Identity-Preserving Portrait Stylization with LoRA-Based Diffusion Models

İdil Görgülü [*1] **Oğuz Kağan Hitit** [*1]

Abstract

Portrait style transfer is the task of generating visually compelling artistic renditions of facial images while preserving the subject's unique identity. While recent diffusion-based approaches, particularly ones that utilize Low-Rank Adaptation (LoRA), have demonstrated significant success in flexible and parameter-efficient stylization, they often fail to preserve essential identity features. This study introduces Identity-Preserving LoRA (IP-LoRA) to improve identity fidelity in portrait stylization. Our method builds upon the strengths of B-LoRA and ConsisLoRA, and integrates an additional identity-preserving loss term into the training of the content LoRA. This loss is computed using embeddings from the ArcFace and DINOv2. We evaluate our model on a broad spectrum of styles and measure its performance by semantic, perceptual, and identity-based scores including DINOv2, CLIP similarity, and DreamSim distance. Our findings show the effective power of embedding-level regularization in closing the gap between visual aesthetics and semantic fidelity in generative models.

1. Introduction

Image style transfer refers to the task of learning the style features of a target image and transferring the style to another image while preserving its semantic content (an Li et al., 2025). The improvements in deep learning frameworks such as Diffusion models have led to the development of high-performing generative methods which are tailored for this purpose. The implementation of style transfer on portrait images is a particularly active area of research and has been widely adopted across various industries. Portrait-style transfer includes the generation of a facial image in a new visual style, such as a cartoon, painting, or sketch, while maintaining the subject's identity-defining characteristics. As generative models gain wider use in creative fields, such as personalized avatar generation, social media effects, digital painting, and entertainment, high-quality stylization has become a rapidly growing demand. Despite notable advances in the diversity and fidelity of stylized outputs, an existing bottleneck issue remains: facial identity preservation. Ensuring that stylized faces remain identifiable—especially in identity-critical applications—is challenging because most existing solutions trade semantic facial attributes to maintain expressive style representations.

Recent advancements in diffusion models, most notably architectures such as Stable Diffusion XL (SDXL) (Podell et al., 2023), have made it possible to produce high-resolution, realistic, and semantically consistent images. These models are the backbone of most modern style transfer pipelines. However, fully fine-tuning such large models is computationally very expensive and memory-intensive. To circumvent this challenge, Low-Rank Adaptation (LoRA) (Hu et al., 2021) has emerged as a quick replacement, enabling task-specific fine-tuning by adding lightweight, low-rank matrices to some layers of a frozen base model otherwise. LoRA not only provides efficient computation, but also presents a modular design, where adapters can be trained for different tasks without modifying the core model.

While LoRA-based methods like B-LoRA (Frenkel et al., 2024) and ConsisLoRA (Chen et al., 2025) have come a long way in style-content disentanglement and consistency, they overlook identity preservation and produce suboptimal results. These models focus mainly on structure or pixel-level reconstruction loss, which is not sufficient to maintain high-level identity features such as jawline, eye shape, and facial expressions. Therefore, the stylized outcomes, although visually engaging, tend to be missing critical features defining the subject's identity.

In this project, we propose IP-LoRA (Identity-Preserving LoRA), a novel framework extending the ConsisLoRA pipeline with identity consistency as the primary objective. Our method incorporates an identity-preserving regularization loss based on facial embeddings derived using Additive Angular Margin Loss for Deep Face Recognition (ArcFace) as a state-of-the-art face recognition model (Deng et al., 2022). We reduce the angular gap between stylized and

[*]Equal contribution [1]Department of Computer Engineering. Correspondence to: İdil Görgülü <igorgulu21@ku.edu.tr>, Oğuz Kağan Hitit <ohitit20@ku.edu.tr>.

COMP547 Deep Unsupervised Learning, Spring 2025.

original face embeddings during content LoRA training to retain essential identity aspects while still enabling rich and diverse stylization.

The overall architecture builds upon SDXL, and the training procedure follows a two-stage approach: a content LoRA is first trained using a hybrid loss with regular diffusion loss and added identity consistency loss parameter; style LoRA is subsequently trained separately with content parameters frozen so that the model only learns about visual details such as texture and color and does not leak any content to the generated output image. The modularity of our approach allows us to switch directly between different content and style adapters at inference, supporting tasks such as exemplar-based style transfer, text-based stylization, and consistent style generation.

Our approach is evaluated both quantitatively, using embedding-based and perceptual similarity metrics including DINOv2, CLIP, and DreamSim Distance, and qualitatively across five different styles and six facial images for content. The results demonstrate that identity-aware regularization yields more faithful, recognizable stylized portraits compared to prior baselines, without sacrificing visual quality. We believe our approach bridges the gap between artistic stylization and semantic fidelity, particularly critical for real-world deployment in identity-sensitive applications.

2. Related Work

This section provides an overview of the early work and key techniques that form the basis of our proposed solution. We specifically cover Low-Rank Adaptation (LoRA) (Hu et al., 2021), B-LoRA (Frenkel et al., 2024), ConsisLoRA (Chen et al., 2025), and ArcFace (Deng et al., 2022), as each contributes uniquely to the construction and motivation of IP-LoRA.

2.1. Low Rank Adaptation (LoRA) (Hu et al., 2021)

LoRA (Low-Rank Adaptation) is a parameter-efficient technique for adapting large pretrained language models (LLMs) to downstream tasks without the need to fine-tune all model parameters. Instead of updating the entire model, LoRA introduces small, trainable low-rank matrices into existing layers—specifically the attention weights of the Transformer architecture—while keeping the original weights frozen.

LoRA significantly reduces the number of trainable parameters—by up to 10,000× compared to full fine-tuning. Despite the reduced parameter count, LoRA matches or outperforms full fine-tuning across various NLP tasks on models such as RoBERTa, DeBERTa, GPT-2, and GPT-3. Another advantage of LoRA is that unlike adapters, LoRA introduces no extra inference time because its low-rank matrices can be merged into the original weights after training. The method reduces VRAM usage and allows task-switching with minimal overhead by swapping in new low-rank matrices.

Instead of updating a full weight matrix W, LoRA models the update as a product of two low-rank matrices $A \in \mathbb{R}^{r\times d}$ and $B \in \mathbb{R}^{d\times r}$, where $r \ll d$, such that the effective weight becomes $W + \alpha BA$, where α is the scaling factor controlling the impact of the adaptation. This reparameterization exploits the observation that weight updates during fine-tuning are typically low-rank in nature. In all of the evaluations and conducted performance comparisons, LoRA achieves comparable or better performance with far fewer parameters.

This study empirically demonstrates that the rank-deficiency of updates holds in practice, and that adapting only parts of the attention mechanism like query and value weights suffices. The subspace analyses performed within the scope of the study show that a very low-rank like 1 or 2 already captures most of the meaningful adaptation directions.

2.2. B-LoRA (Frenkel et al., 2024)

B-LoRA introduces a novel method which enables image stylization by implicitly disentangling the style and content of a single input image using LoRA within the SDXL architecture. In their study, Frenkel et al. perform an empirical analysis on the backbone of SDXL through prompt injection and identify that two specific transformer blocks within the architecture can capture style and content respectively. Therefore, low-rank residual weights for the 4th and 5th blocks of SDXL are optimized, where the 4th block captures the image's content (semantic structure and shape), and the 5th captures the style (color, texture, artistic patterns). The finetuning performed on these two blocks, named as B-LoRAs, is achieved using a single image and prompt "A [v]". Overall, this setup prevents overfitting and is much more lightweight compared to conventional full-model fine-tuning.

Since B-LoRA preserves the original model weights and trains only a small set of parameters, it is efficient and generalizable. Once trained, the learned LoRA weights for style and content can be independently reused, combined, or swapped across different images or prompts to enable flexible applications such as image-based style transfer, text-guided stylization, and consistent style generation, and these can be performed without the need for any further optimization or retraining.

B-LoRA presents a number of benefits over competing stylization methods in the sense that it eliminates the requirement for additional optimization in the process of combining new pairs of style and content, unlike previos LoRA-based generative models. It surpasses baselines such as Style-

Drop and StyleAligned in both qualitative and quantitative metrics, including DINO feature similarity and a large user study. The method demonstrates excellent style transfer with improved content preservation, higher similarity of style, and approximately 70% lower storage requirements. Despite all of these improvements, the model still has some shortcomings: color can become too seriously entwined with style, possibly harming identity conservation; background elements may unintentionally influence the style representation; and performance may decline in complex scenes.

2.3. ConsisLoRA (Chen et al., 2025)

ConsisLoRA proposes a method to improve style-content disentanglement and generation consistency in LoRA-based fine-tuning of large diffusion models by augmenting the training process with regularization losses to explicitly enforce content preservation and style consistency.

The architecture of ConsisLoRA is largely the same as B-LoRA, using separate LoRA modules for content and style. However, ConsisLoRA augments the training of the content LoRA with additional perceptual and consistency objectives to improve generalization and fidelity.

Traditional diffusion models and B-LoRA rely on ϵ-prediction to estimate the added noise. This loss metric focuses on low-level details and often ignores global structure, therefore, it is suboptimal for style transfer. On certain portions of the training process, ConsisLoRA replaces ϵ-prediction with x_0-prediction, which estimates the original image instead of noise. Predicting the original image is considerably more effective, as it emphasizes high-level semantic consistency, thereby ensuring better preservation of both content and style. Further details and mathematical formulations about these loss function will be introduced in Section 3: The Approach.

ConsisLoRA adopts a two-stage training strategy, in contrast to the joint optimization used in B-LoRA. In the first stage, only the Content LoRA is trained. This begins with 500 steps using the standard ϵ-prediction loss, which estimates the noise added during the forward diffusion process. The training then continues for 1000 steps using an x_0-prediction loss, which compares the denoised output to the original clean image. After this stage, the Content LoRA weights are frozen to prevent further updates and possible content leakages. In the second stage, the Style LoRA is trained for 1000 steps independently using the x_0-prediction loss once again. This separation permits the model to learn style-specific features independently without disrupting existing content structure.

Despite its improvements over B-LoRA, ConsisLoRA does not explicitly incorporate any identity-aware metric into the training process. This limits its ability to maintain facial identity, especially under strong style transformations. Our method, IP-LoRA, directly addresses this limitation by adding identity-preserving constraints during content LoRA training.

2.4. ArcFace (Deng et al., 2022) and Insightface (Guo et al., 2021)

Additive Angular Margin Loss for Deep Face Recognition (ArcFace) is a novel loss function designed to improve the discriminative power of deep face recognition models. ArcFace rectifies some limitations of earlier loss functions like softmax via an additive angular margin penalty with a clear geometric interpretation on a hypersphere. The penalty directly optimizes the geodesic distance between features and class centers, leading to improved intra-class compactness and inter-class separation.

The broader usefulness of Arcface lies in its pretrained embedding models, which have become widely adopted for identity verification and similarity measurement. We take advantage of these pretrained ArcFace models as identity encoders in our work. Rather than training an identity model from scratch, we rely on the public implementations provided by InsightFace (Guo et al., 2021) with various backbone architectures. These models provide a general robustness across a wide demographic and pose variations as they are trained on big, curated face databases.

In our Identity-Preserving LoRA, we specifically use the buffalo_l/w600k_r50 ArcFace variant (Guo et al., 2021). The model uses a ResNet-50 backbone trained from the MS1MV3 dataset of over 600,000 identities. It produces 512-dimensional face embeddings that reside on a hypersphere. This makes the model a great fit for cosine similarity comparison tasks. The embeddings are extremely sensitive to face identity and invariant to variation in expression, lighting, and to a certain extent, style. We compute the cosine similarity between the ArcFace embeddings of the source and target images in our training pipeline. Identity preservation loss obtained from this is then added to the diffusion loss to guide the Content LoRA module. This embedding-level regularization enables our model to retain crucial identity features such as eye shape, jawline, and facial structure even under dramatic style transformations.

By including ArcFace's pretrained identity embeddings within our style framework, we not only obtain target style matching in the output but also retain high-fidelity fine-grained identity features. ,

2.5. DINOv2 as an Identity Encoder (Oquab et al., 2023)

DINOv2 is a self-supervised vision transformer (ViT) model trained using knowledge distillation without labels, achiev-

ing state-of-the-art performance across a wide range of visual tasks. Unlike supervised face recognition models such as ArcFace, DINOv2 learns powerful visual features purely from large-scale image data by enforcing invariance across augmented views of the same image. This makes DINOv2 very effective at capturing semantic and structural information in a generalizable way.

In our work, we evaluate DINOv2 as an alternative identity embedding model to ArcFace for computing the identity preservation loss during Content LoRA training. Specifically, we use the publicly available pretrained DINOv2 ViT-Small model. This model produces 384-dimensional image embeddings, and these embeddings are extracted for both the original and stylized images. Similarly to our ArcFace implementation, the comparison is made using cosine similarity. The resulting loss encourages the stylized output to remain semantically close to the original content in the DINOv2 feature space.

Although it is not explicitly trained for facial recognition, DINOv2 embeddings capture high-level semantic and structural cues, and this is helps to achieve identity preservation through cosine similarity of features. This makes DINOv2 pre-trained models strong candidates for tasks requiring general semantic consistency, especially when fine-grained identity details are less critical or when ArcFace may overfit to specific facial distributions.

In our experiments, substituting ArcFace with DINOv2 in the IP-LoRA pipeline yields qualitatively similar improvements in identity preservation, though with slightly lower content alignment scores in some cases if we use the same identity loss weight as we did in ArcFace-based loss. If we employ a higher identity loss weight, DINOv2 performs very similarly to ArcFace. Its performance in our identity-aware stylization task demonstrates that large self-supervised models can serve as viable identity encoders in generative diffusion pipelines.

3. The Approach

3.1. Implementation of B-LoRA

To establish a reliable baseline, we implemented the B-LoRA method. Although an official implementation was available, it could not be used due to package version incompatibilities. Therefore, we reimplemented the approach from scratch, using the original paper and its inference scripts as references. In our implementation, we followed the adapter placement scheme described in the original work, inserting the Content LoRA at transformer block position W^{04} and the Style LoRA at position W^{05}.

The training logic in B-LoRA involves jointly optimizing both the Content and Style LoRA modules in a single-stage training loop. The model is trained using the standard ϵ-prediction loss, defined as:

$$\mathcal{L}_\epsilon = \mathbb{E}_{\mathbf{x}_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2\right], \tag{1}$$

where $\mathbf{x}_t$ is the noisy input at timestep t, and ϵ_θ is the noise predicted by the model. We trained the model for 1000 steps using this loss. A single training takes approximately 12 minutes on a single A100 GPU.

3.2. Implementation of ConsisLoRA

ConsisLoRA was also implemented from scratch, as there was no official code release from the original authors. We based our implementation on the methodology described in the paper and available inference scripts. The adapter positions were kept consistent with those used in B-LoRA, placing the Content LoRA at transformer block position W^{04} and the Style LoRA at W^{05}.

Unlike B-LoRA, ConsisLoRA follows a staged training procedure. In the first stage, only the Content LoRA is trained, beginning with 500 steps using the ϵ-prediction loss as in Equation (1), followed by 1000 steps using the x_0-prediction loss. The x_0-prediction loss is defined as:

$$\mathcal{L}_{x_0} = \mathbb{E}_{\mathbf{x}_0,\epsilon,t}\left[\|\mathbf{x}_0 - \hat{\mathbf{x}}_\theta(\mathbf{x}_t, t)\|^2\right], \tag{2}$$

where $\hat{\mathbf{x}}_0$ is the predicted denoised sample at timestep t. Once the Content LoRA is fully trained, its parameters are frozen. In the second stage, the Style LoRA is trained independently for 1000 steps using only the x_0-prediction loss $\mathcal{L}_{x_0}$. This sequential training strategy enhances the model's ability to learn style-specific parameters while preserving the integrity of the previously optimized content representations. A full ConsisLoRA training process takes approximately 30 minutes to complete.

3.3. Implementation of IP-LoRA

We proposed Identity Preserving LoRA (IP-LoRA) as our novel extension, building upon the sequential training strategy of ConsisLoRA. In IP-LoRA, we retain the same LoRA adapter positions as B-LoRA, with the Content LoRA inserted at transformer block position W^{04} and the Style LoRA at W^{05}. The training process mirrors that of ConsisLoRA, with the addition of an identity preservation loss during the Content LoRA training stage.

The Content LoRA is trained for a total of 1500 steps, where the first 500 steps use a combination of the ϵ-prediction loss (Equation (1)) and the identity preservation loss. The remaining 1000 steps continue training with the x_0-prediction loss (Equation (2)), still combined with the identity loss.

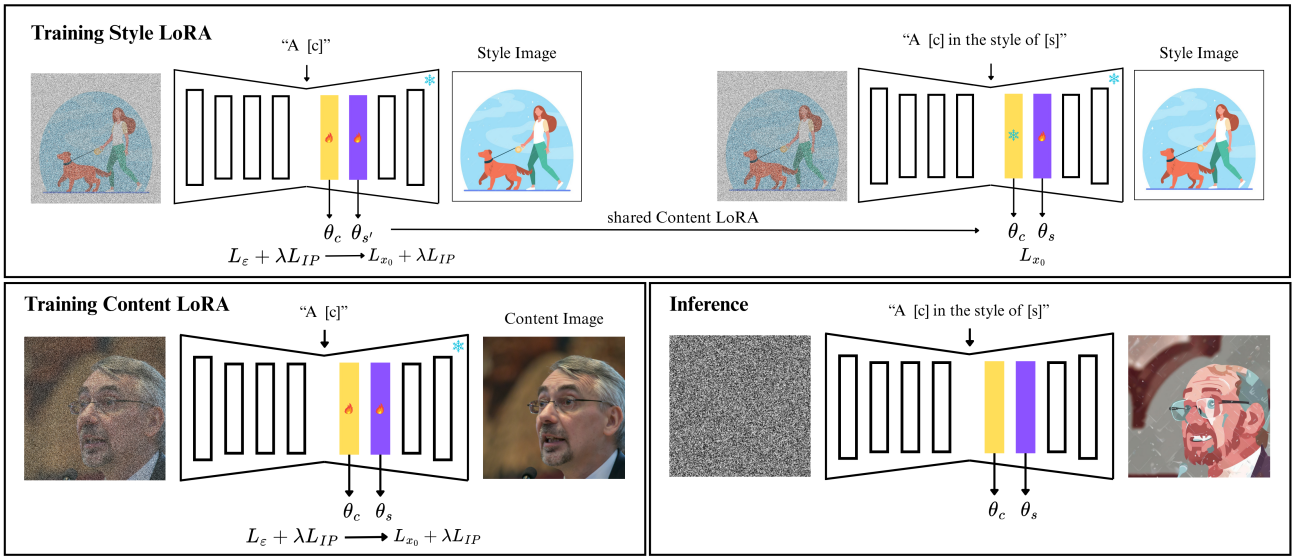


Figure 1. IP-LoRA Training and Inference Pipeline: The Content LoRA is trained with identity-preserving loss using prompts like "A [c]". The Style LoRA is trained separately with "A [c] in the style of [s]" while freezing the Content LoRA. At inference, both adapters are combined to generate stylized images.

After the Content LoRA is trained, its parameters are frozen. The Style LoRA is then trained for 1000 steps using only the x_0-prediction loss, allowing for flexible stylistic modulation without altering the preserved content.

To preserve identity during stylization, we introduce an additional loss term based on cosine similarity. Specifically, we extract deep image embeddings using a pretrained ArcFace model or DINOv2 from both the original and stylized images. For ArcFace, there are several embedding model variants. We used the buffalo_l/w600k_r50 model, which combines a ResNet-50 backbone with training on 600,000 curated identities from the MS1MV3 dataset.

The identity preservation loss is computed as:

$$\mathcal{L}_{\text{IP}} = 1 - \cos\left(f(\mathbf{x}_0), f(\hat{\mathbf{x}}_0)\right), \tag{3}$$

where $f(\cdot)$ denotes the embedding function and $\cos(\cdot, \cdot)$ represents cosine similarity. This term is added to the main training loss during the Content LoRA training process. The overall loss used in training is given by:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + \lambda \cdot \mathcal{L}_{\text{IP}}, \tag{4}$$

where $\mathcal{L}_{\text{main}}$ refers to either $\mathcal{L}_\epsilon$ or $\mathcal{L}_{x_0}$, depending on the training step, and λ is a hyperparameter that balances the influence of the identity loss.

The value of the hyperparameter λ plays a crucial role in balancing identity preservation and generative fidelity. In our experiments, we determined an appropriate λ value by first measuring the average magnitude of the original training loss (without the identity term) across steps. We then scaled the identity loss such that its expected magnitude was roughly comparable to the average base loss. This normalization ensured that the identity-preserving loss contributed a meaningful signal during optimization without overwhelming the diffusion objective. Also, we use a default rank of r = 64 for all LoRA adapters unless otherwise stated.

3.4. Training and Inference

During training, the LoRA adapters are conditioned on structured instance prompts that facilitate disentanglement of content and style representations. Specifically, the Content LoRA is trained with prompts of the form "A [c]", where [c] denotes the class or category of the image (e.g., "a dog", "a person"). The Style LoRA, in contrast, is trained with extended prompts of the form "A [c] in the style of [s]", where [s] denotes the target style (e.g., "Van Gogh", "low-poly", "pixel art") and [c] represents the content category. During Style LoRA training, [c] was set to a generic token such as "image" to prevent the model from learning content-specific biases.

At inference time, our architecture supports multiple generation modes, depending on which LoRA modules are active:

- **Image-to-Image Style Transfer:** In this mode, both the Content and Style LoRA are active. The Content LoRA is extracted from a reference content image, and the Style LoRA from a reference style image—each trained independently. The input prompt follows the

format "A [c] in the style of [s]", where [c] corresponds to the class token used during Content LoRA training and [s] to the style identifier used during Style LoRA training. The model generates an output that preserves the structure and identity of the content image while adopting the learned visual characteristics of the style image.

- **Text-Guided Style Transfer with Content Preservation:** In this mode, only the Content LoRA is active. Given the prompt "A [c] in the style of [s]", where [c] matches the Content LoRA's training class and [s] is any textual style, the model generates identity-consistent outputs stylized according to the prompt.
- **Text-Guided Content Generation with Style Preservation:** In this mode, only the Style LoRA is active. Given the prompt "A [c] in the style of [s]", where [s] corresponds to the trained style and [c] is any content, the model generates outputs with consistent style while adapting to new textual content.

This modular inference design enables flexible control over stylization and content preservation by toggling LoRA adapters at runtime. Since the adapters are lightweight and independently trained, content or style modules can be swapped plug-and-play without retraining the full model. This makes the system well-suited for applications like personalized image synthesis, multi-style generation, and controlled image editing.

3.5. Evaluation

To evaluate our approach, we trained a total of 11 LoRA adapters: 6 for content and 5 for style. The content images were randomly selected from the FFHQ dataset. The style images were taken from the official implementation of the B-LoRA study, using the same artistic references to maintain consistency with prior work. For each method—B-LoRA, ConsisLoRA, and our proposed IP-LoRA—we trained separate LoRA adapters for all 11 images.

We assessed the models on their image-to-image style transfer performance. For that, we constructed all possible combinations of content and style LoRAs. Specifically, we plugged the Content LoRA of each of the 6 content images with the Style LoRA of each of the 5 style images, resulting in $6 \times 5 = 30$ unique content-style pairings. For each pairing, we generated 4 samples during inference to account for sampling variance, yielding a total of 120 generated images per method.

We focused our evaluation on two core dimensions:

- **Content Alignment:** how well the generated image retains the identity and structure of the original content image.
- **Style Alignment:** how faithfully the generated image reflects the appearance and texture of the style reference.

To quantify these properties, we employed the following metrics:

- **CLIP Similarity:** We computed CLIP image embeddings for both the original and generated images. Content alignment was evaluated by computing cosine similarity between the reference content image and the generated image embeddings. Style alignment was evaluated by comparing the generated image with the style reference image.
- **DINOv2 Similarity:** We applied the same evaluation protocol using DINOv2, a self-supervised vision transformer known for its strong visual feature representations. Content alignment was measured as the cosine similarity between DINOv2 embeddings of the reference content and generated images, while style alignment was measured as the cosine similarity between the style reference image and generated images.
- **DreamSim Distance:** We computed DreamSim distances between the generated image and both the reference content and style images. DreamSim is a perceptual similarity metric that aggregates activations from multiple layers of CLIP to capture fine-grained visual similarity beyond high-level semantics.

4. Experimental Results

4.1. Quantitative Results

Tables 1 and 2 present the average content and style alignment scores, respectively, across 30 image-to-image style transfer combinations per method.

Among all methods, B-LoRA achieves the highest content alignment scores across all three metrics. This suggests that training content and style LoRAs jointly allows for better structural preservation in the generated outputs.

IP-LoRA variants, which introduce identity-preserving loss during Content LoRA training, slightly improve upon ConsisLoRA in content alignment. For example, IP-LoRA (ArcFace, λ=25) yields higher CLIP and DINOv2 similarities than ConsisLoRA. This indicates that the identity preservation loss contributes positively to maintaining facial features and structural integrity. However, the improvement is relatively modest numerically, suggesting that current embedding-based metrics may not be sufficiently sensitive to identity-level changes in facial content.

Table 1. Average Content Similarity Metrics

Model	CLIP ↑	DINO ↑	DreamSim ↓
B-LoRA	**0.653**	**0.551**	**0.512**
ConsisLoRA	0.592	0.457	0.618
IP-LoRA$_{\text{ArcFace, }\lambda=10}$	0.586	0.471	0.608
IP-LoRA$_{\text{ArcFace, }\lambda=25}$	0.594	0.473	0.602
IP-LoRA$_{\text{DINOv2, }\lambda=10}$	0.585	0.471	0.606

In contrast, ConsisLoRA outperforms all methods in style alignment. IP-LoRA variants, while close in score, consistently perform slightly worse in style alignment compared to ConsisLoRA. This trade-off reveals an important limitation: while adding identity-preserving loss improves content fidelity, it slightly constrains the Style LoRA's ability to fully express stylistic features.

This trade-off becomes evident when comparing ConsisLoRA and IP-LoRA. The introduction of identity loss helps retain facial identity but appears to limit the extent of style transformation, likely due to the optimization objective pulling generated outputs closer to the content image in embedding space.

Table 2. Average Style Similarity Metrics

Model	CLIP ↑	DINO ↑	DreamSim ↓
B-LoRA	0.657	0.466	0.527
ConsisLoRA	**0.736**	**0.516**	**0.427**
IP-LoRA$_{\text{ArcFace, }\lambda=10}$	0.736	0.503	0.434
IP-LoRA $_{\text{ArcFace, }\lambda=25}$	0.731	0.502	0.437
IP-LoRA $_{\text{DINOv2, }\lambda=10}$	0.727	0.502	0.436

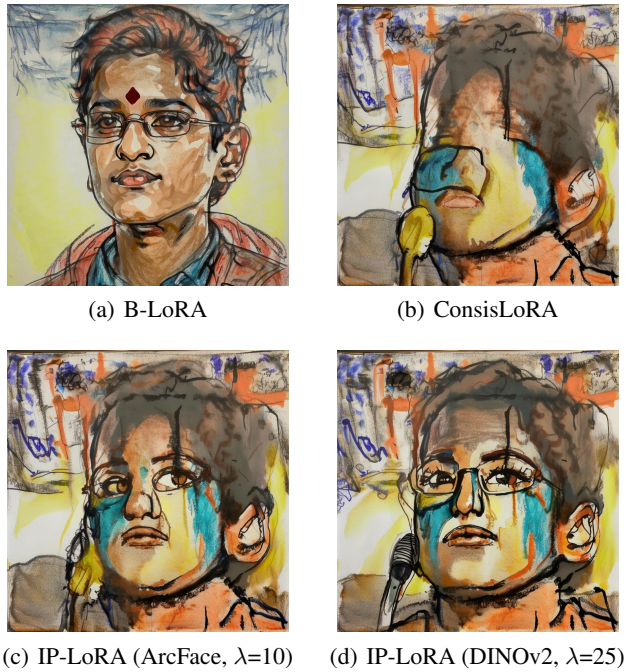
4.2. Qualitative Results

To further examine how well our model achieved identity preservation beyond quantitative metrics, we present qualitative comparisons across four stylization methods: B-LoRA, ConsisLoRA, IP-LoRA with ArcFace Loss, and IP-LoRA with DINOv2 Loss. Figures 2 and 3 display model outputs for two distinct prompts, each stylized using a different style and content images. These examples allow direct visual comparison of identity fidelity and stylistic consistency across methods. The original style and content reference images can be seen in the Appendix of this study, where we display inference results for every possible style-content pair in our dataset.

In Figure 2, the prompt is "An Indian Teen in the style of Drawing." The B-LoRA output (a) maintains facial geometry very well but lacks expressive detail and stylization finesse when compared to ConsisLoRA and IP-LoRA. However, it would be appropriate to state that it is the most visually pleasing among all four inference results. ConsisLoRA (b) improves stylistic abstraction, but severely compromises the facial structure, resulting in a distorted and less recognizable portrait. In contrast, both variants of IP-LoRA (c) using ArcFace loss and (d) using DINOv2 loss produce more semantically faithful results. Notably, IP-LoRA (DINOv2, λ=25) offers the most visually balanced output: it preserves defining features such as the eyes, glasses, and facial contour while also achieving strong artistic expression in the drawing style.

Figure 3 presents results for the prompt "A Chinese Man in the style of Painting." Again, B-LoRA (a) offers a good base identity retention but appears less expressive in terms of style. ConsisLoRA (b) applies vibrant styling, but causes substantial facial distortion and the generated image's identity is almost impossible to recognize. IP-LoRA (ArcFace, λ=10) (c) achieves improved identity retention while adapting well to the painting style. The other implementation of IP-LoRA (DINOv2, λ=25) (d) goes further by offering a clearer expression and better eye structure, with the identity more faithfully reconstructed compared to all baselines.

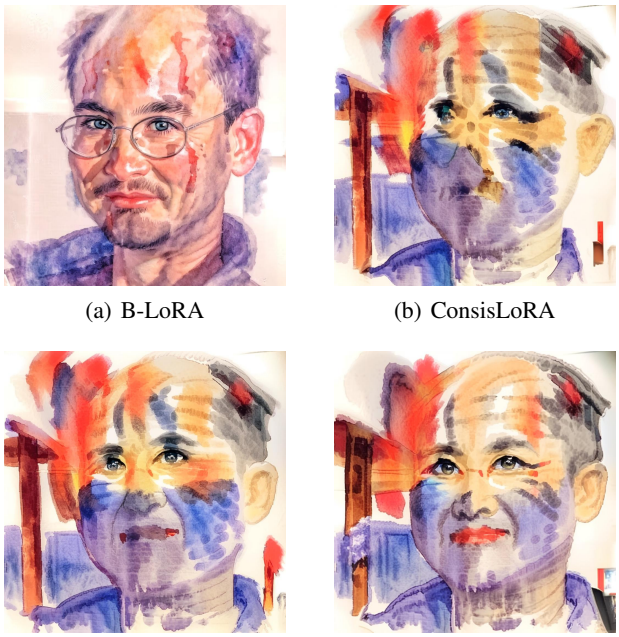
Overall, these visual comparisons validate the impact of embedding-level regularization during Content LoRA training. IP-LoRA consistently enhances identity preservation while maintaining the stylistic expressiveness of ConsisLoRA. Although these findings are somewhat consistent with our quantitative metrics, as the numerical changes are very small, we believe there is a significant need to explore alternative evaluation metrics that can better capture the extent of identity preservation.



(a) B-LoRA (b) ConsisLoRA

(c) IP-LoRA (ArcFace, λ=10) (d) IP-LoRA (DINOv2, λ=25)

Figure 2. Inference Results for the prompt: *"An Indian Teen in the style of Drawing."*



(a) B-LoRA (b) ConsisLoRA

(c) IP-LoRA (ArcFace, λ=10) (d) IP-LoRA (DINOv2, λ=25)

Figure 3. Inference Results for the prompt: *"A Chinese Man in the style of Painting."*

4.3. Limitations

While IP-LoRA demonstrates improved identity preservation in stylized portraits, several limitations remain.

First, our evaluation relies primarily on embedding-based and perceptual similarity metrics—CLIP, DINOv2, and DreamSim—which are not explicitly designed for identity preservation tasks. These models capture broad semantic and visual alignment, but often overlook fine-grained facial structure critical to identity. As a result, improvements that are clearly perceptible to human observers may not be reflected quantitatively. This highlights a gap in current evaluation protocols for identity-aware generative tasks.

Second, our method uses a fixed weighting parameter λ to balance the identity preservation loss against the diffusion objective. While this static approach is effective, it may not adapt optimally across different training phases. A dynamic λ scheduling mechanism, based on training loss behavior or identity confidence, could provide better balance between preserving content fidelity and enabling expressive stylization.

These limitations point toward future work on more adaptive loss weighting and improved identity-sensitive evaluation metrics.

5. Conclusions

In this paper, we introduced IP-LoRA, a novel extension of LoRA-based diffusion pipelines for portrait stylization with identity preservation. Building upon B-LoRA and ConsisLoRA, our method introduces an identity-aware regularization loss derived from ArcFace or DINOv2 embeddings into Content LoRA training. The additional supervision encourages the model to maintain fine-grained facial details and structural information vital to human identity while still allowing for the generation of abundant stylistic variety using independently trained Style LoRAs.

Through a set of controlled experiments on 30 content-style pairs and multiple stylization modes, we demonstrated that IP-LoRA improves content alignment consistently over existing baselines. Although there is a small trade-off in style alignment scores, qualitative results reveal that our method performs perceptually better in identity preservation—particularly in facial regions—without undermining the aesthetic quality of the stylization.

Our findings point to the inability of current embedding-based metrics like CLIP, DINOv2, and DreamSim to capture identity-specific fidelity, and suggest that more specialized evaluation protocols or human studies may be needed for fair benchmarking for such applications.

For future work, one promising direction is the dynamic adjustment of the identity loss weight λ during training. Instead of keeping it fixed, λ could be adaptively modified, raised or lowered, depending on whether the overall training loss stays below certain threshold. This would allow the model to better balance identity preservation and stylization quality throughout training. Additionally, extending the dataset to include a wider variety of reference images, especially with diverse facial poses and lighting conditions, could could possibly further increase the model's overall generalizability to real-world instances and resilience in terms of identity preservation at different viewing angles.

Overall, IP-LoRA offers a promising step toward more semantically faithful, identity-aware stylization systems and paves the way for future work in responsible and user-centric generative design.

References

an Li, H., Wang, L., and and, J. L. A review of deep learning-based image style transfer research. *The Imaging Science Journal*, 73(4):504–526, 2025. doi: 10.1080/13682199.2024.2418216. URL https://doi.org/10.1080/13682199.2024.2418216.

Chen, B., Zhao, B., Xie, H., Cai, Y., Li, Q., and Mao, X. Consislora: Enhancing content and style consistency for lora-based style transfer, 2025. URL https://arxiv.

org/abs/2503.10614.

Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., and Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, October 2022. ISSN 1939-3539. doi: 10.1109/tpami.2021.3087709. URL http://dx.doi.org/10.1109/TPAMI.2021.3087709.

Frenkel, Y., Vinker, Y., Shamir, A., and Cohen-Or, D. Implicit style-content separation using b-lora, 2024. URL https://arxiv.org/abs/2403.14572.

Guo, J., Deng, J., An, X., Yu, J., and Gecer, B. Insightface: 2d and 3d face analysis project. https://github.com/deepinsight/insightface, 2021.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. URL https://arxiv.org/abs/2307.01952.

Acknowledgements

Oğuz Kağan Hitit implemented the SDXL model backbone, and developed the training pipelines for B-LoRA, ConsisLoRA, and IP-LoRA from scratch.

İdil Görgülü implemented the inference pipelines, designed and implemented the integration of ArcFace and DINOv2 embeddings, wrote the evaluation scripts, and conducted the research and model selection for identity encoders.

Both authors contributed equally to the presentations and written reports. Training and inference runs were carried out collaboratively in parallel.

The full implementation of IP-LoRA, including training scripts, model checkpoints, and evaluation pipelines, is available at: https://github.com/idil-gorgulu/identity-preserving-lora-style-transfer

Appendix: Full Inference Results

This appendix presents the complete stylization results for all content-style pairings across each evaluated method. Each figure is organized as a grid where rows represent different content images and columns represent various artistic styles.

Figure 4. Full inference results across all content-style pairings using **B-LoRA**.

Figure 5. Full inference results across all content-style pairings using **ConsisLoRA**.

Figure 6. Full inference results across all content-style pairings using **IP-LoRA (ArcFace Loss,** $\lambda = 10$).

Figure 7. Full inference results across all content-style pairings using **IP-LoRA (DINOv2 Loss,** $\lambda = 25$**)**.