EVALUATION OF PARAMETER-EFFICIENT FINE-TUNING METHODS FOR PERSONALIZED TEXT-TO-IMAGE GENERATION

A Thesis Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCES

in Computer Engineering

by
Kutay ÖZBAY

July 2026
İZMİR

We approve the thesis of **Kutay ÖZBAY**

Examining Committee Members:

Prof. Dr. Yalın BAŞTANLAR
Department of Computer Engineering, İzmir Institute of Technology

Prof. Dr. Alptekin TEMIZEL
Graduate School of Informatics, Middle East Technical University

Assist. Prof. Dr. M. Zübeyir ÜNLÜ
Department of Electrical and Electronics Engineering, İzmir Institute of Technology

8 July 2026

Prof. Dr. Yalın BAŞTANLAR
Department of Computer Engineering
İzmir Institute of Technology

Prof. Dr. Onur DEMIRÖRS
Head of the Department of
Computer Engineering

Prof. Dr. Mehtap EANES
Dean of the Graduate School of
Engineering and Sciences

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my thesis advisor, Prof. Dr. Yalın Baştanlar, for his guidance and support throughout this research. His feedback shaped the direction of this work in ways I did not always anticipate but consistently appreciated.

I am thankful to the members of my examining committee, Prof. Dr. Alptekin Temizel and Assist. Prof. Dr. M. Zübeyir Ünlü, for serving on the jury and for the time and care they devoted to evaluating this thesis.

I am grateful to Anıl Erdem Ulusoy and the HubX AILab team for their collaboration on the LoRA+ID work that began this thesis. I owe particular thanks to Yunus Emre for his guidance and support throughout the PEFT evaluation study, and for providing the H100 infrastructure on which its experiments were run.

I also want to thank my colleagues at İzmir Institute of Technology for the conversations, the shared frustrations, and the occasional encouragement at the right moments.

Finally, to my family: thank you for your patience. This took longer than any of us expected.

ABSTRACT

EVALUATION OF PARAMETER-EFFICIENT FINE-TUNING METHODS FOR PERSONALIZED TEXT-TO-IMAGE GENERATION

Parameter-Efficient Fine-Tuning (PEFT), and Low-Rank Adaptation (LoRA) in particular, has become the standard way to personalize large-scale text-to-image models such as Stable Diffusion XL. It is efficient and widely used, but it carries a persistent weakness: when a model is fine-tuned on a specific person using only text prompts, the adapter often loses facial identity as soon as the generation prompt moves to a new style, context, or composition. The model learns what the training images looked like, not who the person is. This thesis addresses that problem and, in doing so, uncovers a second one in how personalization is evaluated.

The first contribution is **LoRA+ID**, a training framework that integrates ArcFace facial identity embeddings directly into the LoRA training loop, so the adapter learns identity-discriminative features rather than visual statistics alone. The second is an **extended benchmark dataset**, DreamBooth-Ext, covering object and human-face subjects with structured context prompts that test generalization rather than memorization. The third is the **Concept Fidelity Score (CFS)**, a metric that measures how well an adapter stays contained to its trigger through two probes – far-concept drift and compositional capacity – and so detects the concept drift that standard metrics such as FID, CLIP, and DINOv2 similarity cannot see. The fourth is **APEX** (Aggregate Personalization Evaluation indeX), a composite index that folds fourteen complementary metrics from six families into a single geometric-mean score and enables one of the first holistic comparisons of personalization methods spanning PEFT fine-tuning, inference-time adapters, and image-to-image generation. Together these contributions form a single arc: a better training method, the data to test it on, a metric to measure what matters, and a framework to compare everything fairly.

ÖZET

KİŞİSELLEŞTİRİLMİŞ METİNDEN GÖRÜNTÜYE ÜRETME İÇİN PARAMETRE VERİMLİ İNCE AYAR YÖNTEMLERİNİN DEĞERLENDİRİLMESİ

Parametre Verimli İnce Ayar (PEFT) ve özellikle Düşük Sıralı Uyarlama (LoRA), Stable Diffusion XL gibi büyük ölçekli metinden görüntüye üretim modellerini kişiselleştirmenin standart yöntemi haline gelmiştir. Verimli ve yaygın olarak kullanılmaktadır; ancak kalıcı bir zayıflığı vardır: bir model yalnızca metin komutları kullanılarak belirli bir kişi üzerinde ince ayar yapıldığında, oluşturulan adaptör üretim komutu yeni bir stil, bağlam veya kompozisyona geçer geçmez yüz kimliğini kaybedebilmektedir. Model, kişinin kim olduğunu değil, eğitim görüntülerinin nasıl göründüğünü öğrenmektedir. Bu tez söz konusu sorunu ele almakta ve bu süreçte kişiselleştirmenin değerlendirilmesine ilişkin ikinci bir sorunu da ortaya koymaktadır.

İlk katkı, ArcFace yüz kimliği gömülü vektörlerini doğrudan LoRA eğitim döngüsüne entegre eden bir eğitim çerçevesi olan **LoRA+ID**'dir. İkinci katkı, nesne ve insan yüzü öznelerini kapsayan ve ezberleme yerine genelleme yeteneğini sınayan yapılandırılmış bağlam komutları içeren **genişletilmiş bir kıyaslama veri kümesi** olan DreamBooth-Ext'tir. Üçüncü katkı, bir uyarlamanın tetikleyici belirtecine ne kadar iyi sınırlı kaldığını – uzak-kavram kayması ve bileşimsel kapasite olmak üzere iki sonda aracılığıyla – ölçen ve böylece FID, CLIP ve DINOv2 benzerliği gibi standart metriklerin göremediği kavram kaymasını tespit eden bir metrik olan **Kavram Sadakat Skoru (CFS)**'dur. Dördüncü katkı ise altı aileden gelen on dört tamamlayıcı metriği tek bir geometrik ortalama skorunda birleştiren ve bildiğimiz kadarıyla PEFT ince ayarı, çıkarım zamanı adaptörleri ve görüntüden görüntüye üretimi kapsayan ilk bütünsel karşılaştırmalardan birini mümkün kılan bileşik değerlendirme indeksi olan **APEX** (Aggregate Personalization Evaluation indeX)'tir. Bu katkılar birlikte tutarlı bir bütün oluşturmaktadır: daha iyi bir eğitim yöntemi, onu test edecek veri, önemli olanı ölçecek bir metrik ve her şeyi adil biçimde karşılaştıracak bir çerçeve.

PREFACE

This thesis grew from a practical problem. While building image personalization pipelines at HubX, it became clear that existing LoRA fine-tuning methods struggled to reliably preserve a person's identity when the generation prompt moved far from the training conditions. The first contribution of this thesis – LoRA+ID – was a direct attempt to fix that.

Trying to properly validate LoRA+ID, however, revealed a second problem: the evaluation tools available were not up to the task. Metrics like FID and CLIP similarity could not tell whether a fine-tuned model had genuinely learned a concept or had simply memorized the training images. This led to the evaluation contributions that make up the second half of the thesis.

Part of the work on LoRA+ID was published at the 2025 Innovations in Intelligent Systems and Applications Conference (ASYU 2025) (Özbay, Ulusoy, and Baştanlar 2025). The evaluation framework, dataset extension, and Concept Fidelity Score are original contributions developed at İzmir Institute of Technology.

TABLE OF CONTENTS

LIST OF FIGURES x

LIST OF TABLES xiii

LIST OF ABBREVIATIONS xvi

CHAPTER 1. Introduction 1

1.1. What LoRA Does and Where It Falls Short 1

1.2. A Second Problem, Found While Solving the First 2

1.3. Four Contributions 3

1.4. Organization 5

CHAPTER 2. Related Work 6

2.1. Text-to-Image Diffusion Models 6

2.2. Parameter-Efficient Fine-Tuning for Personalized Generation 8

2.2.1. DreamBooth 8

2.2.2. Textual Inversion and Custom Diffusion 9

2.2.3. LoRA 9

2.2.4. LoRA Variants 10

2.3. Identity-Conditioned Generation Without Fine-Tuning 12

2.4. Image-to-Image and Instruction Editing 13

2.5. Forgetting and Concept Drift 14

2.6. Evaluation of Personalized Generation 16

CHAPTER 3. The DreamBooth-Ext Dataset 18

3.1. Motivation and Design 18

3.2. Dataset Composition 19

3.2.1. Extended Object Subjects 20

3.2.2. Human Face Subset 21
3.3. Evaluation Prompt Structure 22

CHAPTER 4. LoRA+ID: Identity-Conditioned LoRA Training 23
4.1. Why Text-Only Conditioning Is Not Enough 23
4.2. Method 24
4.2.1. Face Embedding Extraction 24
4.2.2. Embedding Fusion 25
4.2.3. Implementation 26
4.3. Experimental Setup 27
4.3.1. Hyperparameter Search 27
4.3.2. The Four Pipelines 28
4.3.3. Dataset and Metrics 29
4.4. Results and Discussion 30

CHAPTER 5. Evaluation Framework: CFS and APEX 35
5.1. The Problem with Existing Evaluation 35
5.2. Evaluation Metrics 36
5.3. Concept Fidelity Score 38
5.4. APEX: Composite Evaluation Index 40

CHAPTER 6. Experiments and Results 43
6.1. Methods Evaluated 43
6.2. Experimental Setup 44
6.2.1. Hyperparameter Configuration 44
6.2.2. Implementation Details 45
6.2.3. Generation Protocol 46
6.3. Quantitative Results 46
6.4. CFS: Catching What Standard Metrics Miss 49
6.5. Qualitative Analysis 51
6.6. APEX Rankings and Cross-Paradigm Comparison 52

CHAPTER 7. Conclusion 56

7.1. What Was Shown 56
7.2. Limitations 58
7.3. Ethical Considerations 59
7.4. Future Work 60

REFERENCES 62

LIST OF FIGURES

Figure **Page**

Figure 3.1. Sample subjects from DreamBooth-Ext. *Top:* existing subjects from the base DreamBooth set. *Middle:* objects photographed for this work. *Bottom:* Creative Commons public-figure face subjects. 21

Figure 4.1. The LoRA+ID training framework. Text and face embeddings are fused through a linear projection and injected into the cross-attention layers of the denoising UNet. The identity pathway (blue) was used only at inference in FaceID; in LoRA+ID it is active during training, so the LoRA weights themselves become identity-aware. 25

Figure 4.2. Concept accuracy versus face similarity for the baseline LoRA grid search over rank $\in \{4, 8, 16, 32\}$ and steps $\in \{50, 100, 150, 250\}$. Lower ranks and fewer steps favor prompt adherence (concept accuracy) at the cost of identity; the configurations that balance both cluster around rank 16–32 and 100–150 steps. 27

Figure 4.3. FaceID scale sweep over $\{0.2, 0.4, 0.6, 1.0\}$, plotting concept accuracy against face similarity. Lower scales (0.2) keep concept accuracy high but capture identity weakly; higher scales (0.6, 1.0) lock onto the reference face at the cost of prompt adherence. Scale 0.4 gives the best balance and is used throughout. 28

Figure 4.4. FaceID + Baseline LoRA across rank, step, and scale. Lower scales (0.2) achieve higher concept accuracy, while higher scales (0.4, 0.6) improve face similarity at a small cost to concept fidelity; the best-balanced configuration is around rank 16, 150 steps, scale 0.4. 31

Figure 4.5. FaceID + (LoRA+ID) across rank, step, and scale. The same trade-off holds — higher scales and steps raise face similarity while slightly lowering concept accuracy — but every configuration sits at higher face similarity than the corresponding Baseline-LoRA sweep in Figure 4.4, showing that identity-aware training shifts the frontier rather than only sliding along it. 32

Figure 4.6. Comparison of all four pipelines across face similarity and concept accuracy. FaceID + (LoRA+ID) (green) consistently reaches the highest face similarity while keeping competitive concept accuracy; Base LoRA (red) maximizes concept accuracy but loses identity; FaceID zero-shot (blue) preserves identity at low accuracy; FaceID + Base LoRA (orange) is a balanced trade-off. 33

Figure 4.7. Zoomed-in comparison between FaceID + Baseline LoRA (orange) and FaceID + (LoRA+ID) (green), each point annotated with its rank (R), step (S), and scale (Sc). FaceID + (LoRA+ID) consistently reaches higher face similarity while holding concept accuracy competitive, demonstrating the effect of injecting identity embeddings during LoRA training..... 34

Figure 4.8. Qualitative comparison under the prompt "a cinematic photograph of a man riding a motorcycle, wearing sunglasses and a leather jacket, outdoors, sunny weather, dynamic angle." (a) reference; (b) Base LoRA follows the prompt but loses identity; (c) FaceID holds identity but flattens the style; (d) FaceID + Base LoRA improves both; (e) FaceID + (LoRA+ID) preserves identity while satisfying the prompt most convincingly..... 34

Figure 6.1. **CFS-ext examples.** Each column compares base-model and personalized PEFT generations under the same unrelated prompt and random seed. Although the personalized trigger token is absent, adaptation can alter concepts unrelated to the trained subject. CFS_{ext} measures this deviation as DINOv2 cosine drift between the base and PEFT outputs. 51

Figure 6.2. **CFS-VQA examples.** Top row shows the reference subjects. Bottom row shows compositional generations from prompts requesting the personalized subject together with a second related entity. Flash drive: "a photo of USB flash drive together with an external hard drive"; Grey sloth: "a photo of TOK together with a stuffed elephant"; Stapler: "a photo of TOK together with a tape dispenser"; Antonio and Alexandria: "a photo of TOK standing together with a different person"; Bear plushie: "a photo of TOK together with a stuffed rabbit". The flash-drive, Antonio, and Alexandria examples illustrate compositional failures where the second requested entity is absent or not clearly distinguishable; the grey-sloth, stapler, and bear-plushie examples successfully render both requested entities and achieve higher CFS_{VQA} scores..... 52

Figure 6.3. **Personalization across paradigms.** Top row: object personalization (Wolfie). Bottom row: face personalization (Barack Obama). Representative PEFT, image-to-image, and inference-time methods are shown under fixed prompts and seeds..... 53

LIST OF TABLES

Table		Page
Table 3.1.	DreamBooth-Ext dataset composition.	20
Table 4.1.	The four experimental pipelines compared in the LoRA+ID study. All share the same SDXL base model, seed, prompt list, and generation configuration.	29
Table 5.1.	Evaluation metrics grouped by objective. Arrows indicate whether higher (↑) or lower (↓) values are preferred.	36
Table 6.1.	PEFT training configurations. All methods train for 500 steps, batch size 1, gradient clipping at norm 1.0, and a constant learning rate of 1×10^{-4}, with a DreamBooth-style objective and no prior-preservation loss.	45
Table 6.2.	Generation settings for the inference-time identity adapters and image-to-image models. Common to all: seed 42, bfloat16, 1024×1024, 4 images per prompt.	46
Table 6.3.	Subject fidelity. DINOv2 and CLIP-I are pooled across face and object subjects (weighted by subject count) and computed per backbone; FaceID is a faces-only metric. PuLID† is evaluated on faces only, so its DINOv2 and CLIP-I are over faces and not directly comparable to the pooled rows. Best per column in **bold**.	47
Table 6.4.	Comparison by prompt adherence metrics. Best per column in **bold**.	48
Table 6.5.	Comparison by visual quality and human preference metrics. PickScore is reported on its native cosine scale. Best per column in **bold**.	49
Table 6.6.	Style and diversity quality metrics, averaged over all subjects. Style Loss is scaled by 10^4. Best per column in **bold**.	50

Table 6.7. CFS metrics for PEFT methods, averaged over subjects. CFS_{ext} measures far-concept drift (lower is better), while CFS_{VQA} measures compositional capacity (higher is better). Training-free methods do not modify model weights and have no drift component. Best per column in **bold**. 50

Table 6.8. Computational cost feeding the efficiency term of APEX, reported as SDXL/FLUX.1. Columns are trainable parameter count (M), adapter size on disk (MB), training time per subject (s), and warm per-image generation time (s/img); all are complete for both backbones. The efficiency score additionally uses the backbone parameter count (2.6 B for SDXL, 12 B for FLUX.1-dev), so a cost is defined for every PEFT method. 53

Table 6.9. APEX rankings. F, P, R, Y, D, S are the six normalized family scores (fidelity, prompt adherence, preference quality, style, diversity, stability; higher is better), APEX their geometric mean, and Final $APEX_E$ the cost-discounted score. Each metric is normalized to $[\epsilon, 1]$ by a fixed reference range (Equation 5.3, ϵ=0.01). The heavy rule separates the weight-modifying PEFT methods (above), whose stability is measured, from the training-free methods (below), which take the ideal stability value by construction. Best Final APEX and $APEX_E$ in **bold**..... 54

LIST OF ABBREVIATIONS

AdaLoRA	Adaptive Low-Rank Adaptation
APEX	Aggregate Personalization Evaluation indeX (proposed)
CC	Creative Commons
CFS	Concept Fidelity Score (proposed)
CLIP	Contrastive Language-Image Pretraining
CMMD	CLIP Maximum Mean Discrepancy
DDPM	Denoising Diffusion Probabilistic Model
DINOv2	Self-Distillation with No Labels, version 2
DiT	Diffusion Transformer
FFHQ	Flickr-Faces-HQ Dataset
FID	Fréchet Inception Distance
GPU	Graphics Processing Unit
LoCon	Low-Rank Convolution
LoHA	Low-Rank Hadamard Adaptation
LoKr	Low-Rank Kronecker Adaptation
LoRA	Low-Rank Adaptation
LoRA+ID	Identity-Conditioned Low-Rank Adaptation (proposed)
LPIPS	Learned Perceptual Image Patch Similarity
MLP	Multi-Layer Perceptron
PEFT	Parameter-Efficient Fine-Tuning
SDXL	Stable Diffusion XL
SHiRA	Sparse High-Rank Adapters
SVD	Singular Value Decomposition
ViLT	Vision-and-Language Transformer
ViT	Vision Transformer
VQA	Visual Question Answering

CHAPTER 1

INTRODUCTION

Text-to-image models have changed what it means to generate an image. Systems like DALL·E (Ramesh et al. 2021), Imagen (Saharia et al. 2022), and Stable Diffusion (Rombach et al. 2022) produce photorealistic images from a text description in seconds. Stable Diffusion XL (SDXL) (Podell et al. 2023) pushed this further with higher resolution and stronger prompt understanding, and transformer-based architectures such as FLUX.1 (Black Forest Labs 2024) have continued that trajectory. These models are strong on general prompts. The harder problem is making them faithful to a *specific* person or object — generating not merely "a man riding a motorcycle" but *this particular man* riding a motorcycle, recognizably, across arbitrary prompts.

This is the personalization problem. It adapts a pretrained model to a subject from a handful of reference images, and it sits behind most subject-driven applications, from product photography to consistent character art and avatar generation. Yet despite years of progress, it remains unsolved in a satisfying way: the methods that adapt quickly tend to be brittle, the methods that are robust tend to be expensive, and the tools we use to tell the two apart are weaker than the methods themselves. This thesis is about both halves of that sentence — building a better adaptation method, and building the evaluation needed to know whether it, or any other method, actually works.

1.1. What LoRA Does and Where It Falls Short

Retraining every weight of a backbone for each new subject is impractical. Modern text-to-image models hold billions of parameters, and keeping a full copy per subject does not scale. The standard answer is Low-Rank Adaptation (LoRA) (Hu et al. 2022). Given a small set of reference images, LoRA fine-tunes the model's attention layers using a unique trigger token, and the model learns to associate that token with the subject's appearance. LoRA freezes the backbone and trains an adapter that is usually well under one percent of the weights; methods of this kind train quickly, store compactly, and are the workhorse

of modern personalization. A trained adapter fits in a few hundred megabytes and works well enough for many use cases.

The problem surfaces under distribution shift. A LoRA trained on a person's face renders "a professional headshot of TOK" faithfully, but prompts such as "an oil painting of TOK in a Renaissance style" or "a cinematic shot of TOK on a motorbike in the rain" cause the identity to drift: the face changes. The model has learned the visual statistics of the training images rather than a robust representation of who the person actually is.

This happens because LoRA conditions on text alone. The reconstruction loss during training pushes the adapter toward accurately reproducing the training images under the training prompts, but there is no signal telling the model what is specifically *identity* about those images. Everything the model learns about the subject's face is implicit, inferred from the reconstruction objective. When the prompt shifts far from the training distribution, the implicit cues are exactly what the model lets go of first.

This raises a question that motivated the first part of this thesis: *Can LoRA training benefit from facial identity embeddings — not just at inference, as most identity adapters do, but also during training?* If a dedicated identity signal were folded into the optimization, the adapter would have an explicit anchor for who the person is, separate from how the training photographs happened to look.

1.2. A Second Problem, Found While Solving the First

Answering that question required running evaluations. And running evaluations revealed a second problem: the metrics available for evaluating personalization are not good enough to support the comparison we wanted to make.

FID (Heusel et al. 2017) works at the dataset level — it compares distributions of images and cannot tell you whether a *specific* person was generated correctly. CLIP similarity measures whether an image matches the text, not whether the face matches the reference. DINOv2 features are richer and capture subject appearance, but they do not cleanly separate identity drift from style drift: a stylized but faithful portrait and a photorealistic but wrong-identity portrait can land at similar distances. None of the standard metrics can answer the question that actually matters: *did this fine-tuned model learn the concept, or did it memorize the training images while losing the ability to*

generalize?

There is a specific failure mode that no existing metric catches. After fine-tuning on a single subject — a particular backpack, say — the model can become biased toward generating that subject even when prompted for something unrelated, like a dog. We call this *concept drift*: the unintended influence of a personalization adapter on concepts it was never asked to learn. The term is used here by analogy. Unlike its established meaning in streaming-data learning, where the input distribution shifts over time, it denotes the collateral forgetting that a *single* adaptation induces on concepts the model already knew. A well-regularized method should leave those concepts intact. But you cannot measure whether a method does this with FID or CLIP. Stability against concept drift is one axis of personalization quality among several, and current benchmarks rarely measure it — which means the evaluation gap is as consequential as the methods gap, and it cannot be closed by reporting yet another identity or prompt-alignment number.

There is a second, more practical difficulty as well. Work on personalization has piled up metrics for identity, prompt following, aesthetic quality, and diversity, and few papers agree on which to report. Ranking methods across a dozen unrelated numbers is awkward, and no single method tends to lead on all of them at once. There is no widely adopted way to combine these axes into a single, reproducible ranking, so comparisons stay partial and method selection stays a matter of taste.

1.3. Four Contributions

This thesis makes four contributions, each one a direct response to what the previous step uncovered.

1. **LoRA+ID.** A training framework that integrates ArcFace (Deng et al. 2019) facial identity embeddings into the LoRA fine-tuning loop. A lightweight MLP projects the 512-dimensional identity embedding into the model's token space, where it is concatenated with the text tokens and passed through the existing cross-attention layers. No architectural changes to the backbone are needed. The adapter learns to associate the trigger token with both visual appearance and identity-discriminative features, improving fidelity under novel prompts. LoRA+ID was published at ASYU 2025 (Özbay, Ulusoy, and Baştanlar 2025) and is presented in Chapter 4.

2. **DreamBooth-Ext, an extended benchmark dataset.** We grow the DreamBooth benchmark to 60 subjects: 30 from the original set, 10 new object subjects we photographed ourselves, and 20 human faces sourced under Creative Commons licenses so the full set can be released.[1] Every subject is paired with structured context prompts that place it in novel scenes, so the benchmark measures whether a method generalizes beyond the training views rather than merely reconstructs them. The dataset is described in Chapter 3.

3. **Concept Fidelity Score (CFS).** A metric that measures whether personalization stays contained to its target, through two complementary probes: *far-concept drift*, whether an unrelated concept is degraded after adaptation; and *compositional capacity*, whether the model can still place the personalized subject alongside a second instance. This sets CFS apart from forgetting metrics in continual personalization, which track concepts the model *was* trained on earlier in a task sequence; CFS instead isolates the collateral damage a *single* adaptation inflicts on concepts the model was never asked to learn. CFS is defined in Chapter 5.

4. **APEX (Aggregate Personalization Evaluation indeX).** A composite score that combines fourteen metrics across six families — fidelity, prompt adherence, preference quality, style consistency, diversity, and stability — through a scoring formulation designed to balance these axes while preventing any single metric from dominating the ranking. APEX gives one reproducible ranking in place of many disconnected scores, and, to our knowledge, it provides one of the first holistic comparisons across all three personalization paradigms — PEFT fine-tuning, inference-time identity adapters, and image-to-image models — under a single unified protocol. APEX is defined in Chapter 5 and applied in Chapter 6.

The thesis, in short, is about building a better training method and then building the tools needed to know whether it actually works. We use the resulting framework to benchmark ten methods across the three personalization paradigms; with five PEFT methods each run on two backbones (SDXL and FLUX.1-dev), the results tables span fifteen model–backbone configurations in all. In doing so, we created an evaluation benchmark that covers the different aspects of these methods' capabilities — from subject

1. The code, the DreamBooth-Ext dataset, and the APEX scoring script will be released publicly.

fidelity and prompt adherence to perceptual quality, diversity, and the stability of the model's broader knowledge.

1.4. Organization

The remainder of this thesis is organized as follows. Chapter 2 reviews related work: diffusion-based text-to-image models, PEFT methods, inference-time identity adapters, and the current state of personalization evaluation. Chapter 3 describes the DreamBooth-Ext dataset and its prompt design. Chapter 4 presents LoRA+ID, the identity-conditioned training method. Chapter 5 introduces the evaluation framework, defining the Concept Fidelity Score and the APEX index. Chapter 6 reports the full evaluation across the ten methods and two backbones, including the concept-drift and APEX results. Chapter 7 discusses what the results mean, the honest limitations of the study, and where the work should go next.

CHAPTER 2

RELATED WORK

This chapter surveys the technical landscape that the rest of the thesis builds on and reacts to. It moves from the generative backbones that make personalization possible, through the three families of personalization methods this thesis benchmarks, and finally to the two questions that motivate the contributions: how do we know when personalizing a model has quietly broken something else, and how do we summarize the many competing requirements of personalization into a single comparable verdict? Each section closes by stating how the work reviewed there connects to the dataset, the LoRA+ID method, or the evaluation framework developed in later chapters.

2.1. Text-to-Image Diffusion Models

The architecture underlying most modern text-to-image generation is the latent diffusion model (Rombach et al. 2022). Instead of running the diffusion process directly in pixel space — which is computationally expensive — a pretrained encoder compresses images into a lower-dimensional latent space, and the diffusion model learns to denoise within that space. A pretrained decoder reconstructs the image from the denoised latent. This design made high-resolution generation practical, and it is the reason a personalization adapter can be trained on a single GPU in minutes rather than days: the heavy perceptual compression has already been paid for by the autoencoder, and adaptation only has to reshape the denoiser's behavior inside the compact latent space.

DDPM (Ho, Jain, and Abbeel 2020) established the core denoising objective: given a clean image x_0, add Gaussian noise at a timestep t to get x_t, train a neural network to predict the added noise, and iteratively denoise from random noise at inference. The conditioning on text is handled through cross-attention layers, where the model attends to CLIP (Radford et al. 2021) or T5 text embeddings at each denoising step. This cross-attention pathway is not an incidental detail; it is the precise point where most personalization methods intervene. Subject-driven fine-tuning reshapes the projections that

feed these attention layers, inference-time adapters splice an additional image-conditioned attention stream alongside the text one, and the prompts a benchmark uses to probe a personalized model are ultimately routed through exactly this mechanism. Understanding where conditioning enters the network is therefore prerequisite to understanding where, and how, personalization can fail.

SDXL (Podell et al. 2023) extended this recipe with a substantially larger UNet, *dual* CLIP text encoders whose embeddings are concatenated for richer conditioning, and a second-stage refinement model for high-resolution output. SDXL became a de facto standard for community personalization because it strikes a practical balance: strong image quality, a mature ecosystem of adapter tooling, and a UNet whose cross-attention blocks are convenient targets for low-rank adaptation. More recently, FLUX.1 (Black Forest Labs 2024) replaced the UNet entirely with a Diffusion Transformer (DiT) architecture, applying self-attention jointly over image patches and text tokens rather than confining text to dedicated cross-attention layers. This joint-attention design improves prompt following and global coherence, but it also changes the geometry of where an adapter must act: there is no longer a tidy set of cross-attention key/value projections to single out, and the modules are larger. These DiT models at the multi-billion-parameter scale represent the current state of the art in generation quality, yet they are correspondingly harder to fine-tune efficiently, and an adapter that behaves well on SDXL does not automatically transfer its behavior to FLUX.1.

That last observation is one of the reasons this thesis evaluates every parameter-efficient method on *both* SDXL and FLUX.1-dev. A method's fidelity, its prompt adherence, and especially its tendency to disturb unrelated concepts can depend as much on the backbone's architecture as on the adaptation algorithm itself. Holding the adaptation fixed while swapping the backbone — and vice versa — is the only way to separate the contribution of the method from the contribution of the model it is grafted onto, and it is why the benchmark in Chapter 6 reports five PEFT methods across two backbones as ten distinct model–backbone configurations rather than as five method scores — which, together with the five training-free methods, makes fifteen evaluated configurations in all.

2.2. Parameter-Efficient Fine-Tuning for Personalized Generation

Subject-driven personalization asks a pretrained generator to reproduce a specific instance — this particular dog, this particular person's face, this particular backpack — from only a handful of reference images, while still obeying arbitrary text prompts that place the subject in novel contexts. The tension is intrinsic: the model must memorize enough to be faithful, but generalize enough to remain controllable. The methods in this section all attack that tension by adjusting how much of the model is allowed to move during adaptation.

2.2.1. DreamBooth

DreamBooth (Ruiz et al. 2023) is the reference approach for subject-driven personalization. A unique identifier token (e.g., TOK) is paired with a subject's reference images, and the *entire* model is fine-tuned to associate that token with the subject's appearance. A class-preservation loss — computed using images generated by the base model for the general class — prevents the model from forgetting what dogs, chairs, or people look like in general. This loss is among the earliest explicit acknowledgments, inside the personalization literature, that adapting to one subject can corrupt the model's broader knowledge; DreamBooth calls the failure *language drift* and counteracts it by interleaving generic class images into training. The detail matters for this thesis because that prior-preservation loss is fundamentally a *remedy* applied during training and never a *measurement* reported afterward — a distinction the chapter on forgetting returns to at length.

DreamBooth produces good results, but it is expensive. Fine-tuning all weights means storing a complete model copy per subject (several gigabytes), training times are long, and the very small reference sets make overfitting likely, which in turn makes the language-drift problem worse rather than better. These constraints pushed the community toward parameter-efficient alternatives that keep the personalization signal small and isolable. Everything that follows in this section can be read as a different answer to the same question DreamBooth posed: how little of the model can we move and still capture the subject?

2.2.2. Textual Inversion and Custom Diffusion

Two methods bracket the design space before low-rank adaptation enters it. **Textual Inversion** (Gal et al. 2023) is the lightest-touch answer of all: it freezes the entire model and optimizes a single new word embedding — a pseudo-token — so that the subject is captured purely in text-embedding space. Nothing in the network's weights changes, which makes the adapter a few kilobytes, but the same restraint caps how much fine detail a single learned vector can hold, and identity tends to soften under prompts far from training. **Custom Diffusion** (Kumari et al. 2023) moves slightly further into the network, fine-tuning only the key and value projections of the cross-attention layers — the small slice of weights through which text conditioning enters — and adds a mechanism for composing several learned concepts in one image, either by joint training or by closed-form merging of independently trained concepts. Together they mark the two endpoints that LoRA later interpolates between: embedding-only adaptation on one side and partial-weight fine-tuning on the other, both still relying on the reconstruction objective alone to decide what about the subject is worth keeping.

2.2.3. LoRA

LoRA (Hu et al. 2022) was introduced for language models but transferred naturally to diffusion models. For a weight matrix $W \in \mathbb{R}^{m\times n}$, instead of updating W directly, LoRA parameterizes the update as a product of two low-rank matrices:

$$W' = W + \Delta W = W + BA, \quad B \in \mathbb{R}^{m\times r},\ A \in \mathbb{R}^{r\times n} \tag{2.1}$$

where $r \ll \min(m, n)$. Only A and B are trained; W stays frozen. For a rank of 16, this reduces the number of trainable parameters from mn to $16(m + n)$ — typically a 10–100× reduction. The low-rank constraint is not merely an efficiency trick; it is also an implicit regularizer. By forcing the update to live in a low-dimensional subspace, LoRA limits how far the model can stray from its pretrained behavior, which is part of why a small adapter can be faithful to a subject without catastrophically rewriting the network.

In practice, LoRA modules are inserted into the query, key, value, and output projections of the UNet's cross-attention layers, precisely the conditioning pathway described

earlier. The scaling factor α controls the magnitude of the update, usually set equal to r, and acts as a knob that trades fidelity against preservation of the base model. The resulting adapter is a small file (typically 2–5 MB for SDXL at rank 16), which makes distribution and deployment practical and explains why LoRA, rather than full DreamBooth, became the workhorse of modern personalization: it trains quickly, stores compactly, and can be shared and composed.

The limitation is the one that motivates the LoRA+ID contribution of this thesis. LoRA conditions entirely on text. There is no mechanism that tells the adapter what is specifically identity-preserving about the training images; the model must infer this implicitly from the reconstruction objective alone. That inference is brittle. It holds when the test prompt resembles the training conditions and degrades as the prompt deviates — a new pose, an unusual background, a stylization — because nothing in the training signal ever isolated identity from the incidental appearance of the reference photographs. Chapter 4 addresses exactly this gap by conditioning the low-rank update on an explicit identity embedding.

2.2.4. LoRA Variants

Several extensions of LoRA have been proposed, each exploring a different inductive bias for the weight update. They share LoRA's premise — keep the update small and the base model frozen — but disagree on what shape that small update should take, and those differences turn out to matter for both fidelity and stability.

LoHA (Yeh et al. 2024) replaces the single matrix product with a Hadamard product of two low-rank factorizations, which can express patterns of variation that a plain BA update cannot capture at the same parameter budget. **LoKr** (Yeh et al. 2024) instead uses a Kronecker product, giving the update a structured, block-wise form that scales differently with the layer dimensions, and **LoCon** (Yeh et al. 2024) extends adaptation beyond the attention projections into the convolutional layers, which matters in UNet backbones where much of the spatial structure is carried by convolutions rather than attention. **AdaLoRA** (Q. Zhang et al. 2023) treats rank as something to be allocated rather than fixed: it starts at a high rank, scores the importance of each singular direction during training, and prunes toward a target budget, concentrating trainable capacity in

the layers and directions where it is most needed. **SHiRA** (Bhardwaj et al. 2024) departs from low-rank decomposition altogether, applying *sparse* high-rank updates directly to the original weight matrices — only a small fraction of entries change, but they may change anywhere, trading the low-rank inductive bias for a sparsity-based one.

A parallel line of work constrains the update to be *orthogonal* rather than low-rank. Orthogonal Finetuning (**OFT**) (Qiu et al. 2023) multiplies the frozen weights by a learned orthogonal matrix, preserving the pairwise angles between neurons — the model's hyperspherical energy — which stabilizes subject-driven diffusion fine-tuning, and **BOFT** (Liu et al. 2024) makes this affordable by factorizing that rotation into a product of sparse butterfly matrices, reaching a dense orthogonal transform at a fraction of the parameters. **HRA** (Yuan, Liu, and Xu 2024) arrives at a similar place through a chain of Householder reflections, bridging the low-rank and orthogonal regimes. Two further variants refine the low-rank update itself: **DeLoRA** (Bini, Girrbach, and Akata 2025) decouples the *direction* of the update from its *magnitude* and bounds the latter, making adaptation far less sensitive to learning rate and training length, while **FouRA** (Borse et al. 2024) learns the update in the Fourier domain with an input-dependent adaptive rank, explicitly targeting the diversity collapse and near-duplicate "data copying" that vanilla LoRA exhibits at high adapter strength. These variants are mapped here to chart the breadth of the design space; they are points of reference rather than members of the benchmark in Chapter 6.

These variants are usually compared on subject fidelity and prompt adherence, and on those axes they trade places depending on backbone and subject. What unites them is a blind spot rather than a strength: every one of them learns the subject implicitly, through the reconstruction loss alone, and none of them is normally accompanied by a measurement of whether the resulting adapter has acquired a stable, transferable concept or has merely memorized the training distribution. The five PEFT methods this thesis benchmarks — LoRA, LoHA, LoKr, AdaLoRA, and SHiRA — are chosen precisely to span this range of inductive biases (dense low-rank, Hadamard, Kronecker, adaptive-rank, and sparse high-rank) so that the evaluation framework is stressed against genuinely different adaptation geometries rather than minor variants of one.

Identity-conditioned variants form the exception to the implicit-learning rule. **LoRA+ID** (Özbay, Ulusoy, and Başţanlar 2025) feeds ArcFace identity embeddings into

the adapter during training, conditioning the low-rank update on the subject's face rather than leaving identity to emerge from reconstruction. LoRA+ID is the author's own contribution, presented in full in Chapter 4; it is discussed here as the motivating exception that shows the implicit-learning limitation is not inevitable. Because it is a contribution under study rather than an external baseline, it is analyzed on its own terms in Chapter 4 and is not re-run as one of the adapters in the cross-method benchmark of Chapter 6.

2.3. Identity-Conditioned Generation Without Fine-Tuning

A parallel approach to personalization avoids per-subject fine-tuning altogether. Instead of training an adapter for each new subject, these methods train, once, an encoder-and-adapter pair that can inject identity conditioning at inference time from a single reference image. The appeal is immediate: no optimization at deployment, no per-subject storage, and personalization in a single forward pass. The cost, as this section argues, is that the conditioning is bolted on after the model has finished learning, which places these methods in a fundamentally different regime from fine-tuning and complicates any head-to-head comparison.

IP-Adapter (Ye et al. 2023) introduces a decoupled cross-attention pathway that processes image features separately from text features. A ViT (Dosovitskiy et al. 2020) encoder produces visual embeddings that are fused into the generation through a lightweight adapter trained on large image–text corpora. The decoupling is the key idea: rather than competing with the text prompt for the same attention budget, the image condition is given its own attention stream whose output is added back in, so prompt control and image control can coexist. **FaceID** (Ye et al. 2024), a specialized variant aimed at human subjects, replaces the general-purpose ViT encoder with ArcFace (Deng et al. 2019) to extract identity-specific 512-dimensional embeddings that are far more discriminative of facial identity than CLIP- or ViT-style features. An MLP projects these embeddings into the CLIP text space, where they are concatenated with the text tokens before entering the cross-attention layers. The use of ArcFace embeddings here is the same insight that LoRA+ID carries into the fine-tuning regime — that an explicit face-recognition signal captures identity more reliably than a generic visual encoder — which is why this thesis also uses ArcFace as its face-fidelity metric in Chapter 5.

InstantID (Q. Wang et al. 2024) strengthens spatial control by combining facial keypoints, visual embeddings, and identity embeddings, injecting the resulting signals through both a decoupled cross-attention branch and a ControlNet-style branch that conditions on facial geometry. This yields tighter control over pose and facial structure, at the cost of additional architectural machinery and a heavier inference pipeline. **PuLID** (Z. Guo et al. 2024) attacks a different failure mode of inference-time identity injection — the tendency of strong identity conditioning to override the base model's style and degrade prompt following — by adding a contrastive alignment objective during adapter training that keeps the identity pathway from disturbing the model's stylistic behavior, so that identity guidance and prompt control interfere less.

The shared limitation of these methods is structural: they do not modify the model's weights. Identity conditioning is applied *after* the model has finished learning, and the network's internal representations remain agnostic to the specific subject. This produces consistent, attractive results for simple prompts but tends to break down under semantic diversity or strong stylization, where the frozen backbone has no learned notion of the subject to fall back on. It also means these methods cannot be compared to fine-tuning approaches naively: their adaptation operates in a completely different regime, they leave the base weights untouched and therefore induce no weight-level drift in unrelated concepts, and any fair benchmark has to acknowledge this asymmetry rather than paper over it. This is one of the central reasons a *unified* evaluation protocol was necessary, and it is why the benchmark in this thesis evaluates fine-tuning, inference-time identity adapters, and image-to-image editing under one common set of subjects, prompts, and metrics.

2.4. Image-to-Image and Instruction Editing

A third route to personalization sidesteps both per-subject training and identity-specific adapters: it regenerates or edits a reference image directly, relying on the base model's priors to carry the subject into a new context. The approach traces back to SDEdit (Meng et al. 2022), which partially noises an input image and then denoises it toward a target description, trading off how much of the original is preserved against how much the prompt is allowed to change, and to InstructPix2Pix (Brooks, Holynski, and

Efros 2023), which trains a model to follow a natural-language edit instruction applied to a source image. Both established that a reference image plus a text instruction can substitute for explicit personalization in many cases, with zero subject-specific optimization.

Modern systems push this much further. FLUX.2 (Black Forest Labs 2025) regenerates from one or more reference images with strong fidelity to the depicted subject, and Qwen-Image-Edit (Qwen Team 2025) edits a reference image to satisfy an instruction while preserving the parts that should not change. These models carry very strong priors and produce high-quality, prompt-faithful images with no per-subject training whatsoever, which makes them attractive and increasingly competitive with trained adapters. But the very thing that makes them convenient — that they never learn a persistent concept and instead transform a reference on the fly — raises the question the rest of this thesis is built to answer: does reference-based regeneration actually deliver *faithful* and, crucially, *stable* personalization across diverse prompts, or does it merely produce a pleasing image that resembles the reference? Settling that requires a benchmark that puts these editing models on the same footing as fine-tuning and adapter methods, which is exactly what Chapter 6 does by including image-to-image models as one of its three evaluated paradigms.

2.5. Forgetting and Concept Drift

The risk that personalizing on one subject quietly damages everything else is not new, but it is rarely measured. DreamBooth (Ruiz et al. 2023) already noticed it, called it language drift, and patched it with a prior-preservation loss. That loss is a fix, not a measurement: it pushes against forgetting during training but never reports how much forgetting would have occurred, nor how much remains. The term *concept drift* is borrowed here by analogy and used in a sense unlike its established meaning in streaming-data learning, where it denotes a shift in the input distribution over time. In this thesis it denotes the collateral forgetting that a *single* adaptation induces on concepts the model already knew and was never asked to touch — a static, one-shot side effect of personalization rather than a gradual drift driven by changing data.

The unlearning literature documents the same collateral damage from the opposite direction. Concept *erasure* edits a model to forget a concept rather than learn one, and there the side effects are by now well established: erasing one concept bleeds into semantically

neighboring ones (Saha et al. 2025); dedicated unlearning benchmarks score precisely how much a target edit harms unrelated concepts (Moon et al. 2025); and an erased concept can even resurface once the model is later fine-tuned on unrelated data, showing that the edit was a surface suppression rather than a true removal (Suriyakumar et al. 2025). These results confirm, from the erasure side, that local edits to a generative model have non-local consequences — but they study removing concepts, not adding them.

Personalization is the same problem in reverse. It *adds* a concept instead of removing one, and by the same mechanism — a localized weight edit driven by a tiny data set — it can perturb concepts far from the one being learned. Yet this side of the phenomenon has gone almost entirely unmeasured. Within personalization, forgetting has so far been quantified only in the *continual* setting, where the concern is sequential rather than collateral. ConceptGuard (Guo and Jin 2025) defines forgetting metrics that track how learning a new subject degrades subjects that were *explicitly learned earlier* in a sequence, and Mining Your Own Secrets (Jha et al. 2025) regularizes training so the model retains its zero-shot classification ability, leveraging the observation that a diffusion model is itself an implicit classifier. Both care about preserving capability, and both are valuable, but neither offers a metric for the drift a single adapter induces on concepts the model already knew and that were never part of the personalization process at all — ConceptGuard measures degradation of previously *taught* subjects, and Mining Your Own Secrets is a training method rather than an evaluation.

This is the gap the thesis closes. The Concept Fidelity Score introduced in Chapter 5 treats containment as something to quantify rather than a side effect to hope for: it measures whether adding a subject leaves the rest of the model intact. It does so with two probes — a far-concept drift probe ($\mathrm{CFS}_{\mathrm{ext}}$, lower is better) that checks whether unrelated concepts have shifted after adaptation, and a compositional-capacity probe ($\mathrm{CFS}_{\mathrm{VQA}}$, higher is better) that checks whether the personalized model can still compose the subject into varied scenes. Together they convert the long-acknowledged but unmeasured risk of personalization-induced forgetting into a reported number, and they supply the stability term that the evaluation literature reviewed next has consistently lacked.

2.6. Evaluation of Personalized Generation

A good personalization method has to do several things at once: reproduce the target concept faithfully, follow the text prompt, produce visually convincing images, remain diverse rather than collapsing to a single output, and accomplish all of this without disrupting the model's general capabilities. The central problem of this thesis's evaluation framework is that no single existing metric captures all of these, and that the metrics in common use were largely designed for adjacent tasks and repurposed here, often beyond what they were built to measure.

FID (Heusel et al. 2017) was designed for unconditional generation. It measures a distributional distance between sets of images in Inception feature space, which is a sensible proxy for overall realism but is structurally insensitive to per-identity correctness: two methods with identical FID can differ wildly in whether they generate the *right* subject, because FID compares distributions, not instances. **CLIP similarity** (Radford et al. 2021) measures image–text alignment and is the standard proxy for prompt adherence, but CLIP embeddings compress visual information too aggressively to resolve fine-grained facial identity, so a high CLIP-T score says little about subject fidelity. **DINO** (Caron et al. 2021) and its successor **DINOv2** (Oquab et al. 2024) provide self-supervised features that are far richer for instance-level comparison and that correlate better with human judgments of subject consistency, which is why DINO-style similarity has become a standard fidelity metric for personalization. But even DINOv2 similarity measures fidelity on a *static* set of generations; it does not measure concept *stability* — whether the subject is rendered correctly across genuinely diverse prompts, or whether the adapter has overfit to its training distribution and only looks right under familiar conditions.

DreamBench++ (Peng et al. 2024) pushed personalization evaluation forward by replacing crude similarity scores with VQA-based assessment and human preference judgments, yielding a more faithful picture of whether a generation depicts the intended subject in the intended scene. It is a clear improvement over FID-only evaluation, but it still scores static generation quality rather than concept stability, and it leans on a costly, hard-to-reproduce multimodal judge. Learned human-preference models address a different axis again: PickScore (Kirstain et al. 2023), ImageReward (Xu et al. 2023), HPSv2 (Wu et al. 2023), and its successor HPSv3 (Ma et al. 2025) all predict which

image a human would prefer, and are valuable for capturing aesthetic and instruction-quality nuances that no similarity metric reaches. Yet by construction they are blind to identity: a beautifully rendered image of the *wrong* person earns a high preference score. Holistic evaluation efforts such as HEIM (Lee et al. 2023) and UniGenBench++ (Y. Wang et al. 2025) assemble many such axes into broad scorecards for text-to-image models, and they make clear how multidimensional the problem is — but they report a panel of numbers rather than a single, stability-aware verdict for subject-driven personalization, and they are not designed around the personalization-specific question of collateral drift.

The picture that emerges is one of complementary but individually incomplete metrics: fidelity scores that ignore prompt adherence, prompt scores that ignore identity, preference scores blind to both, and none of them aware of stability. This is the precise gap the evaluation framework of this thesis is built to close. The Aggregate Personalization Evaluation indeX (APEX) introduced in Chapter 5 (Section 5.4) combines fourteen metrics organized into six families — fidelity, prompt adherence, preference quality, style consistency, diversity, and stability — into a single comparable score via an equal-weight geometric mean over normalized family scores, so that a method cannot buy a high overall rank by excelling on one axis while failing another. The Concept Fidelity Score of Section 5.3 supplies the stability family that every benchmark above omits. Together, APEX and CFS turn the scattered, partial evaluation practice surveyed here into a unified protocol that scores fidelity, prompt adherence, quality, diversity, and — to our knowledge, among the first in personalization evaluation to do so — concept stability under one roof.

CHAPTER 3

THE DREAMBOOTH-EXT DATASET

The evaluation in this thesis rests on a single dataset that has to do two jobs at once. First, it must support genuine few-shot fine-tuning of personalization methods, including the identity-aware training studied in Chapter 4, which means several reference images per subject rather than one. Second, it must let many methods be compared fairly across both object and human-face categories, under prompts that test whether a method has learned a transferable concept rather than memorized a view. No off-the-shelf resource satisfies both requirements together, so this chapter introduces **DreamBooth-Ext**, an extended subject set and evaluation protocol built specifically for the benchmark in Chapters 5 and 6.

3.1. Motivation and Design

Two reference resources dominate subject-driven generation, and they were built for different purposes. The **DreamBooth** dataset (Ruiz et al. 2023) is a *training* resource: it provides 30 subjects with 4–6 images each, captured from several viewpoints and under different conditions, so that a model can learn what a subject looks like as a three-dimensional object rather than as one frozen photograph. **DreamBench++** (Peng et al. 2024) is an *evaluation* resource: it offers a substantially larger and more diverse pool of concepts, which is attractive for breadth, but it supplies only a *single* image of each subject.

That single-image design is the deciding factor for this work. Every method benchmarked here is fine-tuned on a handful of images per subject, and fine-tuning on one image overfits almost immediately: the model collapses onto that exact pose, crop, and lighting, and there is then nothing left to measure generalization against, because the only ground truth available is the image the model was trained to reproduce. A benchmark assembled from single images would therefore conflate two very different things — memorizing a picture and learning a subject — and would systematically reward the former. This thesis instead builds on DreamBooth, whose multi-image subjects support

real few-shot fine-tuning, and extends it.

Two further considerations reinforce the choice. DreamBench++ scores generations with a proprietary multimodal model acting as the judge. That is expensive to run at the scale of fifteen model–backbone configurations and is difficult to reproduce exactly, since the judge can change underneath the benchmark; the framework in this thesis instead relies on open, deterministic metrics that anyone can recompute (Chapter 5). And, more fundamentally, no existing protocol — DreamBench++ included — probes for *concept drift* on subjects *unrelated* to the one being learned, which is precisely the failure mode the Concept Fidelity Score (CFS) metric is designed to expose. A benchmark that only ever shows the model its own trained subject can never reveal what that training quietly broke elsewhere.

The DreamBooth subject set, useful as it is, has three specific limitations for the questions asked here:

- **No human identities.** The original subjects are objects, pets, and toys. Human faces are the most demanding and most practically important personalization target, and they are the entire focus of the identity-conditioned method in Chapter 4; a benchmark without them would leave the central case untested.

- **No standardized evaluation prompts.** DreamBooth ships subjects but no fixed prompt protocol, so reported results across papers are rarely comparable. A reproducible benchmark needs a frozen, per-subject prompt list.

- **Insufficient category diversity for compositional probing.** CFS's compositional probe asks a personalized model to place its trained subject alongside a second, related instance from the same broad category. Constructing those pairings, and having enough related concepts to compose against, needs more variety within each category than the original thirty subjects densely provide.

3.2. Dataset Composition

These three gaps are addressed by **DreamBooth-Ext**, an extended dataset of **60 subjects** organized into three categories, summarized in Table 3.1 with representative subjects shown in Figure 3.1. The 30 original DreamBooth subjects are retained un-

Table 3.1. DreamBooth-Ext dataset composition.

Category	Subjects	Images/Subject
DreamBooth subjects	30	4–6
New objects (ours)	10	5
Human faces (CC)	20	5
Total	**60**	–

changed, so that results remain comparable to prior work, and are augmented with 10 newly photographed objects and 20 Creative Commons human-face subjects.

Every subject, regardless of category, is evaluated with *four* per-subject context prompts. Each prompt places the subject in a new situation — a different background, action, attribute, or viewpoint — and so tests whether the method has acquired a movable concept rather than a memorized picture. For example, the `backpack` subject is queried with prompts such as "a photo of TOK backpack sitting on a wooden bench in a park" and "a photo of TOK backpack worn by a hiker on a mountain trail," where TOK denotes that subject's unique trigger token. Crucially, prompts are written *per subject* rather than drawn from one template shared across the whole dataset, so each subject is exercised in scenes that actually make sense for it: a backpack is worn and carried, a teapot is poured and placed on a table, a person is photographed in plausible settings. A single global template would force unnatural pairings (a perfume bottle "riding a bicycle") that confound prompt adherence with semantic nonsense, and this per-subject design avoids that.

3.2.1. Extended Object Subjects

Ten new object subjects were photographed specifically for this work, with five images of each taken under varied backgrounds and lighting. Capturing the images directly, rather than scraping them from the web, has a deliberate practical payoff: the photos carry no licensing restrictions, so they can be released as part of the benchmark without any attribution or redistribution concerns. The ten objects also widen the texture, material, and shape diversity of the original DreamBooth set — glossy, matte, transparent, fabric, and molded surfaces are all represented — which matters both for subject fidelity and for the compositional probing CFS performs, pairing a trained object with a related



Figure 3.1. Sample subjects from DreamBooth-Ext. *Top:* existing subjects from the base DreamBooth set. *Middle:* objects photographed for this work. *Bottom:* Creative Commons public-figure face subjects.

one from the same category. The objects span everyday categories: a LEGO set, a cap, a perfume bottle, a stapler, a speaker, a thermos, a USB drive, a pair of shoes, a computer mouse, and a wallet.

3.2.2. Human Face Subset

The face subset contains 20 human subjects, balanced as 10 male and 10 female, each with five reference images. They are widely photographed public figures, assembled through web image searches restricted to Creative Commons usage rights, choosing individuals for whom several openly licensed photographs are available. The choice of public figures over private individuals is intentional and ethical as much as practical: relying on imagery in which public figures are already extensively and openly photographed keeps the consent footprint of the benchmark small, since no new photography of private people is introduced and only material already published under permissive licenses is used.

Sourcing faces this way — rather than from datasets such as FFHQ (Karras, Laine, and Aila 2019) or CelebA, whose redistribution and consent terms are restrictive — is what allows the benchmark to be released with per-image attribution, something most

personalization benchmarks built on those corpora simply cannot offer. For every face image, the source URL and the specific license type are released as part of the benchmark metadata, so that the provenance of each reference is fully traceable and independently verifiable. Like the other categories, each face subject is paired with four evaluation prompts that probe identity under new contexts away from the training distribution, so that the subset tests recontextualized identity rather than reconstruction of a familiar portrait.

3.3. Evaluation Prompt Structure

The per-subject context prompts are written to push the subject away from its training views along one or more semantic axes: a changed *background*, a new *action* or interaction with the surrounding scene, a modified *attribute*, or a different photographic or artistic *style*. Spreading the four prompts of each subject across these axes is what turns the benchmark from a reconstruction test into a generalization test: a method that has merely memorized its five training images can reproduce them, but it cannot reliably place the subject on a mountain trail, hand it to a hiker, or render it in an unfamiliar style while keeping it recognizable.

These prompts play a deliberate dual role, and this is the design decision that ties the dataset to the metrics that follow. First, they are the standard evaluation prompts against which subject fidelity and prompt adherence are measured — did the generated image keep the subject's identity, and did it follow the requested scene. Second, the same set of subjects becomes the basis for the stability probes that CFS introduces in Chapter 5. There, with the adapter loaded but its trigger token removed, the model is prompted with an entirely *unrelated* concept and compared against the base model, exposing drift on material the adaptation never touched; and, with the trigger present, the subject is asked to share the frame with a second, related instance to test whether it can still compose. Both probes exercise the collateral concept drift that ordinary fidelity-and-adherence benchmarks never look for. Designing one prompt set to serve both the conventional fidelity axis and the stability axis keeps the evaluation internally consistent and avoids tuning a method to two unrelated prompt distributions.

CHAPTER 4

LORA+ID: IDENTITY-CONDITIONED LORA TRAINING

Before turning to the broad evaluation that occupies the rest of this thesis, we present a focused contribution of our own: *LoRA+ID*, an identity-conditioned variant of LoRA training for human-face personalization, published at ASYU 2025 (Özbay, Ulusoy, and Baştanlar 2025). LoRA+ID grew out of a recurring frustration with text-only LoRA, namely that an adapter which reproduces a subject's face perfectly under the training prompt can lose that same face the moment the prompt shifts toward an unfamiliar style. The method is a separate piece of work from the benchmark presented in later chapters; it is not one of the methods evaluated by our framework. We include it here because it motivates, in a concrete and self-contained way, the central tension that the rest of the thesis sets out to measure: the trade-off between holding a subject's identity and following a prompt that pulls that identity into new territory.

4.1. Why Text-Only Conditioning Is Not Enough

Standard LoRA training optimizes a text-conditioned denoising objective. Writing it informally, the adapter weights are fit by minimizing

$$\mathcal{L}_{\text{LoRA}} = \mathbb{E}_{x,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_{\text{text}})\|^2\right], \tag{4.1}$$

where z_t is the noisy latent at timestep t, $\epsilon \sim \mathcal{N}(0, I)$ is the sampled noise, ϵ_θ is the (LoRA-adapted) denoiser, and c_{text} is the CLIP text embedding of the training prompt. The supervision is entirely reconstructive: the adapter learns to regenerate the training images when given the training prompt, and nothing in the objective ever names the subject's identity as a quantity to be preserved.

This is sufficient as long as the prompt stays close to the training distribution. The adapter has effectively memorized that the trigger token predicts a particular distribution of colors, textures, and facial features under familiar phrasing. But the objective gives it no anchor for what *should remain invariant* about the subject when the surrounding

description changes. When a user asks for "an oil painting of TOK in a Renaissance style" or "TOK as a cyberpunk character at night," the adapter must extrapolate, and it has only reconstruction statistics to extrapolate from. Identity drifts under prompt shift precisely because identity was never explicitly represented during training; it was only ever an emergent by-product of fitting pixels.

Inference-time identity adapters such as FaceID attack the symptom rather than the cause. They inject an ArcFace identity embedding into the cross-attention layers at generation time, nudging each denoising step toward the reference face. This helps, but the underlying weights remain identity-agnostic: they were trained without ever seeing an identity signal, so the embedding has to fight an adapter that does not know what to do with it. The guidance steers generation but cannot retroactively teach the adapter the relationship between an identity and its visual appearance across styles.

LoRA+ID closes this loop. We supply identity embeddings *during* training, fusing them with the text conditioning so that the adapter is optimized against a joint signal. Informally, the objective becomes

$$\mathcal{L}_{\text{LoRA+ID}} = \mathbb{E}_{x,t,\epsilon}\left[\left\|\epsilon - \epsilon_\theta(z_t, t, c_{\text{joint}})\right\|^2\right], \tag{4.2}$$

where c_{joint} is a conditioning vector that carries both the prompt and the subject's identity. The change is small in implementation but consequential in spirit: identity stops being something the model must infer from reconstruction and becomes something it is told, every step, so the LoRA weights themselves become identity-aware.

4.2. Method

4.2.1. Face Embedding Extraction

For each training subject we extract a single, fixed identity descriptor with a frozen ArcFace network (Deng et al. 2019). Given a reference portrait x_{ref}, ArcFace returns a 512-dimensional embedding,

$$\mathbf{e}_{\text{id}} = \mathcal{F}_{\text{ArcFace}}(x_{\text{ref}}) \in \mathbb{R}^{512}, \tag{4.3}$$

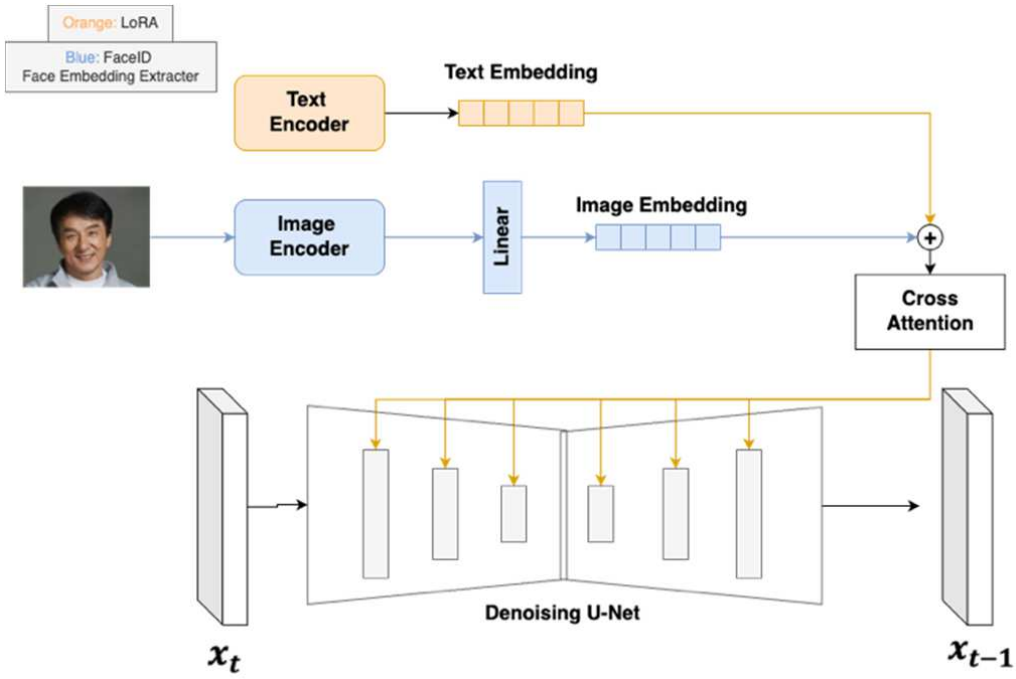


Figure 4.1. The LoRA+ID training framework. Text and face embeddings are fused through a linear projection and injected into the cross-attention layers of the denoising UNet. The identity pathway (blue) was used only at inference in FaceID; in LoRA+ID it is active during training, so the LoRA weights themselves become identity-aware.

that lives in a face-recognition space where cosine distance corresponds to perceived identity difference. These embeddings are computed once and never updated; they act as the identity anchors against which the adapter is trained. Because ArcFace is frozen, the identity signal stays stable and well-conditioned throughout optimization rather than co-adapting with the LoRA weights.

The training prompts are not hand-written. To keep the pipeline scalable and to avoid biasing the text conditioning toward any particular phrasing, we caption every training image automatically with BLIP-2 (Li et al. 2023). Each training sample is therefore a triple: an image, its BLIP-2 caption, and the subject's ArcFace identity embedding. Figure 4.1 shows how these ingredients come together.

4.2.2. Embedding Fusion

The 512-dimensional ArcFace vector and the CLIP text tokens live in different spaces, so we bridge them with a lightweight projection. A small MLP maps the identity

embedding to a set of k identity tokens of the model's hidden width d,

$$\mathbf{t}_{\text{id}} = \text{MLP}(\mathbf{e}_{\text{id}}) \in \mathbb{R}^{k \times d}, \tag{4.4}$$

which are then concatenated with the text tokens to form the joint conditioning,

$$c_{\text{joint}} = \text{Concat}(c_{\text{text}},\ \mathbf{t}_{\text{id}})\ . \tag{4.5}$$

This joint sequence replaces the text-only conditioning in the cross-attention layers, and the denoiser is trained against $\mathcal{L}_{\text{LoRA+ID}}$ as written above.

The mechanism is deliberately identical to the one FaceID uses at inference: identity tokens are concatenated with text tokens and consumed by the same cross-attention pathway. The only, and crucial, difference is *when* it happens. In FaceID the fusion is applied at generation time over frozen weights; in LoRA+ID it is applied during training, so the adapter is optimized in the presence of the identity signal and learns to cooperate with it rather than to be overridden by it.

We keep the architectural footprint as small as possible. The MLP projection is trained jointly with the LoRA weights, but the UNet itself is left untouched: no extra cross-attention blocks, no new attention heads, no structural surgery. The only components that did not already exist in a standard LoRA pipeline are the frozen ArcFace encoder, the trained MLP projection, and the token concatenation. This restraint matters for fairness and reproducibility, because it means any improvement we observe can be attributed to the identity conditioning rather than to added model capacity.

4.2.3. Implementation

All experiments build on Stable Diffusion XL 1.0 (Podell et al. 2023) as the base model. We train at 512×512 resolution in FP16 precision to keep the per-subject cost low. The LoRA modules adapt the query, key, value, and output projections of the UNet cross-attention layers, where text and identity conditioning enter the network and identity awareness is most useful. Optimization uses 8-bit AdamW at a constant learning rate of 1×10^{-4}, no warmup, and gradient checkpointing to fit in memory. Every run uses a fixed seed so that configurations differ only in the variable under study.

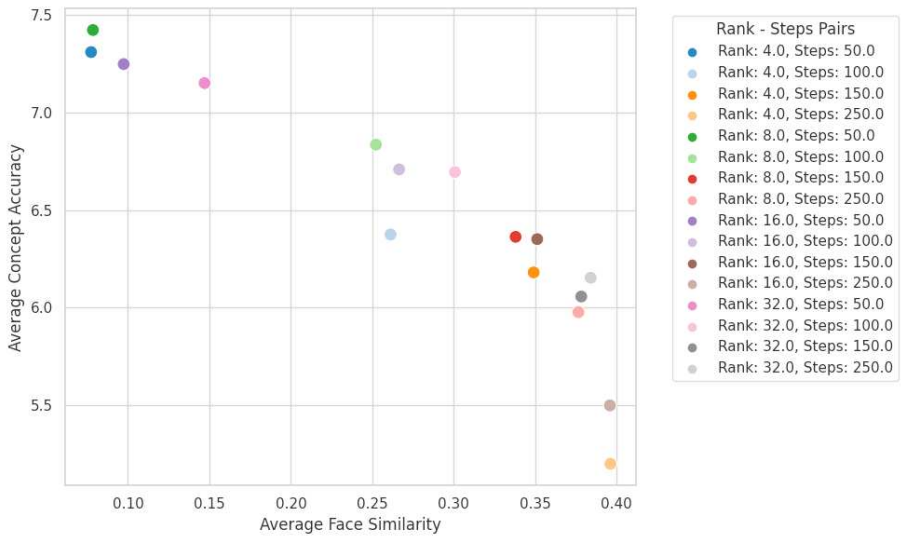


Figure 4.2. Concept accuracy versus face similarity for the baseline LoRA grid search over rank $\in \{4, 8, 16, 32\}$ and steps $\in \{50, 100, 150, 250\}$. Lower ranks and fewer steps favor prompt adherence (concept accuracy) at the cost of identity; the configurations that balance both cluster around rank 16–32 and 100–150 steps.

4.3. Experimental Setup

4.3.1. Hyperparameter Search

The two most influential knobs in LoRA personalization are the adapter rank and the number of training steps, so we began with a grid search over rank $\in \{4, 8, 16, 32\}$ and steps $\in \{50, 100, 150, 250\}$ on the baseline (text-only) LoRA configuration. The outcome, plotted in Figure 4.2, makes the central trade-off visible. Lower ranks and fewer steps bias the adapter toward prompt adherence: the model stays flexible and follows the text, but the face is only loosely captured. Higher ranks and more steps push in the opposite direction, locking onto the subject's identity at the expense of responding to the prompt. The configurations that balance the two cluster around rank 16 to 32 with 100 to 150 steps, and we adopt this region for all subsequent experiments.

A second hyperparameter governs the inference-time identity adapter: the FaceID scale, which controls how strongly the identity embedding pulls the generation toward the reference face. We swept this over $\{0.2, 0.4, 0.6, 1.0\}$, with the result shown in Figure 4.3. At scale 1.0 the model overfits to the face, ignoring the prompt and producing near-replicas of the reference regardless of the requested style, which collapses concept accuracy. At

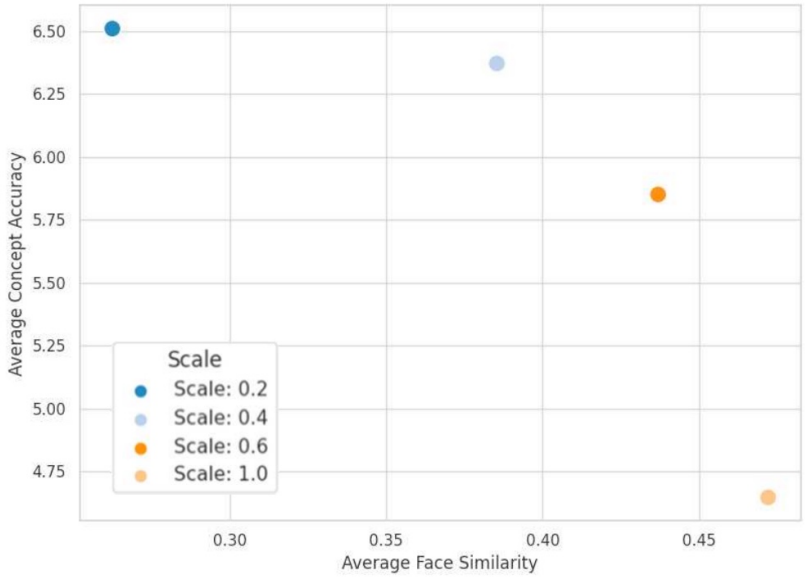


Figure 4.3. FaceID scale sweep over $\{0.2, 0.4, 0.6, 1.0\}$, plotting concept accuracy against face similarity. Lower scales (0.2) keep concept accuracy high but capture identity weakly; higher scales (0.6, 1.0) lock onto the reference face at the cost of prompt adherence. Scale 0.4 gives the best balance and is used throughout.

the low end the identity guidance is too weak to hold the face. The plot makes the trade-off monotonic: as the scale rises the face similarity climbs while concept accuracy falls. A scale of 0.4 was consistently the best-balanced choice, providing enough identity pull to preserve the subject while leaving the prompt free to shape the rest of the image, and we use it throughout.

4.3.2. The Four Pipelines

To isolate exactly what identity conditioning *during training* contributes, as opposed to identity guidance applied only at inference, we compare four pipelines. They are summarized in Table 4.1 and span the natural progression from a plain text-only adapter, through inference-only and decoupled combinations, to our jointly trained method. Crucially, all four share the same SDXL base model, the same fixed seed, the same prompt list, and the same generation configuration, so that any measured difference is attributable to the conditioning strategy alone.

The four rows form a clean ablation. *Baseline LoRA* is the text-only reference. *FaceID (zero-shot)* uses no subject-specific training at all, relying entirely on inference-

Table 4.1. The four experimental pipelines compared in the LoRA+ID study. All share the same SDXL base model, seed, prompt list, and generation configuration.

Configuration	LoRA Training	Inference Conditioning
Baseline LoRA	Text prompts only	Prompt
FaceID (zero-shot)	No LoRA training	Prompt + Face Image
FaceID + Baseline LoRA	Text prompts only	Prompt + Face Image
FaceID + (LoRA+ID)	Text prompts + face embeddings	Prompt + Face Image

time identity injection. *FaceID + Baseline LoRA* stacks inference-time identity guidance on top of an adapter that was trained without any identity awareness, so the two halves operate independently. *FaceID + (LoRA+ID)*, our method, trains the adapter *with* the identity embedding and then also applies FaceID at inference, so the two pathways were optimized to reinforce one another.

4.3.3. Dataset and Metrics

The study uses eight human identities, balanced four male and four female. For each identity we collect eight training images and four additional reference portraits; the portraits are reserved for extracting the ArcFace embedding used at inference, keeping the identity signal at generation time separate from the images the adapter was trained on. Evaluation is carried out over eighteen style-diverse prompts per subject, deliberately ranging from photographic to painterly to stylized, so that the prompts stress the identity-versus-style trade-off rather than staying comfortably near the training distribution.

We report two complementary metrics. **Face Similarity** is the ArcFace cosine similarity between the generated image and the reference identity, computed with InsightFace (J. Guo et al. 2022); it measures how well the subject's identity survives generation. **Concept Accuracy** is a ViLT (Kim, Son, and Kim 2021) text-image retrieval score measuring how well the generated image matches the requested prompt; it measures prompt adherence. The two metrics are intentionally in tension, and the goal of LoRA+ID is to push the achievable frontier between them outward rather than to win on one at the cost of the other.

4.4. Results and Discussion

The grid-search analysis earlier in this chapter already confirmed its central premise: a text-only adapter cannot maximize identity retention and prompt adherence at the same time, because without an explicit identity signal the two objectives are antagonistic and any gain on one is paid for on the other. The four-pipeline comparison then asks whether supplying that signal during training moves the frontier. We report results through the scatter plots of Figure 4.6 and the qualitative grid of Figure 4.8; we describe the trends these figures show rather than tabulating point estimates.

Before comparing the four pipelines head to head, it is instructive to see how the two FaceID-combined methods behave *internally*, across their own rank, step, and scale settings. Figure 4.4 sweeps these for **FaceID + Baseline LoRA**: lower scales (e.g. 0.2) generally win on concept accuracy, while higher scales (0.4, 0.6) favor identity retention but slightly compromise concept fidelity, and the best-balanced point sits around rank 16, 150 steps, scale 0.4. Figure 4.5 repeats the sweep for our **FaceID + (LoRA+ID)** method: the same broad trade-off holds — higher scales and more steps push face similarity up at a mild cost to concept accuracy — but the entire cloud of configurations sits at noticeably higher face similarity than the Baseline-LoRA sweep, the first sign that identity-aware training shifts the achievable frontier rather than merely moving along it. Both sweeps confirm that the conclusion is not an artifact of one lucky hyperparameter choice: the advantage is visible across the configuration grid.

Reading across the pipelines, a consistent story emerges. **Baseline LoRA** sits toward high concept accuracy but low face similarity: it follows the prompts well and renders the requested styles convincingly, yet under stylistically divergent prompts the subject's identity slips away, exactly the failure mode that motivated the method. **FaceID zero-shot** occupies the opposite corner. It preserves identity strongly but at the cost of overriding the prompt; in our scale sweep, once the FaceID scale reaches 0.6 or above the model ignores the prompt's style and reproduces the reference face, so concept accuracy is low. **FaceID + Baseline LoRA** lands in between and clearly beats either alone, since LoRA contributes prompt adherence and FaceID contributes identity. But the two are decoupled: the adapter was trained with no notion of identity, so the pathways coexist rather than cooperate, and the combination does not fully realize the potential of either.

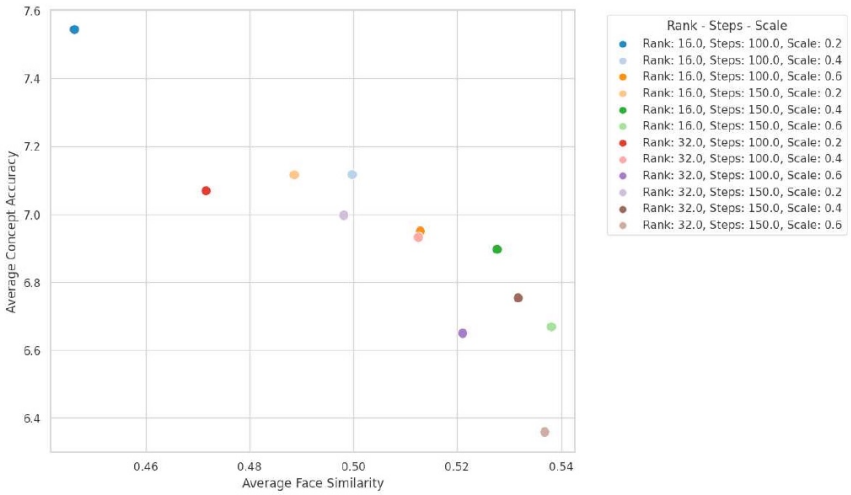


Figure 4.4. FaceID + Baseline LoRA across rank, step, and scale. Lower scales (0.2) achieve higher concept accuracy, while higher scales (0.4, 0.6) improve face similarity at a small cost to concept fidelity; the best-balanced configuration is around rank 16, 150 steps, scale 0.4.

FaceID + (LoRA+ID), our method, consistently reaches the highest face similarity while keeping concept accuracy competitive with the other FaceID-based pipelines. We are candid about the magnitude: because this pipeline and FaceID + Baseline LoRA both apply FaceID at inference, the gap between them is incremental rather than dramatic. What makes it meaningful is that the gap is *consistent* across subjects and configurations rather than the product of a few favorable cases, and that it is clearest precisely where it matters most, under the prompts that diverge furthest from the training images. There, the adapter trained with identity awareness has learned to map identity onto visual features in a way that aligns naturally with the FaceID conditioning, so the two pathways reinforce each other instead of pulling apart.

Figure 4.7 isolates exactly this head-to-head, zooming into the region where FaceID + Baseline LoRA (orange) and FaceID + (LoRA+ID) (green) overlap and annotating each point with its rank, step, and scale. Read configuration by configuration, the green points sit consistently to the right of their orange counterparts — higher face similarity at matched concept accuracy — which is the same incremental-but-systematic gain the full comparison shows, now made legible at the level of individual settings. The proposed method does not win on every single point, but it wins on most, and it never trades away a meaningful amount of concept accuracy to do so.

The qualitative example in Figure 4.8 makes the same trade-off legible at a glance.

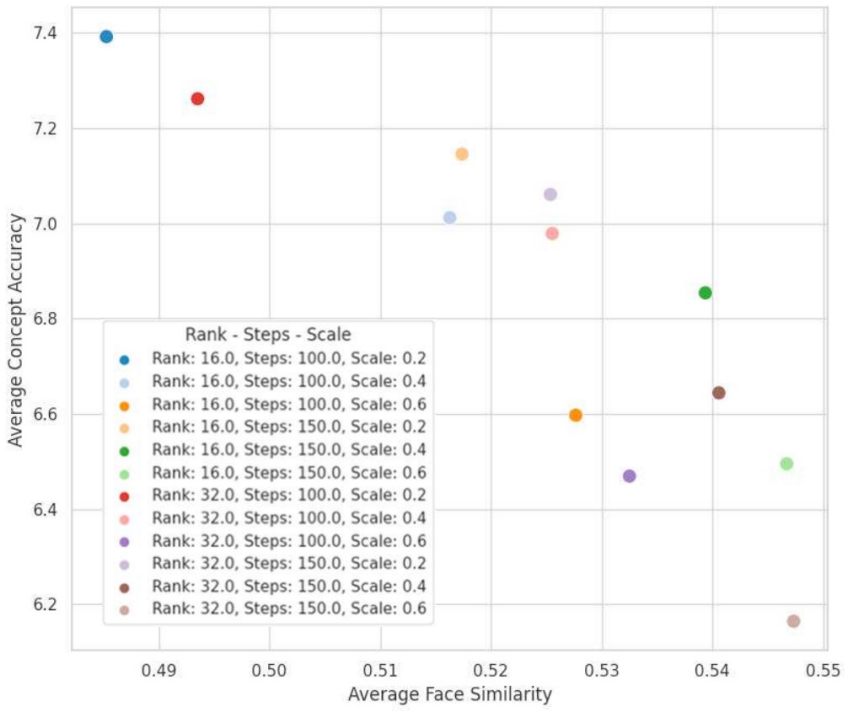


Figure 4.5. FaceID + (LoRA+ID) across rank, step, and scale. The same trade-off holds — higher scales and steps raise face similarity while slightly lowering concept accuracy — but every configuration sits at higher face similarity than the corresponding Baseline-LoRA sweep in Figure 4.4, showing that identity-aware training shifts the frontier rather than only sliding along it.

Under a demanding cinematic prompt, Baseline LoRA (b) obeys the prompt, the motorcycle, sunglasses, and leather jacket are all present, but the man is no longer recognizably the reference subject. FaceID alone (c) keeps the face but flattens the requested cinematic styling toward a generic portrait. The decoupled combination (d) recovers much of both. Our jointly trained pipeline (e) is the one that most convincingly satisfies the prompt while still preserving identity, which is the behavior the scatter plots quantify.

A secondary finding is worth recording. LoRA+ID tends to converge at lower step counts than the baseline: because the adapter is handed an explicit identity signal rather than having to infer identity from reconstruction statistics alone, the learning signal is richer and the subject is captured sooner. In practice this means the identity-conditioned adapter reaches a usable face-similarity level earlier in training, which partly offsets the modest extra cost of the frozen ArcFace encoder and the small MLP projection.

Taken together, these results support a simple but central claim that the remainder of this thesis will revisit at scale. Identity preservation and prompt adherence trade off

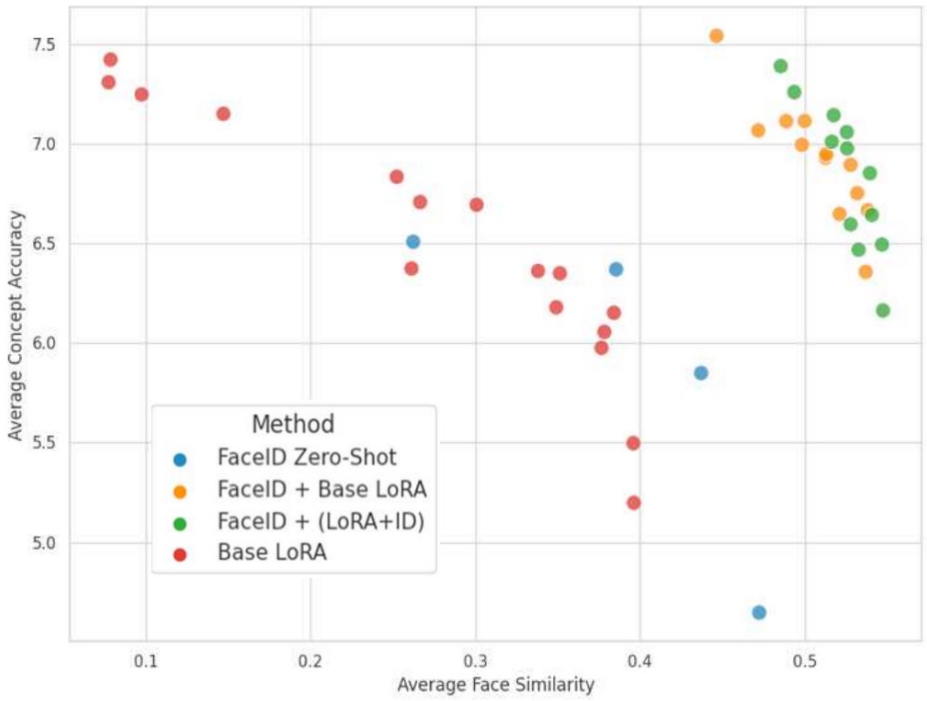


Figure 4.6. Comparison of all four pipelines across face similarity and concept accuracy. FaceID + (LoRA+ID) (green) consistently reaches the highest face similarity while keeping competitive concept accuracy; Base LoRA (red) maximizes concept accuracy but loses identity; FaceID zero-shot (blue) preserves identity at low accuracy; FaceID + Base LoRA (orange) is a balanced trade-off.

against each other, and the most effective way to relax that trade-off is to make identity an explicit part of the training signal rather than an emergent accident of reconstruction. LoRA+ID is one concrete instance of that idea. The benchmark developed in the following chapters generalizes the same concern, measuring how a wide range of parameter-efficient methods navigate this and related tensions across many subjects and backbones.

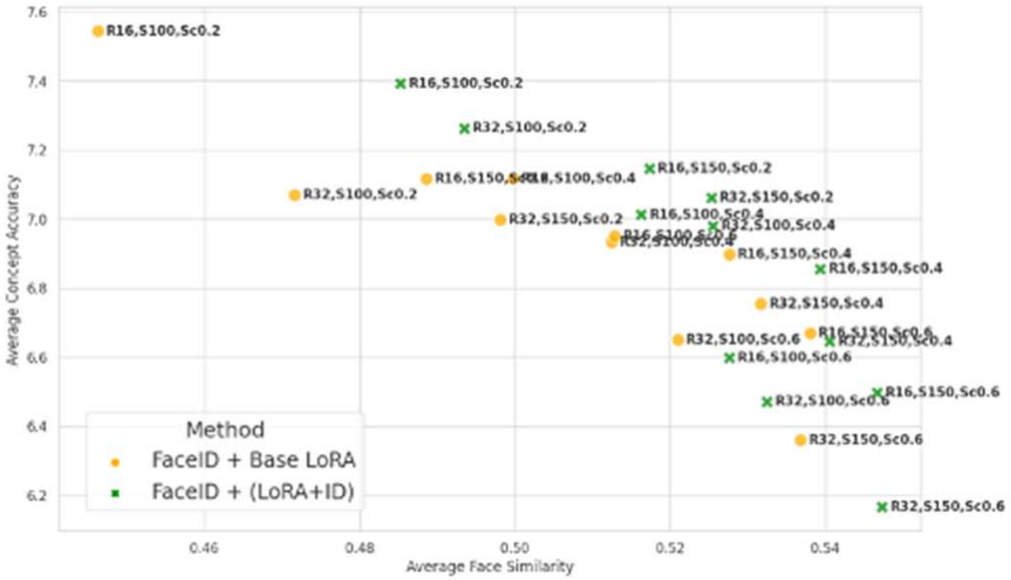


Figure 4.7. Zoomed-in comparison between FaceID + Baseline LoRA (orange) and FaceID + (LoRA+ID) (green), each point annotated with its rank (R), step (S), and scale (Sc). FaceID + (LoRA+ID) consistently reaches higher face similarity while holding concept accuracy competitive, demonstrating the effect of injecting identity embeddings during LoRA training.



(a) Input (b) Base LoRA (c) FaceID (d) FaceID+BaseLoRA (e) FaceID+(LoRA+ID)

Figure 4.8. Qualitative comparison under the prompt “a cinematic photograph of a man riding a motorcycle, wearing sunglasses and a leather jacket, outdoors, sunny weather, dynamic angle.” (a) reference; (b) Base LoRA follows the prompt but loses identity; (c) FaceID holds identity but flattens the style; (d) FaceID + Base LoRA improves both; (e) FaceID + (LoRA+ID) preserves identity while satisfying the prompt most convincingly.

CHAPTER 5

EVALUATION FRAMEWORK: CFS AND APEX

5.1. The Problem with Existing Evaluation

When the personalization methods studied in this thesis were first compared against one another, the same uncomfortable pattern kept recurring. A method would post strong numbers on the standard fidelity metrics — a high CLIP reference similarity, a high DINOv2 match against the training photographs — and yet, when the generated images were laid out and inspected by eye, it was obvious that the model had overfit. Under prompts close to the training distribution it produced the right subject convincingly; under novel prompts the subject's identity drifted, the requested attributes were ignored, or, more insidiously, generations that had nothing to do with the trained subject came out subtly contaminated by its features. None of the metrics in routine use registered any of this.

This is not a question of poor calibration that a threshold adjustment could fix. It is a structural gap in what the standard metrics are built to measure. Reference-based fidelity scores ask a single question — how closely does this generation resemble the training subject? — and a model that has memorized its few training images answers that question extremely well precisely because it has overfit. The very behavior we want to penalize inflates the metric we would use to detect it. What is missing is any measurement of what the adaptation did to everything *outside* the training distribution: whether learning one subject quietly degraded the model's prior knowledge of unrelated concepts, and whether the personalized subject can still share a scene with anything else. We use the term *concept drift* for this collateral effect — the forgetting that a single adaptation induces on concepts the model already knew, in the sense defined in Chapter 2.

Two new instruments are introduced in this chapter to close that gap. The first, the Concept Fidelity Score (CFS), measures drift directly: it compares a personalized model against its own base model on concepts the adaptation never touched, and quantifies how far the two have diverged. The second, the Aggregate Personalization Evaluation indeX (APEX), folds fourteen metrics across six families – with CFS supplying the stability

Table 5.1. Evaluation metrics grouped by objective. Arrows indicate whether higher (↑) or lower (↓) values are preferred.

Family	**Metrics**
Fidelity	DINOv2 Identity (↑), FaceID (↑), CLIP-I (↑)
Prompt Adherence	CLIP-T (↑), VQA Prompt (↑)
Preference Quality	PickScore (↑), ImageReward (↑), HPSv3 (↑)
Style Consistency	Style Loss (↓), LPIPS (↓)
Diversity	Vendi Score (↑), CMMD (↓)
Stability	CFS_{ext} (↓), CFS_{VQA} (↑)

family – into a single composite score, built so that a weakness on any one axis stays visible rather than being averaged away. Together they let the benchmark of Chapter 6 reward methods that learn a subject *stably*, not merely methods that learn it loudly.

5.2. Evaluation Metrics

Personalized generation is inherently multi-objective. A successful method must preserve the target identity, follow the prompt, produce visually convincing outputs, maintain stylistic consistency, generate diverse samples, and avoid degrading concepts unrelated to the one it was trained on. No single scalar captures all of these at once, and optimizing any one of them in isolation tends to sacrifice the others. We therefore evaluate six metric families spanning fourteen metrics, summarized in Table 5.1. The first five families are drawn from established practice; the sixth, Stability, is the contribution of this work and is developed in Section 5.3.

Fidelity. The fidelity family asks how faithfully a generation reproduces the target subject. DINOv2 Identity (Oquab et al. 2024) is the cosine similarity between DINOv2 features of a generated image and the subject's reference images. Because DINOv2's self-supervised features encode object structure and appearance without leaning on class labels, it is a strong, label-agnostic identity signal for both object and human subjects, and we treat it as the primary fidelity metric. CLIP-I compares CLIP (Radford et al. 2021) image embeddings of the generation and the reference; it is softer and more semantic than DINOv2, rewarding overall appearance agreement even when fine detail differs. These two metrics apply to every subject. FaceID adds a complementary signal for human faces alone: it is the cosine similarity between ArcFace (Deng et al. 2019) recognition

embeddings of the generated and reference faces, resolving biometric identity at a precision that the general-purpose visual encoders do not reach.

Prompt Adherence. A faithful subject rendered in the wrong scene is still a failure, so the second family measures whether the generation obeys its prompt. CLIP-T (Radford et al. 2021) is the cosine similarity between the generated image and the prompt text in CLIP space; it is a reliable coarse signal but saturates quickly and is weak on compositional and attribute detail. VQA Prompt (Lin et al. 2024) targets exactly that weakness: a vision–language model is queried about the attributes and object relations that the prompt specifies, and the score reflects how many of them actually hold in the image. It therefore exposes binding and attribute errors that CLIP-T averages away.

Preference Quality. Reproducing a subject correctly does not guarantee an attractive image, and learned human-preference models capture aesthetic and perceptual quality that geometric similarity misses. Because any single preference model carries its own training biases, we average three. ImageReward (Xu et al. 2023) and PickScore (Kirstain et al. 2023) are reward models trained on large pools of human pairwise choices, and HPSv3 (Ma et al. 2025) is a wide-spectrum human-preference predictor. All three are higher-is-better, and combining them dampens the idiosyncrasies of any one.

Style Consistency. Personalization should not silently repaint a subject in a new texture or palette. Style Loss (Gatys, Ecker, and Bethge 2016) measures this as the distance between Gram matrices of VGG features — the classic texture-and-color statistic from neural style transfer — and is lower-is-better. LPIPS (R. Zhang et al. 2018) is a learned perceptual distance calibrated to human judgments; we read it as perceptual deviation from the reference and interpret it alongside diversity, since greater legitimate output variety also raises it.

Diversity. A model that has overfit often collapses onto a single memorized view, which a fidelity metric happily rewards. The diversity family guards against this. The Vendi Score (Friedman and Dieng 2023) is the effective number of distinct images in a generated set, computed from the eigenvalues of a similarity kernel; a low value flags mode collapse. CMMD (Jayasumana et al. 2024) is a maximum mean discrepancy between generated and reference images in CLIP feature space; unlike FID it remains unbiased at the small per-subject sample sizes that the benchmark protocol produces.

Stability. The sixth family is the proposed Concept Fidelity Score, and it is the one

that addresses the gap described in the previous section: it measures how personalization affects the model when it is prompted with concepts that were never part of the adaptation. We develop it in full in Section 5.3.

5.3. Concept Fidelity Score

CFS is built around a single idea: *containment*. A well-behaved adapter should leave the rest of the model alone. When its trigger token is absent from the prompt, the personalized model should behave like the untouched base model; when the trigger is present, the subject should still be willing to share the canvas with other content rather than overwhelming it. CFS turns these two expectations into two complementary probes that, between them, evaluate whether an adaptation has stayed local to its target or has bled out into concepts the training never mentioned.

Both probes need a way to quantify how far personalization has pushed a generation. We use DINOv2 ViT-L/14 embeddings $\mathcal{D}(\cdot)$ for this and define drift as a cosine distance,

$$\delta(x, y) = 1 - \cos\big(\mathcal{D}(x), \mathcal{D}(y)\big) \tag{5.1}$$

where larger values indicate greater deviation between the two images x and y. The measurement is only meaningful if the two images differ *because of the adapter* and not because of sampling noise, so every comparison is made between paired generations that share an identical random seed. With the seed held fixed, the base and personalized models would produce pixel-identical outputs if the adapter changed nothing; any residual drift is therefore attributable to the adaptation itself.

Far-concept drift ($\mathrm{CFS}_{\mathrm{ext}}$). The first probe asks whether learning the subject quietly damaged the model's ability to render things far away from it. The adapter is loaded at full inference scale, but its trigger token is omitted from the prompt; instead the prompt describes an entirely unrelated concept drawn from a fixed pool of sixty categories spanning animals, vehicles, foods, landscapes, instruments, and other everyday things. To keep the probe fixed and reproducible across runs, each subject is assigned its unrelated concept deterministically through a hash of the subject identifier, so the same subject always receives the same far concept. The base model and the personalized model then render this prompt at the same seed, yielding x and y, and $\mathrm{CFS}_{\mathrm{ext}}$ is their

drift under Equation 5.1. A low δ is desired: it means that, away from its trigger, the personalized model still draws the world as the base model would, so the adaptation has had little collateral effect. Because training-free adapters and image-to-image methods never modify the base weights, they cannot drift by construction, and $\mathrm{CFS}_{\mathrm{ext}}$ is therefore reported only for the weight-modifying PEFT methods.

Compositional capacity ($\mathrm{CFS}_{\mathrm{VQA}}$). The second probe asks the opposite question. Here the trigger *is* present, and the prompt deliberately requests the personalized subject together with a second, distinct instance of the same broad category — for example a personalized object placed next to a related object, or a personalized person standing beside a different person. An adapter that has collapsed into always painting its own identity fails this test: the second instance simply never appears, or it appears wearing the trained subject's features. The score is computed by subject type. For faces we use ArcFace recognition embeddings (InsightFace `buffalo_l`). Given two detected faces with embeddings f_i and f_j, we take

$$\mathrm{CFS}_{\mathrm{VQA}} = 1 - \cos(f_i, f_j), \tag{5.2}$$

so that two genuinely distinct identities score near one — the desired outcome — while a duplicated identity, the signature of collapse, scores near zero.

For object subjects, and for the face cases where fewer than two faces are detected and Equation 5.2 cannot be evaluated, we instead use Qwen3-VL-8B-Instruct, an open vision–language model, as a judge. It receives the generated image together with the generation prompt and is asked whether the image contains *both* the personalized subject and the second, different entity that the prompt requested. It rates the outcome on a five-level scale running from complete failure — only one of the two, or neither, is present — to full compliance, where both appear clearly as separate objects. The five levels map onto the values $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, and the assigned value is used as the $\mathrm{CFS}_{\mathrm{VQA}}$ score. Unlike the drift probe, $\mathrm{CFS}_{\mathrm{VQA}}$ is higher-is-better: a high score means the adapter can still place its learned subject in a populated scene rather than collapsing the whole image back into the thing it was trained on. The two probes thus measure stability from two directions — $\mathrm{CFS}_{\mathrm{ext}}$ from the absence of the trigger, $\mathrm{CFS}_{\mathrm{VQA}}$ from its presence — and together they form the Stability family of Table 5.1.

It is worth being explicit about what CFS does and does not do. It says nothing

about how well the model reproduces its target subject; that is the job of the fidelity family. CFS measures what happened to everything else. The two dimensions are independent: a method can pair high target fidelity with either low or high drift, depending on how aggressively it fine-tunes and how well it regularizes. High fidelity at low drift is the ideal, an adapter that learned its subject without disturbing the model around it; high fidelity at high drift signals a method that bought its subject at the cost of the model's general knowledge.

5.4. APEX: Composite Evaluation Index

The fourteen metrics paint a complete picture, but a practitioner choosing among methods needs to collapse that picture into a single comparable number. The danger in doing so is that a flat average lets a strong axis hide a weak one: a method with superb fidelity and collapsed diversity can average out to the same score as a balanced one. APEX is designed to prevent exactly this. It aggregates the metrics *multiplicatively*, by geometric means, so that a low value on any single axis pulls the whole index down and stays visible, and it then applies a small discount for computational cost.

Normalization. The fourteen metrics live on wildly different scales, so each must first be mapped onto a common $[\epsilon, 1]$ range before it can be multiplied with the others. A natural choice would be to min–max scale each metric across the methods under comparison, but that makes every score depend on which other methods happen to be in the set and stretches even negligible differences to fill the whole range. We instead normalize each metric independently against a *fixed reference range*, which keeps each normalized value absolute and comparable across method sets. For a metric M with raw value x and reference range $[\ell_M, u_M]$,

$$\tilde{M} = \min\Big(\max\Big(\frac{x - \ell_M}{u_M - \ell_M},\ \epsilon \Big),\ 1 \Big), \qquad \epsilon = 0.01, \tag{5.3}$$

and lower-is-better metrics (CFS_{ext}, Style Loss, LPIPS, CMMD) instead use $\tilde{M} = \min\big(\max\big((u_M - x)/(u_M - \ell_M),\ \epsilon\big),\ 1\big)$, so that after normalization higher always means better. The small floor $\epsilon = 0.01$ is essential to a multiplicative aggregate: it keeps a worst-in-class metric from collapsing an entire geometric mean to zero and erasing every other axis a method might have earned. Wherever a metric has a theoretical bound we use that bound directly,

and only where one does not exist do we fall back on an empirically observed operating range. Most metrics carry natural ranges: the cosine scores (DINOv2, FaceID, CLIP-I, CLIP-T, PickScore) and the bounded scores (VQA, LPIPS, CFS_{ext}, CFS_{VQA}) lie in $[0, 1]$, the Vendi score in $[1, N]$ with $N = 4$ generated images, and CMMD in $[0, 2]$. Only three metrics are effectively unbounded, so for these we fix operating ranges — HPSv3 in $[0, 15]$, ImageReward in $[-2, 2]$, and Style Loss (scaled by 10^4) in $[0, 3]$ — and release them together with the scoring code so the normalization is fully auditable.

Family scores. With every metric on a common higher-is-better scale, the six families of Section 5.2 are each aggregated into a single score by an equally weighted geometric mean of their members. In the expressions below each metric name denotes its normalized value from Equation 5.3, with the lower-is-better metrics already inverted:

$$F_{\text{face}} = (\text{DINOv2} \cdot \text{FaceID} \cdot \text{CLIP-I})^{1/3}, \qquad F_{\text{object}} = (\text{DINOv2} \cdot \text{CLIP-I})^{1/2},$$
$$R = (\text{PickScore} \cdot \text{ImageReward} \cdot \text{HPSv3})^{1/3}, \qquad P = (\text{CLIP-T} \cdot \text{VQA})^{1/2},$$
$$Y = (\text{StyleLoss} \cdot \text{LPIPS})^{1/2}, \qquad D = (\text{Vendi} \cdot \text{CMMD})^{1/2},$$
$$S = (\text{CFS}_{\text{ext}} \cdot \text{CFS}_{\text{VQA}})^{1/2}.$$

The fidelity family F needs special handling because FaceID is defined only for human faces. We therefore evaluate the face form F_{face} and the object form F_{object} separately and pool them into one per-method fidelity score by their equally weighted geometric mean over the subject types a method covers; a face-only method uses the face form alone. The stability score S likewise combines the two CFS probes of Section 5.3: the inverted far-concept drift and the compositional-capacity score, so that both poor containment and identity collapse pull S down.

Overall APEX. The headline quality score is the geometric mean of the six family scores, weighting none above another:

$$\text{APEX} = (F \cdot P \cdot R \cdot Y \cdot D \cdot S)^{1/6} \tag{5.4}$$

The geometric form is what gives APEX its character: a method that excels on fidelity and preference quality but collapses on stability or diversity cannot hide that collapse behind its strengths, because the weak factor multiplies straight through the mean. This is precisely the behavior the flat-average critique that opened this section called for.

Efficiency discount. Two methods of equal quality are not equally attractive if one costs an order of magnitude more to train and serve, so APEX folds in a modest cost term. We summarize computational cost from four logged signals: adapter storage s (in MB) and training time t_{tr} per subject on the training side, and the backbone parameter count p_{gen} and per-image generation time t_{gen} on the inference side. Because these span several orders of magnitude, we compress sizes and counts with a logarithm and times with a square root, normalize each across methods with min–max scaling $\nu(\cdot)$, and average them into a single cost score,

$$c = \tfrac{1}{4}\Big[\nu(\log(1+s)) + \nu(\sqrt{t_{\text{tr}}}) + \nu(\log p_{\text{gen}}) + \nu(\sqrt{t_{\text{gen}}})\Big], \qquad c \in [0,1]. \tag{5.5}$$

Cost should shape the ranking without ever overruling quality, so it enters as a bounded discount rather than as another factor in the geometric mean,

$$\text{APEX}_E = \text{APEX}\,(1 - \lambda c), \qquad \lambda = 0.05, \tag{5.6}$$

with a small coefficient $\lambda = 0.05$ that lets cost move a method's final score by at most five percent. Because the discount is defined through per-subject training cost, APEX_E is reported only for the PEFT methods; the inference-time adapters and image-to-image models perform no per-subject training and are compared on APEX alone.

A final point of interpretation concerns the training-free methods. Because they never alter the base weights, they cannot drift, and they take the ideal stability value $S = 1.000$ by construction. This is a genuine structural advantage rather than an earned one, and it means the stability axis discriminates only among the weight-modifying PEFT methods. We keep this firmly in mind when reading the results of Chapter 6: the APEX scores of the training-free methods should be understood as an upper bound, inflated by a free pass on the very axis that the rest of this framework was built to measure. To keep the whole index auditable, we release the raw per-metric values together with the APEX scoring script, so that any set of methods can be re-scored and the rankings reproduced.

CHAPTER 6

EXPERIMENTS AND RESULTS

This chapter puts the dataset of Chapter 3 and the metrics of Chapter 5 to work. It evaluates ten personalization methods spanning three paradigms on DreamBooth-Ext, reports their behavior family by family, and then collapses the fourteen metrics into the single APEX index so that whole methods, rather than isolated scores, can be ranked against one another. The recurring theme is that no single number tells the truth on its own: the families disagree often enough that the disagreements themselves are the finding, and CFS in particular is designed to catch the collateral damage that standard fidelity metrics reward. The chapter closes with an honest account of where the measurements are incomplete and of which margins are small enough to be read as effective ties.

6.1. Methods Evaluated

Ten personalization methods are evaluated, grouped into three paradigms that differ in how the subject is injected into the generator.

PEFT methods train a small per-subject adapter while the backbone stays frozen, and they are run on *both* backbones, SDXL 1.0 (Podell et al. 2023) and FLUX.1-dev (Black Forest Labs 2024): LoRA, LoHA, LoKr, AdaLoRA, and SHiRA. With five methods over two backbones this gives ten model–backbone configurations, the core of the benchmark. These methods are the workhorse of modern personalization—they train quickly, store compactly, and reconstruct the subject from a handful of reference images—but they also modify the weights, which is precisely why their stability is worth measuring.

Inference-time adapters require no per-subject training and inject features from a reference image during sampling, run here on SDXL: IP-Adapter (Ye et al. 2023), applied to all subjects, and PuLID (Z. Guo et al. 2024), a human-identity adapter applied only to the face subjects, where face detection succeeds.

Image-to-image models take the reference image directly as a visual conditioning signal with no subject-specific optimization: FLUX.2 Klein 4B, FLUX.2 Klein 9B, and

Qwen-Image-Edit 2511. They are the newest and largest systems in the comparison and serve as a strong upper reference for visual quality.

This three-way taxonomy lets us compare not only generation quality and identity fidelity but also the practical trade-offs in training cost, inference flexibility, and what adaptation does to the rest of the model. LoRA+ID (Chapter 4) is the author's own published method and is treated as the motivating exception to reconstruction-only training; it is discussed there rather than re-benchmarked as one of the adapters here.

6.2. Experimental Setup

6.2.1. Hyperparameter Configuration

Rather than tuning each method on our benchmark, we hold the adapter configuration fixed across methods and both backbones: the low-rank methods (LoRA, LoHA, LoKr) use rank 16 and $\alpha = 16$, AdaLoRA initializes at rank 16, and every method trains at a constant learning rate of 1×10^{-4}, so that differences in the results reflect the methods themselves rather than a configuration search. The configurations are summarized in Table 6.1. For fairness, all PEFT methods also share one training budget rather than per-method schedules—500 steps, batch size 1, gradient clipping at norm 1.0, a DreamBooth-style objective (Ruiz et al. 2023), and no prior-preservation loss—and are trained independently for each subject.

The adapted modules differ by backbone. On SDXL, every method targets the query, key, value, and output projections of the UNet cross-attention layers, leaving the feed-forward layers untouched. On FLUX.1-dev, whose backbone is a diffusion transformer, the adapters target the attention projections of both single- and double-stream blocks together with their feed-forward layers. The method-specific settings follow each method's paper: AdaLoRA initializes at rank 16 and prunes to a target rank of 8 by SVD, and SHiRA trains a sparse high-rank mask rather than a low-rank factorization.

Table 6.1. PEFT training configurations. All methods train for 500 steps, batch size 1, gradient clipping at norm 1.0, and a constant learning rate of 1×10^{-4}, with a DreamBooth-style objective and no prior-preservation loss.

Method	Rank	α	LR	Method-specific
SDXL backbone (1024^2*, AdamW)*				
LoRA	16	16	1e-4	–
LoHA	16	16	1e-4	Hadamard
LoKr	16	16	1e-4	Kronecker
AdaLoRA	16→8	16	1e-4	SVD pruning
SHiRA	16	–	1e-4	sparse mask
FLUX.1-dev backbone (512^2*, 8-bit AdamW)*				
LoRA	16	16	1e-4	–
LoHA	16	16	1e-4	Hadamard
LoKr	16	16	1e-4	Kronecker
AdaLoRA	16→8	16	1e-4	SVD pruning
SHiRA	16	–	1e-4	sparse mask

6.2.2. Implementation Details

All PEFT training ran on a single NVIDIA H100 80 GB GPU, with two base models, SDXL 1.0 and FLUX.1-dev, both in bfloat16. Each adapter is trained independently per subject using a DreamBooth-style objective and no prior-preservation loss; on SDXL the subject is bound to a rare trigger token, whereas the FLUX runs condition on the subject name directly. SDXL trains at 1024×1024 with standard AdamW and gradient checkpointing disabled; FLUX.1-dev trains at 512×512 with 8-bit AdamW, gradient checkpointing, logit-normal timestep weighting, and guidance distilled to a scale of 1.0. Both use batch size 1, gradient clipping at norm 1.0, a constant learning rate of 1×10^{-4}, and 500 steps.

The inference-time adapters and image-to-image models are not trained, but each still requires its own generation configuration, which we fix per method (Table 6.2). All of them use seed 42, bfloat16 precision, 1024×1024 resolution, and four images per prompt. IP-Adapter is applied to all subjects, while PuLID is applied only to the human subjects, where face detection succeeds.

Table 6.2. Generation settings for the inference-time identity adapters and image-to-image models. Common to all: seed 42, bfloat16, 1024×1024, 4 images per prompt.

Method	Steps	Guidance	Other
FLUX.2 Klein 4B	28	3.5	–
FLUX.2 Klein 9B	28	3.5	–
Qwen-Image-Edit 2511	40	1.0	true-CFG 4.0
IP-Adapter	30	7.5	scale 0.9
PuLID (SDXL)	30	7.0	identity scale 0.8

6.2.3. Generation Protocol

For the PEFT methods, images were generated with 30 DDPM (Ho, Jain, and Abbeel 2020) steps, guidance scale 7.5, and a fixed seed scheme (seed + image index) so that every run is reproducible and comparable across methods at a shared seed. Per subject and method, four images were generated per prompt across four context prompts, yielding sixteen generated images per subject and, over the 60 subjects, 960 generations per method. Metric scores are first averaged over images, then over prompts, then over subjects within each category, so that no single noisy generation dominates a method's reported number; dispersion across subjects is retained where available. The CFS stability probes follow the two-component protocol of Section 5.3: with the adapter loaded but the trigger token removed, a hash-assigned concept from a fixed external pool measures far-concept drift against the base model at a shared seed, and a separate trigger-present prompt requests the subject alongside a second entity to probe compositional capacity. Each probe is evaluated with two prompts at two fixed seeds (two images per probe) for every subject, run for each of the five PEFT methods on both backbones, with the base and personalized models sharing the same seeds so that the two outputs differ only because of the adapter.

6.3. Quantitative Results

We report each metric family in turn—fidelity (Table 6.3), prompt adherence (Table 6.4), preference quality (Table 6.5), and style and diversity (Table 6.6)—before turning to stability (Section 6.4) and the overall APEX ranking (Section 6.6). Throughout,

Table 6.3. Subject fidelity. DINOv2 and CLIP-I are pooled across face and object subjects (weighted by subject count) and computed per backbone; FaceID is a faces-only metric. PuLID[†] is evaluated on faces only, so its DINOv2 and CLIP-I are over faces and not directly comparable to the pooled rows. Best per column in **bold**.

Method	DINOv2↑	FaceID↑	CLIP-I↑
PEFT on SDXL/FLUX.1			
LoRA	0.660/0.624	0.481/0.507	0.854/0.846
LoHA	0.477/0.666	0.289/0.596	0.781/0.868
LoKr	0.523/0.637	0.305/0.529	0.800/0.838
AdaLoRA	0.561/0.661	0.382/0.595	0.824/0.852
SHiRA	0.531/0.663	0.326/0.598	0.812/0.863
Image-to-image			
FLUX.2 Klein 4B	0.616	0.595	0.822
FLUX.2 Klein 9B	0.605	0.633	0.823
Qwen-Image-Edit 2511	0.641	**0.665**	0.848
Inference-time adapters			
IP-Adapter (SDXL)	**0.676**	0.343	**0.880**
PuLID (SDXL)[†]	0.514	0.549	0.831

PuLID is evaluated only on the 20 face subjects, so it is reported for completeness but excluded from the best-per-column (**bold**) comparisons.

Subject fidelity needs more than one metric, and the apparent leader makes the point. IP-Adapter tops both appearance metrics, DINOv2 and CLIP-I, but this reflects how it works rather than how well it personalizes: it copies features directly from the reference image, and an image-to-reference similarity score rewards that copying even when little has been re-rendered. Figure 6.3 shows this, and the other tables confirm it—the same method falls far down on FaceID, on prompt adherence, and on the composite, so its appearance lead does not translate into genuine identity or controllability. This is exactly why we report DINOv2, CLIP-I, and FaceID together: the appearance metrics and the biometric one disagree, and FaceID—available only because DreamBooth-Ext adds real human subjects—catches identity that surface similarity misses. Past this case the methods sit close together, and the clearest separation comes from the backbone, with FLUX-based PEFT in most cases surpassing its SDXL counterpart, as expected from a larger and more recent model. Strong appearance fidelity is therefore necessary but not sufficient: it says nothing about prompt following or about what adaptation does to untrained concepts.

Prompt adherence is deliberately a level playing field. Every method is evaluated

Table 6.4. Comparison by prompt adherence metrics. Best per column in **bold**.

Method	CLIP-T ↑	VQA ↑
PEFT on SDXL/FLUX.1		
LoRA	0.323/0.328	0.922/0.920
LoHA	0.324/0.328	0.904/0.909
LoKr	0.328/0.328	0.912/0.907
AdaLoRA	**0.331**/0.327	0.914/0.911
SHiRA	0.317/0.329	0.927/0.918
Image-to-image		
FLUX.2 Klein 4B	0.324	0.935
FLUX.2 Klein 9B	0.326	**0.943**
Qwen-Image-Edit 2511	0.316	0.904
Inference-time adapters		
IP-Adapter (SDXL)	0.262	0.617
PuLID (SDXL)	0.337	0.921

with the same short, generic prompts rather than long, engineered descriptions, and this is a design choice rather than a convenience: detailed prompts would reward whichever model carries the strongest text encoder—typically the most recent ones—and would measure prompt engineering instead of personalization. Keeping the prompts simple isolates the personalization mechanism, and under that control most methods follow the instruction about equally, so this axis separates them little. The image-to-image models hold a slight edge, unsurprising as the newest systems with the strongest language grounding. The informative case is the exception: IP-Adapter collapses, because conditioning on the whole reference image crowds out the text—a failure a fidelity-only evaluation would reward instead of penalize. The two prompt metrics divide the labor: CLIP-T saturates and barely moves, while VQA, by checking attributes and relations explicitly, is what exposes that collapse, so we report both.

Human preference is governed by the backbone, not the adaptation method. The image-to-image models lead, inheriting the aesthetics of large, preference-tuned backbones. The more useful signal is internal to PEFT: FLUX adapters sit well above their SDXL counterparts, and because three independently trained reward models agree on that gap, it reflects a genuine quality difference rather than one model's bias—which is why we average three rather than trust any single one. The inference-time adapters fall in between and expose a limit of preference scoring: IP-Adapter scores poorly overall yet still edges

Table 6.5. Comparison by visual quality and human preference metrics. PickScore is reported on its native cosine scale. Best per column in **bold**.

Method	PickScore ↑	HPSv3 ↑	ImageReward ↑
PEFT on SDXL/FLUX.1			
LoRA	0.214/0.231	4.796/11.219	0.01/1.36
LoHA	0.218/0.229	3.341/11.205	-0.28/1.28
LoKr	0.221/0.231	3.870/11.513	-0.04/1.39
AdaLoRA	0.219/0.228	4.067/11.280	0.08/1.27
SHiRA	0.220/0.230	5.457/11.228	0.38/1.34
Image-to-image			
FLUX.2 Klein 4B	0.238	12.011	1.38
FLUX.2 Klein 9B	0.250	12.143	1.41
Qwen-Image-Edit 2511	**0.255**	**13.172**	**1.54**
Inference-time adapters			
IP-Adapter (SDXL)	0.202	3.956	0.68
PuLID (SDXL)	0.221	4.030	0.52

the SDXL PEFT methods on one reward model, a reminder that these models reward polished, identity-consistent images even when the prompt is ignored.

Style and diversity separate methods only weakly, and are best read as a secondary axis. Style match tracks the backbone almost monotonically: the FLUX-backbone and image-to-image methods adhere most closely, the SDXL PEFT methods least. Diversity is where reporting two metrics pays off, because they need not agree: Qwen produces the most internally varied samples (the highest Vendi score), yet internal variety and closeness to the real-image distribution are different properties that a single number would conflate. The SDXL adapters and IP-Adapter, by contrast, stay closest to the real images, posting the lowest CMMD values. LPIPS hardly moves across methods, so we keep it as a consistency check rather than a discriminator.

6.4. CFS: Catching What Standard Metrics Miss

This is where CFS earns its place, because it disagrees with the fidelity metrics rather than echoing them. AdaLoRA on FLUX is among the most faithful adapters, yet it drifts the most on unrelated concepts; LoHA on SDXL drifts least, precisely because it commits less to the subject. A fidelity-only ranking would therefore reward the method that does the most collateral damage—the blind spot CFS is built to expose. The two probes

Table 6.6. Style and diversity quality metrics, averaged over all subjects. Style Loss is scaled by 10^4. Best per column in **bold**.

Method	StyleLoss ↓	LPIPS ↓	Vendi ↑	CMMD ↓
PEFT on SDXL/FLUX.1				
LoRA	1.156/0.801	0.603/0.635	1.390/1.341	**0.152**/0.313
LoHA	1.797/0.716	0.646/0.622	1.431/1.329	0.312/0.268
LoKr	1.131/0.713	0.643/0.639	1.401/1.320	0.299/0.308
AdaLoRA	1.139/0.650	0.640/0.615	1.370/1.335	0.274/0.265
SHiRA	1.146/0.670	0.634/0.623	1.402/1.328	0.262/0.279
Image-to-image				
FLUX.2 Klein 4B	0.876	**0.555**	1.420	0.329
FLUX.2 Klein 9B	0.653	0.569	1.385	0.344
Qwen-Image-Edit 2511	**0.599**	0.559	**1.457**	0.252
Inference-time adapters				
IP-Adapter (SDXL)	1.065	0.595	1.421	0.199
PuLID (SDXL)	1.042	0.635	1.425	0.295

Table 6.7. CFS metrics for PEFT methods, averaged over subjects. CFS_{ext} measures far-concept drift (lower is better), while CFS_{VQA} measures compositional capacity (higher is better). Training-free methods do not modify model weights and have no drift component. Best per column in **bold**.

Method	CFS_{ext} ↓	CFS_{VQA} ↑
PEFT on SDXL/FLUX.1		
LoRA	0.127/0.120	0.409/0.498
LoHA	**0.081**/0.159	0.509/0.375
LoKr	0.201/0.172	0.531/0.425
AdaLoRA	0.189/0.230	**0.562**/0.386
SHiRA	0.187/0.130	0.494/0.429

are complementary by design: the adapter with the strongest compositional capacity is not the one with the least drift, so neither probe subsumes the other. Containment is thus a distinct axis of personalization quality—a method can reproduce its subject almost perfectly and still quietly reshape everything else the model knew.

The two probes pull apart in instructive ways. The far-concept drift column (CFS_{ext}) flags how much an adapter disturbs concepts it was never meant to touch: LoHA on SDXL is the most contained (0.081) while AdaLoRA on FLUX disturbs the most (0.230), and crucially these are not the same methods that lead on fidelity. Figure 6.1 makes the effect concrete: each column pairs a base-model generation with the personalized



Figure 6.1. **CFS-ext examples.** Each column compares base-model and personalized PEFT generations under the same unrelated prompt and random seed. Although the personalized trigger token is absent, adaptation can alter concepts unrelated to the trained subject. CFS_{ext} measures this deviation as DINOv2 cosine drift between the base and PEFT outputs.

PEFT generation under the same unrelated prompt and seed, and even with the trigger token absent the adapted model visibly alters content it should have left alone. The compositional probe (CFS_{VQA}) is an orthogonal axis, asking whether a personalized model can still place its subject beside a second, related entity. Figure 6.2 shows both outcomes: in the failure cases the second requested entity is absent or indistinct—the model collapses back onto the subject it overfit—while in the successes both entities are rendered cleanly and the score rises accordingly. Read together, the two columns show that low drift and strong composition are different virtues that no single fidelity number can capture.

6.5. Qualitative Analysis

The quantitative tables compress hundreds of generations into a handful of averages; Figure 6.3 shows what those averages look like image by image, under fixed prompts and seeds, for one object subject and one face subject across representative methods from each paradigm.

The object example (top row) asks each method to keep the plush toy's specific appearance while obeying a fixed prompt, and the methods differ visibly in how much of that identity survives the re-rendering. The face example (bottom row) is the more



Figure 6.2. **CFS-VQA examples.** Top row shows the reference subjects. Bottom row shows compositional generations from prompts requesting the personalized subject together with a second related entity. Flash drive: "a photo of USB flash drive together with an external hard drive"; Grey sloth: "a photo of TOK together with a stuffed elephant"; Stapler: "a photo of TOK together with a tape dispenser"; Antonio and Alexandria: "a photo of TOK standing together with a different person"; Bear plushie: "a photo of TOK together with a stuffed rabbit". The flash-drive, Antonio, and Alexandria examples illustrate compositional failures where the second requested entity is absent or not clearly distinguishable; the grey-sloth, stapler, and bear-plushie examples successfully render both requested entities and achieve higher CFS_{VQA} scores.

revealing of the two: inference-time identity injection holds the likeness most tightly—IP-Adapter copies the reference appearance almost verbatim—while the reconstruction-only PEFT adapters drift further from the reference, the qualitative counterpart to the FaceID gap in Table 6.3. The image-to-image models occupy a middle ground, producing the most polished images of the row while preserving identity better than the trained adapters but less literally than the inference-time injection. The figure therefore mirrors the quantitative story: appearance copying buys raw similarity at the cost of controllability, and no single column is best on every count.

6.6. APEX Rankings and Cross-Paradigm Comparison

Read across families, the central finding is that no method wins everywhere—which is exactly what APEX is built to surface. Qwen leads the quality composite and SHiRA on FLUX is the strongest PEFT method, but the family columns show why the single number alone would mislead: Qwen's lead rests partly on the ideal stability score that training-free methods receive by construction, and the PEFT leaders are separated by



Figure 6.3. **Personalization across paradigms.** Top row: object personalization (Wolfie). Bottom row: face personalization (Barack Obama). Representative PEFT, image-to-image, and inference-time methods are shown under fixed prompts and seeds.

Table 6.8. Computational cost feeding the efficiency term of APEX, reported as SDXL/FLUX.1. Columns are trainable parameter count (M), adapter size on disk (MB), training time per subject (s), and warm per-image generation time (s/img); all are complete for both backbones. The efficiency score additionally uses the backbone parameter count (2.6 B for SDXL, 12 B for FLUX.1-dev), so a cost is defined for every PEFT method.

Method	Params(M)	Size(MB)	Train(s)	Gen(s/img)
PEFT on SDXL/FLUX.1				
LoRA	5.81/44.83	22.3/86.0	251/362	2.18/2.28
LoHA	11.61/89.65	44.6/172.0	327/414	2.30/3.06
LoKr	0.72/1.74	3.1/3.5	293/401	2.24/2.97
AdaLoRA	11.62/44.83	44.6/86.0	285/220	2.26/2.44
SHiRA	17.42/44.83	66.6/428.0	241/411	1.83/2.83

margins too small to call a winner. APEX's role is not to crown one method but to make the trade-offs legible—fidelity, preference, and style favor the FLUX backbone, while SDXL keeps a slight edge in stability and diversity. The efficiency-adjusted score then folds in cost, leaving LoRA the most efficient choice on each backbone. Reporting the composite beside its families is the honest presentation: one number for a quick ranking, six for the reasons behind it.

Table 6.8 collects the cost terms that feed the efficiency discount. The adapter sizes, training times, and generation times are complete for both backbones. On FLUX.1-dev, LoKr is the lightest adapter at 1.74 M trainable parameters and 3.5 MB on disk, SHiRA the heaviest on disk at 428 MB, and AdaLoRA the fastest to train at 220 s per subject; on SDXL the adapters are smaller and quicker overall, with LoKr again the most

Table 6.9. APEX rankings. F, P, R, Y, D, S are the six normalized family scores (fidelity, prompt adherence, preference quality, style, diversity, stability; higher is better), APEX their geometric mean, and Final APEX_E the cost-discounted score. Each metric is normalized to $[\epsilon, 1]$ by a fixed reference range (Equation 5.3, ϵ=0.01). The heavy rule separates the weight-modifying PEFT methods (above), whose stability is measured, from the training-free methods (below), which take the ideal stability value by construction. Best Final APEX and APEX_E in **bold**.

Method	F	P	R	Y	D	S	APEX	APEX_E
PEFT on SDXL								
LoRA	0.685	0.546	0.325	0.494	0.347	0.598	0.481	**0.476**
LoHA	0.536	0.541	0.276	0.377	0.348	0.684	0.439	0.431
LoKr	0.565	0.547	0.304	0.472	0.337	0.651	0.462	0.458
AdaLoRA	0.612	0.550	0.314	0.473	0.326	0.675	0.472	0.465
SHiRA	0.578	0.542	0.363	0.476	0.341	0.634	0.476	0.472
PEFT on FLUX.1								
LoRA	0.681	0.549	0.525	0.517	0.310	0.662	0.525	**0.507**
LoHA	0.724	0.546	0.520	0.536	0.308	0.562	0.517	0.492
LoKr	0.689	0.545	0.531	0.525	0.300	0.593	0.515	0.496
AdaLoRA	0.717	0.546	0.519	0.549	0.311	0.545	0.516	0.502
SHiRA	0.724	0.550	0.524	0.541	0.307	0.611	0.526	0.501
Image-to-image								
FLUX.2 Klein 4B	0.686	0.550	0.544	0.561	0.342	1.000	0.583	–
FLUX.2 Klein 9B	0.686	0.554	0.556	0.581	0.326	1.000	0.585	–
Qwen-Image-Edit 2511	0.718	0.534	0.583	0.594	0.365	1.000	**0.604**	–
Inference-time adapters (SDXL)								
IP-Adapter	0.671	0.402	0.330	0.511	0.355	1.000	0.503	–
PuLID	0.617	0.557	0.334	0.488	0.348	1.000	0.519	–

compact at 0.72 M trainable parameters and 3.1 MB, and SHiRA the fastest to generate at 1.83 s per image. The cost score c of Equation 5.5 combines adapter size, training time, per-image generation time, and the backbone parameter count, and is therefore defined for every PEFT method on both backbones. Because the SDXL backbone is far smaller than FLUX.1-dev (2.6 B parameters against 12 B) and its adapters are lighter and faster, the SDXL configurations incur a smaller cost than their FLUX.1-dev counterparts, and the efficiency-adjusted APEX_E in Table 6.9 reflects this. APEX_E is reported only for the PEFT methods, since the training-free methods perform no per-subject training and have no comparable cost. Because the discount coefficient is small ($\lambda = 0.05$), the gap between APEX and APEX_E stays small and the quality composite carries most of the ranking.

Three conditions shape how Table 6.9 should be read. First, the cost discount is defined for every PEFT method on both backbones (Table 6.8), but APEX_E is reported

only for the PEFT rows, since the training-free methods perform no per-subject training and have no comparable cost. Second, the training-free methods—both image-to-image models and the inference-time adapters—receive the ideal stability value $S = 1.000$ by construction, because they never modify the base weights and so have no drift to measure; their APEX is thus an *upper bound* rather than a measured-stability score, and the gap to the PEFT methods, whose S is genuinely penalized, should be discounted accordingly. Third, the heavy rule in the table marks exactly this divide between measured and assumed stability.

With those conditions in mind, the rankings tell a consistent story. Qwen-Image-Edit leads the quality composite (APEX = 0.604), helped by that free stability term — but the lead does not depend on it: even setting stability aside and comparing on the other five families alone, the image-to-image models still outscore every PEFT adapter, so the free $S = 1.000$ inflates the margin without manufacturing the ordering. The two FLUX.2 Klein models follow. Among the methods whose stability is actually measured, SHiRA and LoRA on FLUX sit at the top of the PEFT field (APEX = 0.526 and 0.525), but their separation is well inside the ~ 0.01 noise band and is best read as a tie rather than a strict ordering; the firmer statement is that the FLUX backbone dominates SDXL essentially everywhere—every FLUX PEFT row outscores its SDXL twin on the composite. The family columns explain why a single number would mislead: SDXL keeps a real edge on stability (S) and a slight one on diversity (D), while FLUX wins fidelity, preference, and style, so the composite hides genuine disagreements. After the efficiency discount, LoRA is the most efficient choice on each backbone ($\text{APEX}_E = 0.476$ on SDXL, 0.507 on FLUX), reflecting its small adapter and fast training. The headline is not a single winner but a map of trade-offs: no method leads on every axis, and the six family scores are what let a practitioner choose by priority rather than by a single ranking. Several of the margins are admittedly small — the FLUX LoRA, LoHA, and AdaLoRA cluster falls within about 0.01 on the composite — and differences of this size should be read as effective ties rather than as a strict ranking. To keep the index auditable on this point, the raw per-subject scores and the scoring script are released, so that the composite can be recomputed under alternative family weights, normalization ranges, or cost coefficients by anyone who wishes to test how far a given ordering depends on these design choices.

CHAPTER 7

CONCLUSION

This thesis began with one problem and, in the course of trying to solve it rigorously, uncovered another. The first problem was concrete: LoRA-based fine-tuning adapts a diffusion model to a subject quickly and cheaply, but it does not hold onto facial identity once the generation prompt drifts away from the training context. That observation motivated LoRA+ID, an identity-conditioned training method that injects ArcFace embeddings into the adaptation signal so that the trigger token learns identity-discriminative features rather than the surface statistics of a handful of reference photographs. Validating LoRA+ID against its competitors, however, exposed the second problem: the evaluation infrastructure for PEFT-based personalization was not good enough to support the comparisons we wanted to make. Subject-fidelity scores rewarded methods that overfit; no metric measured the collateral forgetting that aggressive adaptation induces; and there was no single, principled way to weigh fidelity, prompt adherence, perceptual quality, diversity, and stability against one another, let alone across the three very different families of personalization methods. The remainder of the work — the DreamBooth-Ext benchmark, the Concept Fidelity Score (CFS), and the APEX composite index — grew directly out of that gap. The narrative arc is therefore a single one: a fidelity problem led to a method, and evaluating the method honestly led to a measurement framework.

7.1. What Was Shown

LoRA+ID is effective and efficient. Conditioning LoRA training on facial identity embeddings, rather than on reconstruction loss alone, gives the adapter a richer and more targeted learning signal. By projecting ArcFace identity features into the model's token space and training under joint text–identity conditioning, LoRA+ID learns to bind the trigger token to what makes a face that particular face, instead of to the incidental lighting, pose, and background of the training images. The published results show that this yields better face similarity under diverse, out-of-context prompts and faster convergence than

a standard LoRA baseline, at no additional inference cost (Özbay, Ulusoy, and Baştanlar 2025). The full method, its identity-injection pipeline, and its ablations are described in Chapter 4.

DreamBooth-Ext is a releasable, structured benchmark. Evaluating personalization fairly requires data that tests generalization rather than memorization, and that can be shared without redistributing faces under contested licenses. DreamBooth-Ext contributes a fully releasable, Creative Commons human-face subset paired with a self-photographed, license-free object subset, together with structured context prompts that deliberately push each subject into novel scenes, actions, and styles. Each image is distributed with its source URL and license, so the benchmark can be reconstructed and audited. The dataset, its construction, and the prompt design are detailed in Chapter 3.

CFS measures concept drift directly. Existing metrics can tell us how faithfully a method reproduces its target subject, but they are silent on the cost of that fidelity: adapting hard to one subject can quietly corrupt the model's prior knowledge of related concepts. CFS makes that corruption observable and reproducible. Through two probes — far-concept drift (CFS_{ext}, lower is better) and compositional capacity (CFS_{VQA}, higher is better) — it turns an abstract worry about forgetting into a quantity that can be reported, compared, and improved. The metric is simple by design, but it captures something real and method-dependent, as the per-method results in Chapter 5 make clear.

APEX unifies the three paradigms. The final contribution is an aggregate index that brings PEFT fine-tuning, inference-time adapters, and image-to-image editing into a single comparison. APEX, the Aggregate Personalization Evaluation indeX, scores each method across six metric families — fidelity, prompt adherence, preference quality, style consistency, diversity, and stability — and combines them with an equal-weight geometric mean rather than a flat average, so that a method cannot mask a catastrophic weakness in one family with strength in another. A small efficiency discount further rewards methods that achieve their quality at lower cost. The construction of CFS and APEX is given in Chapter 5, and the full benchmark in Chapter 6. Across ten methods and sixty subjects, the central finding is a negative one in the best scientific sense: no single paradigm dominates. PEFT methods, inference-time adapters, and image-to-image models each win on some dimensions and lose on others, and the trade-offs only become legible when the entire metric suite is read together, not as a single headline number.

7.2. Limitations

These conclusions are real, but they are bounded, and it is worth being candid about where.

The LoRA+ID results in the published paper rest on **only eight identities** (Özbay, Ulusoy, and Başţanlar 2025). That is enough to demonstrate the effect, but it is a small sample, and the method's advantage should be read as evidence rather than as a settled, broadly quantified margin. DreamBooth-Ext's face subset of **twenty subjects** improves on this and is the basis for the framework's identity experiments, yet twenty public figures still offer limited coverage of age, ethnicity, and the varied imaging conditions that real deployments encounter.

The **efficiency term is a deliberately light proxy** rather than a precise cost model. It folds four signals—adapter size, training time, per-image generation time, and backbone parameter count—into a single discount, with generation time measured at each backbone's native operating resolution (1024^2 for SDXL, 512^2 for FLUX.1-dev) rather than at a shared hardware configuration. Because the discount coefficient is small ($\lambda = 0.05$), APEX is best read as a quality index lightly adjusted for cost, not as a benchmark of raw computational efficiency.

The **stability family applies only to methods that modify weights**. CFS measures the collateral drift that weight modification induces, so inference-time adapters and image-to-image models—which leave the base weights untouched—cannot drift and take the ideal value by construction. This is the intended behavior of the metric rather than an artifact: it states explicitly that these methods gain stability structurally, not by earning it on a contested axis. The framework keeps this transparent by reporting every metric family alongside the composite, and the central finding that no paradigm dominates rests on that per-family pattern, which holds however the stability term is assigned.

APEX is a transparent, reproducible index rather than a model fitted to data. Its choices are deliberate and made in the open: equal family weights, so no axis is privileged by a hidden agenda, and fixed reference ranges—the metrics' theoretical bounds where these exist and their observed operating ranges otherwise—so each normalized score is absolute and independent of which methods happen to be in the comparison. Because the raw scores and the scoring script are released, any reader can re-weight or re-range

the composite and reproduce the result. In practice the ordering tracks the per-metric tables rather than producing surprising reversals, which is exactly the behavior a sound aggregate should show: APEX is an auditable summary of explicit priorities, built so that its conclusions can be checked against the evidence they rest on rather than taken on trust.

Finally, CFS measures drift in the feature space of DINOv2 (Oquab et al. 2024), a strong, label-agnostic encoder whose features capture structural and semantic content—the dimension along which collateral concept damage is most visible, and exactly what the far-concept probe is meant to detect. Like any feature-based metric, CFS inherits its encoder's emphasis: change that lives almost entirely in fine-grained texture or high-frequency detail registers less strongly than structural change. CFS is therefore best understood as a reproducible drift proxy over structural–semantic content, and pairing it with a texture-sensitive encoder is a natural refinement. This bounds what CFS claims to measure without affecting the drift it does detect, which is large and clearly method-dependent in the results of Chapter 6.

7.3. Ethical Considerations

Personalized generation of human faces carries real risks, and the benchmark was built with those risks in mind. The face subset of DreamBooth-Ext contains **only photographs of widely photographed public figures** released under Creative Commons licenses; it contains no private individuals. Each image is distributed with its source URL and license, so its provenance can be checked. Restricting the subset to public figures keeps the consent footprint small, since these photographs already circulate broadly, and it avoids the redistribution concerns attached to large face corpora.

It must nonetheless be stated plainly that a **Creative Commons license covers the copyright of a photograph, not any biometric or identity consent** of the person depicted. DreamBooth-Ext is intended as an instrument for evaluation research on personalization methods, not as a deployable identity-generation product. Personalization is dual-use: the same techniques that this thesis evaluates can be turned to producing non-consensual or deceptive imagery of real people. The benchmark is released to study and strengthen the evaluation of these methods, with the explicit request that it be used only for research. For any downstream deployment built on methods like these, we recommend explicit consent

checks for depicted individuals, provenance signals such as watermarking on generated images, and active monitoring for misuse.

7.4. Future Work

Several directions follow naturally from what has, and has not, been done here.

Quantifying the robustness of the rankings. The most immediate work is to stress-test the index numerically on the released scores — jittering the family weights, sweeping the normalization floor and cost coefficient, and applying paired significance tests to the close FLUX cluster — so that the stability of the rankings is demonstrated quantitatively rather than argued from the per-metric tables.

DiT architectures. LoRA+ID was designed for the UNet attention of SDXL. Diffusion transformers such as FLUX.1-dev use joint image–text attention with a different internal structure, and adapting the identity-injection strategy to that setting — where text and image tokens are processed together rather than through cross-attention — is a substantial and natural extension.

Multi-concept personalization. The present work handles a single subject at a time. Generating a scene that contains two specific people, or a specific person interacting with a specific object, raises a harder problem of compositional identity binding, where multiple learned concepts must coexist without bleeding into one another. Neither the method nor the current evaluation framework addresses this, and both would need to be extended to do so.

Learned metric weights. The APEX family weights are set by hand. A promising alternative is to learn them from human preference judgments, fitting the composite so that its ordering matches what people actually prefer rather than what the designer assumes they should. This would turn APEX from a principled-but-manual index into a calibrated one, and would let the robustness analysis be framed around a learned reference rather than a chosen one.

Video generation. The evaluation framework extends naturally to video models, where identity must remain stable not only across prompts but across frames. Temporal identity consistency — does a personalized subject stay the same person as the camera and action move — is a dimension that image-only metrics cannot see, and CFS-style drift

measurement could be a useful starting point for it.

Broader identity coverage. Finally, both LoRA+ID and DreamBooth-Ext would benefit from a larger and more diverse identity set, with greater age range, broader demographic representation, and more varied reference-image conditions. Expanding coverage is the surest way to confirm that the findings reported here generalize beyond the specific subjects on which they were measured, and to guard against evaluation conclusions that hold only for a narrow slice of faces.

REFERENCES

Bhardwaj, Kartikeya, Nilesh Prasad Pandey, Sweta Priyadarshi, Viswanath Ganapathy, Shreya Kadambi, Rafael Esteves, Shubhankar Borse, et al. 2024. "Sparse High Rank Adapters (SHiRA)." In *Advances in Neural Information Processing Systems (NeurIPS).* ArXiv:2406.13175.

Bini, Massimo, Leander Girrbach, and Zeynep Akata. 2025. "DeLoRA: Decoupling Angles and Strength in Low-rank Adaptation." In *International Conference on Learning Representations (ICLR).* ArXiv:2503.18225.

Black Forest Labs. 2024. *FLUX.1.* https://blackforestlabs.ai/announcing-black-forest-labs/. Open-weight rectified-flow text-to-image model; code at https://github.com/black-forest-labs/flux.

———. 2025. *FLUX.2: Frontier Visual Intelligence.* https://blackforestlabs.ai.

Borse, Shubhankar, Shreya Kadambi, Nilesh Prasad Pandey, Kartikeya Bhardwaj, Viswanath Ganapathy, Sweta Priyadarshi, Risheek Garrepalli, Rafael Esteves, Munawar Hayat, and Fatih Porikli. 2024. *FouRA: Fourier Low Rank Adaptation.* ArXiv:2406.08798.

Brooks, Tim, Aleksander Holynski, and Alexei A Efros. 2023. "InstructPix2Pix: Learning to Follow Image Editing Instructions." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Caron, Mathilde, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. "Emerging Properties in Self-Supervised Vision Transformers." In *Proceedings of the IEEE/CVF International Conference on Computer Vision,* 9650–9660.

Deng, Jiankang, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 4690–4699.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *arXiv preprint arXiv:2010.11929.*

Friedman, Dan, and Adji Bousso Dieng. 2023. "The Vendi Score: A Diversity Evaluation Metric for Machine Learning." *Transactions on Machine Learning Research (TMLR).*

Gal, Rinon, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion." In *International Conference on Learning Representations (ICLR).* ArXiv:2208.01618.

Gatys, Leon A, Alexander S Ecker, and Matthias Bethge. 2016. "Image Style Transfer Using Convolutional Neural Networks." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Guo, Jia, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. 2022. "Sample and Computation Redistribution for Efficient Face Detection." In *International Conference on Learning Representations (ICLR).* SCRFD; arXiv:2105.04714.

Guo, Zinan, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. 2024. "PuLID: Pure and Lightning ID Customization via Contrastive Alignment." In *Advances in Neural Information Processing Systems (NeurIPS).* ArXiv:2404.16022.

Guo, Zirun, and Tao Jin. 2025. "ConceptGuard: Continual Personalized Text-to-Image Generation with Forgetting and Confusion Mitigation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* ArXiv:2503.10358.

Heusel, Martin, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium." In *Advances in Neural Information Processing Systems,* vol. 30.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel. 2020. "Denoising Diffusion Probabilistic Models." *Advances in Neural Information Processing Systems* 33:6840–6851.

Hu, Edward J, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhen Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. "LoRA: Low-Rank Adaptation of Large Language Models." In *International Conference on Learning Representations.*

Jayasumana, Sadeep, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. 2024. "Rethinking FID: Towards a Better Evaluation Metric for Image Generation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* CMMD; arXiv:2401.09603.

Jha, Saurav, Shiqi Yang, Masato Ishii, Mengjie Zhao, Christian Simon, Muhammad Jehanzeb Mirza, Dong Gong, Lina Yao, Shusuke Takahashi, and Yuki Mitsufuji. 2025. "Mining Your Own Secrets: Diffusion Classifier Scores for Continual Personalization of Text-to-Image Diffusion Models." In *International Conference on Learning Representations (ICLR).* ArXiv:2410.00700.

Karras, Tero, Samuli Laine, and Timo Aila. 2019. "A Style-Based Generator Architecture for Generative Adversarial Networks." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 4401–4410.

Kim, Wonjae, Bokyung Son, and Ildoo Kim. 2021. "ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision." *arXiv preprint arXiv:2102.03334.*

Kirstain, Yuval, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. "Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation." In *Advances in Neural Information Processing Systems (NeurIPS).*

Kumari, Nupur, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. "Multi-Concept Customization of Text-to-Image Diffusion." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* ArXiv:2212.04488.

Lee, Tony, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, et al. 2023. "Holistic Evaluation of Text-to-Image Models." In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track.* ArXiv:2311.04287.

Li, Junnan, Dongxu Li, Caiming Xiong, and Steven Hoi. 2023. “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models.” *arXiv preprint arXiv:2301.12597.*

Lin, Zhiqiu, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. “Evaluating Text-to-Visual Generation with Image-to-Text Generation.” In *European Conference on Computer Vision (ECCV).* VQAScore; arXiv:2404.01291.

Liu, Weiyang, Zeju Qiu, Yao Feng, Yuliang Xiu, Yuxuan Xue, Longhui Yu, Haiwen Feng, et al. 2024. “Parameter-Efficient Orthogonal Finetuning via Butterfly Factorization.” In *International Conference on Learning Representations (ICLR).* BOFT; arXiv:2311.06243.

Ma, Yuhang, Yunhao Shui, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. 2025. “HPSv3: Towards Wide-Spectrum Human Preference Score.” In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).* ArXiv:2508.03789.

Meng, Chenlin, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. “SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations.” In *International Conference on Learning Representations (ICLR).*

Moon, Saemi, Minjong Lee, Sangdon Park, and Dongwoo Kim. 2025. “Holistic Unlearning Benchmark: A Multi-Faceted Evaluation for Text-to-Image Diffusion Model Unlearning.” In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).* ArXiv:2410.05664.

Oquab, Maxime, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2024. “DINOv2: Learning Robust Visual Features without Supervision.” *Transactions on Machine Learning Research (TMLR).*

Özbay, Kutay, Anıl Erdem Ulusoy, and Yalın Baştanlar. 2025. “LoRA+ID: Enhancing Identity Preservation in Generative Models through Face-Conditioned LoRA Training.” In *2025 Innovations in Intelligent Systems and Applications Conference (ASYU).* IEEE. https://doi.org/10.1109/ASYU67174.2025.11208451.

Peng, Yuang, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. 2024. “DreamBench++: A Human-Aligned Benchmark for Personalized Image Generation.” *arXiv preprint arXiv:2406.16855.*

Podell, Dustin, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Peng, and Robin Rombach. 2023. “SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis.” *arXiv preprint arXiv:2307.01952.*

Qiu, Zeju, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. 2023. “Controlling Text-to-Image Diffusion by Orthogonal Finetuning.” In *Advances in Neural Information Processing Systems (NeurIPS).* OFT; arXiv:2306.07280.

Qwen Team. 2025. “Qwen-Image Technical Report.” *arXiv preprint arXiv:2508.02324.*

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. “Learning Transferable Visual Models From Natural Language Supervision.” *arXiv preprint arXiv:2103.00020.*

Ramesh, Aditya, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. “Zero-Shot Text-to-Image Generation.” *arXiv preprint arXiv:2102.12092.*

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. “High-Resolution Image Synthesis with Latent Diffusion Models.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 10684–10695.

Ruiz, Nataniel, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. “DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 22500–22510.

Saha, Shaswati, Sourajit Saha, Manas Gaur, and Tejas Gokhale. 2025. “Side Effects of Erasing Concepts from Diffusion Models.” In *Findings of the Association for Computational Linguistics: EMNLP.* ArXiv:2508.15124.

Saharia, Chitwan, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding.” *Advances in Neural Information Processing Systems* 35:36479–36494.

Suriyakumar, Vinith M., Rohan Alur, Ayush Sekhari, Manish Raghavan, and Ashia C. Wilson. 2025. “Unstable Unlearning: The Hidden Risk of Concept Resurgence in Diffusion Models.” In *International Conference on Learning Representations (ICLR) Workshop (DATA-FM).* ArXiv:2410.08074.

Wang, Qixun, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. 2024. “InstantID: Zero-shot Identity-Preserving Generation in Seconds.” *arXiv preprint arXiv:2401.07519.*

Wang, Yibin, Zhimin Li, Yuhang Zang, Jiazi Bu, Yujie Zhou, Yi Xin, Junjun He, et al. 2025. “UniGenBench++: A Unified Semantic Evaluation Benchmark for Text-to-Image Generation.” *arXiv preprint arXiv:2510.18701.*

Wu, Xiaoshi, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. “Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis.” *arXiv preprint arXiv:2306.09341.*

Xu, Jiazheng, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. “ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation.” In *Advances in Neural Information Processing Systems.*

Ye, Hu, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. “IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models.” *arXiv preprint arXiv:2308.06721*.

———. 2024. *IP-Adapter-FaceID: Face ID Embedding for Identity-Preserving Generation*. https://huggingface.co/h94/IP-Adapter-FaceID. Model release in the IP-Adapter project.

Yeh, Shih-Ying, Yu-Guan Hsieh, Zhidong Gao, Bernard B W Yang, Giyeong Oh, and Yanmin Gong. 2024. “Navigating Text-To-Image Customization: From LyCORIS Fine-Tuning to Model Evaluation.” In *International Conference on Learning Representations (ICLR)*. ArXiv:2309.14859.

Yuan, Shen, Haotian Liu, and Hongteng Xu. 2024. “Bridging the Gap between Low-rank and Orthogonal Adaptation via Householder Reflection Adaptation.” In *Advances in Neural Information Processing Systems (NeurIPS)*. HRA; arXiv:2405.17484.

Zhang, Qingru, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. “AdaLoRA: Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning.” In *International Conference on Learning Representations*.

Zhang, Richard, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. “The Unreasonable Effectiveness of Deep Features as a Perceptual Metric.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.