Avatar Project with SDXL: Comparing DreamBooth-LoRA, LoRA, and Textual Inversion

Anh Dao

May 8, 2025

Abstract

In this project, I conduct a comprehensive benchmark of three personalization techniques for Stable Diffusion XL: (1) DreamBooth-LoRA, (2) LoRA adapters, and (3) Textual Inversion—implemented in three codebases (Diffuser, Kohya, Cog). Using a 24-image dataset of a single subject (`<itay>`), I tune hyperparameters (LoRA ranks: {4,8,16,32,64}; TI vectors: {1,2,4,8,16}), evaluate at 10-epoch intervals on 10 positive and 1 negative prompt, and measure CLIP-T and ArcFace similarity. Then I present convergence curves for each method, summary tables of optimal settings, bar charts of best metrics, and qualitative examples for a key prompt. Finnaly, I choose the best model setup then doing finetuning with (`<irit>`).

1 Introduction

Personalized text-to-image generation enables creation of custom avatars and brand assets. Parameter-efficient methods such as LoRA adapters and Textual Inversion offer trade-offs between memory footprint, convergence speed, and output fidelity. I compare three techniques across five training configurations to identify optimal approaches for small-scale personalization tasks.

2 Dataset & Prompts

2.1 Training Data

- 24 high-resolution face images of a single subject stored in `data/itay/`.
- Diverse poses and lighting conditions to ensure coverage.

2.2 Inference Prompts

Table 1 lists the 10 positive prompts generated by **ChatGPT** and 1 negative prompt used for quantitative evaluation.

Table 1: Summary of Inference Prompts

ID	Prompt Description
0	Professional headshot: frontal face, suit, blurred glass background
1	Preppy fashion portrait: old-money style, country club aesthetic
2	Neon-lit rooftop: bomber jacket, city glow, film-noir mood
3	Monochrome studio: high-contrast B&W profile
4	Golden-hour countryside: wheat field at sunset
5	Rembrandt-style studio: black turtleneck, moody shadows
6	Dramatic jungle: tribal paint, misty waterfall backdrop
7	Cozy indoor lifestyle: reading a book in Scandinavian interior
8	Dark academia library: vintage wooden table, tungsten lights
9	Rainy-window café: raindrop blur, streetlight bokeh
N	Negative: asian, back view, multiple heads, bad anatomy, nsfw, etc.

3 Training Configuration

Table 2 summarizes training settings for each method.

Table 2: Training Configurations

Method	Codebase	Hyperparameters	Epoch Range	Save Interval	Replicates
DreamBooth-LoRA	Diffuser	Rank {4, 8, 16, 32, 64}	10–600	Every 10	1
LoRA	Diffuser	Rank {4, 8, 16, 32, 64}	10–600	Every 10	1
Textual Inversion	Diffuser	Vectors {1, 2, 4, 8, 16}	10–300	Every 10	5
LoRA	Kohya	Rank {4, 8, 16, 32, 64}	10–300	Every 10	1
LoRA	Cog	Rank {4, 8, 16, 32, 64}	10–600	Every 10	1

4 Evaluation Metrics

To quantify both prompt-fidelity and identity-fidelity of the generated avatars, I use two complementary scores:

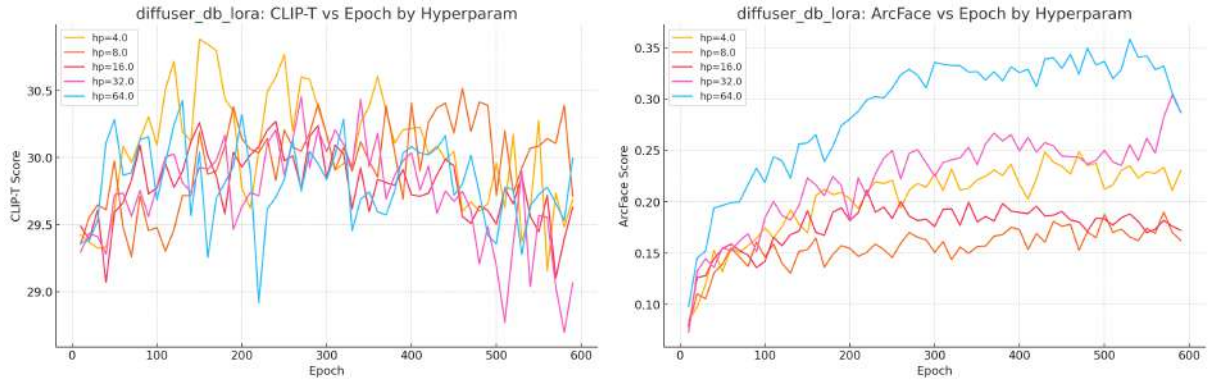
- **CLIP-T Score.** Measures semantic alignment between each generated image and its text prompt via OpenAI's CLIP model. Higher values indicate that the image more faithfully depicts the prompt. I compute this with the `torchmetrics` package using the `openai/clip-vit-large-patch14` checkpoint.
- **ArcFace Similarity.** Quantifies how closely a generated image preserves the subject's identity by comparing its facial embedding to those of the training set using the ArcFace model. Embeddings are extracted via the `insightface` library, and cosine similarity is used as the metric. Higher similarity reflects stronger identity preservation.

For each hyperparameter setting and epoch checkpoint, I generate images at two fixed seeds across ten positive prompts. CLIP-T and ArcFace scores are computed per image, then averaged over all prompts and seeds to yield the reported convergence curves.

5 Quantitative Results

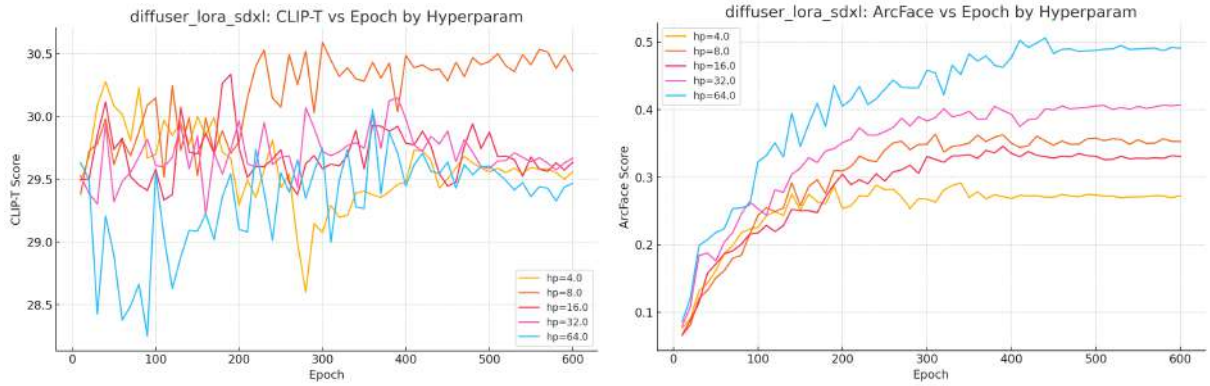
5.1 Convergence Curves per Method

Figures 1–5 show CLIP-T and ArcFace similarity versus epoch for each method across its hyperparameter sweep.



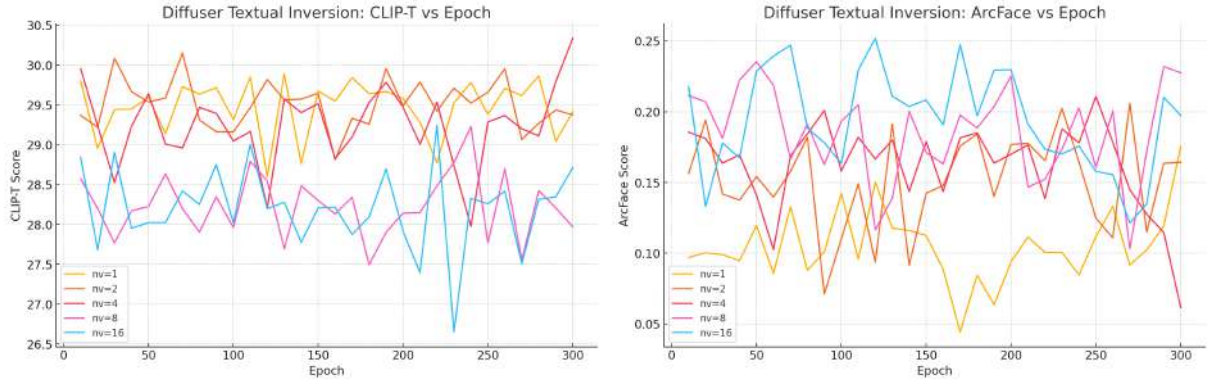
(a) CLIP-T (b) ArcFace

Figure 1: DreamBooth-LoRA (Diffuser) convergence over hyperparameters.



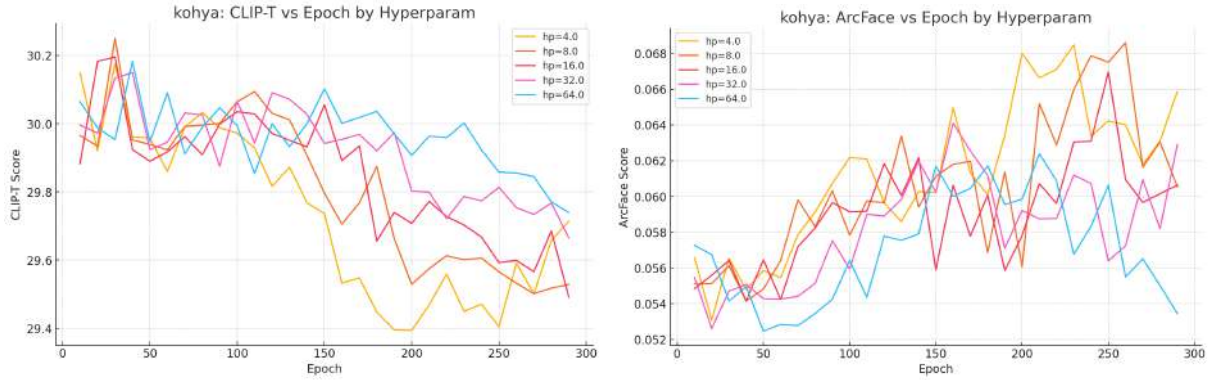
(a) CLIP-T (b) ArcFace

Figure 2: LoRA adapters (Diffuser) convergence.



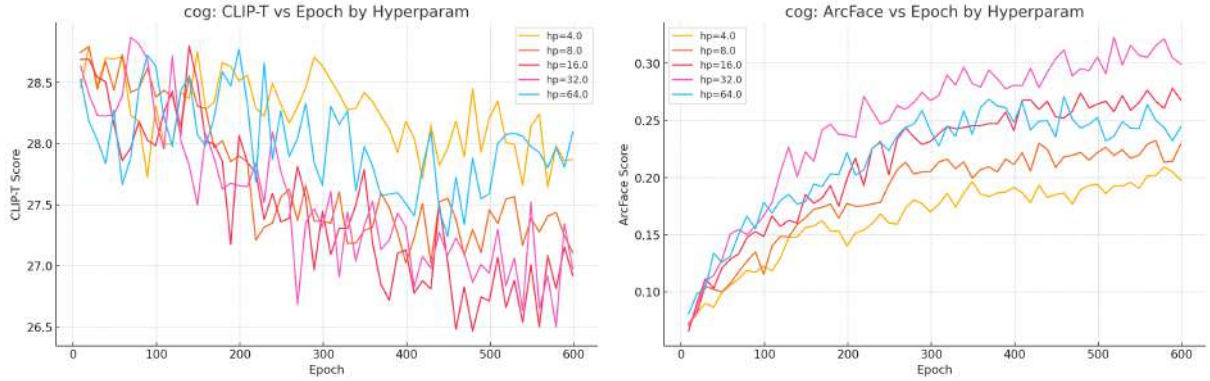
(a) CLIP-T (b) ArcFace

Figure 3: Textual Inversion (Diffuser) convergence.



(a) CLIP-T (b) ArcFace

Figure 4: LoRA adapters (Kohya) convergence.



(a) CLIP-T (b) ArcFace

Figure 5: LoRA adapters (Cog) convergence.

5.2 Optimal Hyperparameter Summary

Table 3 summarizes each method's optimal hyperparameter (based on average ArcFace) and its corresponding average CLIP-T and ArcFace scores. Table 4 lists the epoch at which the peak ArcFace occurred for that setting.

Table 3: Optimal Hyperparameter and Average Metrics

Method	Hyperparameter	Avg. CLIP-T	Avg. ArcFace
DreamBooth-LoRA (Dfr)	rank = 8	30.85	0.29
LoRA-SDXL (Dfr)	rank = 8	29.75	0.41
Textual Inversion	nv = 4	28.95	0.19
LoRA (Kohya)	rank = 4	29.45	0.07
LoRA (Cog)	rank = 32	26.84	0.32

Table 4: Peak-Epoch Results for Optimal Settings

Method	Hyperparam	Epoch	CLIP-T	ArcFace
DreamBooth-LoRA	rank = 8	530	29.28	0.36
LoRA-SDXL	rank = 8	440	29.54	0.51
Textual Inversion	nv = 16	120	28.20	0.25
LoRA (Kohya)	rank = 4	230	29.45	0.07
LoRA (Cog)	rank = 32	519	26.84	0.32

6 Qualitative Results

6.1 Prompt 1: Professional Headshot

Detail prompt 1: (wide shot) analog modelshoot 8k close-up LinkedIn profile picture of <itay>, frontal face, looking at camera, professional suit, upper body, blurred glass building background, crisp details, neutral expression, photorealistic, high-resolution, sharp focus on eyes, ambient cinematic lighting, hyperrealistic, masterpiece, best quality, ultra-detailed
Best-epoch outputs (seed 6969 above, seed 6970 below):

(a) DB-LoRA (b) LoRA-SDXL (c) Textual Inversion (d) Kohya-LoRA (e) Cog-LoRA

(f) DB-LoRA (g) LoRA-SDXL (h) Textual Inversion (i) Kohya-LoRA (j) Cog-LoRA

Figure 6: Prompt 1 qualitative comparison at each method's best epoch.

6.2 Prompt 4: Golden-hour countryside

Detail prompt 4: (golden hour portrait) <itay> in a wheat field at sunset, frontal face, looking at camera, wearing white linen shirt and khaki trousers, gentle breeze, backlighting, shallow depth of field, romantic countryside vibe, photorealistic, ultra-sharp, masterpiece, best quality, ultra-detailed
Best-epoch outputs (seed 6969 above, seed 6970 below):

(a) DB-LoRA (b) LoRA-SDXL (c) Textual Inversion (d) Kohya-LoRA (e) Cog-LoRA

(f) DB-LoRA (g) LoRA-SDXL (h) Textual Inversion (i) Kohya-LoRA (j) Cog-LoRA

Figure 7: Prompt 4 qualitative comparison at each method's best epoch.

7 Additional <irit> Training

I fine-tuned the LoRA-SDXL model (Diffuser) with rank = 8 at epoch = 440 on 15 images of the subject <**irit**> Below are qualitative outputs for two new prompts, each rendered with seeds 6969 and 6970.

Prompt 1: "a professional photo portrait of <iris> holding flowers, model photoshoot, professional photo, white background, Amazing Details, Best Quality, Masterpiece, dramatic lighting highly detailed, analog photo, overglaze, 80mm Sigma f/1.4 or any ZEISS lens"

(a) Seed 6969

(b) Seed 6970

Prompt 2: "portrait of <iris> wearing a business suit, model photoshoot, professional photo, white background, Amazing Details, Best Quality, Masterpiece, dramatic lighting highly detailed, analog photo, overglaze, 80mm Sigma f/1.4 or any ZEISS lens"

(c) Seed 6969

(d) Seed 6970

Figure 8: Qualitative examples for two new prompts on the *<irit>* dataset, using LoRA-SDXL (rank = 8, epoch = 440).