Ostbayerische Technische Hochschule Amberg-Weiden
Department of Electrical Engineering, Media and Computer Science

PERSONALIZED IMAGE GENERATION WITH STABLE DIFFUSION: PYTORCH LORA FINE-TUNING APPROACH

Student

Tran Thai Duc Hieu

t.tran@oth-aw.de

Supervised by

Prof. Dr. Tatyana Ivanovska

January 18, 2026

1 Introduction

Recent advances in diffusion models, particularly Stable Diffusion [1], have demonstrated remarkable capabilities in text-to-image generation. However, generating highly personalized images of specific subjects or celebrities remains challenging, as it requires the model to learn and preserve identity-specific features while remaining generalizable to diverse prompts and contexts.

A popular modern approache to this problem have relied on UI-based tools such as Kohya and A1111 WebUI, which abstract away the underlying PyTorch configuration details. While these tools offer user accessibility, they limit customization and reproducibility for research purposes.

This work presents a PyTorch-based implementation achieving personalized image generation performance comparable to existing community tools and tutorials. By implementing Low-Rank Adaptation (LoRA) fine-tuning from first principles, this study provides a transparent, modular, and extensible framework for personalization that facilitates reproducible research and systematic experimentation.

The objectives of this work are threefold:

1. Provide a reproducible, open-source PyTorch implementation for Stable Diffusion personalization
2. Systematically study the impact of caption structure and other hyperparameters on model performance
3. Establish quantitative evaluation metrics (LPIPS and FID) for assessing generation quality and identity preservation

2 Related work

2.1 Diffusion models and Stable Diffusion

Diffusion models have emerged as a powerful class of generative models, achieving state-of-the-art results in image synthesis [2]. Stable Diffusion [1] builds upon this foundation by combining diffusion models with latent space representations, enabling efficient high-quality image generation. The architecture comprises three main components: a text encoder (CLIP) [3], a UNet-based denoising network, and a variational autoencoder (VAE) [4].

2.2 Low-rank adaptation (LoRA)

Full model fine-tuning requires updating billions of parameters, which is computationally expensive and prone to overfitting on small datasets. Low-Rank Adaptation (LoRA) [5] addresses this by introducing trainable low-rank decomposition matrices into the network. Instead of updating the weight matrix W, LoRA decomposes the weight update as:

$$\Delta W = BA^T$$

where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{k \times r}$, with $r \ll \min(d, k)$ (typically $r = 8$ or $r = 16$). This dramatically reduces the number of trainable parameters while maintaining model expressiveness. LoRA is particularly effective for Stable Diffusion personalization, where only the attention layers in the UNet are typically adapted.

2.3 Personalized image generation

Personalized image generation aims to teach the model to recognize and generate specific subjects (e.g., a particular person or object). Common approaches include:

- **Textual Inversion:** Learns a new token embedding to represent a subject [6], requiring only 3-5 reference images but producing less flexible results.
- **DreamBooth:** Uses subject-specific identifiers combined with class-preservation loss to achieve high-fidelity personalization [7].
- **LoRA Fine-Tuning:** Adapts attention layers with low-rank updates, balancing efficiency and quality. This approach is implemented in popular community tools like Kohya and A1111 WebUI.

2.4 Evaluation metrics

Quantitatively evaluating personalized image generation requires metrics that capture both identity preservation and overall synthesis quality:

- **LPIPS (Learned Perceptual Image Patch Similarity):** Measures perceptual similarity between generated and training reference images using features from a pre-trained deep network (typically AlexNet) [8]. Lower values indicate better identity preservation. Benchmark targets: < 0.15 (excellent), $0.15 - 0.25$ (good), > 0.25 (weak).
- **FID (Frechet Inception Distance):** Measures statistical distance between generated and reference image feature distributions using the Inception-v3 network [9]. Lower values indicate better overall generation quality and diversity. Benchmark targets: < 10 (excellent), $10 - 30$ (good), $30 - 50$ (moderate), > 50 (poor).

2.5 Community tools and motivation for PyTorch implementation

UI-based tools such as Kohya and A1111 WebUI have enabled broader adoption of LoRA fine-tuning through user-friendly interfaces and pre-configured settings, and are accompanied by practical community guides detailing SD 1.5 workflows [10], [11]. However, these tools typically obscure implementation details, limiting research flexibility and code reproducibility. This work implements an equivalent PyTorch-based pipeline that follows established community configurations while providing fully interpretable, modular code suitable for research and controlled experimentation.

3 Dataset preparation and experimental setup

3.1 Data collection and preprocessing

A dataset comprising 151 images of a target subject (Jennie) was assembled to evaluate personalized generation capabilities. The dataset preparation process involved two primary steps:

1. **Image collection:** High-quality reference images of the target subject were curated from public sources.
2. **Image preprocessing:** All images were standardized to 512×512 pixels while preserving the original aspect ratio to maintain visual fidelity. Images exhibiting excessive occlusion or low perceptual quality were excluded.

3.2 Caption strategies and dataset versions

A keyword-based captioning paradigm was employed across all experiments. This approach relies on learning unique identifiers and descriptive tokens to achieve personalization, contrasting with traditional fine-tuning that updates all model weights. To systematically evaluate the effect of caption structure on model performance, three dataset versions with varying levels of attribute annotation were created:

- **Version 1 (v1):** Single unique trigger word "J3NN13". This identifier was deliberately chosen to be a non-dictionary word, ensuring the model associates this specific token exclusively with the target subject. All captions in this version consist solely of this trigger word, establishing a minimal baseline where the model must learn identity purely through visual association without additional contextual cues.
- **Version 2 (v3):** Detailed comma-separated keyword tags combining the trigger word "J3NN13" with multiple attribute descriptors: expressions (e.g., "wink", "teeth-smiling"), camera framing ("close-up", "straight-look"), appearance details ("long hair", "strap top"), and context ("wall background", "bright"). Example: "J3NN13, wink, close-up, straight-look, make-up, tongue-out, braiding hair". This maximizes attribute diversity and provides rich contextual information alongside the trigger word.
- **Version 3 (v4):** Simplified comma-separated keywords with fewer attributes per image, focusing on essential identity and contextual features. Example: "J3NN13, wink, close-up, make-up, tongue-out, braiding hair" or "J3NN13, close-up, wall background". This balances attribute specificity with annotation simplicity.

4 Experimental setup and implementation

This section documents the iterative experimental methodology employed to optimize personalized image generation. Rather than presenting a single configuration, the evolution of the approach across multiple experiments is described, emphasizing key findings and design decisions. Detailed results for all experiments (excluding Experiment 0) are presented below.

4.1 Initial baseline experiments (experiments 0)

Following community tutorials [10] recommending 1500-4500 training steps, we established initial baselines to understand training dynamics:

Table 1: Experiment 0 configuration

Parameter	Value
Base Model	Stable Diffusion v1.5 (runwayml/stable-diffusion-v1-5)
Dataset	151 images with v1 captions
LoRA Rank (r)	32
LoRA Alpha (α)	16
Batch Size	2
Training Epochs	100

Results: Although the configuration follows established community guidelines from popular Kohya and Automatic1111 tutorials, the results did not achieve the expected performance. Visual quality was poor, with weak identity preservation and significant visual artifacts. This outcome indicated that the training duration was insufficient.

4.2 Realistic base model experiments (experiments 1-3)

Vanilla Stable Diffusion v1.5 lacks photorealistic capabilities necessary for high-quality celebrity personalization. Consequently, a community-fine-tuned model (**SG161222/Realistic_Vision_V5.1_noVAE**) [12] specializing in realistic portrait generation was employed for all subsequent experiments.

Table 2 specifies the common training configuration used consistently across experiments 1–3. For each experiment, dataset versions 1, 2, and 3 respectively were used to investigate the effect of caption structure on model performance.

Table 2: Common training configuration for experiments 1–3

Parameter	Value
Base Model	SG161222/Realistic_Vision_V5.1_noVAE
Dataset Size	151 images
LoRA Rank (r)	32
LoRA Alpha (α)	16
LoRA Dropout	0.1
Target Modules	to_q, to_k, to_v, to_out.0
Batch Size	2
Learning Rate	5×10^{-5}
Training Epochs	300
Inference model	SG161222/Realistic_Vision_V5.1_noVAE

However, for each experiment, we use dataset versions 1 to 3 respectively.

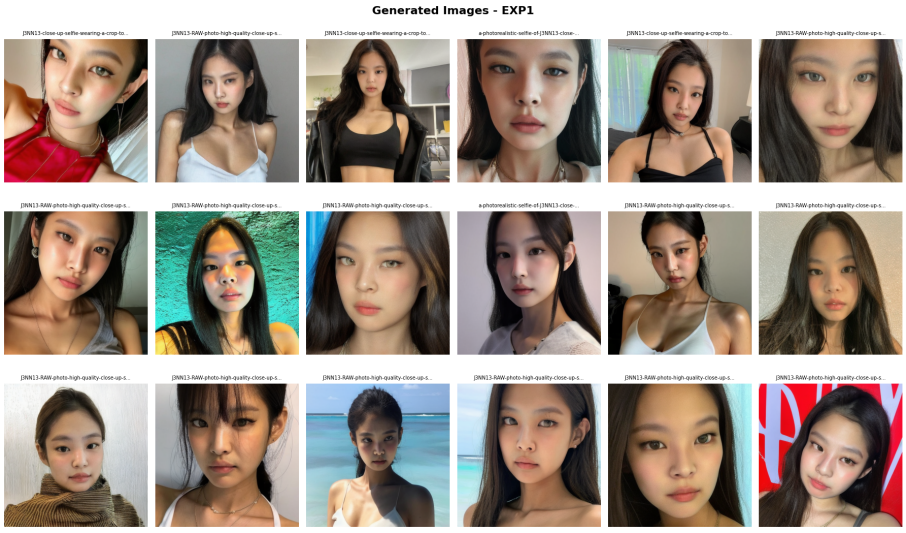

Figure 1: Experiment 1 samples

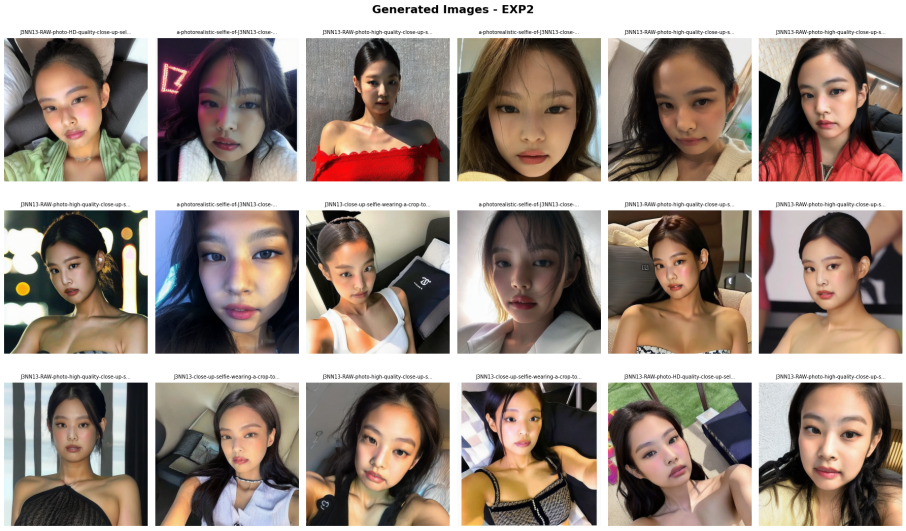

Figure 2: Experiment 2 samples

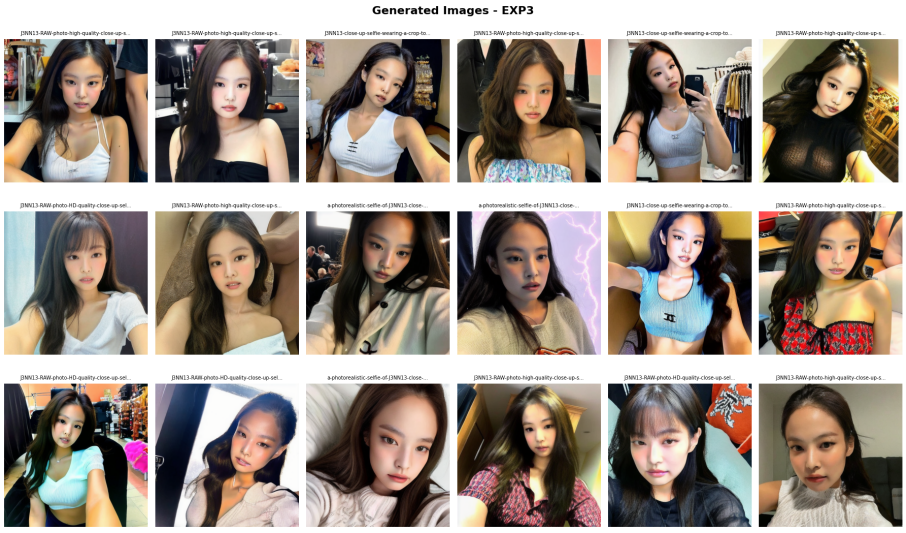

Figure 3: Experiment 3 samples

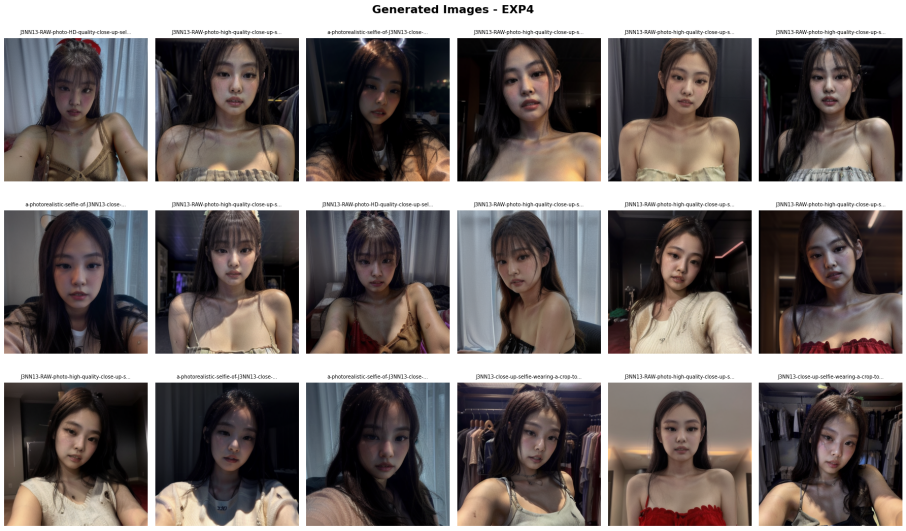

Figure 4: Experiment 4 samples

4.3 Transferred inference experiment (experiment 4)

Experiments 1–3 did not reach the quality reported in community tutorials. To assess the effect of the inference model, we separate training from inference: LoRA weights are trained on Stable Diffusion v1.5 and later applied to the Realistic Vision model at inference time. This cross-model transfer evaluates whether features learned during training can improve generation quality when used with a higher-fidelity, photorealistic base model.

Table 3: Training configuration for experiment 4: cross-model transfer

Parameter	Value
Base Model (Training)	runwayml/stable-diffusion-v1-5
Dataset Size	151 images
LoRA Rank (r)	32
LoRA Alpha (α)	16
LoRA Dropout	0.1
Target Modules	to_q, to_k, to_v, to_out.0
Batch Size	2
Learning Rate	5×10^{-5}
Training Epochs	300
Base Model (Inference)	SG161222/Realistic_Vision_V5.1_noVAE

4.4 Comparison across all experiments

Figure 5 summarizes LPIPS and FID across all experiments. Visual inspection indicates that Experiments 1 and 4 produce the most natural samples. Quantitatively, Table 4 shows Experiment 4 achieving the lowest LPIPS of **0.752**. An additional cross-model settings reported in Table 5. Further analysis is provided in the Discussion section.

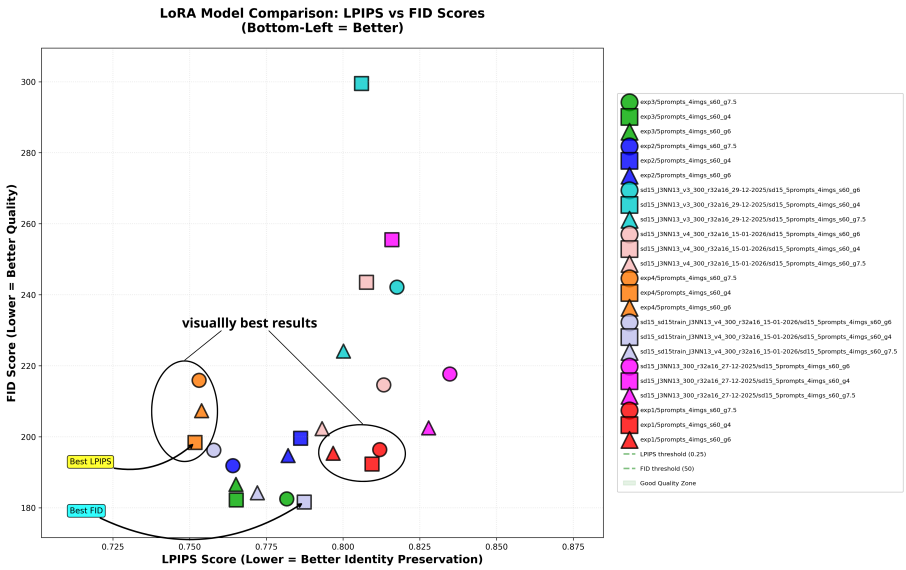

Figure 5: LPIPS and FID metrics across all experiments

Table 4: Summary of all experiments

Run	Base model	Infer. model	Caption type	Guidance	**LPIPS**	**FID**
Experiment 1	R.Vision	R.Vision	v1	4.0	0.809	192.3
Experiment 1	R.Vision	R.Vision	v1	6.0	0.797	195.4
Experiment 1	R.Vision	R.Vision	v1	7.5	0.812	196.4
Experiment 2	R.Vision	R.Vision	v2	4.0	0.786	199.6
Experiment 2	R.Vision	R.Vision	v2	6.0	0.782	194.7
Experiment 2	R.Vision	R.Vision	v2	7.5	0.764	191.9
Experiment 3	R.Vision	R.Vision	v3	4.0	0.765	**182.2**
Experiment 3	R.Vision	R.Vision	v3	6.0	0.765	186.6
Experiment 3	R.Vision	R.Vision	v3	7.5	0.782	182.5
Experiment 4	R.Vision	SD v1.5	v3	4.0	**0.752**	198.4
Experiment 4	R.Vision	SD v1.5	v3	6.0	0.754	207.3
Experiment 4	R.Vision	SD v1.5	v3	7.5	0.753	215.9

Table 5: Additional experiments

Run	Base model	Infer. model	Caption type	Guidance	LPIPS	FID
Experiment 1	R.Vision	SD v1.5	v1	4.0	0.816	255.5
Experiment 1	R.Vision	SD v1.5	v1	6.0	0.835	217.7
Experiment 1	R.Vision	SD v1.5	v1	7.5	0.828	202.5
Experiment 2	R.Vision	SD v1.5	v3	4.0	0.806	299.5
Experiment 2	R.Vision	SD v1.5	v3	6.0	0.818	242.2
Experiment 2	R.Vision	SD v1.5	v3	7.5	0.800	224.1
Experiment 3	R.Vision	SD v1.5	v4	4.0	0.808	243.5
Experiment 3	R.Vision	SD v1.5	v4	6.0	0.813	214.6
Experiment 3	R.Vision	SD v1.5	v4	7.5	0.793	202.3
Experiment 4	SD v1.5	SD v1.5	v4	4.0	0.787	181.6
Experiment 4	SD v1.5	SD v1.5	v4	6.0	0.758	196.2
Experiment 4	SD v1.5	SD v1.5	v4	7.5	0.772	184.2

4.5 Evaluation protocol

For each experiment, the following evaluation procedure was applied:

1. Image generation: 5 distinct text prompts were used to generate 4 images per prompt, yielding 20 generated images total. Inference parameters were set to 60 diffusion steps with guidance scale $\lambda \in \{4.0, 6.0, 7.5\}$.

2. LPIPS computation: Learned Perceptual Image Patch Similarity was computed by comparing each generated image to randomly selected reference images from the training set using the AlexNet feature extractor.

3. FID computation: Fréchet Inception Distance was calculated between feature distributions of the 20 generated images and 151 reference training images using Inception-v3.

4. Qualitative analysis: Generated samples were visually assessed for identity consistency, artifact presence, and adherence to text prompts.

Note: The relatively small number of generated images (20 total) for evaluation may introduce variance in metric estimates. Future work should employ larger evaluation sets (100+ images) for more robust and statistically reliable metric computation.

4.6 Discussion

4.7 Key findings

1. **Training duration:**

   Experiment 0 with 100 epochs demonstrated insufficient convergence. Extending training to 300 epochs (approximately 22,800 training steps with batch size 2) proved necessary for model adaptation.

   If the generated images still exhibit poor quality (many artifacts, object edges is too sharp, background inconsistencies or it feels like a painting) or weak identity preservation, consider increasing the number of training epochs further (e.g., 400-500 epochs)

   The samples from Experiment 2 and Experiment 3 show low quality at 300 epochs, indicating that even longer training may be required for optimal results.

2. **Caption strategy trade-offs:**

   - Version 1 (single trigger word):
     Minimal information, requiring purely visual identity learning. Surprisingly, this simple approach can achieve reasonably good results with sufficient training, see Fig. 6.

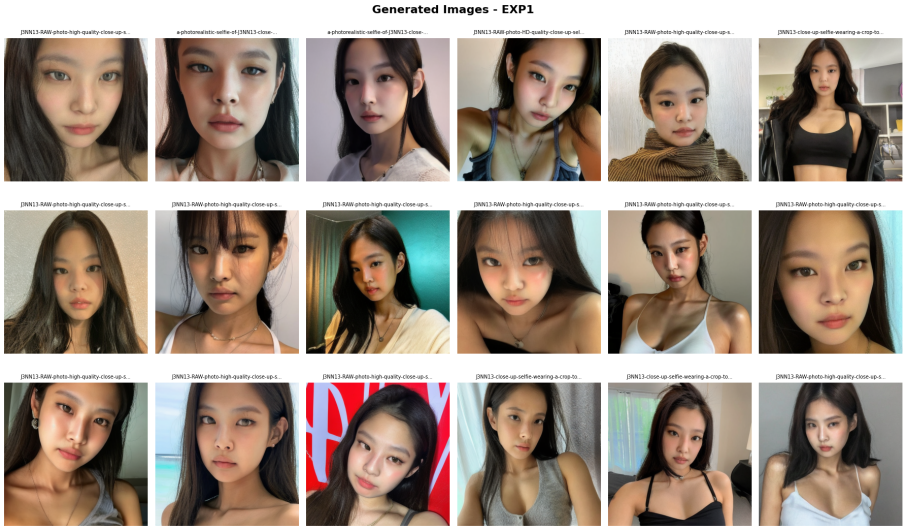

Figure 6: Best samples from Experiment 1

The weakness of this approach is that it can only generate good results of close-up portraits or features that are very similar to the training images. The model lacks contextual understanding, leading to failures when generating diverse poses, backgrounds, or full-body shots.

   - Version 2 (detailed attribute tags):

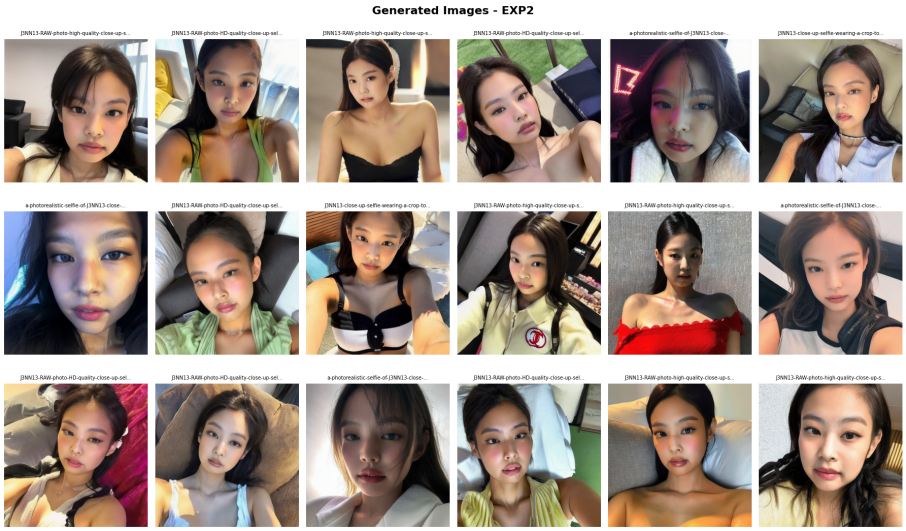

Figure 7: Best samples from Experiment 2

- Version 3 (simplified attribute tags): Optimal balance between annotation simplicity and attribute diversity. Good results already achieve in Experiment 4, see Fig. 8.

3. **Cross-model transfer:**
Experiment 4 demonstrated that training LoRA weights on Stable Diffusion v1.5 and applying them to the Realistic Vision model during inference can enhance identity preservation (lowest LPIPS). The general base model (SD v1.5) appears to force the LoRA to learn stronger identity features and information, which then benefit from the photorealistic capabilities of the Realistic Vision model at inference time. But this improvement in LPIPS comes at the cost of overall image quality (higher FID), likely due to domain mismatch between training and inference models.

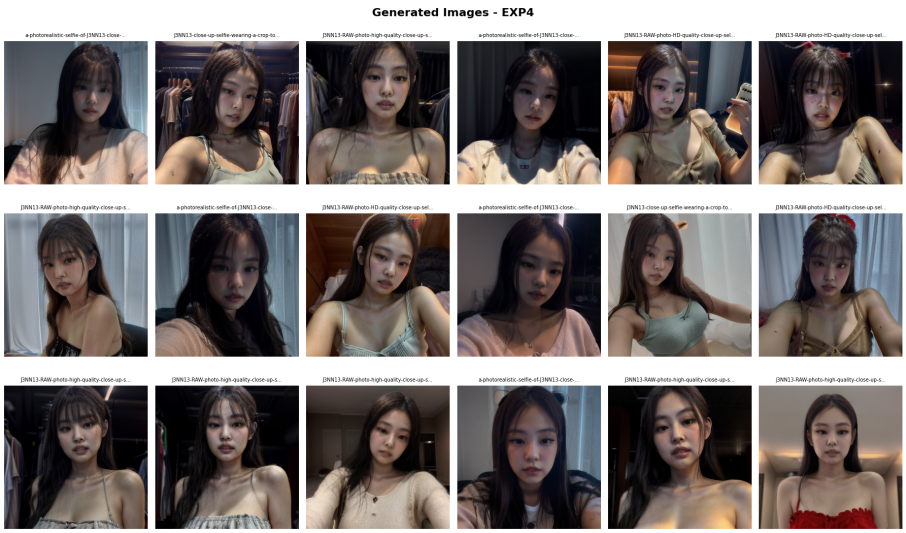

Figure 8: Best samples from Experiment 4

Notice that the style of Realistic Vision model (the overall theme and lighting is dark and moody) is still dominant in the generated images

4. **Inference prompts:**
Carefully chosen keywords in inference prompts significantly influence generation quality. Including descriptive words like “HD quality”, “good lighting”, “RAW photo”, or “4K” noticeably improves realism and quality. A practical tip: investigate the training data and captions used to train your base model. Reusing similar keywords and phrasing patterns from the model’s training set can further enhance results, as the model is more familiar with these descriptors.

5. **Guidance scale:**
   A guidance scale is a hyperparameter that controls the trade-off between how closely the generated image follows the text prompt versus how much creativity/diversity is allowed.
   Higher guidance scales force the model to obey the prompt more strictly, which can improve relevance but may also introduce unwanted artifacts if words in the prompt are ambiguous or unfamiliar to the model during training.

   This bad effect of high guidance scale can be seen clearly in Experiment 3 and Experiment 4, where increasing the guidance scale to 7.5 leads to more artifacts, distorted body parts and unnarual textures and skin tones.

6. **Negative prompts:**
   Carefully crafted negative prompts offer an effective way to enhance generation quality and structural correctness. By specifying undesired features, the model learns to actively avoid them, resulting in improved image structure and coherence.

   Across all experiments, asymmetric or misaligned facial features appeared regularly in generated images. Adding negative prompts like "asymmetric eyes", "crossed eyes", or "distorted face" consistently reduced these artifacts. Similarly, negative prompts addressing anatomical issues (e.g., "wrong anatomy", "malformed hands", "distorted body parts") noticeably improved overall generation quality. This approach complements positive prompts and guidance scale adjustments, providing an additional mechanism for fine-tuning output quality.

4.8 Limitations and future work

While this work establishes a solid foundation for personalized image generation using LoRA fine-tuning, several limitations and avenues for future research remain:

- **Natural sentence captions vs keyword-based captions:**
  The current approach relies exclusively on keyword-based captions. Future research should explore natural language descriptions (e.g., "J3NN13 is smiling at the camera with long black hair") to assess whether more contextual, grammatically complete captions improve identity learning and generation flexibility. Natural sentences may provide richer semantic relationships between attributes.

- **Extended training for complex caption structures:**
  Experiments 2 and 3, which employed more detailed caption versions, showed suboptimal results at 300 epochs. Increasing training duration (e.g., 500–1000 epochs) for datasets with comprehensive attribute annotations may allow the model to fully leverage the additional contextual information and achieve better identity preservation with enhanced controllability.

- **Facial features and hand generation:**
  Despite negative prompt adjustments, eyes and hands remain challenging to generate accurately. Asymmetric eyes, misaligned pupils, and malformed fingers appeared across experiments. Future work should investigate targeted training techniques (e.g., weighted loss on facial regions, specialized hand datasets) or post-processing refinement methods to address these persistent anatomical issues.

- **Post-processing and refinement workflows:**
  Integrating post-processing techniques such as face restoration models (e.g., CodeFormer, GFPGAN), inpainting tools, or ControlNet-based refinement could systematically improve generation quality. Establishing automated pipelines that combine LoRA generation with targeted post-processing steps would enhance practical applicability while maintaining identity consistency.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *arXiv preprint arXiv:2112.10752*, 2022.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *arXiv preprint arXiv:2006.11239*, 2020.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, *et al.*, "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.14030*, 2021.

[4] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[5] E. J. Hu, Y. Shen, P. Wallis, *et al.*, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[6] R. Gal, Y. Alaluf, Y. Atzmon, G. Pavlakos, A. Shamir, and S. Gur, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint arXiv:2208.01618*, 2022.

[7] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," *arXiv preprint arXiv:2208.12242*, 2022.

[8] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[9] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

[10] P. Arbués. "Training a lora of your face with stable diffusion 1.5." Practical guide covering dataset preparation, captioning, Kohya training configuration, and inference workflow for SD 1.5 LoRA. (Oct. 24, 2024), [Online]. Available: `https://www.pelayoarbues.com/notes/Training-a-LoRa-of-your-face-with-Stable-Diffusion-1.5` (visited on 01/18/2026).

[11] __2_. "Celebrity lora training guide (consolidated) – sd 1.5." Overview of SD 1.5 LoRA training including dataset curation, tagging, Kohya settings, and model selection best practices. (Apr. 20, 2024), [Online]. Available: `https://civitai.com/articles/5010/celebrity-lora-training-guide-consolidated-sd-15` (visited on 01/18/2026).

[12] SG_161222, *Realistic vision v5.1 novae*, Hugging Face, Photorealistic Stable Diffusion model checkpoint based on SD 1.5, 2023. [Online]. Available: `https://huggingface.co/SG161222/Realistic_Vision_V5.1_noVAE` (visited on 01/18/2026).