively preserves compositional disentanglement for up to three personalized concepts, scaling to four or more introduces partial latent overlap due to the high-dimensional reasoning limits of fixed-resolution latent spaces. We provide a detailed analysis and qualitative examples of these failure cases in Appendix E.

6 SUMMARY

In this work, we revisit the identity–composition trade-off in personalized text-to-image generation and discover that this trade-off is driven by two distinct mechanisms: identity underfitting and latent-space drift, amplified in layout-free, multi-subject prompts. To systematically balance the two mechanisms, we propose a teacher–student scheme that incorporates explicit composition-preserving regularizers to anchor the student to the teacher’s semantic and spatial priors, alongside a standard reconstruction objective. By better preserving the compositional capability of the foundation model while learning new concepts, our approach effectively reduces latent-space drift and enables robust layout-free multi-subject composition without requiring external spatial guidance.

Acknowledgements

This work was supported by Samsung Electronics, Youlchon Foundation, National Research Foundation of Korea (NRF) grants (RS-2021-NR05515, RS-2024-00336576, RS-2023-0022663), and the Institute for Information & Communication Technology Planning & Evaluation (IITP) grants (RS-2022-II220264, RS-2024-00353131) funded by the Korean government.

References

- Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization. *ACM Transactions on Graphics (TOG)*, 2023.
- Moab Arar, Andrey Voynov, Amir Hertz, Omri Avrahami, Shlomi Fruchter, Yael Pritch, Daniel Cohen-Or, and Ariel Shamir. Palp: Prompt aligned personalization of text-to-image models. In *Proceedings of the ACM SIGGRAPH Conference and Exhibition on Computer Graphics and Interactive Techniques in Asia (SIGGRAPH ASIA)*, 2024.
- Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- 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 (NeurIPS)*, 2023.
- Yusuf Dalva, Hidir Yesiltepe, and Pinar Yanardag. LoRAShop: Training-free multi-concept image generation and editing with rectified flow transformers. In *The Thirtieth Annual Conference on Neural Information Processing Systems*, 2026.
- Ganggui Ding, Canyu Zhao, Wen Wang, Zhen Yang, Zide Liu, Hao Chen, and Chunhua Shen. FreeCustom: Tuning-free customized image generation for multi-concept composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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 *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.
- Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 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 (ICCV)*, 2023.
- Shaozhe Hao, Kai Han, Shihao Zhao, and Kwan-Yee K Wong. VICO: Plug-and-play visual condition for personalized text-to-image generation. *arXiv:2306.00971*, 2023.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.
- Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Identity decoupling for multi-subject personalization of text-to-image models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.
- Xuhui Jia, Yang Zhao, Kelvin C. K. Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv:2304.02642*, 2023.
- 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 (CVPR)*, 2023.
- Gihyun Kwon and Jong Chul Ye. Tweediemix: Improving multi-concept fusion for diffusion-based image/video generation. In *International Conference on Learning Representations*, 2025.
- Gihyun Kwon, Simon Jenni, Dingzeyu Li, Joon-Young Lee, Jong Chul Ye, and Fabian Caba Heilbron. Concept weaver: Enabling multi-concept fusion in text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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 (NeurIPS)*, 2023a.
- Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee.

- Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b.
- Zhizhong Li and Derek Hoiem. Learning without forgetting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.
- Yaoli Liu, Yao-Xiang Ding, and Kun Zhou. FreeFuse: Multi-subject lora fusion via auto masking at test time. *arXiv:2510.23515*, 2025.
- Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv:2310.04378*, 2023.
- Tuna Han Salih Meral, Enis Simsar, Federico Tombari, and Pinar Yanardag. Contrastive test-time composition of multiple LoRA models for image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.
- Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024.
- Jisu Nam, Heesu Kim, DongJae Lee, Siyoon Jin, Seungryong Kim, and Seunggyu Chang. DreamMatcher: Appearance matching self-attention for semantically-consistent text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.
- 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. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv:2204.06125*, 2022.
- Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- 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 (CVPR)*, 2022.
- Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fit-Nets: Hints for thin deep nets. *arXiv:1412.6550*, 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 (CVPR)*, 2023.
- 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 (CVPR)*, 2024.
- Suho Ryu, Kihyun Kim, Eugene Baek, Dongsoo Shin, and Joonseok Lee. Towards scalable human-aligned benchmark for text-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.
- Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

- Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *Proceedings of the ACM Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 2023.
- Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. p+: Extended textual conditioning in text-to-image generation. *arXiv:2303.09522*, 2023.
- Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. MS-Diffusion: Multi-subject zero-shot image personalization with layout guidance. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.
- 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 (ICCV)*, 2023.
- Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Decoupled textual embeddings for customized image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024.
- 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 (IJCV)*, 2024.
- Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- 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 (ICCV)*, 2023.
- Yasi Zhang, Peiyu Yu, and Ying Nian Wu. Object-conditioned energy-based attention map alignment in text-to-image diffusion models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024a.
- Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, and Zhongliang Jing. SSR-encoder: Encoding selective subject representation for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.
- Chenyi Zhuang, Ying Hu, and Pan Gao. MAGNET: We never know how text-to-image diffusion models work, until we learn how vision-language models function. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Preserving Compositionality for Robust Multi-Subject Personalization in Text-to-Image Generation (Supplementary Material)

Sangho Lee^{*1}

Eugene Baek^{*1}

Suho Ryu¹

Dongsoo Shin¹

Joonseok Lee^{†1}

¹Seoul National University, Seoul, Korea

A EXPERIMENTAL DETAILS

Details on Baseline Implementation. We evaluate our approach against existing fine-tuning methods, including DreamBooth [Ruiz et al., 2023], Textual Inversion [Gal et al., 2023], Custom Diffusion [Kumari et al., 2023], MuDI [Jang et al., 2024], TweedieMix [Kwon and Ye, 2025], FastComposer [Xiao et al., 2024], and MS-Diffusion [Wang et al., 2025]. We use the *official or widely adopted third-party implementations* for these baselines and configure them as follows:

- **DreamBooth** [Ruiz et al., 2023]: Implemented with Hugging Face Diffusers; batch size 1, learning rate 5×10^{-4} , 400 steps.
- **Textual Inversion** [Gal et al., 2023]: Hugging Face Diffusers implementation; token embedding trained for 3,000 steps with learning rate 0.005.
- **DETEX** [Wei et al., 2024]: We use the *official repository* and adopt the default settings.
- **FastComposer** [Xiao et al., 2024]: We use the *official repository* and follow the authors’ default settings.
- **MS-Diffusion** [Wang et al., 2025]: We use the *official repository* and follow the authors’ default settings. As the official implementation provides a layout-free option with all-zero bounding boxes, we adopt this setting.
- **Custom Diffusion** [Kumari et al., 2023]: Hugging Face Diffusers community implementation; 250 steps, batch size 1, learning rate 1×10^{-5} ; regularization images generated with DDIM.
- **MuDI** [Jang et al., 2024]: We use the *official repository* and follow the authors’ default settings. Unless otherwise noted, we evaluate in the *layout-free (text-only)* regime to match our protocol (i.e., no external boxes/poses/masks), keeping the optimizer, learning-rate schedule, and loss coefficients as provided by the authors.
- **TweedieMix** [Kwon and Ye, 2025]: We use the *official repository* and adopt the default configuration for multi-concept fusion. For layout-free evaluation, we disable any optional spatial guidance and use the recommended adapter ranks/fusion weights from the repo; other hyperparameters are left at their defaults.

All other baselines are evaluated with the authors’ official codebases when available; if multiple implementations exist, we prefer the official repository and default configurations to ensure fairness and reproducibility.

Details on Evaluation Metrics. To ensure reproducibility, we provide the specific implementation details for the metrics used in our experiments:

- **Backbone Architectures:** For all CLIP-based metrics ($T_{\text{CLIP}}, I_{\text{CLIP}}$), we employ OpenAI CLIP ViT-L/14 (openai/clip-vit-large-patch14) checkpoint. For I_{DINO} , we use DINOv2 [Oquab et al., 2024] ViT-B/14 (dinov2_vitb14) model. All models are used in their pre-trained states without further fine-tuning.

^{*}Equal Contribution

[†]Corresponding author

^{*}Equal Contribution

[†]Corresponding author

- **Object-level Evaluation:** Following the MuDI protocol [Jang et al., 2024], object-level crops for multi-concept scenarios are obtained using an off-the-shelf segmentation model to isolate each concept’s region before computing cosine similarity.
- **Embedding Normalization:** All embeddings are ℓ_2 -normalized prior to calculating the cosine similarity, ensuring that the scores are bounded between $[-1, 1]$.

B COMPUTATIONAL COST

We compare the computational efficiency of our method with the baselines in Tab. I. DreamMatcher, FastComposer, and MS-Diffusion do not require additional training time as they leverage existing baselines (*e.g.*, DreamBooth or Stable Diffusion). Their inference time varies depending on the underlying baselines, so they are not separately measured. MuDI and Tweedie are fine-tuned on SDXL, and their resolution is set to 1024×1024 to match the SDXL standard, whereas other models operate on the 512×512 resolution. This difference should be considered when comparing computational costs. We conduct all experiments on four NVIDIA A100 HBM2e 80GB PCIe GPUs.

Method	Training Time (hrs)	Inference Time (s) per 1 image
DreamBooth	0.5	1.2
Textual Inversion	0.5	1.2
DETEX	0.75	1.5
MuDI	1	20
Tweedie	1	50
Custom Diffusion	1.5	1.0
Ours (SDv2.1)	0.8	1.2
Ours (SDXL)	1.2	5.0

Table I: Computational cost comparison. Our method achieves fast training and competitive inference efficiency.

C PROMPTS IN OUR EXPERIMENTS

Prompts in our Experiments. We design our prompts to evaluate compositionality across four categories: **Place**, **Attribute**, **Outfit**, and **Style**. Each prompt contains placeholders (`{ }`), which are replaced with the personalized concepts during evaluation.

Place:

- A photo of { } and { } in the jungle / in the snow / on the beach / on a cobblestone street
- A photo of { } and { } on top of pink fabric / on a wooden floor / on a purple rug in a forest
- A photo of { } and { } with a city / mountain / blue house / wheat field / tree with autumn leaves / the Eiffel Tower in the background
- A photo of { } and { } floating on water / on green grass with sunflowers / on a dirt road / on a white rug

Attribute:

- A photo of red { } and purple { }
- A photo of purple { } and green { }
- A gold / bronze / paper sculpture of { } and { }

Outfit:

- A photo of { } and { } wearing a bowtie / top hat / scarf / santa hat / crown
- A photo of { } and { } in an astronaut outfit / firefighter outfit

Style:

- An oil painting of {} and {} in Baroque style
- A painting of {} and {} in the style of Van Gogh
- A watercolor painting of {} and {} in Art Nouveau style
- A charcoal sketch of {} and {} with intricate shading
- A minimalist line art drawing of {} and {}

Multi-concept Subject Pairs. To evaluate the ability to handle multiple personalized subjects, we curate a dataset consisting of 15 diverse subject pairs. These pairs test both identity preservation and the prevention of identity leakage between distinct subjects. The specific combinations used in our multi-concept experiments are detailed in Tab. II.

Concept 1	Category 1	Concept 2	Category 2
backpack	backpack	dog6	dog
dog	dog	robot_toy	toy
can	can	monster_toy	toy
cat2	cat	dog3	dog
bear_plushie	stuffed animal	robot_toy	toy
backpack	backpack	cat	cat
pink_sunglasses	glasses	dog2	dog
teapot	teapot	rc_car	toy
wolf_plushie	stuffed animal	duck_toy	toy
berry_bowl	bowl	clock	clock
shiny_sneaker	sneaker	candle	candle
red_cartoon	cartoon	dog8	dog
fancy_boot	boot	grey_sloth_plushie	stuffed animal
colorful_sneaker	sneaker	poop_emoji	toy
dog5	dog	monster_toy	toy

Table II: Multi-concept subject pairs and their categories used in our evaluation.

D USER STUDY DESIGN

We provide an example of user-study questions in Fig. I.

E EXTENDED ANALYSIS OF LIMITATIONS

In this section, we provide further discussion on the boundary conditions of our proposed method, focusing on attribute binding and latent space scalability, as illustrated in Fig. II.

Attribute Binding Issues. Our model’s performance is fundamentally tied to the teacher model’s prior knowledge. Fig. II(a) illustrates cases where prompt complexity exceeds the teacher’s inherent capacity. When multiple subjects are assigned specific colors (*e.g.*, “purple backpack and green plushie”), color bleeding or swapped assignments can occur. This indicates that while our method preserves the teacher’s structural integrity, it cannot rectify existing linguistic-visual misalignments in the foundation model [Rassin et al., 2023].

Scalability to Multiple Concepts. To evaluate the scalability of our approach, we visualize the latent space using t-SNE for an increasing number of personalized tokens, as shown in Fig. II(b). The semantic grouping remains distinct and reasonably disentangled for up to three concepts. However, scaling to four concepts introduces partial overlap in feature activations. This bottleneck suggests a capacity limit in fixed-resolution latent manifolds when representing multiple high-dimensional identities simultaneously.

From below concept images and prompt, which one is the best to reflect both identity details and prompt

Prompt: Watercolor painting of (left) dog and (right) backpack

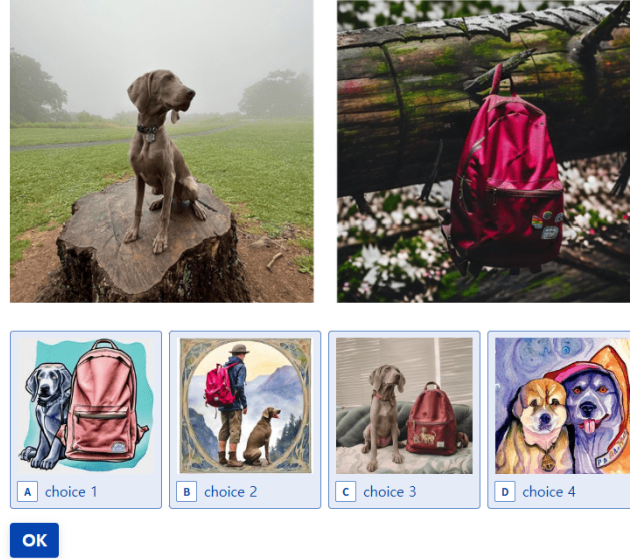


Figure I: An example of our user study question.

F ADDITIONAL QUALITATIVE RESULTS

To further demonstrate effectiveness of our method, we present additional qualitative comparisons with baselines in Fig. III-VI and Fig. VII. In each generated image, we mark with three icons indicating whether it faithfully contains the personalized objects (*e.g.*, can or toy) and the textual prompt. Our approach generates coherent multi-concept images without explicit layout guidance, preserving both concept distinction and spatial consistency where baselines often show concept blending or omission. Each image includes icons for target concepts and the prompt: green indicates accurate alignment, red indicates an error.

While previous figures have focused on multi-subject personalization (two personalized concepts), we also evaluate our framework on scenarios combining a single personalized concept (v_1) with native objects (*e.g.*, a dog or a beach).

As shown in Fig. VII, our approach generates coherent images without explicit layout guidance, preserving both the identity of the personalized concept and the semantic attributes of native objects. Our method maintains high spatial consistency and concept distinction, whereas baselines often exhibit identity leakage or fail to render the native objects correctly according to the prompt.

G ADDITIONAL ABLATION STUDIES

Effect of Distillation Strength. Fig. VIII complements Tab. 3 by illustrating how the distillation weight α controls the identity-composition balance on a prompt “photo of V^* and dog”. With $\alpha=0$, reconstruction-only training drifts from the teacher manifold, yielding concept omission (dog missing) and entangled shapes. A moderate range (0.7-1.1) restores spatial binding and prompt adherence, while preserving V^* identity—consistent with the Scene Fidelity peak in Tab. 3. When α becomes too large, however, the student over-mimics the teacher, slightly dulling the fine details of V^* even though compositionality remains strong. Token-wise cross-attention maps (dog on the top and the GT reference object on the bottom) corroborate this trend; attention becomes localized and non-overlapping in the moderate regime, evidencing reduced alignment remapping and fewer blending errors.

Intermediate Layer Selection. Beyond the distillation strength, we further investigate the choice of supervised U-Net blocks used for feature distillation and cross-attention alignment. Table III compares different layer configurations, including encoder, bottleneck, and decoder blocks. The results are measured on a subset of multi-concept pairs for design analysis and are therefore not directly comparable to the full benchmark results in the main paper. Overall, restricting supervision to

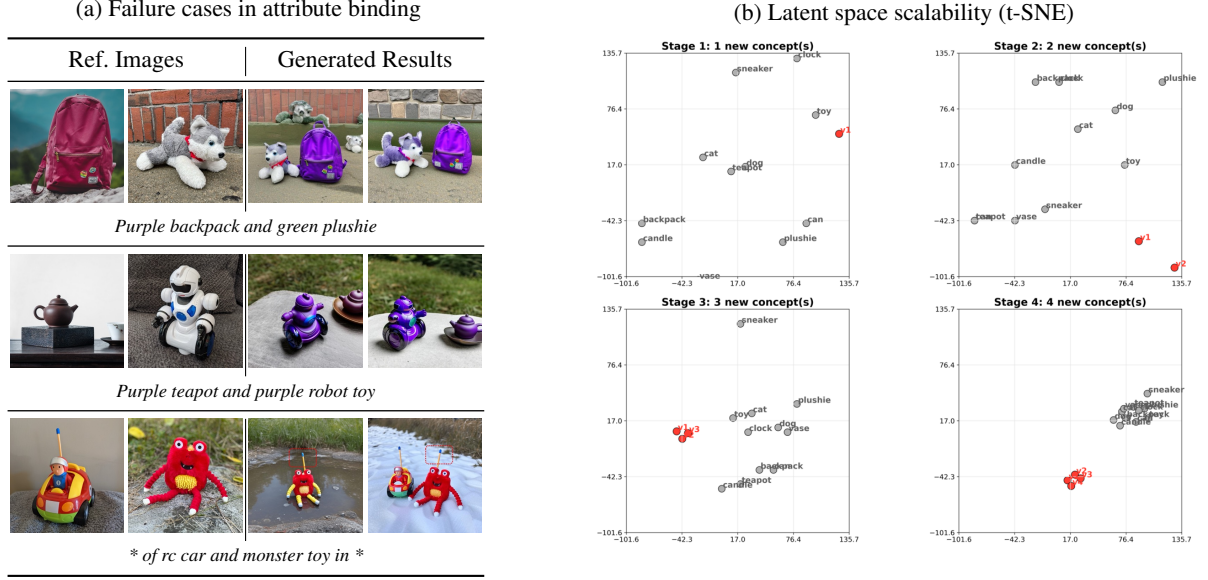


Figure II: Analysis of limitations. (a) Occasional failures in attribute binding inherited from the teacher model’s prior. (b) Scalability analysis showing increased latent overlap as the number of concepts grows to four.

Table III: **Ablation on supervised U-Net layers.** Restricting supervision to the middle block achieves a favorable efficiency-performance trade-off, while larger layer sets provide no consistent gains.

Setting	Object $I_{CLIP} \uparrow$	Object $I_{DINO} \uparrow$	$T_{CLIP} \uparrow$	Comp $I_{CLIP} \uparrow$	Comp $I_{DINO} \uparrow$
Mid-only (Ours)	0.65 ± 0.07	0.47 ± 0.04	0.25 ± 0.03	0.72 ± 0.08	0.56 ± 0.11
down _{0,1,2}	0.65 ± 0.07	0.46 ± 0.04	0.25 ± 0.03	0.72 ± 0.08	0.56 ± 0.14
Mid+down ₂	0.66 ± 0.07	0.47 ± 0.04	0.25 ± 0.03	0.72 ± 0.08	0.55 ± 0.12
Mid+down _{0,1,2}	0.66 ± 0.07	0.46 ± 0.04	0.25 ± 0.03	0.73 ± 0.08	0.56 ± 0.13
up ₁	0.63 ± 0.06	0.41 ± 0.06	0.24 ± 0.02	0.74 ± 0.07	0.52 ± 0.14
Mid+up ₁	0.64 ± 0.06	0.42 ± 0.06	0.24 ± 0.03	0.75 ± 0.08	0.53 ± 0.13
Mid+up _{1,2,3}	0.64 ± 0.06	0.42 ± 0.06	0.25 ± 0.02	0.76 ± 0.07	0.56 ± 0.13

the middle block achieves performance comparable to larger layer sets despite its lowest spatial resolution and smallest computational overhead. While incorporating additional encoder or decoder blocks occasionally improves composition fidelity, these gains are inconsistent and often come at the expense of identity preservation. These observations justify our default choice of applying supervision only to the middle block.

Sensitivity to Layer Weighting. We also examine whether the proposed framework is sensitive to the exact weighting assigned to different supervised layers. Specifically, we compare uniform weighting against early-heavy and late-heavy schemes for both feature distillation and cross-attention alignment. As reported in Table IV, performance remains relatively stable across weighting strategies, indicating that our framework is robust to the precise choice of layer weights. Consistent with the layer-selection analysis, stronger supervision on decoder layers tends to slightly improve composition fidelity while weakening identity preservation, suggesting over-anchoring to the teacher representation.

Ref. Image	Prompt	Tuning-free		Fine-tuning			
		FastComp.	MS-Diff.	Custom Diff.	MuDI	Tweedie	Ours
 v1	Charcoal sketch of						
	Simple lineart of						
	Watercolor painting of						
	Photo of in jungle						
	Photo of in snow						
	Red can and purple toy						
	Painting of Van Gogh style						
	Gold statue of						
	Painting of						
	Photo of floating on the water						
 v2	Charcoal sketch of						
	Simple lineart of						
	Watercolor painting of						
	Photo of in jungle						
	Photo of in snow						
	Red can and purple toy						
	Painting of Van Gogh style						
	Gold statue of						
	Painting of						
	Photo of floating on the water						

Figure III: Comprehensive comparison of multi-subject personalization baselines.

Ref. Image	Prompt	Tuning-free		Fine-tuning			
		FastComp.	MS-Diff.	Custom Diff.	MuDI	Tweedie	Ours
 v1	Bronze statue of						
	Gold statue of						
	Oil painting of						
	Charcoal sketch of						
 v2	Oil painting of Van Gogh style						
	Photo of wearing santa hat						
	Photo of wearing top hat						
	Simple line art of						
	Watercolor painting of						
	Photo of floating on the water						

Figure IV: Comprehensive comparison of multi-subject personalization baselines on the Backpack and Dog dataset.

Table IV: **Sensitivity to layer weighting schemes.** The proposed framework remains robust across different weighting configurations.

Setting	Object $I_{\text{CLIP}} \uparrow$	Object $I_{\text{DINO}} \uparrow$	$T_{\text{CLIP}} \uparrow$	Comp $I_{\text{CLIP}} \uparrow$	Comp $I_{\text{DINO}} \uparrow$
Uniform	0.65 ± 0.07	0.37 ± 0.15	0.25 ± 0.03	0.72 ± 0.08	0.45 ± 0.21
Feat-early-heavy	0.65 ± 0.07	0.34 ± 0.14	0.25 ± 0.02	0.79 ± 0.08	0.50 ± 0.24
Feat-late-heavy	0.66 ± 0.07	0.35 ± 0.15	0.25 ± 0.02	0.79 ± 0.07	0.51 ± 0.23
Attn-early-heavy	0.66 ± 0.07	0.34 ± 0.14	0.25 ± 0.02	0.80 ± 0.07	0.51 ± 0.22
Attn-late-heavy	0.65 ± 0.07	0.34 ± 0.14	0.25 ± 0.03	0.78 ± 0.08	0.51 ± 0.23

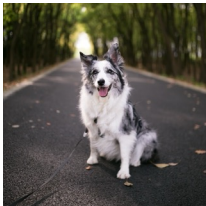
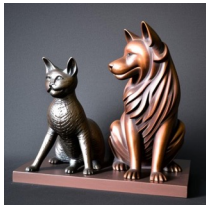
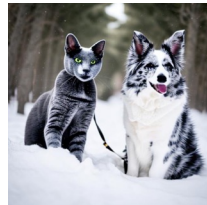
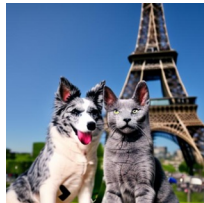
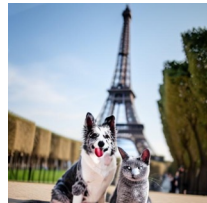
Reference	Prompt	Generated Images (Ours)		
 v1  v2		Style Transfer		
	Charcoal sketch of a dog			
	Van Gogh-style portrait			
	Bronze statue			
	Gold statue			
	Environmental Change			
	At beach			
	In snow			
	In front of Eiffel tower			

Figure V: Qualitative results under various conditions. The first column shows the input reference images.

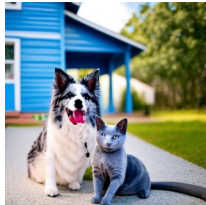
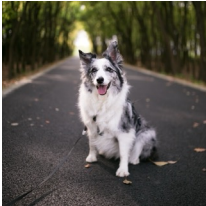
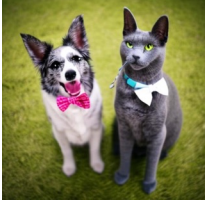
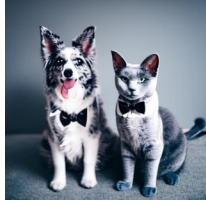
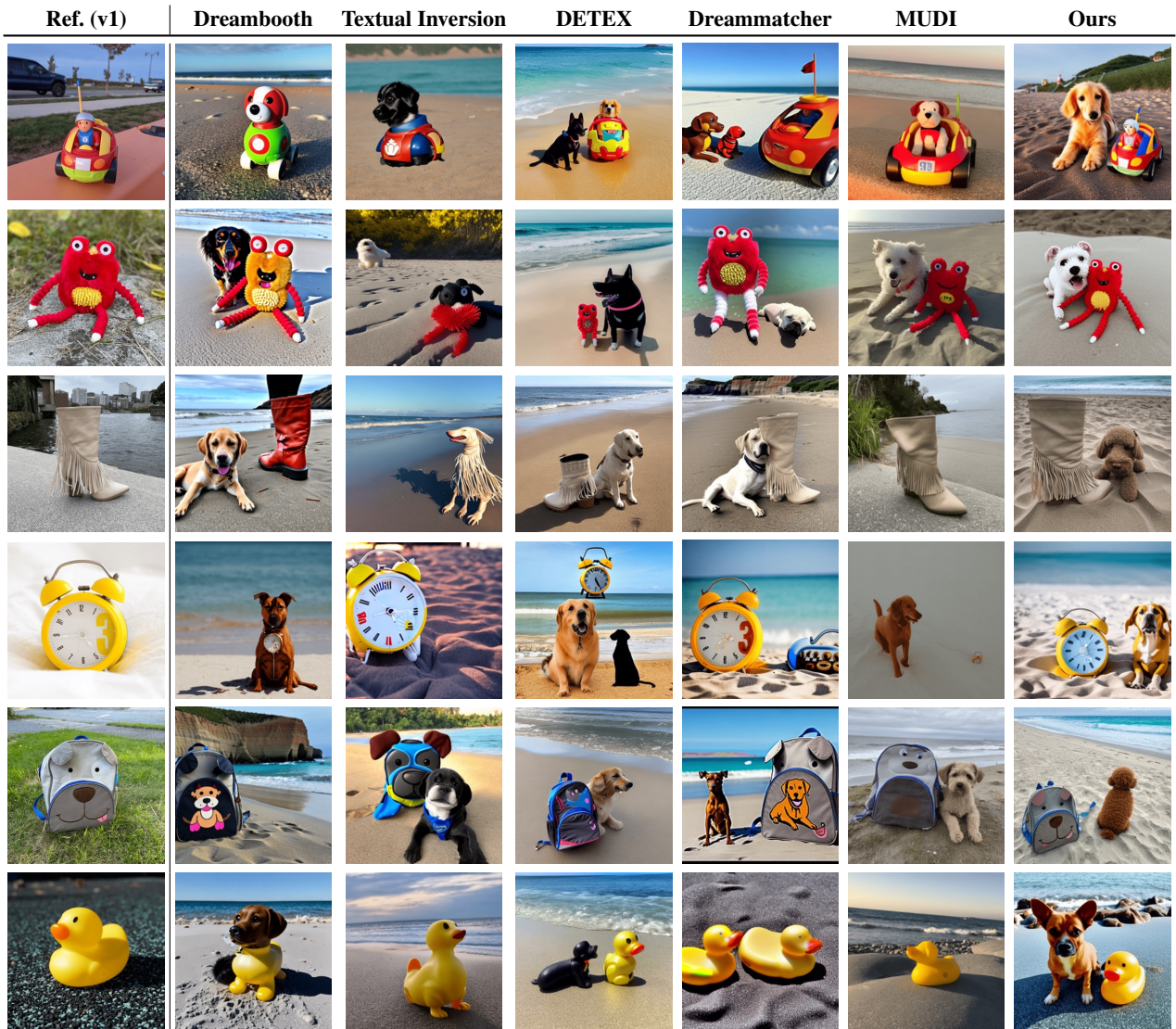
Reference	Prompt	Generated Images (Ours)		
 v1	In front of blue house			
	In front of mountain			
	Environmental Change			
	Photo of wearing santa hat			
	Photo of wearing top hat			
	Photo of wearing astronaut outfit			
 v2	Photo of wearing bow tie			
	Photo of wearing a crown			
	Outfit Change			

Figure VI: Examples of generated images from our model under different conditions. The first column shows the input reference images.



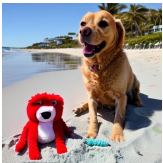
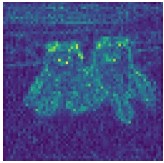
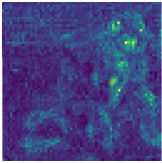
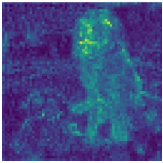
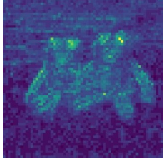
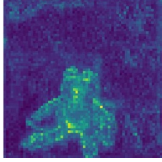
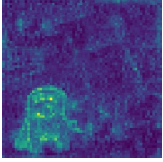
Reference	No Distill ($\alpha=0.0$)	Moderate ($\alpha=0.7$)	Full ($\alpha=2.0$)
			
Cross-Attention			
			
Concept Fidelity	Coherent, but entangled	Faithful rendering	V* fidelity is low
Compositionality	Fused objects	Mostly coherent	Well-integrated
Prompt Adherence	Dog is missing	Matches prompt	V* is less prominent

Figure VIII: **Effect of the distillation weight α .** For the prompt “photo of V* and dog”, small α causes omission and fused objects; moderate α yields faithful, well-integrated scenes; excessive α slightly weakens V* identity. Rows show outputs and tokenwise cross-attention (dog/top, GT object/bottom).