Diff-ID: Identity Consistent Facial Image Generation and Morphing via Diffusion Models

Anonymous authors
Paper under double-blind review

Abstract

Generative diffusion models have revolutionized facial image synthesis, yet robust identity preservation in high resolution outputs remains a critical challenge. This issue is especially vital for security systems, biometric authentication, and privacy sensitive applications, where any drift in identity integrity can undermine trust and functionality. We introduce Diff-ID, a diffusion based framework that enforces identity consistency while delivering photorealistic quality. Central to our approach is a custom 210K image dataset synthesized from CelebA-HQ, FFHQ, and LAION-Face and captioned via a fine tuned BLIP model to bolster identity awareness during training. Diff-ID integrates ArcFace and CLIP embeddings through a dual cross attention adapter within a fine tuned Stable Diffusion UNet. To further reinforce identity fidelity, we propose a pseudo discriminator loss based on ArcFace cosine similarity with exponential timestep weighting. Experiments on held out and unseen faces demonstrate that Diff-ID outperforms state of the art methods in both identity retention and visual realism. Finally, we showcase a unified DDIM based morphing pipeline that enables seamless facial interpolation without per identity fine tuning. We further argue that identity preservation and photorealism should be evaluated jointly rather than in isolation, as high identity similarity alone does not guarantee realistic outputs. To this end, we introduce a unified evaluation metric that combines identity similarity and perceptual realism into a single interpretable score.

1 Introduction

Generative diffusion models have recently emerged as a leading paradigm for high fidelity image synthesis, outperforming earlier techniques such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) in training stability, diversity, and output quality Ho et al. (2020); Rombach et al. (2022). By iteratively refining noisy latent representations through a learned denoising process, diffusion based frameworks can generate hyper realistic images that capture intricate textures, complex structures, and subtle semantic variations.

The success of diffusion models spans diverse applications: from artistic content creation and virtual reality to medical imaging and scientific visualization, where photorealism and anatomical precision are critical Nichol & Dhariwal (2021); Saharia et al. (2022). In the domain of facial image synthesis, diffusion methods have demonstrated impressive visual quality, yet they exhibit limitations when precise control over identity attributes is required. In particular, subtle identity drift, manifesting as changes in facial geometry, expression, or distinguishing features, can undermine the trustworthiness of generated outputs in scenarios demanding strict identity fidelity.

Maintaining robust identity consistency is a multifaceted challenge: models must preserve core identity defining features (for example bone structure, eyes, and facial contours) under varying manipulations (for example expression, pose, lighting). However, most current diffusion methods Liu et al. (2023); Zhang et al. (2023b); Chen et al. (2023); Zhang et al. (2023a) emphasize global attribute changes, such as clothing styles or accessories, and produce avatar like, non photorealistic faces rather than fine grained semantic edits or strict identity preservation.

This identity control gap poses a significant barrier for security sensitive applications, including biometric authentication, forensic analysis, and privacy preserving data generation, where any deviation from the true identity can have severe consequences. Synthetic morphing pipelines, used for adversarial robustness testing and privacy aware dataset augmentation, further demand seamless identity interpolation without per identity retraining or checkpoint swapping.

Beyond methodological motivation, identity consistent and photorealistic face generation is essential for several practical application scenarios. A primary example arises from data protection and privacy regulations such as GDPR, which increasingly restrict the use of real biometric data for training and evaluation. In this context, synthetic but photorealistic identities provide a viable alternative, enabling the development and benchmarking of face recognition systems without exposing real individuals. Crucially, such synthetic identities must preserve realistic facial structure while exhibiting diversity in pose, illumination, resolution, expression, and hairstyle in order to reflect real operational conditions.

A second important scenario is dataset augmentation. Modern face recognition systems require large and diverse training corpora, yet collecting balanced samples across capture conditions and demographics remains challenging. Identity preserving generative models enable controlled augmentation by synthesizing realistic variations of a subject without compromising identity coherence. Finally, identity consistent morphing provides a stepping stone toward morph attack generation and detection, where realistic blends of identities are required to study vulnerabilities of biometric systems. In all these scenarios, the simultaneous preservation of identity and photorealism is not optional but essential, motivating the design of Diff-ID.

To address these challenges, we propose **Diff-ID**, a unified diffusion framework explicitly designed for identity aware facial image synthesis and morphing. Our contributions are fourfold:

1. **Custom Identity Centric Dataset**: We curate a 210K image corpus by blending and captioning CelebA-HQ, FFHQ, and LAION-Face images with a fine tuned BLIP model, ensuring comprehensive coverage of facial variations and semantic contexts.

2. **Dual Cross Attention Adapter**: We develop a lightweight adapter that fuses identity embeddings from ArcFace with semantic embeddings from CLIP within a fine tuned Stable Diffusion UNet, enabling robust identity preservation while maintaining photorealistic visual quality. Our framework does not aim to provide explicit attribute editing or fine grained attribute control.

3. **Pseudo Discriminator Identity Loss**: We introduce an ArcFace based identity loss with exponential dynamic weighting over diffusion timesteps, reinforcing identity coherence throughout the denoising process.

4. **Unified Morphing Pipeline**: Leveraging DDIM inversion and joint embedding plus latent interpolation, Diff-ID performs smooth, identity preserving morphs between arbitrary face pairs without additional model fine tuning or multiple checkpoints.

5. **Unified Identity–Realism Evaluation Metric**: We introduce the Face Image Quality (FIQ) metric, which jointly captures identity preservation and photorealism by coupling ArcFace similarity with Fréchet Inception Distance, enabling more meaningful evaluation of identity guided face generation models.

We validate Diff-ID on both held out and unseen high resolution facial datasets, showing marked improvements in ArcFace similarity scores and Fréchet Inception Distance (FID) compared to state of the art baselines. Our approach sets a new benchmark for identity aware diffusion models and paves the way for practical, security driven applications in facial synthesis and biometric data augmentation.

2 Related Work

Diff-ID builds on two converging lines of research: general diffusion based image synthesis and specialized identity preserving adaptations. We first review foundational text to image diffusion frameworks, then

examine approaches for embedding based identity control, attention mechanisms, adapter modules, and morphing techniques.

2.1 Text to Image Diffusion Models

Text to image diffusion models have rapidly become the dominant paradigm in generative image synthesis by combining iterative denoising processes with powerful text encoders. Early transformer based approaches, such as DALL·E Ramesh et al. (2021), CogView Ding et al. (2021), and Make A Scene Gal et al. (2022), demonstrated rich text image correlations via discrete tokenization, but they incurred substantial compute and latency costs at ultra high resolutions due to quadratic self attention complexity Ramesh et al. (2021).

Diffusion models overcome these bottlenecks by operating in continuous latent spaces and gradually transforming noise into coherent images. Song and Ermon formalized this process via stochastic differential equations, showing that score matching on Gaussian perturbations recovers complex data distributions Song & Ermon (2021). Ho et al.'s denoising diffusion probabilistic models then introduced practical noise schedules for stable, end to end training Ho et al. (2020).

Subsequent advances improved both fidelity and efficiency. GLIDE proposed classifier free guidance, enabling a simple trade off between fidelity and diversity without an external classifier Nichol & Dhariwal (2021). DALL·E 2 adopted a hierarchical two stage diffusion, first generating CLIP embeddings from text, then super resolving back to image space, which further boosted sample quality Ramesh et al. (2022). Imagen achieved state of the art FID on COCO 30k by scaling up text encoders and carefully tuning noise schedules Saharia et al. (2022).

Latent Diffusion Models marked another leap by performing denoising in a compressed latent space, reducing memory and computation while retaining perceptual detail Rombach et al. (2022). Stable Diffusion, an open source latent diffusion model at 512×512 resolution, democratized these techniques and inspired numerous extensions for style, structure, and spatial conditioning Zhang et al. (2023c). Despite these successes in text driven generation, vanilla diffusion pipelines do not include mechanisms to anchor outputs to a specific face identity. As a result, attribute manipulations, such as altering expression or pose, can inadvertently shift identity features, leading to drift. This limitation underscores the need for identity aware diffusion frameworks like Diff-ID.

2.2 Zero Shot Embedding Approaches

Zero shot embedding techniques have recently emerged as an efficient way to steer pretrained text to image diffusion models toward identity preserving outputs without per subject fine tuning. Many methods build on large scale vision language encoders such as CLIP Radford et al. (2021) to extract semantic image representations, and then inject these into the generation process. IP Adapter Liu et al. (2023) uses a decoupled cross attention mechanism that maintains separate query, key, and value projections for text and image embeddings, allowing detailed visual prompts to be fused without updating the backbone model weights. Similarly, InstantID Chen et al. (2023) leverages a frozen face recognition encoder to produce identity embeddings on the fly, which are concatenated with text tokens at each denoising step to guide the model toward the correct subject.

Despite their efficiency, CLIP based zero shot embeddings can lack the fine grained discriminability needed for strict identity retention. In particular, they may fail to capture subtle but crucial facial characteristics, such as jawline contour, eye shape and spacing, nose bridge structure, lip curvature, and overall facial bone structure, leading to identity drift under challenging poses or lighting conditions Chen et al. (2023); Liu et al. (2023). To address this, Diff-ID integrates high precision ArcFace embeddings Deng et al. (2019b), which are explicitly trained with an additive angular margin objective to maximize inter class separability and preserve intra class consistency, into a frozen Stable Diffusion pipeline. By concatenating these embeddings with the model intermediate feature maps and applying a lightweight adapter layer, Diff-ID faithfully retains identity specific attributes, such as precise jawline geometry, characteristic eye proportions, nose bridge angle, and unique lip shape, across diverse generation tasks.

2.3 Decoupled Cross Attention Mechanisms

Decoupled cross attention is a technique that separates the processing of text and image inputs, improving how identity and attribute features are integrated during image generation. Models like IP Adapter Liu et al. (2023) and InstantID Chen et al. (2023) utilize decoupled cross attention to independently attend to textual and visual inputs, aligning generated content more precisely with the text description. This mechanism is particularly beneficial for preserving semantic alignment in the outputs, allowing models to generate contextually relevant images based on text prompts.

However, while decoupled cross attention enhances attribute alignment, it alone does not ensure identity consistency, particularly in models reliant solely on CLIP embeddings. In cases where detailed identity features are critical, this limitation can lead to identity drift and compromise realism. Diff-ID shows that tuning the existing cross attention layers in diffusion models, rather than adding decoupled modules, is sufficient to achieve strong identity consistency and visual realism. While more advanced semantic or accessory control may still require additional modules such as ControlNet, our results indicate that faithful identity preservation and photorealism can be achieved without decoupled attention, which significantly reduces architectural complexity.

2.4 Stacked ID Embeddings for Identity Fidelity

A promising direction for identity preservation in text to image diffusion is the use of stacked identity embeddings, where multiple reference images of the same individual are jointly encoded to form a richer identity representation. PhotoMaker Zhang et al. (2023b) exemplifies this strategy by first extracting per image identity features via a pretrained face encoder, then concatenating these into a single stacked embedding that is injected into each denoising step through a dual cross attention mechanism. By combining class based attributes (for example gender or age) with individual specific embeddings, PhotoMaker maintains consistent identity cues, such as jawline structure, eye spacing, and hairstyle, across variations in pose, expression, and lighting Zhang et al. (2023b).

This multi image fusion leads to substantial gains in identity fidelity. Similar ideas have appeared in related work on multi view face synthesis, where aggregating embeddings from different angles improves three dimensional consistency and realism Deng et al. (2019a). However, the stacked identity approach incurs significant computational cost: training an image encoder from scratch on large scale face datasets (for example MS Celeb 1M Guo et al. (2016)) and managing multiple high dimensional embeddings per subject can slow both training and inference.

In contrast, Diff-ID sidesteps these overheads by leveraging off the shelf ArcFace embeddings Deng et al. (2019b) within a frozen Stable Diffusion backbone Rombach et al. (2022). Rather than stacking many vectors, Diff-ID projects a single, high precision identity vector into the model intermediate feature space via a lightweight adapter. A dual cross attention block then fuses identity and textual attributes in parallel, preserving fine grained features, such as nose bridge angle and lip curvature, without the need for multi view encoding or encoder retraining. This yields comparable identity retention to stacked identity methods while reducing memory footprint and computational latency.

2.5 Diffusion Based Morphing and Synthetic Identity Generation

Interpolation and morphing between identities have long been explored with GANs via latent space operations. Early works such as StyleGAN Karras et al. (2019) demonstrated smooth identity interpolation by spherical linear interpolation between learned style vectors, while subsequent methods such as StyleFlow Abdal et al. (2021) and GANSpace Härkönen et al. (2020) provided fine grained, attribute aware traversals in the GAN latent manifold. However, GAN based interpolation often suffers from mode collapse and fixed model checkpoints per identity or attribute setting.

Diffusion models offer an alternative morphing paradigm. DiffMorpher Lee et al. (2024) performs latent space interpolation by inverting two target images into DDIM embeddings and linearly blending these codes during reverse diffusion. To achieve sharp identity transitions, DiffMorpher relies on per identity low rank

adaptation fine tuning Hu et al. (2021) and multiple diffusion checkpoints, which incurs substantial compute and engineering overhead.

In contrast, Diff-ID natively supports morphing via our unified adapter. During DDIM inversion, we fuse ArcFace identity embeddings Deng et al. (2019b) with CLIP text embeddings Radford et al. (2021) through a dual cross attention block. This semantically conditioned noise injection yields latent trajectories that preserve identity features throughout forward and reverse passes, enabling smooth transitions without external fine tuning or checkpoint swapping. As a result, Diff-ID facilitates not only continuous identity morphing but also large scale synthetic identity generation, privacy preserving face blending, and adversarial robustness testing, all while maintaining high visual quality and identity fidelity.

3 Methodology

3.1 Data Preparation and Curation

To support identity consistent and semantically rich facial generation, we curated a large scale dataset composed of images from three well known high resolution facial datasets: CelebA-HQ, FFHQ, and LAION-Face. All images were resized to a uniform resolution of 512 to ensure consistency during model training.

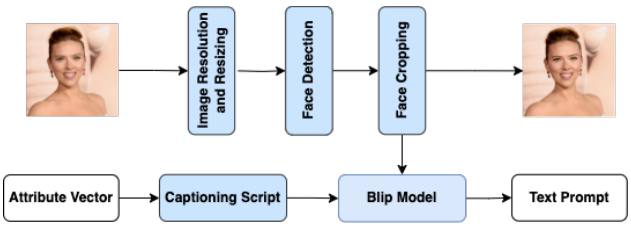


Figure 1: Data pipeline used to curate and caption the identity centric dataset.

Unlike CelebA-HQ, which provides 40 annotated binary attributes per image, FFHQ and LAION-Face lacked descriptive captions. To address this, we developed a custom script that converts CelebA-HQ attribute vectors into structured sentence style captions. We then fine tuned a BLIP model on this annotated subset to generate high quality captions for FFHQ and LAION-Face images.

To maintain data quality and semantic consistency, we removed images with occlusions (for example sunglasses, hats), low quality images, and any samples undetectable by the InsightFace library Ren et al. (2023); Guo et al. (2021); Gecer et al. (2021); An et al. (2022; 2021); Deng et al. (2020a;b); Guo et al. (2018); Deng et al. (2018; 2019c). This rigorous cleaning and captioning process resulted in a dataset of approximately 210,000 images, each paired with a descriptive and unique caption, which enables more expressive and semantically descriptive prompts while maintaining identity fidelity, without providing explicit attribute control.

3.2 Overview of Diff-ID

Diff-ID is a diffusion framework built on a fine tuned Stable Diffusion UNet Rombach et al. (2022) that enforces strict identity consistency in high resolution facial synthesis. At its core is a lightweight dual cross attention adapter that fuses high precision ArcFace identity embeddings Deng et al. (2019b) with CLIP semantic embeddings Radford et al. (2021), anchoring generated outputs to the target identity. This design ensures that identity specific attributes, such as facial bone structure, eye spacing, nose shape, and lip contour, are preserved with high fidelity during denoising, while still allowing CLIP driven semantic prompts to provide broad contextual guidance within identity constraints.

To reinforce identity fidelity throughout the diffusion process, we introduce a pseudo discriminator identity loss: an ArcFace cosine similarity term with exponential timestep weighting that dynamically emphasizes the preservation of identity cues as noise levels decrease.

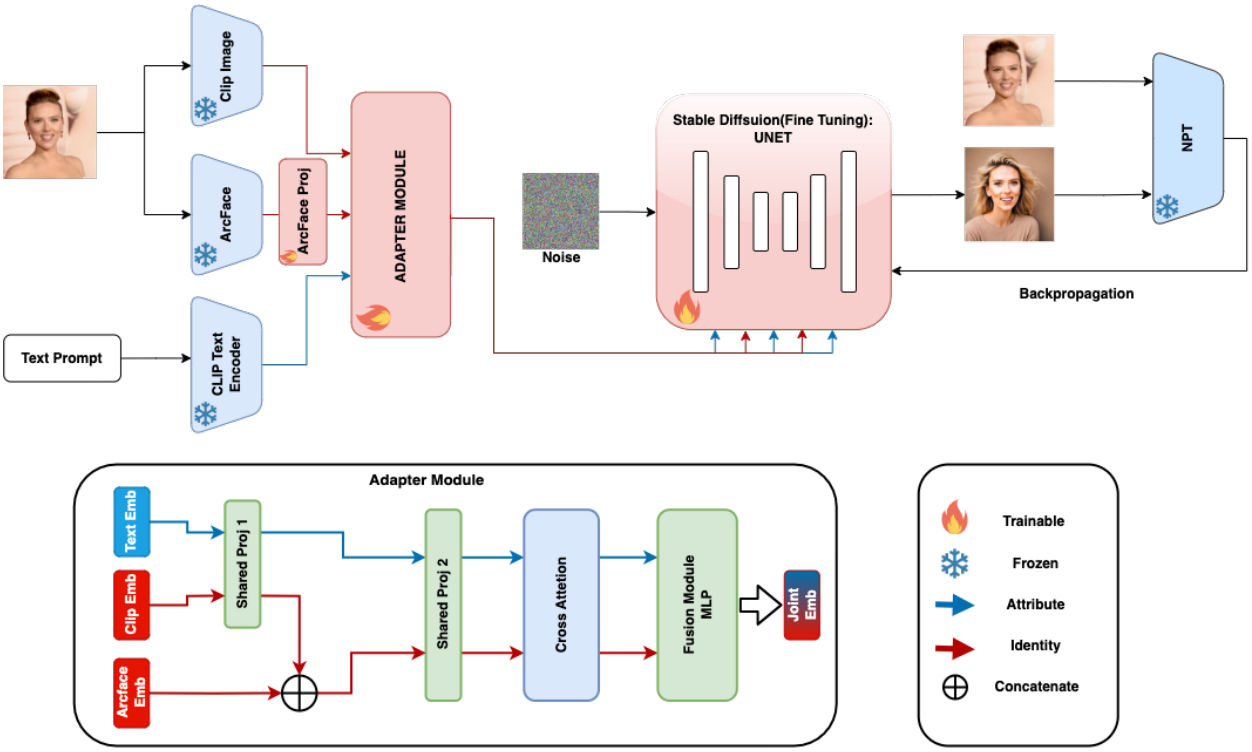


Figure 2: Diff-ID architecture. We extract semantic features from CLIP and identity features from ArcFace, project and fuse them into a unified identity representation, and process them via dual cross attention and a Fusion multilayer perceptron. The resulting embeddings are cross attended into a pre trained Stable Diffusion UNet to guide identity consistent denoising.

3.3 Model Architecture

Diff-ID is built upon a pretrained Stable Diffusion 1.5 UNet architecture, chosen for its effectiveness in high resolution image synthesis within a compressed latent space Rombach et al. (2022). The architecture consists of several key modules designed to retain identity fidelity while maintaining the flexibility needed for morphing and prompt conditioned synthesis, without targeting explicit attribute editing.

3.3.1 Embedding Extraction and Fusion

In many prior models, identity features are extracted using a single embedding source, either via ArcFace (for example in Arc2Face) or through the CLIP image encoder (for example in IP Adapter and InstantID). However, relying solely on one modality can be limiting. ArcFace is highly effective at capturing fine grained, identity specific details that are crucial for face recognition, but it may not fully capture the semantic nuances necessary for aligning with textual descriptions. In contrast, the CLIP image encoder provides robust semantic representations that align well with text inputs, yet it may overlook subtle identity cues that are essential for preserving individual characteristics. To address these complementary strengths and weaknesses, Diff-ID integrates both embedding types, thereby ensuring a more comprehensive representation of identity and semantics.

CLIP Embeddings Let A denote the input text prompt and I the reference image. The text embedding is computed as $e_t = \text{CLIP}_{\text{text}}(A)$ while the image embedding is $e_i = \text{CLIP}_{\text{image}}(I)$. Here, $\text{CLIP}_{\text{text}}$ and $\text{CLIP}_{\text{image}}$ represent the text and image encoding components of CLIP, respectively, with e_t encapsulating the semantic content of the prompt and e_i providing complementary visual context.

ArcFace Embeddings Given the same input image I, the ArcFace model produces a 512 dimensional identity embedding $e_{\text{arc}} = \text{ArcFace}(I)$. This embedding captures unique identity specific features that are critical for maintaining identity fidelity during generation Deng et al. (2019b).

Embedding Projection and Fusion To integrate multi modal information, we first project the ArcFace and CLIP image embeddings into a common latent space. Their projected forms are given by

$$e'_{\text{arc}} = W_{\text{arc}}\, e_{\text{arc}}, \qquad e'_i = W_i\, e_i, \tag{1}$$

where W_{arc} and W_i are learnable projection matrices. Next, we concatenate these projected embeddings to form a combined image representation,

$$e_{\text{img}} = \text{Concat}(e'_{\text{arc}},\, e'_i)\,. \tag{2}$$

To further capture salient identity features, we apply both max pooling and average pooling on e_{img}, yielding $e_{\text{max}} = \text{MaxPool}(e_{\text{img}})$ and $e_{\text{avg}} = \text{AvgPool}(e_{\text{img}})$. We then form the final identity representation by concatenating the individual components:

$$e_{\text{id}} = \text{Concat}\big(e'_{\text{arc}},\, e'_i,\, e_{\text{max}},\, e_{\text{avg}}\big)\,. \tag{3}$$

In parallel, the CLIP text embedding e_t is aligned with the identity modality by passing both e_{id} and e_t through a shared linear projection layer with weight matrix W_s, resulting in

$$\tilde{e}_{\text{id}} = W_s\, e_{\text{id}}, \quad \tilde{e}_t = W_s\, e_t\,. \tag{4}$$

This yields two separate yet aligned branches: the identity branch and the text branch for subsequent processing. Although $\tilde{e}_{\text{id}}$ and $\tilde{e}_t$ share the same dimensionality, we deliberately avoid projecting ArcFace directly into the CLIP semantic space. ArcFace embeddings encode geometric identity under an angular margin objective, whereas CLIP encodes global semantic and contextual cues. Forcing them into a single shared representation collapses this complementarity and empirically degrades identity discriminability. Instead, Diff-ID maintains modality specific projections and fuses them only after cross attention refinement, preserving both fine grained identity cues and semantic alignment.

3.3.2 Dual Cross Attention Mechanism

To effectively integrate and refine the identity and textual information, we propose a dual cross attention mechanism. The motivation behind this design is twofold. First, by enabling bidirectional interactions, the model allows each modality to inform and enhance the other, ensuring that identity features are enriched by semantic cues from the prompt and conversely that semantic features are refined in an identity aware way. Second, by omitting softmax normalization, the mechanism preserves the raw magnitude relationships between features, which can help retain fine grained details that might otherwise be diminished. Importantly, we do not treat this as a dedicated attribute editing module: the goal is to stabilise identity under text conditioning, rather than to offer precise, controllable attribute manipulation.

We compute two cross attention branches jointly using the generic template

$$e = \left(\frac{QK^\top}{\sqrt{d}}\right) V, \tag{5}$$

where d is the embedding dimensionality.

Identity Branch In the identity branch, the identity embedding $\tilde{e}_{\text{id}}$ serves as the query and the CLIP text embedding $\tilde{e}_t$ provides both the key and the value:

$$Q_{\text{id}} = W_q^{\text{id}}\, \tilde{e}_{\text{id}}, \quad K_{\text{id}} = W_k^{\text{id}}\, \tilde{e}_t, \quad V_{\text{id}} = W_v^{\text{id}}\, \tilde{e}_t. \tag{6}$$

The enriched identity representation is then

$$e_{\text{identity}} = \left(\frac{Q_{\text{id}}\, K_{\text{id}}^\top}{\sqrt{d}}\right) V_{\text{id}}. \tag{7}$$

Text Branch In the text branch, the CLIP text embedding $\tilde{e}_t$ acts as the query, while the identity embedding $\tilde{e}_{\text{id}}$ is used as both key and value:

$$Q_{\text{attr}} = W_q^{\text{attr}}\,\tilde{e}_t, \quad K_{\text{attr}} = W_k^{\text{attr}}\,\tilde{e}_{\text{id}}, \quad V_{\text{attr}} = W_v^{\text{attr}}\,\tilde{e}_{\text{id}}, \tag{8}$$

$$e_{\text{attribute}} = \left(\frac{Q_{\text{attr}}\,K_{\text{attr}}^{\top}}{\sqrt{d}}\right) V_{\text{attr}}. \tag{9}$$

3.3.3 Fusion Multilayer Perceptron

Although using a multilayer perceptron for feature fusion is standard practice, our design incorporates a Fusion multilayer perceptron module for an important reason: it enables the network to learn complex, nonlinear interactions between the enriched identity and text features. In doing so, we refine the joint representation in a way that preserves identity specific details while maintaining consistency with the textual description, without attempting explicit fine grained attribute control.

The outputs from the dual cross attention modules, the enriched identity representation e_{identity} and the enriched text representation $e_{\text{attribute}}$, are concatenated to form a unified feature representation:

$$e_{\text{fused}} = \text{Concat}\big(e_{\text{identity}},\, e_{\text{attribute}}\big)\,. \tag{10}$$

This fused embedding is then processed through a Fusion multilayer perceptron block to generate a refined embedding:

$$e_{\text{refined}} = \text{MLP}\big(e_{\text{fused}}\big)\,, \tag{11}$$

where the multilayer perceptron is defined as:

$$\text{MLP}(x) = \sigma\big(W_2\,\sigma(W_1 x + b_1) + b_2\big)\,. \tag{12}$$

Here, W_1 and W_2 are learnable weight matrices, b_1 and b_2 are biases, and σ represents the rectified linear unit activation function.

3.4 Objective Function

Our objective function is carefully designed to balance two essential goals: high quality denoising and robust identity preservation. The denoising loss, adapted from the Stable Diffusion framework, drives the model to effectively remove noise and reconstruct images that follow the target distribution. In parallel, the adversarial identity loss, computed using ArcFace embeddings, ensures that the generated images retain core identity features. Recognizing that identity details are less discernible at higher noise levels, we incorporate an exponential dynamic weighting strategy. This strategy modulates the importance of identity loss throughout the diffusion process, emphasizing identity preservation when the noise is lower and identity information is more reliable.

Denoising Loss Following the Stable Diffusion framework, the model minimizes the difference between the true noise ϵ added during the forward process and the noise predicted by the network $\epsilon_\theta(z_t, t)$. The denoising loss is given by:

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{z_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right] \tag{13}$$

where t is uniformly sampled from $\{1, 2, \ldots, T\}$ and the noisy latent z_t is computed as:

$$z_t = \sqrt{\alpha_t}\,z_0 + \sqrt{1-\alpha_t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{14}$$

Adversarial Identity Loss To ensure that generated images preserve core identity features, ArcFace is employed as a pseudo discriminator. It extracts identity embeddings e_{original} and $e_{\text{generated}}$ from the original and generated images, respectively. The identity loss is defined as

$$\mathcal{L}_{\text{identity}} = 1 - \text{CosineSim}(e_{\text{original}}, e_{\text{generated}}) \tag{15}$$

where CosineSim denotes the cosine similarity between the two embeddings.

Exponential Dynamic Weighting Strategy Since identity features become less discernible at higher noise levels, we weight the identity loss exponentially based on the current timestep. Let $\tau = \frac{t}{T}$ (with t as the current timestep and T as the total number of timesteps) and define the weight as

$$W_t = \exp\big(-k\,\tau\big), \tag{16}$$

with k as a decay hyperparameter controlling the rate of exponential decrease. The weighted identity loss per sample is computed as

$$\mathcal{L}_{\text{identity, weighted}} = \mathbb{E}\Big[W_t \cdot \mathcal{L}_{\text{identity}}\Big]. \tag{17}$$

Intuitively, this schedule aligns the strength of identity supervision with the diffusion dynamics. At early timesteps the latent is dominated by Gaussian noise and identity cues are unreliable, so the identity loss is down weighted. At later timesteps, when the denoiser reconstructs semantically coherent faces, identity discrepancies become meaningful and the loss gradually plays a larger role.

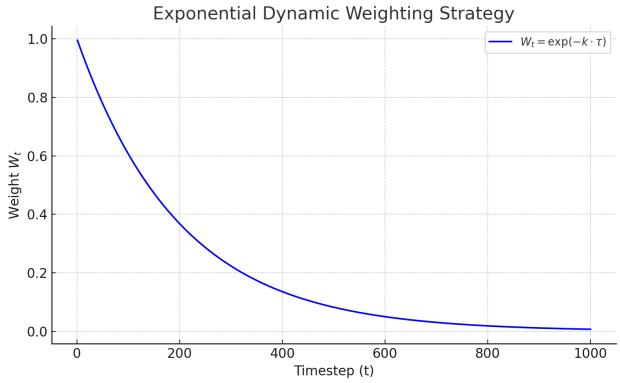


Figure 3: Illustration of the exponential timestep weighting W_t as a function of the normalized timestep τ. Identity supervision is weak at high noise levels and becomes increasingly important as the denoising process progresses.

Overall Training Objective The final training loss is a combination of the denoising loss and the dynamically weighted identity loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{denoise}} + \lambda_{\text{identity}} \cdot \mathcal{L}_{\text{identity, weighted}}, \tag{18}$$

where $\lambda_{\text{identity}} = 0.10$ is a hyperparameter balancing the importance of identity preservation against denoising accuracy.

3.5 Morphology Based Generation

We extend Diff-ID to support identity preserving facial morphing between two source identities without requiring per identity fine tuning. Our method builds on the pretrained dual branch adapter and UNet backbone and leverages deterministic sampling via DDIM to ensure stable and high fidelity outputs. The generation proceeds in three key stages.

Stage 1: Inversion to Latent Space Given two real input images x_1 and x_2, we use DDIM inversion Song & Ermon (2020) to recover their corresponding noise latents z_1^T and z_2^T at the final timestep T of the diffusion process:

$$z^T = \text{DDIMInvert}(x;\theta) = \Phi_\theta^{-1}(x)\,, \tag{19}$$

where θ are the parameters of the pretrained denoising network and Φ_θ^{-1} represents the learned inversion mapping. This step ensures the latent noise preserves both the visual and identity information of each input image in the diffusion space.

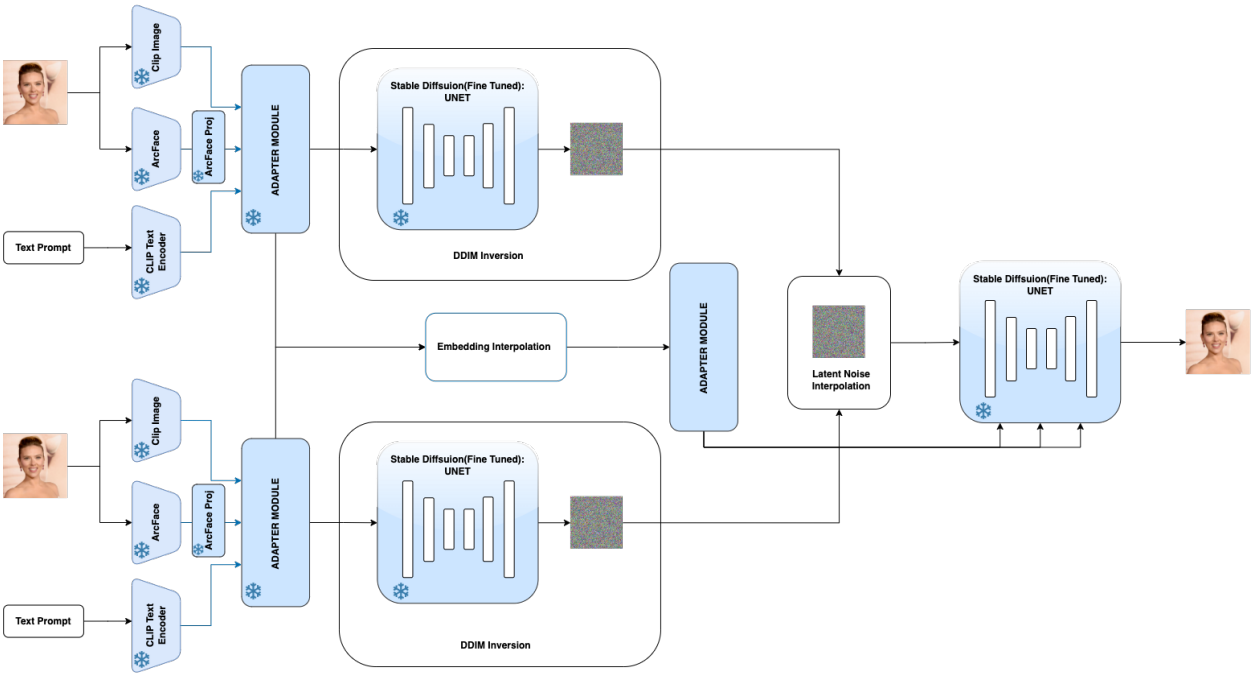


Figure 4: DiffID Morph: identity preserving morphing pipeline built on DDIM inversion and dual embedding interpolation.

Stage 2: Embedding and Latent Interpolation We extract identity embeddings e_1 and e_2 using the pretrained CLIP and ArcFace encoders, as integrated into the Diff-ID dual adapter module. To morph between the two identities, we perform spherical linear interpolation of the embeddings on the unit hypersphere:

$$\begin{aligned} e_{\text{mix}} &= \text{SLERP}(e_1, e_2; \alpha) \\ &= \frac{\sin\big((1-\alpha)\,\omega\big)}{\sin\omega}\, e_1 + \frac{\sin\big(\alpha\,\omega\big)}{\sin\omega}\, e_2, \\ \omega &= \arccos\Big(\frac{e_1 \cdot e_2}{\|e_1\|\|e_2\|}\Big), \quad \alpha \in [0,1] \end{aligned} \tag{20}$$

where α controls the morphing intensity between the two faces. In parallel, we interpolate the latent noise vectors z_1^T and z_2^T using linear interpolation:

$$z_{\text{mix}}^T = (1-\alpha)\, z_1^T + \alpha\, z_2^T\,, \tag{21}$$

which ensures that the structure and fine details of both identities are blended in the denoising trajectory. Combining both embedding and latent interpolation allows for smooth transitions in both semantic identity space and low level generative signal.

Stage 3: Identity Conditioned Sampling Using the interpolated noise vector z_{mix}^T and the blended identity embedding e_{mix}, we perform deterministic DDIM sampling guided by the Diff-ID dual cross attention mechanism:

$$\hat{x} = \text{DDIMSample}\big(z_{\text{mix}}^T;\, e_{\text{mix}},\, \theta\big)\,, \tag{22}$$

producing the final morphed image $\hat{x}$ that exhibits characteristics from both source identities while preserving realism and coherence. Our approach supports continuous control over the morph factor α, which enables a spectrum of identities between x_1 and x_2.

Unlike other methods that require multiple fine tuned checkpoints, for example DreamBooth Ru et al. (2022) or DiffMorpher Lee et al. (2024), or low rank adaptation layers such as low rank adaptation based

fine tuning Hu et al. (2021), our pipeline operates in a single unified model. This results in faster, scalable inference suitable for applications such as identity obfuscation Rathgeb & Busch (2019), biometric robustness testing Ferrara et al. (2014), and photorealistic avatar creation in augmented reality and virtual reality systems via three dimensional morphable models Blanz & Vetter (1999).

3.6 Training Configuration and Hyperparameters

The training setup for Diff-ID involves a mix of frozen and trainable layers, with the following components kept frozen to retain their pretrained feature extraction capabilities:

- **Variational Autoencoder Layers**: Encoder and decoder layers of the variational autoencoder are frozen to maintain the latent to image space transformations without additional fine tuning Rombach et al. (2022).
- **CLIP and ArcFace Models**: Both models are kept frozen, preserving their pretrained strengths in capturing semantic (CLIP) and identity specific (ArcFace) features Deng et al. (2019b).

Diff-ID was trained using the Adam optimizer with a learning rate of 1×10^{-5} and a batch size of 16, employing mixed precision (fp16) for efficiency. Gradient checkpointing was utilized to manage memory consumption, which allows larger batch sizes without exceeding graphics processing unit memory limits. The training was conducted over 1,000,000 steps using four NVIDIA A100 80GB graphics processing units, achieving convergence and stable performance in approximately 72 hours.

4 Evaluation and Results

4.1 Evaluation Framework

To thoroughly assess Diff-ID effectiveness, we employ a two pronged evaluation strategy that combines qualitative inspection with quantitative metrics targeting identity preservation and visual realism.

4.1.1 Qualitative Analysis

Following established practice Liu et al. (2023); Zhang et al. (2023a), we present side by side image grids comparing Diff-ID outputs with those of baseline methods. This visual inspection emphasizes:

- **Identity Fidelity:** The extent to which characteristic facial features are retained.
- **Perceptual Realism:** The absence of unnatural artifacts or distortions.

Because all models are driven by the same BLIP generated descriptive prompts for each identity, the qualitative comparisons also reveal whether semantic attributes (for example hair colour, apparent age cues, coarse expression, and lighting) remain consistent with the caption. Diff-ID is not designed as an explicit attribute editing system, and we do not claim or evaluate fine grained attribute control. In practice, the reciprocal cross attention in Section 3 can still yield some identity preserving, prompt aligned variation in these high level attributes while maintaining geometry and identity defining structure, but this behaviour is incidental rather than a primary objective.

4.1.2 Quantitative Analysis

We evaluate two complementary dimensions of performance.

Identity Preservation To quantify how well a generated image maintains the identity of the target subject, we use the **Face Similarity (FS)** metric. FS is computed as the cosine similarity between the 512-dimensional embeddings extracted by a pretrained ArcFace model Deng et al. (2019b) for each source–generated image pair. Since ArcFace is optimized for face recognition, FS provides a strong and widely

adopted proxy for measuring identity fidelity. However, we note that FS measures identity independently of visual realism or style, and may therefore assign high similarity scores to highly stylized, painted, or otherwise non-photorealistic outputs, provided that the underlying facial geometry is preserved. This behaviour is illustrated in Figure 5, where stylistically diverse images still yield high FS values.

Visual Realism We measure realism using the Fréchet Inception Distance Heusel et al. (2017), which quantifies the statistical divergence between the distributions of generated and real images in the Inception v3 feature space. While the Fréchet Inception Distance captures large scale discrepancies, it can sometimes under represent small, but perceptually significant, facial artifacts.

4.1.3 Face Image Quality Metric

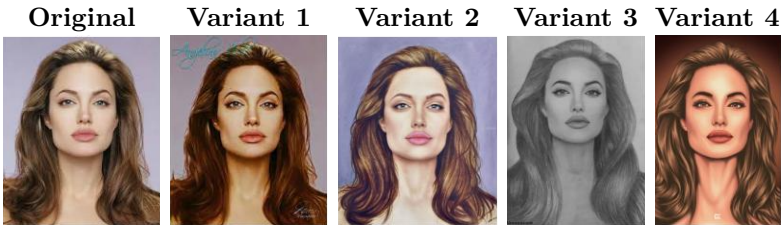


Figure 5: Face Similarity (FS) for one subject under four stylistic variations. FS remains high even for non-photorealistic or stylized outputs, provided facial structure and geometry are preserved.

Identity preservation in generative face models is commonly quantified using Face Similarity (FS), typically computed using embedding-based cosine similarity from recognition models such as ArcFace. FS is expressed on a $[0, 100]$ scale and measures how well the generated face maintains the target identity. However, FS evaluates identity independently of realism or style, meaning that it may assign high similarity scores to images that are heavily stylized, artistic, or non-photorealistic, as long as they preserve the underlying geometric features of the face.

Figure 5 demonstrates this behaviour. Despite significant stylistic variation across the four generated images, FS remains consistently high because the essential identity-defining structure is retained. This highlights an important limitation: while FS is effective for identity comparison, it does not penalize unrealistic textures, artefacts, or deviations from natural image statistics. Consequently, relying solely on FS can be misleading when evaluating the quality of generative models, as it cannot distinguish between realistic and unrealistic renditions of a face.

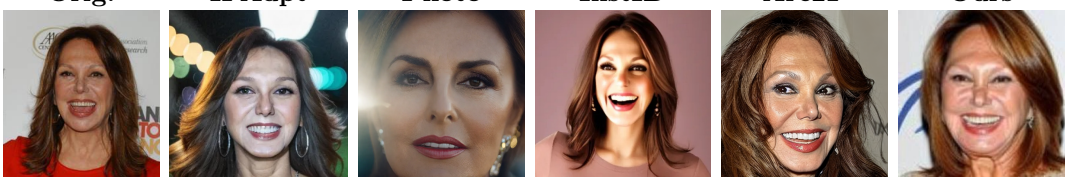
To address this, realism is typically assessed using the Fréchet Inception Distance (FID), which measures distributional similarity between generated images and real face distributions. While FID captures realism effectively, it ignores identity preservation. Thus, FS and FID each measure complementary but independent aspects of generative performance.

To jointly assess both identity fidelity and perceptual realism, we introduce the **Face Image Quality (FIQ)** metric.

To unify identity preservation and realism into a single interpretable measure, we define the FIQ score as the normalized ratio of FS to FID:

$$\mathrm{FIQ} = 100 \cdot \frac{\mathrm{FS}}{\mathrm{FID}}. \tag{23}$$

This formulation enforces a coupled dependency between identity and realism. Higher FS increases FIQ, reflecting better identity alignment, while higher FID (indicating poorer realism) decreases FIQ, penalizing degraded or unrealistic outputs. Although the equation includes a factor of 100, FIQ is an unbounded ratio and should not be interpreted as a percentage.



FS ↑	47.92	0.09	42.82	63.15	81.42
FID ↓	171.7	127.5	129.9	136.3	101.3
FIQ ↑	27.91	0.07	32.96	46.33	80.38

Figure 6: Identity-specific comparison for one subject (ID1) across different models. Metrics (FS, FID, and FIQ) are shown below each result.

By integrating both identity and realism into a single coherent metric, FIQ enables more balanced, comprehensive, and interpretable evaluations of generative face models, capturing the essential interplay between identity fidelity and perceptual quality that underpins Diff-ID.

4.2 Evaluation Settings

For quantitative evaluation, we use two image sets of size 5,000 each. The first is a validation split of 5,000 images sampled from our curated CelebA-HQ, FFHQ, and LAION-Face pool, each paired with a descriptive caption produced by the BLIP based captioning system described in Section 3. The second is an unseen split of 5,000 images drawn from the LFW dataset, again captioned using the same BLIP model to obtain identity aware prompts. All methods, including Diff-ID and the baselines in Table 1, are conditioned on these captions when generating outputs, which ensures a controlled comparison across identity fidelity and realism on both seen and unseen identities.

4.3 Results

4.3.1 Identity Preservation

	Validation Set			Unseen Data		
Model	FS ↑	FID ↓	FIQ ↑	FS ↑	FID ↓	FIQ ↑
IP Adapter	40.53	171.71	23.60	35.90	151.56	23.69
PhotoMaker	33.87	127.47	26.57	29.56	113.50	26.04
InstantID	**75.13**	129.98	57.80	**74.12**	119.39	62.08
Arc2Face	73.71	136.25	54.10	72.51	110.69	65.51
Diff-ID	72.68	**101.31**	**71.74**	71.53	**103.19**	**69.32**

Table 1: Evaluation: Face Similarity (FS), Fréchet Inception Distance (FID), and Face Image Quality (FIQ) on validation and unseen sets. Arrows indicate whether higher (↑) or lower (↓) values are better.

Our comparative image grid in Figure 6 and the extended results in Figure 7 show that Diff-ID preserves critical identity specific details, such as facial bone structure, eye spacing, nose shape, and lip contour, more faithfully than competing methods. IP Adapter and PhotoMaker exhibit the weakest preservation, with noticeable drift in geometric structure and texture. Arc2Face maintains stronger identity cues but suffers from reduced sharpness and consistency across samples.

InstantID achieves the highest raw Face Similarity on both validation (75.13) and unseen data (74.12), as summarized in Table 1, but its outputs often lack finer detail and can overemphasize stylized similarities. Diff-ID ranks closely behind in Face Similarity (72.68 and 71.53) while maintaining much stronger visual fidelity. This trade off becomes clearer when we consider the Face Image Quality score, which jointly accounts for identity similarity and realism via Equation equation 23. Diff-ID attains the best FIQ across both splits

(71.74 on validation and 69.32 on unseen data), which indicates that it preserves identity in a way that is both consistent and photorealistic.

4.3.2 Perceptual Realism

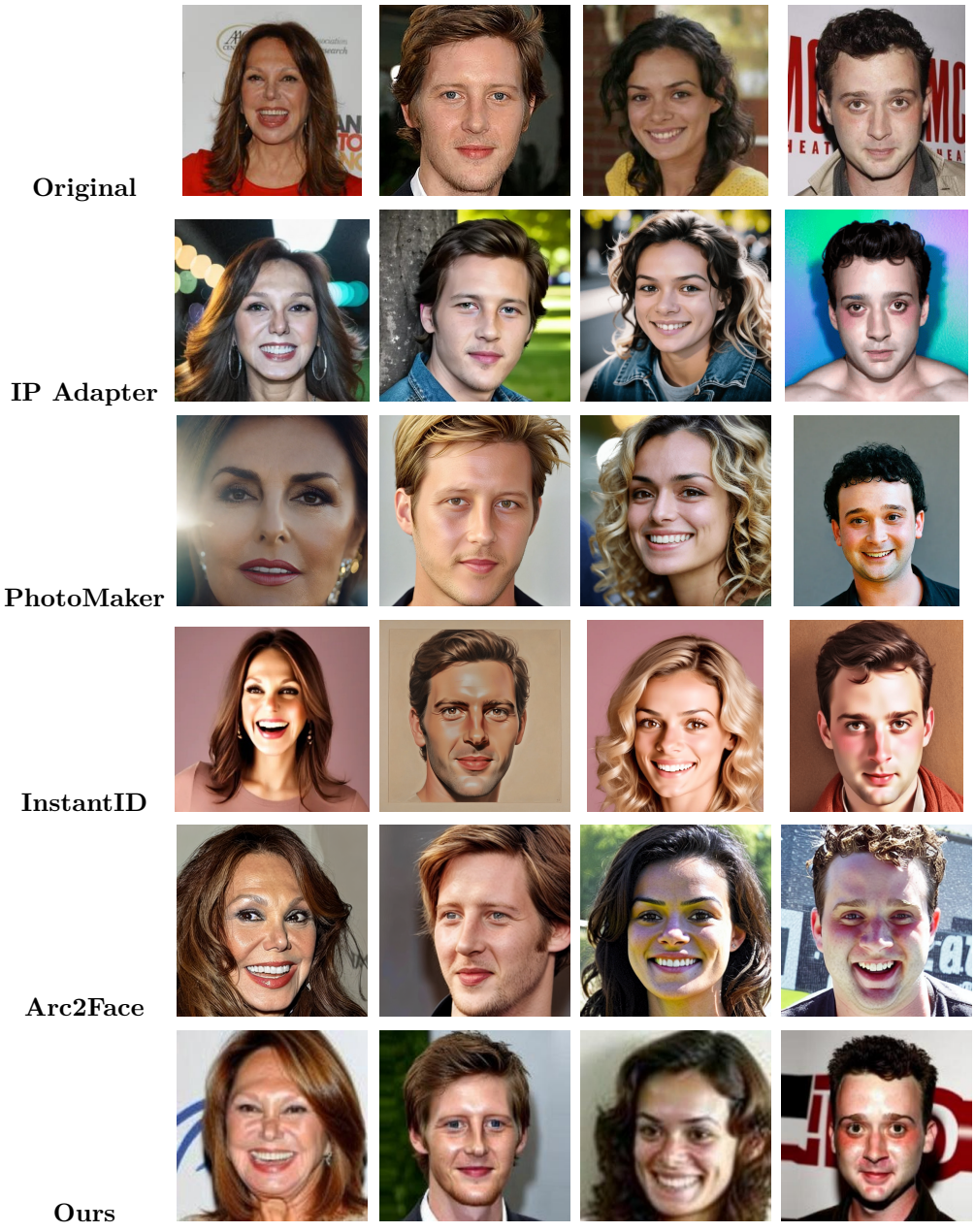


Figure 7: Identity preservation: comparison across IP Adapter, PhotoMaker, InstantID, Arc2Face, and Diff-ID. All methods are driven by the same BLIP generated descriptive prompts for each subject.

Perceptual realism is measured by the Fréchet Inception Distance and incorporated into the FIQ score. Table 1 shows that Diff-ID achieves the lowest Fréchet Inception Distance on both the validation set (101.31) and unseen data (103.19), outperforming IP Adapter, PhotoMaker, InstantID, and Arc2Face. This translates directly into the highest overall FIQ scores, despite a modest drop in Face Similarity compared to InstantID.

In contrast, PhotoMaker attains relatively low Fréchet Inception Distance (127.47 and 113.50) but fails to preserve identity well, which results in weaker FIQ. Arc2Face strikes a more balanced mid range performance, with solid Face Similarity but higher Fréchet Inception Distance than Diff-ID. These trends highlight that Diff-ID not only anchors identity more reliably than most baselines but also produces coherent skin textures, hair details, and backgrounds, avoiding the cartoon like or over smoothed artifacts present in some methods. We note that the Fréchet Inception Distance may overlook small yet perceptually salient flaws, while Face Similarity can over reward stylized similarity. Nonetheless, the superior FIQ of Diff-ID demonstrates that it provides the most reliable trade off between photorealism and identity retention across our benchmarks.

5 Ablation Study

5.1 Morphing

Figure 8 illustrates three morphing strategies: (i) embedding only interpolation of fused identity vectors, (ii) DiffID Morph via linear interpolation, and (iii) DiffID Morph via spherical interpolation.

Embedding only interpolation (top row) produces overly smooth, low detail transitions that often misalign key facial features and lack high frequency texture. By contrast, DiffID linear interpolation (middle row) injects identity and semantic cues at each denoising step, yielding sharp, coherent blends, though some intermediate frames still exhibit mild ghosting of features. DiffID spherical interpolation (bottom row) further regularizes interpolation on the hypersphere, which improves mid point consistency in bone structure, skin texture, and overall likeness.

Although we do not report quantitative FIQ scores for each frame here, the qualitative visual fidelity of DiffID with spherical interpolation aligns with its superior aggregate FIQ in our main evaluation. It strikes the best balance between identity retention and perceptual realism. We acknowledge that pure Face Similarity on individual frames can be inflated by stylized or avatar like artifacts; instead, these visual results confirm that our diffusion based joint interpolation approach produces smooth, photorealistic face morphs without the need for per identity fine tuning or multiple checkpoints.

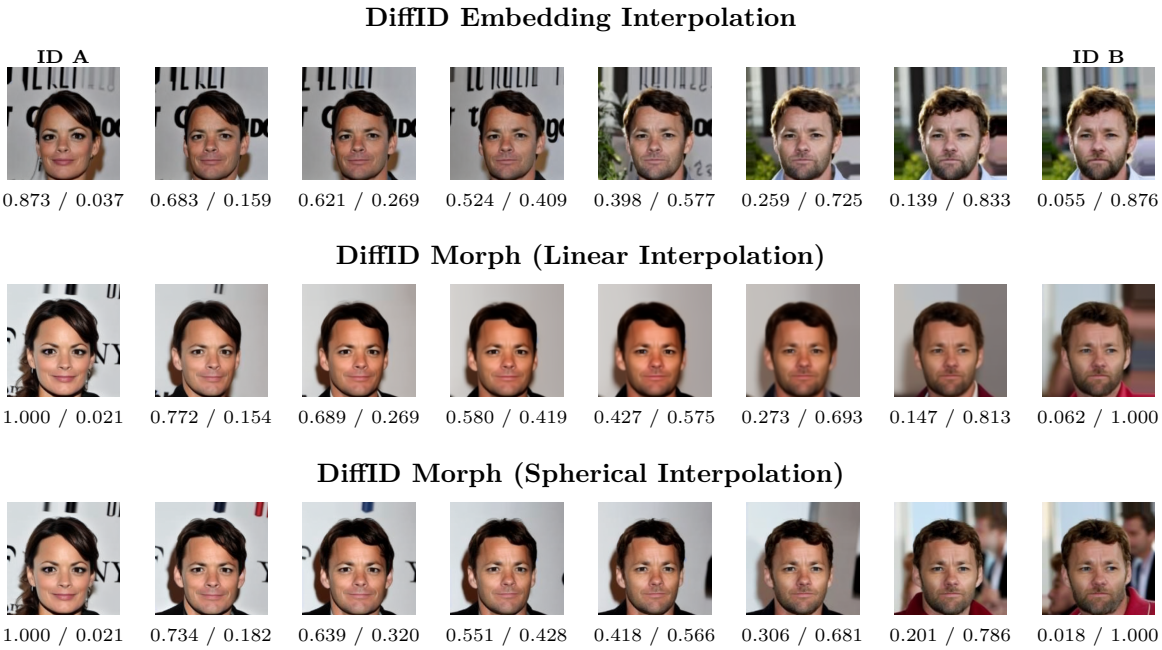


Figure 8: Comparison of morphing methods. Top: embedding space interpolation. Middle: linear interpolation. Bottom: spherical interpolation. Values under each image are shown as (s_A/s_B), where s_A is ArcFace cosine similarity to identity A (left endpoint) and s_B is similarity to identity B (right endpoint).

5.2 Impact of Different Adapter Strategies

Figure 9 visualizes outputs from four adapter variants: DiffID ArcFace, DiffID CLIP, DiffID Joint, and our full dual cross attention plus Fusion multilayer perceptron (DiffID Final). Each row shows two representative identities.

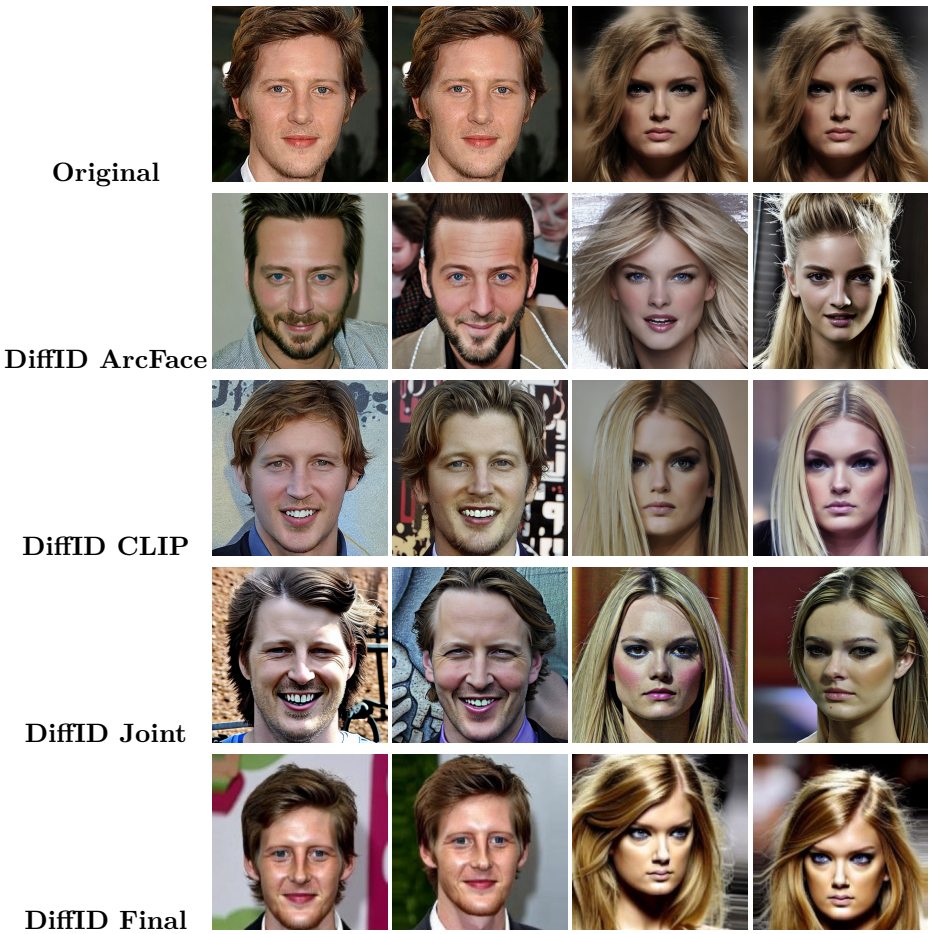


Figure 9: Adapter ablation: identity preservation across variants. DiffID Joint uses simple concatenation of ArcFace and CLIP image embeddings without the reciprocal dual cross attention or Fusion multilayer perceptron refinement, which isolates the contribution of the adapter architecture itself.

Qualitative Observations The ArcFace only adapter (DiffID ArcFace) preserves broad geometry (jawline, eye spacing) but lacks fine texture detail. The CLIP only adapter (DiffID CLIP) recovers richer semantic context and sharper textures but introduces identity drift, for example nose shape and lip contour deviate from the source. The concatenated Joint adapter (DiffID Joint) balances geometry and texture but still shows slight mid face averaging. In this variant, projected ArcFace and CLIP image embeddings are simply concatenated and injected without the reciprocal dual cross attention or Fusion multilayer perceptron refinement. Our final DiffID Final adapter robustly reproduces high frequency details (for example pores and subtle wrinkles) and faithfully maintains bone structure, lip shape, and eye geometry across examples.

Quantitative Insights As shown in Table 2, the ArcFace only adapter scores 48.91, which reflects limited textural fidelity. The CLIP only variant achieves 53.41 by leveraging semantic detail but misaligns fine identity cues. The combined Joint adapter jumps to 65.83, which highlights the benefit of multi modal fusion. Finally, our DiffID Final configuration reaches 72.68, which validates how dual cross attention and the Fusion multilayer perceptron synergistically maximize identity coherence without sacrificing realism.

Model Variant	Face Similarity
DiffID ArcFace	48.91
DiffID CLIP	53.41
DiffID Joint	65.83
DiffID Final	**72.68**

Table 2: Adapter ablation results: Face Similarity comparison across model variants. The final configuration combining ArcFace and CLIP with dual cross attention and a Fusion multilayer perceptron achieves the highest performance.

These ablations confirm that both ArcFace and CLIP embeddings are necessary: ArcFace provides precise identity discrimination, while CLIP enriches semantic and textural context. Their joint integration in our adapter is critical for achieving state of the art identity retention in diffusion based facial synthesis.

5.3 Impact of Identity Loss

Convergence Behavior

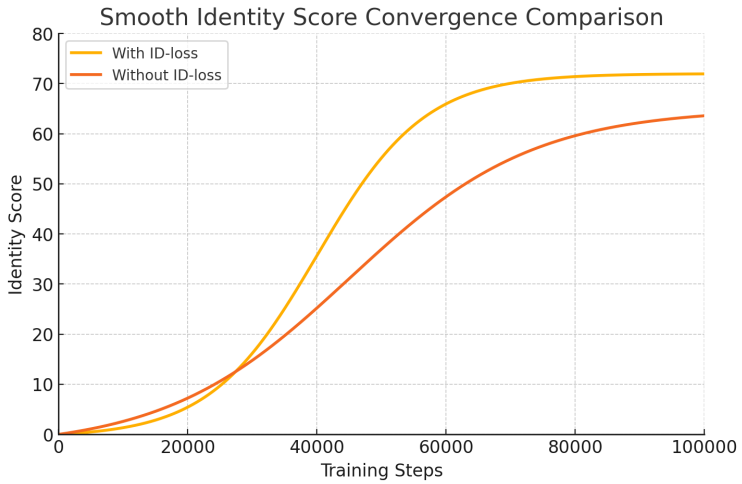


Figure 10: Training dynamics with and without the identity loss term. Exponential timestep weighting concentrates identity supervision on later denoising stages, which stabilizes convergence and improves identity fidelity.

Figure 10 compares training curves for the denoising loss when using our pseudo discriminator identity loss versus a baseline without it. Incorporating the identity term yields faster early convergence: the model quickly learns to preserve coarse identity features and minimizes the combined loss, whereas the baseline drifts more gradually as it struggles to infer identity from reconstruction alone. By emphasizing identity coherence at low noise levels, our weighted identity loss guides the network toward stable identity retention from the outset.

Effect of Identity Loss Weighting

Figure 11 illustrates the impact of varying the identity loss weight λ. At low values ($\lambda = 0.10$), the identity term gently steers the diffusion process, which yields high fidelity identity retention without disturbing the denoising dynamics. As λ increases to values between 0.30 and 0.50, the model over prioritizes identity, which causes artifacts and degraded denoising quality. Features become unnaturally sharp or appear stuck, and background details collapse. These results confirm that a moderate weighting (around 0.10 to 0.20) achieves the best balance between identity fidelity and photorealism, and they validate our choice of $\lambda = 0.10$ in Equation equation 18.

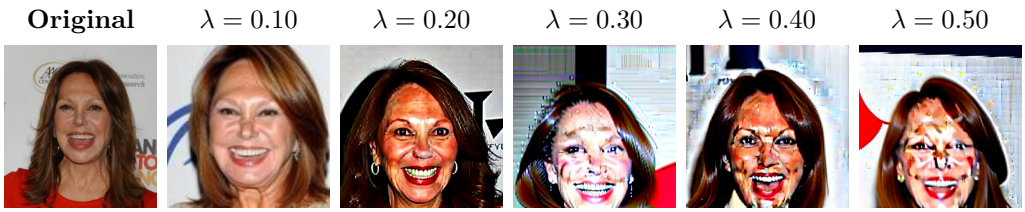


Figure 11: Effect of identity regularization strength (λ) on generated outputs. Moderate values achieve a good balance between identity fidelity and perceptual quality.

6 Limitations and Future Work

While Diff-ID sets a new standard for identity consistent face synthesis and morphing, several limitations remain. First, by relying on frozen CLIP image and text encoders for semantic context, our framework lacks explicit control over non facial elements, such as backgrounds, clothing, and accessories, which remain governed by coarse CLIP embeddings rather than dedicated spatial or attribute specific modules. Second, Diff-ID is optimized for single, frontal face crops; extreme head poses, severe occlusions, or full body scenes can degrade identity fidelity, since the model has not been trained to disentangle complex scene elements or non facial regions. Third, our morphing pipeline employs simple linear and spherical interpolation of latents and identity embeddings, which, while effective for smooth transitions, does not support attribute conditioned or region specific morph trajectories (for example selectively blending expressions or hairstyles).

Although we instantiate Diff-ID on a UNet backbone for comparability with existing identity guided diffusion models, the dual cross attention adapter operates purely on projected token sequences and cross attention blocks. As such, it is architecturally compatible with transformer based diffusion backbones such as DiT. Adapting the method would primarily involve choosing appropriate injection points within DiT attention layers rather than redesigning the adapter itself. Exploring Diff-ID style identity conditioning in DiT architectures is therefore an interesting direction for future work.

Finally, our morphing experiments are designed to probe the identity stability of the generator rather than to evaluate biometric vulnerability. We therefore do not report attack specific metrics such as false match rates, false non match rates, or morph attack detection scores, which depend on external verification systems and protocols. Studying Diff-ID morphs under standardized biometric security and morph attack detection benchmarks is orthogonal to our current focus and remains an important avenue for future research.

Future research can also address the limitations above by integrating ControlNet style spatial conditioning or learned attention masks to separately modulate background and clothing attributes, thereby enriching non facial control. Enhancing the dataset with multi view and occluded face samples, or incorporating three dimensional aware diffusion priors, could improve robustness to pose and occlusion. On the morphing side, developing learned interpolation networks or diffusion paths guided by semantic anchors would enable finer grained, attribute aware transitions, for example expression morphing without altering identity geometry. By extending Diff-ID in these directions, one can build a more generalizable and controllable platform for secure, identity anchored image synthesis and morphing.

7 Conclusion

In this work, we introduced **Diff-ID**, a unified diffusion framework for high resolution facial image generation and morphing that explicitly enforces identity consistency. By fusing ArcFace and CLIP embeddings through a lightweight dual cross attention adapter within a fine tuned Stable Diffusion UNet, and by incorporating a pseudo discriminator identity loss with exponential timestep weighting, Diff-ID achieves state of the art performance in both identity retention and visual realism. Our extensive evaluations on held out and unseen face sets demonstrate superior ArcFace similarity scores and Face Image Quality metrics compared to leading baselines. Furthermore, our DDIM based morphing pipeline delivers smooth, photorealistic face interpolations without per subject fine tuning or multiple checkpoints.

Diff-ID lays the groundwork for practical applications in biometric security, privacy preserving data augmentation, and photorealistic avatar creation. By addressing the identified limitations, such as non facial attribute control, transformer based diffusion backbones, and more sophisticated morph trajectories, future work can further elevate the versatility and robustness of identity anchored diffusion models.

References

Rinon Abdal, Yipeng Qin, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13153–13162, 2021.

Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, and Ying Fu. Partial fc: Training 10 million identities on a single machine. In *ICCVW*, 2021.

Xiang An, Jiangkang Deng, Jia Guo, Ziyong Feng, Xuhan Zhu, Yang Jing, and Liu Tongliang. Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc. In *CVPR*, 2022.

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *ACM SIGGRAPH*, pp. 187–194. ACM, 1999.

Wei Chen et al. Instantid: Fast and accurate identity preservation in text-to-image models. In *arXiv preprint arXiv:2304.XXXX*, 2023.

Jiankang Deng, Anastasios Roussos, Grigorios Chrysos, Evangelos Ververas, Irene Kotsia, Jie Shen, and Stefanos Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *IJCV*, 2018.

Jiankang Deng, Grigorios Chrysos, Evangelos Ververas, Jie Shen, Irene Kotsia, and Stefanos Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. *arXiv preprint arXiv:1905.04150*, 2019a. Available at `https://arxiv.org/abs/1905.04150`.

Jiankang Deng, Jia Guo, Yuxiao Ni, Stefanos Zafeiriou, and Sameer Kumar. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019b.

Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019c.

Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *Proceedings of the IEEE Conference on European Conference on Computer Vision*, 2020a.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020b.

Zizhao Ding, Yunzhi Wang, Yuxiao Dong, Xiaodong Dong, Shengjie Tang, Feng Wu, Xunyu Lu, Gang Liu, Xiong Zhu, and Jian Gao. Cogview: Mastering text-to-image generation via transformers. In *arXiv preprint arXiv:2105.13290*, 2021.

Matteo Ferrara, Giovanni L. Franco, and Davide Maltoni. The magic passport. In *Proc. International Workshop on Biometrics and Forensics (IWBF)*, pp. 1–6. IEEE, 2014.

Yoni Gal, Raphael Kluger, Yair Rubinstein, Panayotis Angelov, and Aravind Narayanan. Make-a-scene: Text-to-image generation with visual scene control. In *arXiv preprint arXiv:2204.06125*, 2022.

Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. Ostec: One-shot texture completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Jia Guo, Jiankang Deng, Niannan Xue, and Stefanos Zafeiriou. Stacked dense u-nets with dual transformers for robust face alignment. In *BMVC*, 2018.

Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714*, 2021.

Yandong Guo, Lei Zhang, Yuxiao Yuan, Jun Liu, Tao Wang, Yu Cai, Jessey Chen, Zhangyang Wang, Zhen Zhang, and Huabin Su. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision (ECCV)*, pp. 87–102, 2016.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 5966–5978, 2021.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9841–9850, 2020.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Jinwoo Lee, Hyun Kim, and Junseok Choi. Diffmorpher: Identity morphing via diffusion models. In *NeurIPS Workshop on Generative Models*, 2024.

Jian Liu et al. Ip-adapter: Learning to adapt pretrained text-to-image models for personalized image editing. In *arXiv preprint arXiv:2303.XXXX*, 2023.

Alexander Quinn Nichol and Prafulla Dhariwal. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *arXiv preprint arXiv:2112.10752*, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, and Mark Chen. Zero-shot text-to-image generation. In *arXiv preprint arXiv:2102.12092*, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Celine Rathgeb and Christoph Busch. A survey on face morphing attacks: Vulnerability exposure in biometric systems. *IEEE Transactions on Information Forensics and Security*, 14(8):2156–2179, 2019.

Xingyu Ren, Alexandros Lattas, Baris Gecer, Jiankang Deng, Chao Ma, and Xiaokang Yang. Facial geometric detail recovery via implicit representation. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *arXiv preprint arXiv:2112.10752*, 2022.

Yatao Ru, Samuel Lombardi, Brendan McMahan, Alice Shih, Xiang Gao, Antoine Miech, and Christoph Feichtenhofer. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Ghasemipour, Aditya Ramesh, Ilya Sutskever, Ruslan Salakhutdinov, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

Yang Song and Stefano Ermon. Denoising diffusion implicit models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Yang Song and Stefano Ermon. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations (ICLR)*, 2021.

Xi Zhang et al. Controlnet: Adding conditional control to diffusion models. In *arXiv preprint arXiv:2301.11368*, 2023a.

Xi Zhang et al. Photomaker: Stacked-id embeddings for enhanced identity preservation in diffusion models. In *arXiv preprint arXiv:2303.XXXX*, 2023b.

Ze Zhang, Ding Zhu, Jorge Martinez, Jian Yang, Yixin Wang, Yi Zhang, and Li Fei-Fei Sun. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023c.