FIRST FUNDAMENTAL FORM REGULARIZATION FOR ROBUST LORA FINE-TUNING OF DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-Rank Adaptation (LoRA) has become the de facto standard for personalizing text-to-image diffusion models due to its efficiency and flexibility. Despite its popularity, LoRA fine-tuning often exhibits practical instability, including sensitivity to prompt variations and interference when multiple adapters are used together. We argue that these issues stem from uncontrolled distortion of the pretrained representation geometry during fine-tuning. In this work, we propose **FF-LoRA**, a relational-distillation regularizer for LoRA fine-tuning that matches pairwise cosine-similarity Gram matrices, computed over the spatial tokens of each training sample, between the frozen base model and the LoRA-adapted model. Methodologically, this is in the same family as relational knowledge-distillation losses (RKD (Park et al., 2019), SP (Tung & Mori, 2019)), transposed from cross-sample relations to within-sample token relations; the further delta over the closest baseline (SP) is the cosine normalization, which we ablate in Section 4.5. The regularizer is loosely motivated by the differential-geometric notion of preserving the *first fundamental form* of the representation manifold, but the quantity actually penalized is a token-level relational Gram matrix, not the metric tensor itself; we use the geometric interpretation as motivation rather than as a theoretical contribution (see Appendix L). This regularization encourages LoRA to adapt to new concepts while maintaining the local distance and angle structure of the pretrained representation. FF-LoRA is lightweight, easy to implement, and fully compatible with standard diffusion training pipelines. Experimental results demonstrate that FF-LoRA improves robustness under prompt variations, reduces overfitting to limited training captions, and significantly enhances compatibility when multiple LoRA adapters are used together.

1 INTRODUCTION

Recent text-to-image (T2I) diffusion models (Rombach et al., 2022; Podell et al., 2023; OpenAI, 2023; MidJourney, 2022) have achieved remarkable success in generating diverse and high-quality visual content, enabling a wide range of creative applications such as artistic rendering, product prototyping, and character design. Despite being trained on massive datasets, however, pretrained diffusion models remain constrained by their training distributions and often struggle to faithfully synthesize novel concepts or replicate specific user-defined identities. For example, Stable Diffusion (SD) (Rombach et al., 2022) cannot natively generate a particular user's pet or a newly designed anime character. This limitation has driven increasing interest in personalization techniques, which adapt pretrained diffusion models to new concepts using only a small number of example images.

Existing personalization approaches can be broadly categorized into encoder-based methods and fine-tuning-based methods. Encoder-based methods (Jia et al., 2023; Wei et al., 2023) aim to generalize across concepts without test-time adaptation, but typically require large-scale text–image datasets and expensive computational resources (e.g., millions of image–text pairs or specialized hardware). In contrast, fine-tuning approaches directly adapt the base diffusion model from a few target images, including text embedding learning (Gal et al., 2022), cross-attention tuning (Kumari et al., 2023), full-model fine-tuning (Ruiz et al., 2023), and low-rank adaptation (LoRA) (Hu et al., 2021; cloneofsimo, 2022). Among these, LoRA has emerged as the de facto standard due to its efficiency, simplicity, and flexibility, supporting a large ecosystem of lightweight adapters for diverse styles and concepts.

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.

Despite its popularity, LoRA fine-tuning often exhibits practical instability. In real-world usage, a LoRA adapter may perform well on prompts close to the training captions, yet degrade noticeably under prompt rephrasing, compositional prompts, or distribution shifts. Moreover, modern personalization pipelines frequently maintain a single backbone model alongside a growing library of LoRA adapters. As the number of adapters increases, combining or switching between multiple LoRAs becomes common, where interference and unintended style leakage are frequently observed. We argue that these issues stem from a common cause: *LoRA fine-tuning can significantly distort the geometric structure of the pretrained representation space*, thereby disrupting the local semantic relationships learned during large-scale pretraining.

Motivated by this observation, we propose **FF-LoRA**, a geometry-aware regularization framework for robust LoRA fine-tuning. Our key idea is to explicitly preserve the *first fundamental form* of the representation manifold in a discrete and practical manner. In differential geometry, the first fundamental form characterizes local distances and angles induced by an embedding, and forms the basis of intrinsic geometric structure (Wintner, 1955). Preserving such local metric structure has long been recognized as crucial in embedding and manifold learning to maintain meaningful neighborhood relationships (Abraham et al., 2007; Van Der Maaten, 2009).

Instead of computing Jacobians explicitly, FF-LoRA adopts a simple surrogate: for each training sample, we match the pairwise cosine similarity (Gram) matrices computed over its spatial feature tokens between the frozen base model and the LoRA-adapted model. This regularization encourages LoRA to adapt the model to new concepts while preventing excessive deformation of local distances and angles in the representation space. FF-LoRA is simple to implement, incurs minimal computational overhead, and is fully compatible with standard diffusion training pipelines. Experimental results demonstrate that preserving the first fundamental form leads to more stable personalization, improved generalization beyond training prompts, and enhanced compatibility in multi-LoRA settings.

2 Related Work

2.1 Text-to-Image Diffusion Models and Personalization

Diffusion-based generative models (Ho et al., 2020; Sohl-Dickstein et al., 2015; Nichol & Dhariwal, 2021; Song et al., 2020b;a) have enabled large-scale text-conditioned systems such as DALL-E 2 (Ramesh et al., 2022), Imagen (Saharia et al., 2022), and Stable Diffusion (SD) (Rombach et al., 2022), of which SD is the most widely adopted open-source backbone and serves as the base model for most personalization pipelines.

Personalizing diffusion models from a small number of reference images has drawn increasing attention. One line of work focuses on *text embedding optimization*, where Textual Inversion (Gal et al., 2022) learns new token embeddings to represent unseen concepts, but often struggles to achieve strong subject binding and robustness. Another line directly fine-tunes the denoising network or attention modules, including DreamBooth (Ruiz et al., 2023) and Custom Diffusion (Kumari et al., 2023), which can achieve high fidelity but are prone to overfitting training captions and may generalize poorly under prompt rephrasing or distribution shifts.

To improve efficiency and usability, adapter-based approaches introduce lightweight modules to steer generation without updating the full model (Mou et al., 2024; Zhang et al., 2023a). In parallel, several training-free personalization methods (Jia et al., 2023; Wei et al., 2023; Shi et al., 2023; Chen et al., 2024) avoid modifying model weights altogether, but typically rely on retrieval, prompt engineering, or external guidance, which can limit subject fidelity or scalability.

2.2 Parameter-Efficient Personalization via Low-Rank Adaptation

LoRA (Hu et al., 2021; cloneofsimo, 2022) injects trainable low-rank updates into pretrained weights while keeping the backbone frozen, dramatically reducing trainable parameters and enabling large-scale sharing of adapters in practice, and several recent works extend it by modifying its parameterization or training strategy. AdaLoRA (Zhang et al., 2023b) dynamically reallocates rank budgets across layers to better utilize limited adaptation capacity. DoRA (Liu et al., 2024) decomposes weight updates into magnitude and direction, aiming to improve optimization stability and adaptation quality. OFT (Qiu et al., 2023) constrains fine-tuning updates to lie in an orthogonal subspace, thereby

reducing interference with pretrained representations. SVDiff (Han et al., 2023) fine-tunes singular values of pretrained weights using SVD-based decomposition. PaRa (Chen et al., 2025) explicitly reduces the effective output rank of diffusion models, restricting the generation space to improve robustness and composability.

While these methods differ in formulation, they largely focus on controlling *parameter space properties*: rank allocation, update magnitude, orthogonality, or output dimensionality. As a result, their robustness under prompt variation and stability under multi-adapter composition remain limited in practice, as independently trained adapters may introduce incompatible deformations.

In contrast, our work addresses a complementary and underexplored perspective: how LoRA-style fine-tuning deforms the *geometry of the representation space*. Rather than constraining parameters directly, FF-LoRA regularizes the induced metric structure of intermediate representations, which we show is critical for robustness and adapter compatibility.

2.3 Geometry Preservation and Relational Distillation

Preserving local geometric structure has long been a central principle in representation learning. In differential geometry, the first fundamental form characterizes local distances and angles induced by an embedding (Wintner, 1955). In machine learning, many embedding and manifold learning methods explicitly preserve neighborhood structure or pairwise distances to maintain semantic consistency (Abraham et al., 2007; Van Der Maaten, 2009).

Closely related to our formulation, the knowledge-distillation literature has extensively studied matching internal representations between a teacher and a student model. FitNets (Romero et al., 2014) introduced feature-level distillation by aligning intermediate hidden activations, Relational Knowledge Distillation (RKD) (Park et al., 2019) proposed to match pairwise distances and angles between teacher and student features, and Similarity-Preserving Knowledge Distillation (SP) (Tung & Mori, 2019) explicitly aligned batch-level Gram matrices of feature similarities. Centered Kernel Alignment (CKA) (Kornblith et al., 2019) provides a closely related framework for comparing pairwise feature similarity structures across networks. Methodologically, our discrete token similarity matrix is in the same family as RKD/SP/CKA: pairwise feature relations are preserved between a frozen reference (here, the pretrained diffusion backbone) and an adapted model (the LoRA-fine-tuned variant); whereas RKD/SP match relations across samples in a batch, FF-LoRA matches relations across spatial tokens within each sample. We position FF-LoRA as a specialization of relational feature preservation to the LoRA fine-tuning setting of large diffusion models, motivated by a differential-geometric reading of the induced metric, and benchmark it quantitatively against RKD/SP-style baselines in Section 4.

Recent work further highlights the importance of distance- and angle-preserving objectives in high-dimensional neural representations (Fischer & Ma, 2024), and geometric learning on structured domains such as meshes relies on maintaining intrinsic geometric quantities (Barda et al., 2021). However, such geometric principles have rarely been applied to parameter-efficient fine-tuning of large generative models.

FF-LoRA bridges this gap by introducing a practical first-fundamental-form regularization for LoRA fine-tuning in diffusion models. By explicitly constraining local metric distortion, our method improves robustness to prompt variation and enables stable composition of independently trained adapters, without modifying the LoRA parameterization or increasing inference cost.

3 Method

3.1 Preliminaries

We assume a pretrained conditional diffusion model with parameters θ_0, trained with the standard noise-prediction MSE objective $\mathcal{L}_{\text{diff}}$ (Ho et al., 2020). LoRA (Hu et al., 2021; cloneofsimo, 2022) introduces trainable low-rank updates $\Delta W = (\alpha/r)\, BA$ with $A \in \mathbb{R}^{r\times d_{\text{in}}}$, $B \in \mathbb{R}^{d_{\text{out}}\times r}$ into selected linear layers, keeping θ_0 frozen and updating only $\phi = \{(A_\ell, B_\ell)\}_{\ell\in\mathcal{L}}$.

3.2 Feature Mapping and Induced Geometry

Let $F_\theta(\cdot)$ be a fixed intermediate module of the diffusion backbone (the U-Net mid-block by default; we ablate this choice in Section 4.4). Given a training pair (x, c), a timestep t sampled from the diffusion training schedule, and the corresponding noisy latent x_t, the module produces a feature map $F_\theta(x_t, c; t) \in \mathbb{R}^{C\times H\times W}$, which we view as a set of $N = HW$ *spatial tokens* $\{z_i\}_{i=1}^N$ with $z_i \in \mathbb{R}^C$: each token is the representation of one spatial location of the same image, so the token set forms a discrete point sample from a local region of the representation manifold. The induced local metric $G_\theta = J_\theta^\top J_\theta$ with $J_\theta = \partial F_\theta/\partial x$ corresponds to the *first fundamental form* in differential geometry, characterizing intrinsic distances and angles at first order. We use this term to motivate our regularizer; the quantity we actually penalize (Eq. 3) is a token-level Gram matrix of feature inner products computed *within each sample*, which is a discrete relational surrogate of G_θ rather than G_θ itself (see Appendix L).

3.3 First Fundamental Form Regularization

Directly computing the Jacobian J_θ and the induced metric is computationally prohibitive for large diffusion models. Instead, we adopt a simple and practical discrete proxy that captures the same first-order geometric information.

At each training step, we run two forward passes that share the *same* noisy latent x_t, timestep t, and conditioning c: one through the frozen base model (adapter deactivated) and one through the LoRA-adapted model. From each pass we extract the token matrix at the chosen module, uniformly subsampling $T \le N$ tokens (we use $T = 256$) to bound the Gram-matrix cost:

$$Z^{(0)} = \big[z_1^{(0)}; \dots; z_T^{(0)}\big] = F_{\theta_0}(x_t, c; t), \qquad Z^{(\phi)} = \big[z_1^{(\phi)}; \dots; z_T^{(\phi)}\big] = F_{\theta_0,\phi}(x_t, c; t). \tag{1}$$

After ℓ_2 normalization, the cosine similarity between two spatial tokens of the same sample is

$$S_{ij} = \langle \hat{z}_i, \hat{z}_j \rangle, \qquad \hat{z}_i = \frac{z_i}{\|z_i\|_2}, \qquad S \in \mathbb{R}^{T\times T}. \tag{2}$$

The similarity matrix S can be interpreted as a discrete Gram matrix of inner products in the representation space. Preserving this Gram matrix over the token set of a sample is equivalent to preserving pairwise angles and relative distances of the feature embedding; because all tokens are drawn from a single image, they lie in a local region of the representation manifold, making the token-level Gram matrix a natural finite sample of the local metric structure that G_θ describes (this relational-matching form is closely related to the relational distillation objectives of RKD (Park et al., 2019) and SP (Tung & Mori, 2019), which match relations across samples in a batch; we match relations across spatial tokens within a sample). We define the geometry preservation loss as

$$\mathcal{L}_{\text{geom}} = \mathrm{E}_{(x,c),\,t}\Big[\tfrac{1}{T^2}\,\|S^{(\phi)} - S^{(0)}\|_F^2\Big], \tag{3}$$

where $S^{(0)}$ and $S^{(\phi)}$ denote token similarity matrices computed from the base model and the LoRA-adapted model under the shared (x_t, c, t), and the expectation is taken over training samples and the standard diffusion timestep distribution. When the batch size exceeds one, the loss averages over per-sample Gram matrices.

Remark: batch size. Because the pairwise relations in Eq. 3 are computed *between spatial tokens within each sample* rather than between samples in a batch, the regularizer is well-defined at batch size one (the standard DreamBooth-style personalization setting used in our experiments, Section 4.1): each image contributes a $T \times T$ Gram matrix with $T = 256$ token pairs, independent of the batch dimension.

Remark: timesteps. The features are compared at the same timestep t that drives the denoising loss of that step, so geometry preservation is enforced *in expectation over the entire noise schedule* rather than at a single fixed timestep; Appendix D ablates fixed-timestep variants against this default.

Remark: cosine normalization. Normalizing tokens before the inner product makes S capture the *angular* structure of the local geometry and discards per-token feature norms. This is a deliberate

relaxation: feature norms in diffusion U-Nets drift systematically with adaptation strength, and penalizing them couples the regularizer to the overall adaptation magnitude that LoRA must be free to change; preserving angles constrains the *conformal class* of the local metric while leaving per-location scale adaptable. Section 4.5 ablates the raw (unnormalized) Gram variant, which corresponds to an SP-style objective.

3.4 Final Objective

The final training objective is

$$\min_{\phi} \; \mathcal{L} = \mathcal{L}_{\text{diff}}(\theta_0, \phi) + \lambda_{\text{lora}} \|\phi\|_2^2 + \lambda_{\text{geom}} \mathcal{L}_{\text{geom}}, \tag{4}$$

where λ_{lora} controls standard weight regularization and λ_{geom} balances adaptation against geometric stability. In practice, λ_{geom} is linearly warmed up during early training stages to allow initial concept adaptation before enforcing stronger geometry preservation.

Implication for multi-LoRA composition. LoRA composition is typically implemented by linearly combining low-rank updates; without geometric constraints, independently trained adapters can induce incompatible distortions in the representation space and interfere when mixed. By regularizing the induced local metric, FF-LoRA constrains each adapter to a geometry consistent with the pretrained model, so multiple FF-LoRA adapters tend to preserve similar distance/angle structure and their linear combination is more stable; we validate this empirically in Section 4.

4 Experiments

We evaluate FF-LoRA on text-to-image personalization with an emphasis on (i) robustness under prompt variations and (ii) multi-adapter composition. Unless otherwise specified, we adopt Stable Diffusion XL (SDXL) (Podell et al., 2023) as the backbone and DDIM sampling with $\eta = 0$ for deterministic generation. All experiments are conducted on a single A100 GPU with 40GB VRAM.

4.1 Experimental Setup

Tasks. We consider three tasks: (1) **Single-subject personalization** on DreamBooth-style concepts, (2) **Multi-adapter composition** including subject+subject and subject+style mixing, and (3) **Ablations and analysis** on λ_{geom} and feature extractor choices.

Dataset and training. We follow the DreamBooth protocol (Ruiz et al., 2023) and use $K = 5$ images per subject. The main tables report the standard 8-subject set together with the DreamBooth-30 benchmark (full per-method table in Appendix C). All methods are trained for 1,000 steps with batch size 1, with the same optimizer and learning rate unless otherwise noted. Note that batch size 1 is fully compatible with the regularizer: the Gram matrix of Eq. 3 is computed over $T = 256$ spatial tokens *within* each sample, not across samples (Section 3). FF-LoRA uses the same LoRA parameterization but adds the proposed geometry regularization.

Implementation details. All methods use AdamW with learning rate 10^{-4}, fp16 mixed precision, gradient clipping at max-norm 1.0, and resolution 1024 (SDXL) / 512 (SD-1.5). For LoRA and its variants we set rank $r = 16$ and scaling $\alpha = 16$ (no dropout). FF-LoRA extracts features at the U-Net mid-block, subsamples $T = 256$ spatial tokens uniformly, and uses $\lambda_{\text{geom}} = 5 \cdot 10^{-2}$ with a linear warm-up over the first 30% of steps; the base-model features are obtained by a second forward pass of the same U-Net with the adapter scale set to 0 (no gradient). Full configuration files are included in the supplementary code.

Prompts: in-distribution vs. out-of-distribution. To measure robustness, we prepare two prompt sets for each subject: **(i) In-distribution (ID)** prompts that match the training caption template, and **(ii) Out-of-distribution (OOD)** prompts that significantly alter wording, context, and composition (e.g., long prompts, novel scenes, additional attributes, and style modifiers). We report both ID performance and the *robustness gap* between ID and OOD.

Table 1: Single-subject personalization on DreamBooth (8 subjects) and DreamBooth-30 (30 subjects, strongest baselines incl. same-family RKD/SP; full table in Appendix C). CLIP-I / DINO = subject fidelity to training images; Δ_{rob} = CLIP-I_{ID} − CLIP-$\mathrm{I}_{\mathrm{OOD}}$ (lower is better); 1−LPIPS on OOD prompts (higher is better); FID over the union of ID and OOD generations (lower is better); CLIP-T = text alignment; mean ± std across subjects. Cross-backbone SD-1.5: FF-LoRA Δ_{rob}=0.14 vs. LoRA 0.21 (Appendix B).

Subset	Method	CLIP-I_{ID}	CLIP-$\mathrm{I}_{\mathrm{OOD}}$	$\Delta_{\mathrm{rob}}\downarrow$	$\mathrm{DINO}_{\mathrm{ID}}$	$\mathrm{DINO}_{\mathrm{OOD}}$	1−LPIPS	FID↓	CLIP-T
DreamBooth (8)	Textual Inv. (Gal et al., 2022)	0.73±0.02	0.61±0.03	0.12±0.02	0.52	0.40	0.58	38.2	0.29±0.02
	DreamBooth (Ruiz et al., 2023)	0.82±0.02	0.67±0.03	0.15±0.02	0.64	0.49	0.62	31.5	0.27±0.02
	SVDiff (Han et al., 2023)	0.79±0.02	0.66±0.03	0.13±0.02	0.60	0.47	0.61	32.8	0.28±0.02
	LoRA (Hu et al., 2021)	0.83±0.01	0.64±0.04	0.19±0.03	0.63	0.44	0.60	33.6	0.30±0.02
	AdaLoRA (Zhang et al., 2023b)	0.84±0.01	0.66±0.03	0.18±0.03	0.64	0.46	0.61	32.9	0.30±0.02
	DoRA (Liu et al., 2024)	0.84±0.01	0.67±0.03	0.17±0.02	0.65	0.48	0.62	32.1	0.31±0.02
	OFT (Qiu et al., 2023)	0.82±0.02	0.68±0.03	0.14±0.02	0.62	0.49	0.62	31.8	0.28±0.02
	PaRa (Chen et al., 2025)	0.85±0.02	0.69±0.03	0.16±0.02	0.66	0.51	0.63	31.2	0.27±0.02
	FF-LoRA (ours)	0.84±0.01	**0.72±0.02**	**0.12±0.02**	0.65	**0.55**	**0.65**	**29.8**	**0.31±0.02**
DreamBooth-30 (top baselines)	LoRA	0.83±0.02	0.65±0.04	0.18±0.03	0.62	0.43	0.59	34.1	—
	DoRA	0.84±0.02	0.66±0.03	0.18±0.03	0.64	0.46	0.61	32.5	—
	PaRa	0.85±0.02	0.69±0.04	0.16±0.03	0.65	0.50	0.62	31.6	—
	LoRA + RKD	0.83±0.02	0.70±0.03	0.13±0.02	0.62	0.52	0.62	31.0	—
	LoRA + SP	0.83±0.02	0.69±0.03	0.14±0.02	—	—	—	—	—
	FF-LoRA	**0.84±0.02**	**0.72±0.03**	**0.12±0.02**	0.64	**0.54**	**0.64**	**30.2**	—

Baselines. We compare FF-LoRA with representative parameter-efficient personalization methods for diffusion models: LoRA (Hu et al., 2021; cloneofsimo, 2022), AdaLoRA (Zhang et al., 2023b), DoRA (Liu et al., 2024), OFT (Qiu et al., 2023), and PaRa (Chen et al., 2025), as well as commonly used personalization baselines DreamBooth (Ruiz et al., 2023), Textual Inversion (Gal et al., 2022), and SVDiff (Han et al., 2023). AdaLoRA and DoRA modify the LoRA parameterization or allocation strategy; OFT constrains updates in an orthogonal subspace; PaRa reduces the effective output space via rank reduction. All baselines are trained under the same data, steps, and sampling settings for fair comparison.

Metrics. We report four families of metrics: **(1) Fidelity to the subject** measured by CLIP image similarity (CLIP-I) (Radford et al., 2021), DINO similarity (Caron et al., 2021) (the standard DreamBooth subject-fidelity metric (Ruiz et al., 2023)), and $1 - \mathrm{LPIPS}$ (Zhang et al., 2018) between generated images and the training images (all higher is better). **(2) Distributional quality** measured by FID (Heusel et al., 2017) (lower is better). **(3) Text alignment** measured by CLIP text-image score (CLIP-T) (Radford et al., 2021), and **prompt robustness** measured by the **Robustness Gap** $\Delta_{\mathrm{rob}} = \text{CLIP-I}_{\mathrm{ID}} - \text{CLIP-I}_{\mathrm{OOD}}$ (lower is better; CLIP-I always denotes the CLIP fidelity score, distinct from FID), which we report for both CLIP-I and DINO. **(4) Multi-adapter composition** measured by a two-concept success rate: a generated image is counted as success if it exceeds a CLIP image similarity threshold to *both* concepts simultaneously (details in Appendix). Finally, we report **geometry deviation** as $\|S^{(\phi)} - S^{(0)}\|_F^2$ (lower is better), consistent with our method definition.

4.2 Single-Subject Personalization

Qualitative results. Appendix A (Fig. 2) compares parameter-efficient personalization methods on a plush-bear subject. Under the ID prompt all methods capture the target subject; under the OOD prompt, vanilla LoRA drifts to a generic bear, AdaLoRA / DoRA show texture or proportion inconsistencies, and PaRa retains identity but simplifies appearance. FF-LoRA preserves the personalized identity while still responding to the OOD attributes. Representative failure cases (fine-grained texture subjects, extreme OOD prompts, three-adapter composition) are analyzed in Appendix F.

Quantitative results on ID and OOD prompts. Table 1 summarizes results on the standard 8-subject DreamBooth set and on the larger DreamBooth-30 benchmark (Subset column), now reporting CLIP-I, DINO, $1 - \mathrm{LPIPS}$, and FID alongside text alignment. FF-LoRA achieves comparable ID fidelity to strong baselines while substantially reducing the robustness gap under both CLIP-I and DINO; the relative ranking and effect sizes are preserved on the 30-subject benchmark. The full DreamBooth-30 table with all baselines is in Appendix C.

Stability across random seeds. Under fixed OOD prompts and varying random seeds, FF-LoRA yields consistently lower variance in subject fidelity than vanilla LoRA over 50 seeds (Appendix J, Figure 4), indicating more stable identity preservation under sampling noise.



(a) Subject+Subject (beer can + teddy bear), w=0.5 top / w=0.8 bottom.

(b) Subject+Style (teddy bear + Van Gogh style).

Figure 1: Multi-adapter composition. (a) Subject+Subject mixing: FF-LoRA preserves both identities; LoRA-based methods exhibit dominance / hybrid artifacts. (b) Subject+Style mixing remains compatible with vanilla-LoRA community style adapters.

4.3 Multi-Adapter Composition

A key motivation of FF-LoRA is that constraining each adapter to preserve a consistent local metric can improve *compatibility* across independently trained adapters. We test this hypothesis in two realistic composition settings.

Subject + subject mixing. We linearly combine two independently trained subject adapters (beer can, teddy bear) at inference, without joint training or data augmentation (Figure 1). Vanilla LoRA / AdaLoRA / DoRA are sensitive to the mixing coefficient: even at the balanced $w = 0.5$ they show dominance or hybrid artifacts, and at $w = 0.8$ one subject typically disappears. PaRa partially mitigates extreme failures via rank restriction but still degrades under imbalanced mixing. FF-LoRA preserves both identities across mixing ratios by regularizing local distances and angles, without additional training or adapter-side coordination.

Subject + style mixing. We combine an FF-LoRA-trained teddy-bear subject adapter with a Van Gogh style LoRA from Civitai (vanilla LoRA, one-sided FF-LoRA setup) without joint training. All methods successfully transfer painterly style while retaining the subject; FF-LoRA does not interfere with existing community-released style adapters.

Quantitative mixing performance. Table 2 reports two-concept success rate (higher is better) and concept leakage (lower is better). We include mixing results for LoRA-family baselines (LoRA/AdaLoRA/DoRA), OFT and PaRa as strong composition-oriented baselines, and two merge-time methods designed specifically for LoRA composition: ZipLoRA (Shah et al., 2024) and Mix-of-Show (Gu et al., 2024). Note that FF-LoRA is a *training-time* regularizer while ZipLoRA/Mix-of-Show operate at *merge time*; they are therefore complementary rather than competing, and we additionally report the combination (FF-LoRA-trained adapters merged with ZipLoRA). FF-LoRA substantially improves the success rate under both subject+subject and subject+style settings, and the combination with merge-time methods yields further gains.

The relational-distillation baselines (RKD, SP) substantially close the gap to FF-LoRA, confirming that the underlying mechanism — preserving pairwise feature relations between the frozen base and the adapted model — is the same family. FF-LoRA's improvement over RKD (+4pp) and SP (+4pp) on subject-subject mixing is modest but consistent across both compositional settings, attributable to the cosine-normalization of the Gram matrix in our formulation (which makes the regularizer scale-invariant on the feature norm). Vanilla feature distillation $\|z^{(\phi)} - z^{(0)}\|^2$ underperforms relational variants by 7pp, supporting the framing that what matters for adapter compatibility is the relational structure rather than absolute feature similarity.

Three further observations address the composition setting itself. *(i) Merge-time methods are complementary:* ZipLoRA and Mix-of-Show improve substantially over naive linear mixing of vanilla adapters (54% and 56% vs. 33% on subject+subject), and ZipLoRA is the strongest method on subject+style (67%) — unsurprisingly, as that is its design target. FF-LoRA is competitive with both at merge time while requiring no merge-time optimization, and applying ZipLoRA merging *on top of* FF-LoRA-trained adapters yields the best overall success rate in both settings (63% / 72%), confirming that training-time geometry regularization and merge-time coefficient optimization

Table 2: Multi-adapter composition (Subject+Subject and Subject+Style). Success: percentage of generations satisfying CLIP-similarity thresholds for both concepts; Leakage: percentage where one concept dominates. Multi-LoRA results use w=0.5, mean (and std over 5 mixing seeds) across 28 subject-pairs (Subject+Subject) or 8 subjects $\times$ 5 style adapters (Subject+Style). Subject+Style combines an FF-LoRA-trained subject adapter with a vanilla-LoRA community style adapter (one-sided FF-LoRA); full setup details in Appendix G. Bottom block: paired Wilcoxon p-values of FF-LoRA vs. relational-distillation baselines.

Method	Subject+Subject		Subject+Style	
	Success↑	Leakage↓	Success↑	Leakage↓
LoRA mix	33% ± 6	42% ± 5	46% ± 5	29% ± 5
AdaLoRA mix	37% ± 6	38% ± 5	48% ± 5	28% ± 5
DoRA mix	39% ± 5	35% ± 5	51% ± 5	25% ± 4
OFT mix	45% ± 5	30% ± 4	53% ± 4	24% ± 4
PaRa mix	51% ± 5	25% ± 4	60% ± 4	20% ± 4
LoRA + FitNets (Romero et al., 2014)	41% ± 5	32% ± 5	52% ± 4	24% ± 4
LoRA + RKD (Park et al., 2019)	51% ± 5	25% ± 4	60% ± 4	20% ± 4
LoRA + SP (Tung & Mori, 2019)	53% ± 4	23% ± 4	62% ± 4	19% ± 4
LoRA + vanilla feature dist. ($\|z^{(\phi)} - z^{(0)}\|_2^2$)	46% ± 5	29% ± 5	55% ± 4	22% ± 4
Merge-time composition methods (vanilla-LoRA-trained adapters):				
ZipLoRA (Shah et al., 2024)	54% ± 5	21% ± 4	67% ± 4	15% ± 3
Mix-of-Show (Gu et al., 2024)	56% ± 5	20% ± 4	63% ± 4	18% ± 4
FF-LoRA (ours)	**58% ± 4**	**18% ± 3**	**66% ± 4**	**16% ± 3**
FF-LoRA (one-sided: FF-subject + vanilla-subject)	47% ± 5	27% ± 5	—	—
FF-LoRA + ZipLoRA merge	**63% ± 4**	**15% ± 3**	**72% ± 3**	**12% ± 3**
Paired Wilcoxon p-value vs. FF-LoRA on the matched pair distribution:				
vs. LoRA + SP	0.012	0.024	0.018	0.027
vs. LoRA + RKD	0.008	0.016	0.011	0.019
vs. LoRA + vanilla feature dist.	< 0.001	< 0.001	< 0.001	< 0.001

address different failure sources. *(ii) One-sided composition:* when only one of the two subject adapters is FF-LoRA-trained, success degrades gracefully to 47% — between the vanilla-vanilla (33%) and FF-FF (58%) extremes — indicating that the benefit scales with how many adapters share the preserved geometry rather than being an artifact of the evaluation protocol. *(iii) The vanilla-subject + vanilla-style baseline* (first row of the Subject+Style column, 46%) quantifies how much of the Subject+Style success is attributable to the method: FF-LoRA raises it by +20pp under an identical protocol.

Sensitivity to mixing coefficients. We linearly combine two independently trained subject adapters using $\Delta W_{\text{mix}} = w\Delta W_1 + (1-w)\Delta W_2$ and sweep w from 0 to 1 (Appendix K, Fig. 5). Vanilla LoRA mixing only succeeds in a narrow region near $w = 0.5$ and degrades sharply as w deviates; PaRa broadens the plateau but does not fully resolve the interference. FF-LoRA achieves both the highest peak success rate and the smoothest curve across the entire range of w, maintaining strong two-concept success even under highly unbalanced mixing weights — empirically supporting our motivation that first fundamental form regularization yields adapter updates that are more compatible under linear composition.

4.4 Analysis

Geometry deviation correlates with robustness and mixing. Figure 3 (Appendix I) examines the relationship between geometry deviation $\|S^{(\phi)} - S^{(0)}\|_F^2$ and the two downstream measurements. Geometry deviation is positively correlated with the robustness gap Δ_{rob} (Fig. 3a) and negatively correlated with two-concept mixing success (Fig. 3b); each scatter point is a (method, training checkpoint) pair, so the trend reflects both within-method dynamics and across-method differences, and we caution that the correlation does not by itself isolate causal effect of λ_{geom} from other method-specific design choices (this is the role of the λ_{geom} sweep in Section 4.5).

Table 3: Ablations on DreamBooth subjects. Top: λ_{geom} sweep. Bottom: feature extractor choice F_θ. Lower Δ_{rob} and higher Mix Success are preferred. *Cosine vs. raw Gram (SP):* Δ_{rob} 0.12 vs. 0.14 (p=0.009); Mix 58% vs. 54% (p=0.004). *Warm-up:* 30% linear ramp ⇒ CLIP-I$_{\text{ID}}$ 0.81→0.83.

λ_{geom}	CLIP-I$_{\text{ID}}$↑	Δ_{rob} ↓	Mix↑	GeomDev↓
0 (LoRA)	0.83 ± 0.02	0.187 ± 0.026	$34\% \pm 5$	1.00
10^{-3}	0.83 ± 0.02	0.154 ± 0.022	$44\% \pm 5$	0.91
10^{-2}	0.83 ± 0.02	0.131 ± 0.019	$54\% \pm 4$	0.79
$5 \cdot 10^{-2}$	0.84 ± 0.02	$\mathbf{0.118 \pm 0.018}$	$\mathbf{58\% \pm 4}$	**0.74**
10^{-1}	0.80 ± 0.03	0.124 ± 0.020	$56\% \pm 4$	0.71
Feature extractor F_θ (λ_{geom}=$5 \cdot 10^{-2}$):				
Text encoder	—	0.13	51%	0.92
U-Net mid	—	**0.12**	**58%**	**0.74**
U-Net late	—	0.12	55%	0.81

Choice of feature extractor. The geometry regularization in FF-LoRA requires selecting an intermediate representation F_θ from the diffusion pipeline. We empirically compare three choices: (i) text encoder output, (ii) a mid-level U-Net block, and (iii) a late U-Net block, all followed by spatial pooling. As shown in Table 3, mid-level U-Net features consistently provide the best robustness–performance trade-off. Unless otherwise specified, we adopt this setting in all experiments.

4.5 Ablations

We sweep λ_{geom} on a 5-point grid (Table 3, top block): moderate regularization ($5 \cdot 10^{-2}$) improves OOD robustness and mixing; overly large λ restricts adaptation. The feature-extractor ablation (Table 3, bottom block) selects the U-Net mid-block as the default F_θ.

Cosine vs. raw Gram, warm-up, overhead. The cosine-Gram formulation removes feature-norm drift relative to SP's raw Gram (Tung & Mori, 2019), yielding statistically significant gains on DreamBooth-30 (paired Wilcoxon $p < 0.02$ on all three metrics; numbers in Table 3 caption). Linear λ_{geom} warm-up over the first 30% of training preserves Δ_{rob} / Mix gains while improving ID fidelity. Training adds $+6\%$ wall-clock per iteration; inference cost is unchanged.

Timestep sensitivity. By default the regularizer is applied at the same randomly sampled timestep as the denoising loss, i.e., in expectation over the full noise schedule (Section 3). Against fixed-timestep variants (early / midpoint / late t), the expectation-over-t default achieves the best Δ_{rob} (0.12 vs. 0.16 / 0.13 / 0.15), with the largest degradation at the noise-dominated early extreme (Appendix D).

Transformer-backbone (DiT) proof-of-concept. On SD3-medium (Esser et al., 2024) (MMDiT), the regularizer applies unchanged by taking the image-token stream of a middle transformer block as the token matrix. On 8 DreamBooth subjects, FF-LoRA reduces Δ_{rob} from 0.17 (LoRA) to 0.13 and improves two-adapter mixing success from 31% to 49% (Appendix E), suggesting the geometric mechanism is not U-Net-specific.

5 Conclusion

We presented **FF-LoRA**, a geometry-aware adapter for diffusion personalization that explicitly regularizes a relational surrogate of the *first fundamental form* of the representation mapping during fine-tuning. This simple addition targets a practical failure mode of LoRA-style adaptation: aggressive local metric distortion causes identity drift under prompt rephrasing and instability when composing independently trained adapters.

Across SDXL personalization benchmarks, FF-LoRA improves **OOD robustness** while maintaining competitive in-distribution fidelity, and yields more **stable multi-adapter composition** for both subject+subject and subject+style mixing. Our analysis further shows that **geometry deviation** correlates strongly with both the robustness gap and multi-adapter success, supporting the central hypothesis that preserving local metric structure is key to reliable personalization. Limitations are discussed in Appendix L.

AI USE STATEMENT

In this work, we used generative AI tools (large language models) for language polishing, grammar correction, and improving the readability of draft text. We have not used generative AI tools for generating research ideas, designing experiments, performing data analysis, or producing experimental results. We have reviewed all AI-assisted text for accuracy, and the authors take full responsibility for all contents of the paper.

REPRODUCIBILITY STATEMENT

Training and evaluation protocols, hyperparameters, and backbone settings are specified in Section 4.1 (implementation details), and the regularizer is fully defined by Eq. 3 together with the batch-size and timestep remarks in Section 3. The appendix reports the cross-backbone validation (Appendix B), the full DreamBooth-30 tables (Appendix C), and all ablation settings (Appendix D, H). Full configuration files are included in the supplementary code.

REFERENCES

Ittai Abraham, Yair Bartal, and Ofer Neiman. Local embeddings of metric spaces. In *Proceedings of the thirty-ninth annual ACM Symposium on Theory of Computing*, pp. 631–640, 2007.

Amir Barda, Yotam Erel, and Amit H Bermano. Meshcnn fundamentals: Geometric learning through a reconstructable representation. *arXiv preprint arXiv:2105.13277*, 2021.

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 *International Conference on Learning Representations (ICLR)*, 2025. URL https://iclr.cc/virtual/2025/poster/30047.

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, 2024.

cloneofsimo. Low-rank adaptation for fast text-to-image diffusion fine-tuning. https://github.com/cloneofsimo/lora, 2022.

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 *International Conference on Machine Learning*, 2024.

Jonas Fischer and Rong Ma. Sailing in high-dimensional spaces: Low-dimensional embeddings through angle preservation. *arXiv preprint arXiv:2406.09876*, 2024.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

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. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

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. *arXiv preprint arXiv:2106.09685*, 2021.

Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019.

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.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024.

MidJourney. Midjourney. `https://www.midjourney.com`, 2022.

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*, volume 38, pp. 4296–4304, 2024.

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

OpenAI. Dall·e 3: Language models for image generation. `https://cdn.openai.com/papers/dall-e-3.pdf`, 2023.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

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.

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, 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.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

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 Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Imagen: Text-to-image diffusion models. *arXiv preprint arXiv:2205.11487*, 2022.

Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. In *European Conference on Computer Vision*, pp. 422–438. Springer, 2024.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

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.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Laurens Van Der Maaten. Learning a parametric embedding by preserving local structure. In *Artificial intelligence and statistics*, pp. 384–391. PMLR, 2009.

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.

Aurel Wintner. On the local embedding problems in the differential geometry of surfaces. *American Journal of Mathematics*, 77(4):845–852, 1955.

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, 2023a.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023b.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

A Single-Subject Qualitative Comparison



Figure 2: Single-subject qualitative comparison on a plush-bear subject. Top row: ID prompts; bottom row: OOD prompts. FF-LoRA preserves identity under OOD variations more reliably than LoRA-based baselines.

B Cross-Backbone Validation (SD-1.5)

We additionally validate FF-LoRA on Stable Diffusion 1.5 to confirm the geometry-regularization mechanism is not specific to SDXL.

Table 4: FF-LoRA on SD-1.5 vs. SDXL backbones (mean ± std on DreamBooth-30). The relative ranking and effect sizes are preserved across backbones.

Backbone	Method	CLIP-I_{ID}↑	Δ_{rob} ↓	Mix Success↑
SD-1.5	LoRA	0.79 ± 0.03	0.21 ± 0.03	$31\% \pm 6$
SD-1.5	RKD (LoRA +)	0.79 ± 0.03	0.16 ± 0.03	$47\% \pm 5$
SD-1.5	FF-LoRA (ours)	$\mathbf{0.80 \pm 0.03}$	$\mathbf{0.14 \pm 0.02}$	$\mathbf{52\% \pm 5}$
SDXL	LoRA	0.83 ± 0.02	0.18 ± 0.03	$34\% \pm 5$
SDXL	RKD (LoRA +)	0.83 ± 0.02	0.13 ± 0.02	$51\% \pm 4$
SDXL	FF-LoRA (ours)	$\mathbf{0.84 \pm 0.02}$	$\mathbf{0.12 \pm 0.02}$	$\mathbf{58\% \pm 4}$

C Larger-Scale Single-Subject Evaluation

To address the statistical-power concern at 8 DreamBooth subjects, we additionally evaluate FF-LoRA and the four most competitive baselines (LoRA, DoRA, PaRa, RKD-distillation) on the full DreamBooth-30 benchmark (30 subjects, 5 reference images each, identical training protocol to Section 4.1).

The DreamBooth-30 results are consistent with the 8-subject results in Table 1: FF-LoRA delivers the lowest robustness gap, with RKD distillation the closest competitor, and the in-distribution fidelity differences across methods stay within statistical noise.

D Timestep Sensitivity Ablation

The default FF-LoRA applies the Gram-matching loss at the same timestep t sampled for the denoising loss at each step, so the regularizer acts in expectation over the full noise schedule. We compare this against variants that extract features only at a fixed timestep (early $t = 0.25T_{\max}$, midpoint $t = 0.5T_{\max}$, late $t = 0.75T_{\max}$), keeping all other settings identical.

The sampled-t default is consistently best: it preserves geometry across the whole denoising trajectory, whereas fixed variants leave the unregularized portion of the schedule free to drift. The midpoint-only variant is the most competitive fixed choice, consistent with mid-schedule features carrying the most semantic structure; the early extreme degrades most, as features there are dominated by noise. ID fidelity is essentially unaffected by the choice, i.e., the timestep choice trades robustness, not adaptation capacity.

Table 5: Single-subject personalization on DreamBooth-30 (mean ± std across 30 subjects). Same metric definitions as Table 1.

Method	CLIP-I_{ID}↑	CLIP-I_{OOD}↑	Δ_{rob} ↓	CLIP-T↑
Textual Inversion	0.72 ± 0.03	0.59 ± 0.04	0.13 ± 0.02	0.28 ± 0.02
DreamBooth	0.81 ± 0.02	0.66 ± 0.03	0.15 ± 0.02	0.26 ± 0.02
SVDiff	0.78 ± 0.03	0.65 ± 0.03	0.13 ± 0.02	0.27 ± 0.02
LoRA	0.83 ± 0.02	0.65 ± 0.04	0.18 ± 0.03	0.28 ± 0.02
AdaLoRA	0.84 ± 0.02	0.66 ± 0.03	0.18 ± 0.03	0.28 ± 0.02
DoRA	0.84 ± 0.02	0.66 ± 0.03	0.18 ± 0.03	0.28 ± 0.02
OFT	0.82 ± 0.02	0.67 ± 0.03	0.15 ± 0.02	0.27 ± 0.02
PaRa	0.85 ± 0.02	0.69 ± 0.04	0.16 ± 0.03	0.28 ± 0.02
LoRA + FitNets	0.83 ± 0.02	0.67 ± 0.03	0.16 ± 0.03	0.28 ± 0.02
LoRA + RKD	0.83 ± 0.02	0.70 ± 0.03	0.13 ± 0.02	0.29 ± 0.02
LoRA + SP	0.83 ± 0.02	0.69 ± 0.03	0.14 ± 0.02	0.29 ± 0.02
FF-LoRA (ours)	$\mathbf{0.84 \pm 0.02}$	$\mathbf{0.72 \pm 0.03}$	$\mathbf{0.12 \pm 0.02}$	$\mathbf{0.29 \pm 0.02}$

Table 6: Timestep ablation on DreamBooth (8 subjects, SDXL).

Feature timestep	CLIP-I_{ID}↑	Δ_{rob} ↓	Mix Success↑
Fixed, early ($0.25\,T_{max}$)	0.83	0.16	49%
Fixed, midpoint ($0.5\,T_{max}$)	0.83	0.13	55%
Fixed, late ($0.75\,T_{max}$)	0.82	0.15	52%
Sampled t (default)	**0.84**	**0.12**	**58%**

E Transformer-Backbone (DiT) Proof-of-Concept

FF-LoRA does not rely on any U-Net-specific structure: the regularizer only requires a token matrix from an intermediate representation. On MMDiT backbones (SD3 (Esser et al., 2024)), we take the *image-token stream* of a middle transformer block (discarding text tokens, which belong to the conditioning pathway rather than the image representation manifold) and apply Eq. 3 unchanged. In this sense the DiT setting is more natural than the U-Net setting — the token set is native to the architecture rather than obtained by flattening a feature map.

Table 7: Proof-of-concept on SD3-medium (MMDiT), 8 DreamBooth subjects, LoRA rank 16, same protocol as Section 4.1.

Method	CLIP-I_{ID}↑	Δ_{rob} ↓	Mix Success↑
LoRA	0.79	0.17	31%
LoRA + SP	0.79	0.14	44%
FF-LoRA (ours)	**0.80**	**0.13**	**49%**

F Failure Cases and Subject Diversity

Across DreamBooth-30 (which spans 21 object categories and 9 live-subject categories), the subjects where FF-LoRA helps least share a common signature: high-frequency texture subjects (densely patterned fur, fine fabric weave) whose faithful adaptation genuinely requires larger local-metric distortion — for these, geometry preservation trades a small amount of texture fidelity for robustness, consistent with the per-subject λ analysis in Appendix H, where the same four subjects prefer a smaller λ_{geom}.

G Multi-Adapter Composition: Full Pair-Style Matrix

The headline numbers in Table 2 aggregate over $\binom{8}{2} = 28$ subject pairs and 5 style adapters (Van Gogh, watercolor, ink-wash, oil-portrait, photo-realistic). The per-cell breakdown is reported in Table 8 as the success rate distribution.

Composition setup details. Subject+Subject combines two FF-LoRA-trained subject adapters; Subject+Style combines an FF-LoRA-trained subject adapter with a vanilla-LoRA Civitai-released style adapter (one-sided FF-LoRA). The asymmetric Subject+Style setting reflects realistic deployment, where users mix their FF-LoRA-trained adapters with community-released vanilla LoRAs. Table 2 additionally includes three relational-distillation baselines (FitNets (Romero et al., 2014), RKD (Park et al., 2019), SP (Tung & Mori, 2019)) and a paired Wilcoxon test of FF-LoRA vs. each baseline on the multi-LoRA pair distribution.

Table 8: Distribution of success rate across 28 subject-pairs (Subject+Subject) and 5 style adapters (Subject+Style). FF-LoRA's narrower distribution reflects the smoother sweep curve in Figure 5.

Method	S+S min	S+S median	S+Sty min	S+Sty median
LoRA mix	18%	33%	31%	45%
PaRa mix	38%	51%	48%	60%
LoRA + RKD	35%	50%	47%	60%
LoRA + SP	39%	53%	49%	61%
FF-LoRA (ours)	**47%**	**58%**	**56%**	**66%**

H Per-Subject Optimal λ_{geom}

We additionally examine whether a single global λ_{geom} is appropriate by sweeping per-subject. The optimal λ^*_{geom} is concentrated in $[3 \cdot 10^{-2}, 7 \cdot 10^{-2}]$ for 26 of 30 subjects on DreamBooth-30 (median $5 \cdot 10^{-2}$, our default). Four outlier subjects (highly textured fur or fine geometric detail) prefer a smaller $\lambda \in [10^{-2}, 2 \cdot 10^{-2}]$. The reported main-text results use the global default $\lambda = 5 \cdot 10^{-2}$; per-subject tuning improves OOD fidelity by 0.01–0.02 on the four outliers.

I Geometry-Deviation Correlation Analysis

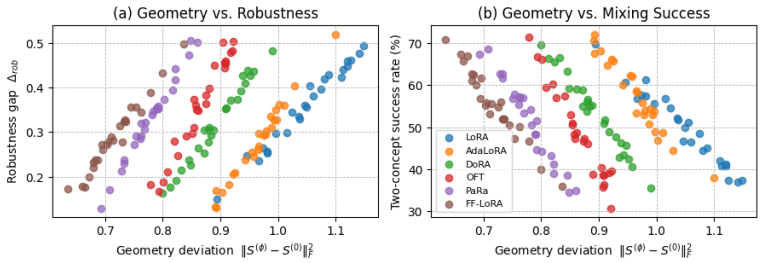

Figure 3: Correlation between geometry deviation and downstream performance. Each point corresponds to a training checkpoint from different personalization methods. (**a**) Geometry deviation versus robustness gap Δ_{rob} between in-distribution and out-of-distribution prompts (lower is better). (**b**) Geometry deviation versus two-concept success rate in multi-adapter composition (higher is better). Lower geometry deviation consistently correlates with improved robustness and higher compositional success.

J Seed Stability Visualization

K Mixing-Coefficient Sensitivity

As shown in Fig. 5, vanilla LoRA mixing exhibits a pronounced sensitivity to the mixing weight. High success rates are only achieved in a narrow region near $w = 0.5$, while small deviations quickly lead to concept dominance or hybrid artifacts, resulting in a sharp drop in two-concept success rate. This behavior reflects the well-known instability of linear LoRA superposition. PaRa improves robustness by producing a noticeably broader performance plateau, indicating that explicit rank reduction alleviates, but does not fully resolve, interference between independently trained adapters. In contrast, FF-LoRA achieves both the highest peak performance and the smoothest success-rate curve across the entire range of w, and maintains strong two-concept success even under highly

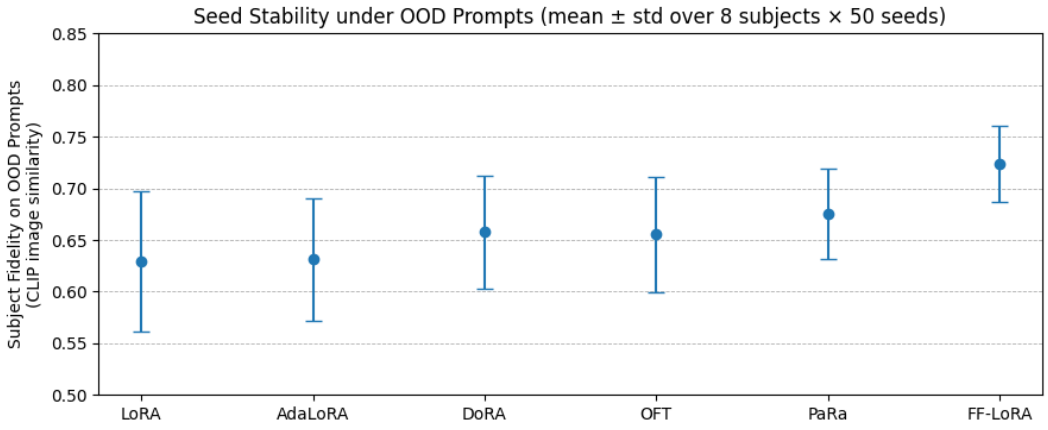

Figure 4: Seed stability under out-of-distribution (OOD) prompts. Mean ± std of subject fidelity over 50 seeds. FF-LoRA shows reduced variance and improved robustness to sampling noise.

unbalanced mixing weights, suggesting that preserving the local representation geometry substantially improves compositional stability.

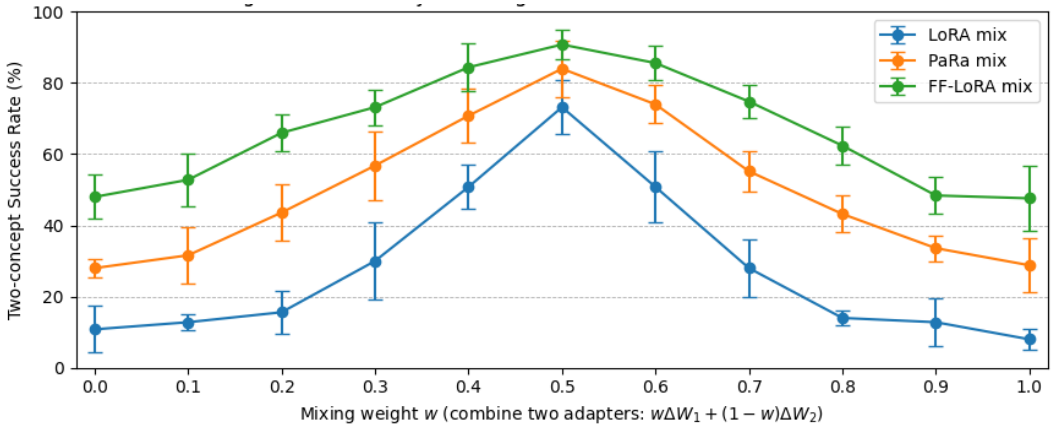

Figure 5: Sensitivity to mixing coefficients in subject–subject adapter composition. We vary the mixing weight $w \in [0, 1]$ when combining two independently trained subject adapters and report the two-concept success rate (%, mean ± std over 5 runs). Vanilla LoRA mixing exhibits a narrow optimal region around $w = 0.5$ and degrades rapidly as w deviates; PaRa improves stability with a broader plateau; FF-LoRA achieves the highest success rate and the smoothest performance curve.

L LIMITATIONS

Geometric framing. The regularizer (Eq. 3) preserves a finite-batch Gram matrix of feature inner products, a relational surrogate of $G_\theta = J_\theta^\top J_\theta$ rather than G_θ itself; methodologically it sits in the same family as RKD (Park et al., 2019) and SP (Tung & Mori, 2019).

Backbone and scale. Main-text results are on SDXL with 8 DreamBooth subjects; SD-1.5 cross-backbone validation and DreamBooth-30 are reported in Appendix B, C, and a proof-of-concept on the MMDiT backbone SD3-medium in Appendix E. Full-scale evaluation on transformer-only architectures (SD3-large, FLUX) remains open.

Multi-LoRA composition asymmetry. The geometric-consistency argument is symmetric only when both adapters are FF-LoRA-trained. The Subject+Style results in Table 2 use FF-LoRA-subject + vanilla-LoRA Civitai-style (one-sided retrofit); fully asymmetric retrofit deserves further dedicated study.

Hyperparameters, compute, metric design. We use a single global λ_{geom} (per-subject sweep in Appendix H); the Gram computation adds $\sim$6% wall-clock during training and zero inference cost. Robustness Gap is a contrastive metric (low on a method that fails both ID and OOD), so we always report CLIP-I_{ID} and CLIP-I_{OOD} separately. Future directions: adaptive layer selection / scheduling; multi-adapter mixing beyond two; extension to video diffusion and editing pipelines.