Harder Interpolation for Faces: DDPM-Based Zero-Shot

Andrea MIELE
EPFL
andrea.miele@epfl.ch

Abstract

Denoising diffusion probabilistic models (DDPMs) have shown strong performance in image generation and editing. We hypothesize that multi-guidance editing in pretrained diffusion models succeeds only for local changes (e.g., adding a smile or slight aging) and fails for full 3D rotations without intermediate examples. We evaluate zero-shot face editing under this hypothesis and find that it handles small edits reliably but cannot produce coherent rotations. Attempts to improve pose control,using TransFusion with LLaMA 3.2, spectral normalization, and classifier-free guidance, yield only modest gains unless near-frontal faces are included. Overall, zero-shot editing handles local changes well but still needs extra supervision for large pose shifts.

1 Introduction

Denoising diffusion probabilistic models (DDPMs) have emerged as a powerful approach for high-quality image generation and editing. Ho et al. (2020) introduced DDPMs as a way to learn complex data distributions by gradually adding and then removing noise during a Markovian forward–reverse process (Ho et al., 2020; Song et al., 2020b). Nichol & Dhariwal (2021) later showed that simple modifications to the DDPM training objective can further improve sample quality and sampling speed. More recent work has demonstrated that diffusion models can perform zero-shot editing by guiding the reverse diffusion process using semantic conditions, without any fine-tuning or paired labels (Deschenaux et al., 2024; Meng et al., 2021). In particular, Deschenaux et al. (2024) showed that a pretrained DDPM can interpolate between two disjoint attribute manifolds, such as "neutral face" and "smiling face", by applying multi-guidance at inference time. Classifier-free guidance, which mixes conditional and unconditional score estimates, has also become a standard technique to balance sample fidelity and diversity when editing with diffusion models (Ho & Salimans, 2022).

Despite these advances, most zero-shot diffusion editing methods succeed only on tasks that involve small, local changes,such as adding a smile or indicating a moderate age shift,where pixel-level adjustments are limited to localized regions (Meng et al., 2021). It remains unclear whether the same zero-shot interpolation procedure can handle harder, global transformations that require consistent 3D understanding, such as rotating a face from full left to full right. Prior work on face aging via diffusion has shown that age progression typically requires fine-tuning or explicit conditioning, since large texture and shape changes (e.g., wrinkles, hair color) exceed what simple guidance can achieve (Chen & Lathuilière, 2023). Similarly, recent zero-shot portrait view synthesis methods exploit specialized controllers or masked attention to manipulate pose without retraining (Gu et al., 2024), but these approaches rely on additional modules beyond pure DDPM inference. Concurrently, hybrid architectures like TransFusion combine autoregressive and diffusion losses in a single transformer backbone to support multimodal generation, but adapting them for pose control has not been explored under strict compute limits (Zhou et al., 2024).

Preprint. Under review.

In this work, we hypothesize that multi-guidance editing in pretrained diffusion models works reliably only when the desired change involves mostly local pixel shifts (e.g., neutral $\rightarrow$ smile, young $\rightarrow$ old). We further posit that editing fails when a transformation demands a consistent, global 3D rotation (e.g., full left $\rightarrow$ full right pose) unless the model has seen intermediate, near-frontal examples to bridge the gap. In other words, classifier-free guidance can mix internal pose representations only if the endpoint conditions are already close in latent pose space; without any true "middle" samples, smooth global rotations cannot emerge from zero-shot inference alone.

Our contributions are the following:

- **Characterize Zero-Shot Editing Limits.** We show that multi-guidance editing succeeds for local semantic shifts (e.g., expressions, mild aging) but fails for global 3D rotations without intermediate examples (Ho et al., 2020; Deschenaux et al., 2024; Meng et al., 2021; Song et al., 2020b; Nichol & Dhariwal, 2021).
- **Evaluate Pure Zero-Shot on Age and Pose.** We demonstrate that applying pure zero-shot guidance,using a age classifier for "young $\rightarrow$ old" and a classifier for "left $\rightarrow$ right", produces realistic minor age changes but yields artifacts for large age gaps and incoherent rotations, confirming that direct gradient-based guidance is insufficient for these tasks (Chen & Lathuilière, 2023; Ho & Salimans, 2022; Meng et al., 2021; Gu et al., 2024).
- **Introduce Auxiliary Techniques for Pose Control.** To bridge the gap for full 3D rotations, we try to TransFusion with pretrained LLaMA 3.2 (Grattafiori et al., 2024) weights (which does not converge under our compute constraints) and revisit multi-guidance by (1) including unlabeled, near-frontal images in the training splits and trying on a smaller gap (front and extreme left images), (2) applying spectral normalization during inference, and (3) using classifier-free guidance to interpolate between frontal and rotated embeddings. These tweaks, especially classifier-free guidance on the small reduced gap, produce some plausible rotations even without near-frontal examples, though performance still improves when such examples are available, indicating that while CFG can bridge certain pose gaps, fully zero-shot global rotations remain challenging (Deschenaux et al., 2024; Ho & Salimans, 2022; Zhou et al., 2024).

2 Background

2.1 Denoising Diffusion Probabilistic Models

Denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020; Dhariwal & Nichol, 2021) define a forward noising process that gradually adds Gaussian noise to a clean image x_0 over T timesteps. At each timestep t, the forward distribution is

$$q(x_t \mid x_{t-1}) = \mathcal{N}\big(x_t; \sqrt{1-\beta_t}\, x_{t-1},\, \beta_t \mathbf{I}\big),$$

where $\{\beta_t\}$ is a variance schedule. During training, one samples $t \sim \{1, \ldots, T\}$ uniformly and optimizes a noise prediction network ϵ_θ to match the true noise ϵ:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{x_0,\epsilon,t}\Big[\big\|\epsilon - \epsilon_\theta(x_t, t)\big\|_2^2\Big] \quad \text{where} \quad x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad \bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s).$$

At inference time, one starts from $x_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively denoises using ϵ_θ to recover a clean sample x_0.

2.2 Conditional Generation via Classifier Guidance and Classifier-free guidance

A simple way to steer a pretrained DDPM toward a specific attribute c is *classifier guidance* (Dhariwal & Nichol, 2021). Given a pretrained classifier $p_\phi(c \mid x_t)$, one can approximate the gradient of the log-posterior $\nabla_{x_t} \log p_\theta(x_t \mid c)$ by

$$\nabla_{x_t} \log p_\phi(c \mid x_t) \;\approx\; \nabla_{x_t} \log p_\theta(x_t \mid c) \;-\; \nabla_{x_t} \log p_\theta(x_t),$$

so that each reverse step is modified as

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\Big(x_t - \tfrac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\Big) \;+\; \gamma\, \nabla_{x_t} \log p_\phi(c \mid x_t),$$

where $\gamma > 0$ is the *guidance weight*. By increasing γ, the model's output is pushed more strongly toward samples the classifier assigns high probability for class c.

Alternatively, *classifier-free guidance* (CFG) (Ho & Salimans, 2022) trains the diffusion model to predict both conditional and unconditional noise, $\epsilon_\theta(x_t, t \mid c)$ and $\epsilon_\theta(x_t, t)$. At inference, one interpolates between these two predictions:

$$\hat{\epsilon}_\theta(x_t, t \mid c) = (1 + w)\, \epsilon_\theta(x_t, t \mid c) - w\, \epsilon_\theta(x_t, t),$$

where w is a guidance scale. CFG has the advantage of not requiring a separate classifier and often yields more stable attribute control (Nichol et al., 2021).

2.3 Zero-Shot Interpolation by Gradient Blending

Deschenaux et al. (2024) introduced a *zero-shot interpolation* strategy that blends classifier gradients from two terminal attributes c_A and c_B without retraining. Denote by $\nabla_x \log p_\phi(c \mid x_t)$ the classifier gradient for class c. At each diffusion step t, one computes a blended gradient

$$g_t = (1 - \lambda_t)\, \nabla_{x_t} \log p_\phi(c_A \mid x_t) \;+\; \lambda_t\, \nabla_{x_t} \log p_\phi(c_B \mid x_t),$$

where $\lambda_t \in [0, 1]$ is a schedule (e.g., linearly increasing from 0 to 1). The sample x_t is then updated by adding a scaled version of g_t, effectively guiding the diffusion trajectory from c_A toward c_B. This approach requires no additional training once ϵ_θ and p_ϕ are available, making it *zero-shot*.

In our work, we apply this method first to *age progression* (with $c_A =$ "young" and $c_B =$ "old") and then to *face rotation* (with $c_A =$ "left profile" and $c_B =$ "right profile"). We use a fixed guidance weight (30) and a linear λ_t schedule over 4000 steps, following Deschenaux et al. (2024).

2.4 Spectral Normalization for Stable Guidance

Spectral normalization (SN) (Miyato et al., 2018) constrains the Lipschitz constant of convolutional layers by dividing each weight matrix W by its largest singular value $\sigma_{\max}(W)$. When applied to the *guidance classifier* p_ϕ, SN helps keep classifier gradients bounded, which can reduce artifacts in regions outside the training distribution. Although Deschenaux et al. (2024) reported smoother interpolations with SN applied to p_ϕ, we find in Section 7 that SN on the U-Net backbone yields limited visual improvement for large geometry changes.

2.5 TransFusion Architecture

TransFusion (Zhou et al., 2024) augments a pretrained DDPM's U-Net with a transformer encoder to enable fine-grained semantic control. In brief, TransFusion:

- Embeds text or attribute tokens (e.g., "frontal face," "left profile") via a transformer encoder into a spatial feature map aligned with U-Net activations.
- Uses cross-attention to fuse transformer outputs with intermediate U-Net layers at multiple resolutions.
- Trains both diffusion denoising and cross-attention jointly, often on large text–image corpora.

Because TransFusion requires extensive joint training on image–text data, we instead experiment (Section 6) with injecting pretrained LLaMA 3.2 weights into the transformer encoder.

Together, these background componentsm DDPM formulation, classifier guidance, zero-shot gradient blending, and architectural variants, form the foundation for our experiments on age progression and face rotation.

3 Related Work

3.1 Diffusion Models for Image Synthesis and Editing

Denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020) and related score-based approaches (Song & Ermon, 2019; Song et al., 2020b) have become competitive with GANs for

high-quality image generation. Improved sampling methods (e.g., DDIM (Song et al., 2020a)) and latent diffusion frameworks (Rombach et al., 2022) reduce compute cost while maintaining fidelity. Beyond unconditional synthesis, diffusion models support a variety of editing tasks. SDEdit (Meng et al., 2021) and ILVR (Choi et al., 2021) use conditional denoising to perform inpainting and style transfer without retraining. Classifier-free guidance (CFG) (Ho & Salimans, 2022) enables smooth trade-offs between conditional and unconditional generation, and has underpinned many recent text-to-image systems (Saharia et al., 2022). Other work applies diffusion to super-resolution (Moser et al., 2024) and domain transfer (Li & Yan, 2024). We do not focus on architectural or training variants (e.g., alternative UNet designs or corruptions) since our emphasis is on inference-time guidance of pretrained models. For comprehensive overviews of diffusion architectures and training, see Ulhaq & Akhtar (2022) and Moser et al. (2024).

3.2 Zero-Shot Interpolation and Compositional Editing

Deschenaux et al. (2024) introduced zero-shot interpolation for DDPMs, showing that a model trained only on extreme attribute labels (e.g., clearly smiling vs. non-smiling) can generate intermediate samples by blending classifier gradients at inference. Variations include blending text embeddings (Hu et al., 2023) for instance. Compositional diffusion methods (Kim et al., 2023) fuse multiple classifiers to produce novel attribute combinations (e.g., "smiling + eyeglasses"), and latent inversion approaches (e.g., Imagic (Kawar et al., 2023)) refine real images via per-image optimization. Our work extends the classifier-guided interpolation framework to more complex face attributes,namely, age progression and face rotation,rather than simple expressions. Unlike latent inversion methods, we retain a pure zero-shot inference focus and do not perform any optimization per input image.

3.3 Pose-Aware Face Generation and 3D Priors

In the GAN literature, many methods target explicit 3D control for faces. F-GAN (Nowozin et al., 2016) and Exp-GAN (Lee et al., 2022) combine 3D morphable models (3DMMs) with adversarial training to frontalize or manipulate pose. MOST-GAN (Medin et al., 2022) and CGOF++ (Sun et al., 2023) enforce 3DMM constraints to disentangle shape, expression, and lighting. More recent implicit 3D methods (e.g., EG3D (Chan et al., 2022), LiftedGAN (Shi et al., 2021)) learn tri-plane or NeRF representations that support novel views and relighting. In diffusion, a few works integrate 3D priors: Zhou *et al.* (Liu et al., 2023) embed a 3DMM into the conditional diffusion process for face swapping. These approaches require specialized training pipelines or multi-view data and lie outside our zero-shot inference framework.

By contrast, we attempt zero-shot pose interpolation directly in pixel space with classifier guidance and focus on the failure modes that arise when geometry changes are large. Our results highlight the gap between simple diffusion-based edits and fully 3D-aware generation methods for pose control.

4 Methodology

4.1 Problem Statement and Formal Definitions

We denote our dataset of N face images by

$$D = \{(x_i, z_i)\}_{i=1}^{N}, \quad x_i \in \mathbb{R}^D,\ z_i \in [0,1]^L,$$

where each latent vector z_i encodes L continuous attributes (e.g., age or yaw) under a semi-order: two values $z_i^{(\ell)}, z_j^{(\ell)}$ are considered indistinguishable if $|z_i^{(\ell)} - z_j^{(\ell)}| < \epsilon$. We fix $\delta > 0$ such that the "extreme" intervals $[0, \delta]$ and $[1-\delta, 1]$ do not overlap the "mild" band $[0.5 - \delta,\ 0.5 + \delta]$. Our goal is zero-shot interpolation: training a DDPM only on extreme samples (i.e., indices with $z^{(\ell)} \in [0, \delta] \cup [1-\delta, 1]$ for each attribute) and generating samples whose inferred z-values lie in the mild interval.

4.2 Multi-Guidance Sampling

We employ a product-of-experts sampling density:

$$p_{\Pi}(x) \propto p(x) \prod_{m=1}^{M} p_{\phi}(y_m \mid x)^{\lambda_m},$$

where $p(x)$ is the unconditional DDPM prior and each $p_{\phi}(y_m \mid x)$ is a guidance classifier (FaRL for age; EfficientNet for other attributes). At each reverse diffusion step $t = 1, \ldots, T$ (with $T = 4000$ and cosine noise schedule parameter $s = 0.008$), we blend gradients for two target classes $c_A \to c_B$ via

$$g_t = w\Big[(1-\lambda_t)\nabla_x \log p_{\phi}(c_A \mid x_t) + \lambda_t \nabla_x \log p_{\phi}(c_B \mid x_t)\Big],$$

where λ_t increases linearly from 0 to 1, and $w = 30$ for age (or $w = 45$ for pose). The mean update becomes

$$\tilde{\mu}_t = \mu_{\theta}(x_t, t) + \sigma_t^2 \, g_t,$$

and we sample

$$x_{t-1} \sim \mathcal{N}(\tilde{\mu}_t, \, \sigma_t^2 I).$$

4.3 Extremal Training Set Extraction

All images are downscaled to 64×64. For both CelebA-HQ (Liu et al., 2015) and LFR (Elharrouss et al., 2020), we fine-tune an EfficientNet to predict each binary attribute, retaining images with soft-label confidence above threshold τ_1 (set to achieve perfect precision on a held-out validation set). We then split this pool into five folds, train five EfficientNet models, and keep only images whose minimum ensemble confidence exceeds τ_2. Finally, we select the top-k most extreme samples in each band $[0, \delta]$ and $[1-\delta, 1]$ to form the training set D^*. For LFR yaw, we define "left" as $\geq 30^{\circ}$ and "right" as $\leq -30^{\circ}$, discarding the outer 20% slack to avoid mild poses.

4.4 Model Architecture and Training

Our diffusion model uses a U-Net backbone with three residual blocks per scale and self-attention at feature map resolutions of 32, 16, and 8 channels. We train for 150,000 steps using the Adam optimizer with learning rate 10^{-4} and exponential moving-average decay of 0.9999. A cosine noise schedule with $s = 0.008$ and $T = 4000$ timesteps is used. Sampling employs 250 reverse steps with the multi-guidance update described above.

4.5 Spectral Normalization and Classifier-Free Guidance

To stabilize guidance gradients, we apply spectral normalization to all convolutional weights, dividing each weight by its largest singular value via one power iteration. In parallel, we incorporate classifier-free guidance by randomly dropping the conditional embedding with probability $p_{\text{drop}} = 0.1$ during training. At inference, we blend the conditional and unconditional noise predictions:

$$\hat{\epsilon}_{\theta}(x_t, t \mid c) = (1+w)\,\epsilon_{\theta}(x_t, t \mid c) \; - \; w\,\epsilon_{\theta}(x_t, t),$$

with w matching the multi-guidance weight above.

5 Experiments and Results

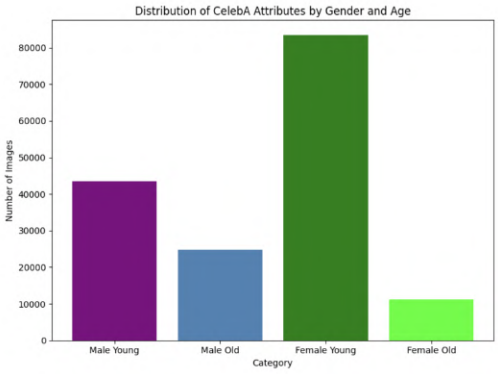
We first want to test the results on a harder task than smile, so we therefore test to interpolate zero-shot on age.

5.1 Zero-Shot Age Progression

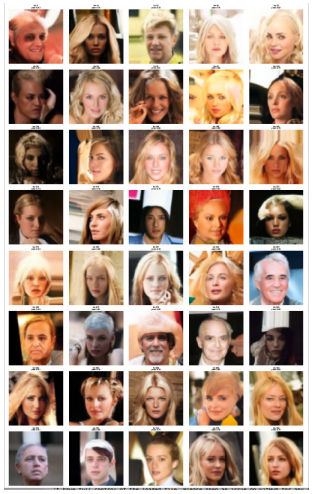
Dataset We base our experiments on the CelebA dataset (Liu et al., 2015), which contains face images each annotated with 40 binary attributes such as "Smiling," "Young," and "Blond Hair" . For inference at 64 × 64 resolution, we downscale all images accordingly. Because CelebA's labels are binary, we first train attribute-specific classifiers on the full dataset and calibrate them via temperature scaling.

Attribute Definition and Data We define two age categories: "young" (age ≤ 30) and "old" (age ≥ 60). From the CelebaHQ dataset, we train an evaluation classifier and take only extremal data points to construct an extremal dataset, as described by Deschenaux et al. (2024).

Dataset imbalance The CelebA-HQ dataset shows a clear imbalance between age and gender groups. In our subset, there are 1 267 male images labeled as "young" (age ≤ 30) and 2 034 male images labeled as "old" (age ≥ 60). In contrast, there are 12 275 female images labeled as "young" and only 128 female images labeled as "old." As a result, the dataset is heavily skewed toward young female faces. From visual inspection (see Figure 1b), men tend to appear older in the photos compared to women. To better understand any potential bias, it would be useful to check the proportion of men and women in each age group during training to see how this imbalance might affect the model. This is what we do and we find an excessive proportion of young female in the dataset, compared to young male (Fig. 1a).



(a) Imbalance of the CelebA dataset regarding age and gender.



(b) There is a bias when generating samples out of the imbalanced CelebA dataset.

Figure 1: An imbalanced dataset results in highly biased images when sampling the DDPM model.

Fixing the dataset imbalance To correct this imbalance, we use our age classifier to select more reliable samples. First, for the "old" category, we keep only those images where the classifier's probability of being old is below 0.4 and the ground-truth label is "old," yielding about 7 000 images. Second, for the "young" category, we keep only images where the classifier's probability of being young is above 0.8 and the ground-truth label is "young," giving around 18 000 images. We then train a second classifier on a different subset (inspired by Deschenaux et al. (2024)) and retain only the images on which both classifiers agree. Finally, we apply oversampling so that neither age category is overrepresented in the final training set.

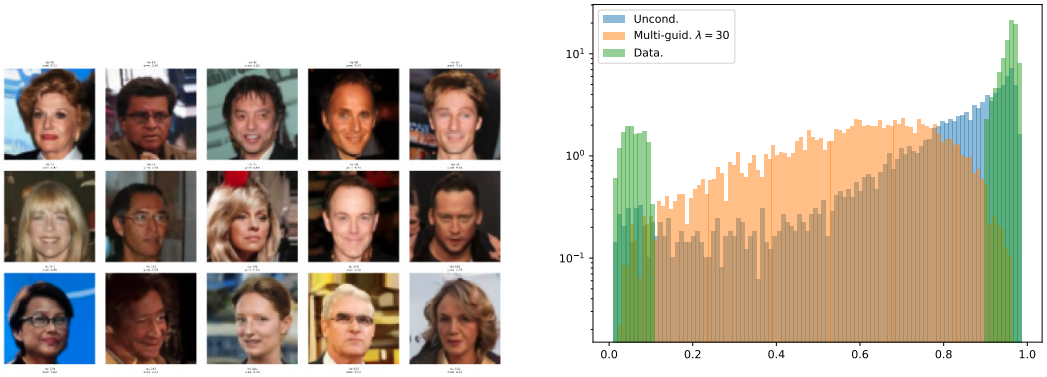
Guidance Schedule and Interpolation Let $\epsilon_\theta(x_t, t \mid c)$ denote the noise prediction network conditioned on attribute c. During inference, we perform zero-shot interpolation between two conditioning signals c_A ("young") and c_B ("old"). At each timestep t, we compute

$$\nabla x_t \;=\; 30\Big[(1-\lambda)\,\nabla_x \log p_\theta\big(x_t \mid \text{young}\big) \;+\; \lambda\,\nabla_x \log p_\theta\big(x_t \mid \text{old}\big)\Big],$$

where $\lambda \in [0, 1]$ is incremented linearly from 0 to 1 over the course of 4000 denoising steps, and the factor 30 is the fixed guidance weight applied to both classes as in Deschenaux et al. (2024). Gradients $\nabla_x \log p_\theta$ are approximated via classifier guidance using a pretrained age-attribute classifier trained on the extremal "young" vs. "old" dataset.

Results We see a good interpolation for using the multi-guidance technique, with a more uniform distribution along the classifier prediction 2b, a sign that the method is at least working correctly. The generated images 2a show a clear change in age, but the editing appears rough. In many cases, the hair and face look like they were patched together, almost as if a younger face was placed onto a silhouette. The face does look younger overall, but the blending is not seamless and resembles an

amateur photoshopping effort. Despite these visible artifacts, the main goal of making the face appear younger is achieved.



(a) Sample grid guidance.

(b) Distribution with $\lambda = 30$.

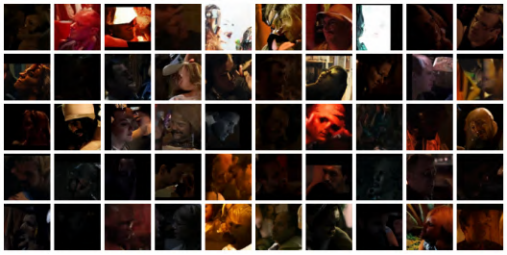
Figure 2: Guidance generation with $\lambda = 30$ with the dataset described in 5.1

5.2 Zero-Shot Face Rotation

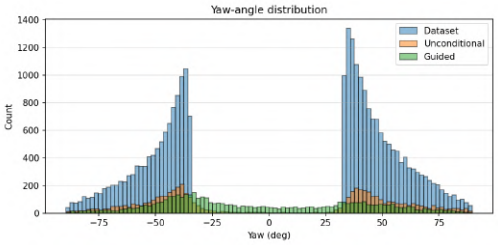
Pose Binning and Data Splits For our first pose-based split, we selected images from the LFR dataset (Elharrouss et al., 2020) according to extreme yaw angles, predicted using yakhyo (2025) (see Figure 3). Specifically, all images with yaw $\geq 30°$ were assigned to the left-facing class, and all images with yaw $\leq -30°$ were assigned to the right-facing class. We further enforced extremity by removing the bottom 20% of left-facing images (i.e., those with the smallest yaws above $25°$) and the top 20% of right-facing images (i.e., those with the largest yaws below $-25°$). After resizing all images to 64×64, this split contains approximately 22000 left-facing and 23000 right-facing samples.

Figure 3: Using yakhyo (2025) to predict the yaw on the LFR dataset, see the blue line.

Guidance Schedule and Interpolation We apply the same zero-shot blending strategy as in Section 5.1, with $c_A =$ "left profile face" and $c_B =$ "right profile face." We linearly increase λ from 0 to 1 over 4000 steps.



(a) Sample grid guidance.



(b) Distribution with $\lambda = 45$.

Figure 4: Guidance generation with $\lambda = 45$ with the dataset described in 5.2.

Results We observe three primary failure modes in 4a:

- *Geometric Inconsistency:* Facial landmarks are not preserved; eyes or nose are displaced (3rd row, 7th column).
- *Texture Artifacts:* Skin textures become smeared or patchy, particularly near occluded regions (e.g., ear edges) (1st row, last column).
- *Pose Collapse:* Instead of smoothly rotating, the model "averages" frontal and profile features, producing an unnatural intermediate face (4th column, 2nd row).
- *Dark silhouette:* We observe dark image with a light face silhouette in it but no special details (5th row, 1st column).

These failures motivate exploration of more specialized architectures (Section 6).

6 Adapting TransFusion with LLaMA 3.2 Weights

6.1 TransFusion Overview

TransFusion Zhou et al. (2024) integrates a transformer encoder into a diffusion framework to improve fine-grained semantic control. The architecture comprises:

- A pretrained U-Net backbone for diffusion denoising.
- A transformer encoder that maps text or attribute embeddings to a spatial feature map aligned with the U-Net's intermediate layers.
- Cross-attention layers that fuse transformer outputs with U-Net activations at multiple scales.

Our aim was to adapt TransFusion for face pose interpolation by conditioning on text as we thought it might give a better and more precise conditioning and information than a classifier. To avoid any issue with dataset contamination, we choose to retrain the image modality and use a pretrained LM backbone to already have a semantic understanding of language.

6.2 LLaMA 3.2 Weight Injection

Motivation We hypothesize that LLaMA 3.2's large language model (LLM) embeddings (Grattafiori et al., 2024) encode richer semantic priors, which could help the transformer encoder interpret pose prompts. Given compute limits, we replace TransFusion's transformer weights with corresponding layers from LLaMA 3.2 (1B parameters). We keep TransFusion's positional embeddings and cross-attention projections, but initialize most self-attention and feed-forward blocks from LLaMA 3.2. To avoid the cost of full fine-tuning, we insert lightweight LoRA adapters into each attention layer and train only those adapters.

Implementation Details

- **Transformer Alignment:** We use TransFusion training repository but import Llama 3.2 architecture from the official repository to ease the loading of the weights

- **Embedding Adaptation:** We project 256-dim pose attribute embeddings into LLaMA 3.2's 768-dim input space via a learned linear layer, so that pose tokens ("`<pose>left`" or "`<pose>frontal`") can enter the LLaMA-initialized encoder.
- **LoRA Layers:** Instead of updating all of LLaMA 3.2's weights, we add LoRA adapters (Hu et al., 2022) (rank 4) to each query and value projection. During training, we freeze all original LLaMA 3.2 parameters and train only the LoRA matrices and the embedding projection.
- **Training Dataset:** We use the LFR subsets described in Section 5.2, with pose labels encoded as text tokens. We train for 15 epochs with a learning rate of 1×10^{-4}, updating only LoRA and projection weights.

6.3 Results and Analysis

Despite training the LoRA adapters for 15 epochs on our ≈50 k face-patch dataset, the LLaMA 3.2–initialized TransFusion model fails to produce coherent rotations. It generates noise or generic denoised faces rather than reflecting pose changes. We identify several factors:

- **Token Count Mismatch.** The original TransFusion model sees up to 0.5–2 trillion tokens (combined text + image) to learn high-fidelity synthesis. In contrast, our training data represents only ≈10 million tokens (image patches + pose tokens), far below the scale needed to override LLaMA 3.2's language priors.
- **Strong Language Bias.** Injecting LLaMA 3 weights imparts a heavy text-centred prior to the transformer. Since we update only small LoRA matrices and the pose projection, the model retains its original LLM attention patterns, optimizing primarily for denoising instead of pose semantics.
- **Limited Adapter Capacity.** The LoRA adapters (rank 4) introduce only a small number of additional parameters. This low-rank update is insufficient to shift the encoder's behaviour from language modeling to nuanced pose manipulation given our dataset size.

Figure 15 shows an example sample after 200 000 training steps: the outputs remain noisy without clear face rotations. Consequently, we conclude that under our compute and data constraints, LoRA-based weight injection from LLaMA 3 does not improve pose control in TransFusion. We therefore return to enhancing multi-guidance diffusion (Section 7).

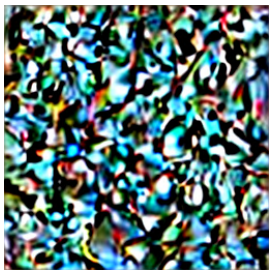


Figure 5: Samples from the TransFusion model with LLaMA 3 weight injection (LoRA adapters) after 200 000 training steps.

7 Enhanced Multi-Guidance: Better data splitting, Spectral Normalization and Classifier-Free Guidance

7.1 Motivation and Overview

Our inability to achieve zero-shot pose interpolation with vanilla multi-guidance indicates that more stable gradient conditioning and more controlled data partitions are necessary (Deschenaux et al., 2024). We therefore incorporate three core enhancements: first, *alternative dataset splits* to enforce clear separation between left-, right-, and frontal-facing images according to the LFR dataset; second,

we apply spectral normalization to the attribute-guidance classifier, rescaling each weight matrix by its largest singular value to enforce a 1-Lipschitz mapping,this smooths the classifier's gradients and yields gradual probability transitions in latent space, enabling the sampler to interpolate into unseen pose regions; and third, *classifier-free guidance* (CFG), whereby the diffusion model is trained jointly on conditional and unconditional objectives, enabling smoother interpolation trajectories between pose embeddings at inference time. Spectral normalization has been shown to stabilize GAN and diffusion gradients by normalizing the largest singular value of convolutional weight matrices at each forward or backward pass. Classifier-free guidance further improves sample fidelity by allowing interpolation between conditional and unconditional score estimates, which is particularly helpful when target attributes (e.g., extreme yaw angles) reside in sparsely covered regions of the training distribution. It is important to note that our dataset splits are still derived from the LFR (Left-Front-Right) pose-invariant face dataset (Elharrouss et al., 2020), which doesn't provide precise yaw annotations for each image, but we construct the yaw annotation using the head-pose-estimation classifier from (yakhyo, 2025), as explained in 5.2.

7.2 Dataset Splitting by Pose Clustering

Instead of random train–test splits, we partition the LFR dataset by yaw angle into three distinct splits, each containing approximately 50–60 k images. The splits are defined as follows:

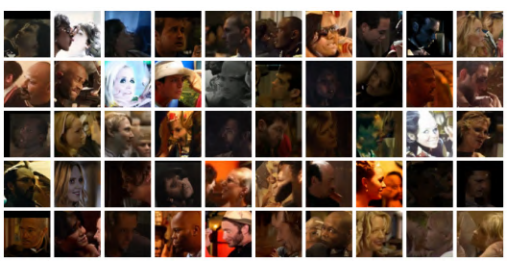
1. A is extreme left + extreme right (like in 5.2).
2. B is extreme left + extreme right + few unlabeled front.
3. C is slightly right (almost front) + extreme left, to reduce the gap between the two extremum of the dataset.

All images are resampled to 64×64 resolution to match our DDPM training configuration. During both training and inference, we use the head-pose-estimation model from yakhyo (2025) to compute predicted yaw angles for each generated sample, enabling quantitative evaluation of pose estimation.

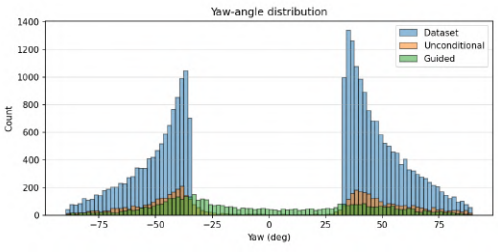
7.2.1 Results

Dataset A After training the model for 100 000 steps, we sample using classifier guidance with $\lambda = 45$. Figure 6b shows that many generated faces follow the "left" and "right" profile labels, but several images already look distorted or warped. The yaw distribution in Figure 6a confirms that guided samples cluster around the extreme poses more strongly than the unconditional samples, yet there remains a gap in the middle yaw values.

We observe three main failure modes in these outputs. First, facial landmarks often become misaligned,eyes or the nose shift unnaturally,indicating the model struggles to maintain geometric consistency when pushed strongly toward opposite profiles. Second, skin textures tend to smear or form patchy artifacts, especially around occluded regions like the ears; this suggests that the classifier gradient is too coarse and overrides the detailed texture priors learned by the DDPM. Third, instead of producing a smooth rotation, the model sometimes "averages" frontal and profile features, yielding blurry, unnatural intermediate faces. This pose collapse likely arises because the two classifier gradients point in conflicting directions, causing the diffusion process to settle on a mean-value blend rather than a clear pose.



(a) Samples generated using classifier guidance, with $\lambda = 45$.



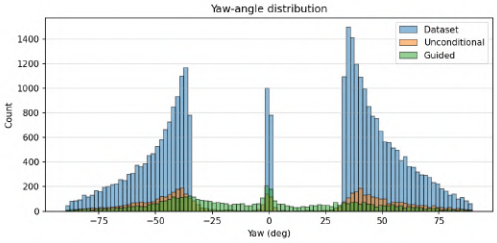
(b) Distribution with $\lambda = 45$.

Figure 6: Guidance generation with $\lambda = 45$ with the dataset A described in 5.2.

Dataset B Even though our pose classifier was never trained on frontal faces, multi-guidance generates realistic front-facing images. In Figure 7a, the distribution of predicted yaw angles has a clear peak around $0°$, showing that many generated samples are frontal. This works because the DDPM training set included a small number of (around 5%) unlabeled frontal images, which the diffusion model learned even though the classifier did not.

Figure 8a shows several high-quality frontal faces produced by multi-guidance. We also see a few slightly turned faces in Figure 8b (slightly left) and Figure 8c (slightly right). Although these intermediate poses are less common, they still look reasonable for some samples.

These results suggest that including a small number of unlabeled frontal faces in the DDPM training set, but not in the classifier, allows multi-guidance to reproduce front-facing images and even fill in some slightly turned poses. In other words, the DDPM has implicitly learned a "frontality" manifold from those few frontal examples, and blending classifier gradients for "left" and "right" is enough to traverse that manifold. The diffusion model's latent space already encodes front-on poses, so multi-guidance can guide samples into that region without ever having seen a labeled frontal class.

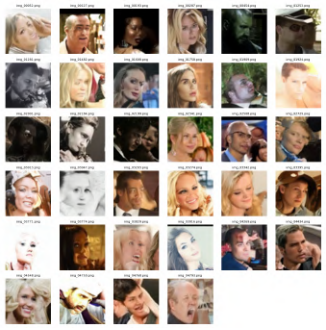


(a) Samples generated using classifier guidance, with $\lambda = 20$.

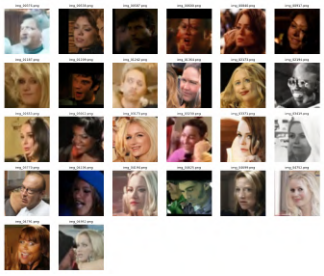
(b) Distributions with $\lambda = 20$.

Figure 7: Guidance generation with $\lambda = 20$ with the dataset described in 5.2.

(a) Quality samples from front, with $\lambda = 20$.



(b) Quality samples from slightly right, with $\lambda = 20$.



(c) Quality samples from slitghly left, with $\lambda = 20$.

Figure 8: Example of quality generation with the dataset B 5.2.

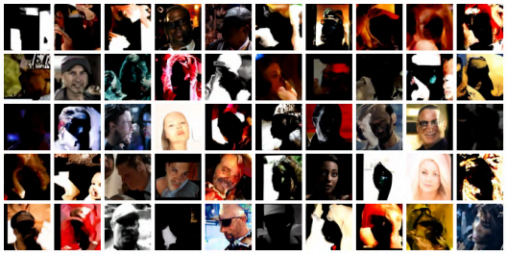
Dataset C Figure 9a shows the outputs when $\lambda = 20$. Most images look like dark, silhouette-shaped blobs, and only a few hint at a slightly turned face. The yaw distribution in Figure 9b confirms that very few samples fall in the intermediate pose.

When $\lambda = 30$ (Figure 10a), almost all images remain blob-like, with almost no facial details. The yaw histogram in Figure 10b still peaks at the extremes, showing few samples in between.

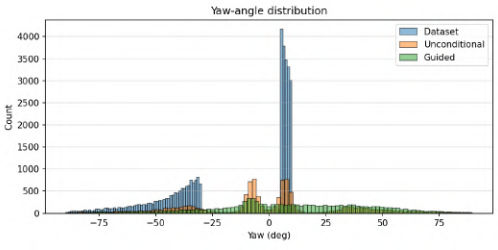
At $\lambda = 45$ (Figure 11a), sample quality drops further: nearly every image is a silhouette blob, and Figure 11b shows that almost no samples bridge the gap.

Figure 12c presents the best examples for $\lambda = 15$, $\lambda = 30$, and $\lambda = 45$. Even with $\lambda = 15$, most are still silhouette blobs, and only a couple hint at a "slightly left" pose. As λ increases in the center and right panels of Figure 12c, almost none of the images show a clear face.

In summary, for Dataset C, multi-guidance usually produces dark silhouettes, and only a handful of outputs suggest a recognizable face,even though the DDPM does capture left–right symmetry in those few flips.

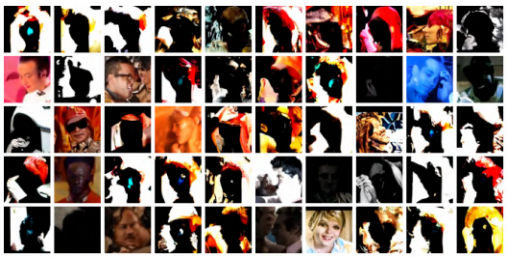


(a) Samples generated using classifier guidance.

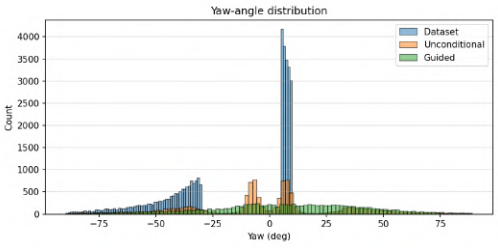


(b) Distribution with $\lambda = 20$.

Figure 9: Guidance generation with $\lambda = 20$ with the dataset described in 5.2.



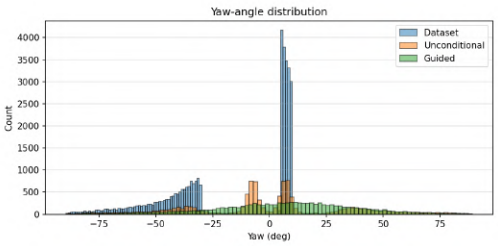
(a) Samples generated using classifier guidance.



(b) Yaw distribution with $\lambda = 30$.

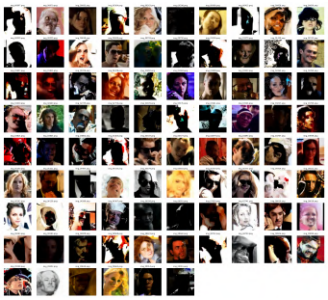
Figure 10: Guidance generation with $\lambda = 30$ with the dataset described in 5.2.

(a) Samples generated using classifier guidance.

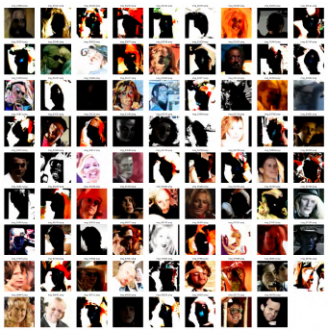


(b) Distribution with $\lambda = 45$.

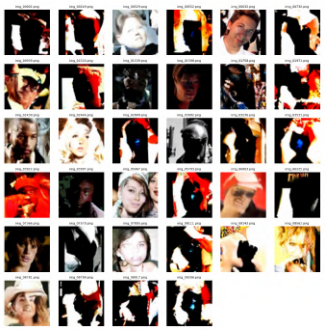
Figure 11: Guidance generation with $\lambda = 45$ with the dataset described in 5.2.



(a)



(b) Distribution with $\lambda = 20$.



(c) Distribution with $\lambda = 20$.

Figure 12: Guidance generation with different weights paramaters: 15, 30 and 45 on the dataset C.5.2.

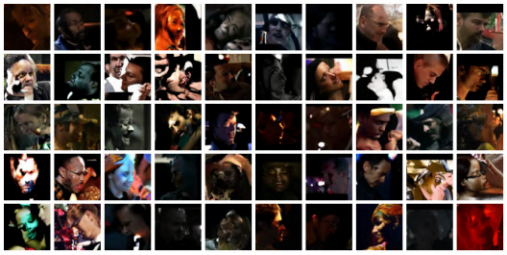
7.3 Spectral Normalization on U-Net Features

To prevent gradient magnitudes from exploding when mixing score estimates for widely separated pose classes, we applied spectral normalization to all convolutional weight matrices in the pretrained U-Net during both training and inference. In the zero-shot interpolation context, (Deschenaux et al., 2024) demonstrate that spectral normalization applied to the guidance classifier can smooth predictions in unseen regions, potentially yielding more coherent intermediate samples. Specifically, during training and inference, for each convolutional layer weight W we compute its largest singular value $\sigma_{\max}(W)$ via one-step power iteration and replace W with $\bar{W} = W/\sigma_{\max}(W)$, constraining the layer's Lipschitz constant to 1 and keeping gradients bounded. Although this regularization theoretically stabilizes the denoising trajectory, in our experiments we did not observe any visually noticeable improvement in pose-interpolation quality.

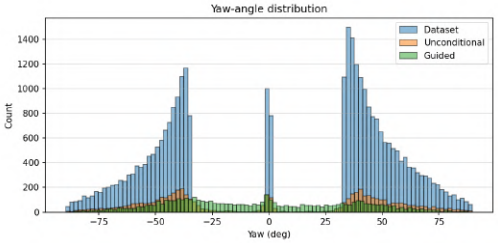
7.3.1 Results

Figure 13a shows sample outputs generated with classifier guidance using spectral normalization (SN), and Figure 13b plots the corresponding pose label distribution for $\lambda = 45$. In practice, applying SN to the U-Net's convolutional layers did not yield consistently better interpolation results. Qualitatively, the faces in Figure 13a still exhibit blurring and distortions similar to the no-SN baseline, and in some cases artifacts appear more pronounced. Quantitatively, the pose distribution in Figure 13b remains wide and multi-modal, indicating that SN fails to concentrate outputs toward intermediate poses more effectively than without SN.

One possible explanation is that SN constrains the Lipschitz constant of each convolutional layer, which helps prevent gradient explosion but does not directly address the semantic gap between extreme poses. Since the classifier gradients remain noisy when extrapolating to unseen intermediate angles, simply bounding weight norms is insufficient to guide the diffusion process toward coherent rotations. In other words, SN stabilizes gradient magnitude but cannot invent missing "in-between" pose information; thus, it provides little to no improvement,and sometimes degrades pose-interpolation quality altogether.



(a) Samples generated using classifier guidance with spectral normalization.



(b) Distribution with $\lambda = 45$.

Figure 13: Guidance generation with $\lambda = 45$ with the dataset described in 5.2.

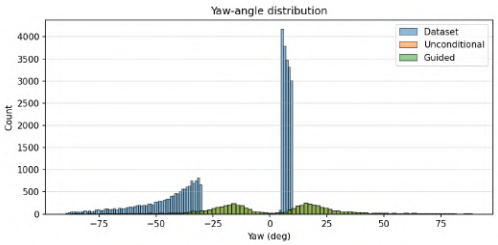
7.4 Classifier-Free Guidance Integration

We train our DDPM on split C using the classifier-free guidance (CFG) framework: at each diffusion timestep t, with probability $p_{\text{drop}} = 0.1$ we replace the pose conditioning label c with a null token $\varnothing$, training the network to predict both conditional and unconditional noise estimates.

Dataset C with Classifier-Free Guidance Figure 14a shows samples generated using classifier-free guidance on Dataset C, and Figure 14b shows the corresponding yaw-angle distribution. Compared to classifier guidance, classifier-free guidance produces much smoother transitions between "extreme left" and "slightly right" poses. In Figure 15a, many faces appear to move gradually from left-profile toward an almost front pose, whereas before most outputs were just dark blobs. The yaw histogram in Figure 14b confirms that guided samples now cover a wider range of yaw angles, including intermediate values that were missing previously. Surprisingly, the model even generates flipped (mirror) versions of faces, indicating it has learned to recreate left–right symmetry rather than relying solely on the classifier's two labels 15b. We also see this in the histogram with the some faces being

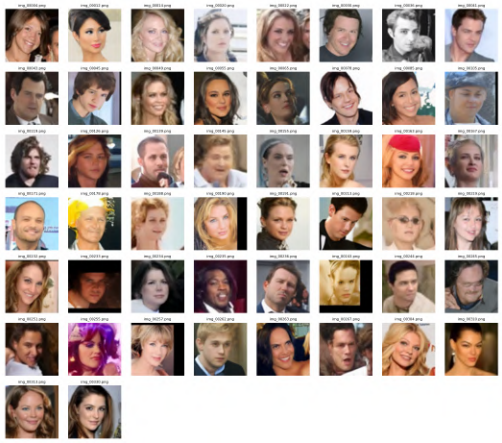
generated with yaws over 20 degrees. We plot several of them (with $20 \leq \text{yaw} \leq 25$) in Figure 15b. Overall, classifier-free guidance allows the DDPM to fill in the gap more effectively and produce higher-quality, more varied poses than classifier guidance alone.

(a) Samples generated using classifier guidance.

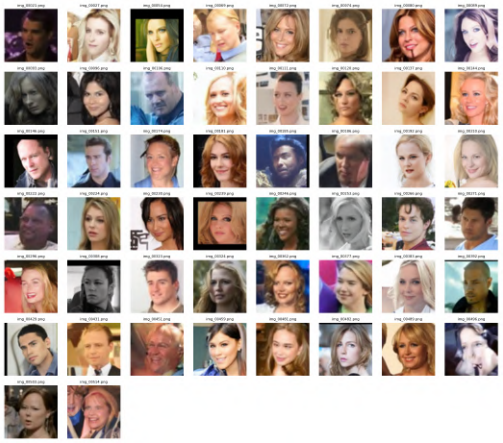


(b) Distribution with $\lambda = 45$.

Figure 14



(a) Quality samples between -25 and -20 degrees of yaw orientation.



(b) Quality samples between +20 and +25 degrees of yaw orientation.

Figure 15

7.5 Synthetic Dataset

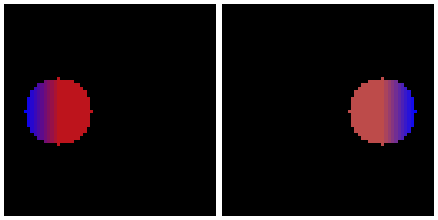


Figure 16: Example of samples included in the synthetic dataset.

We also evaluate our guidance methods on a simple, synthetic dataset (see Figure 16 for visual examples) inspired by Deschenaux et al. (2024). The dataset comprises 15000 64×64 images evenly split between two classes ("left" vs. "right"). Each image contains a single disc whose horizontal position, vertical position, hue (red tints can vary in average by 10), and gradient start have a epsilon variance. This controlled setting allows us to precisely assess each method's ability to interpolate between the two endpoint modes.

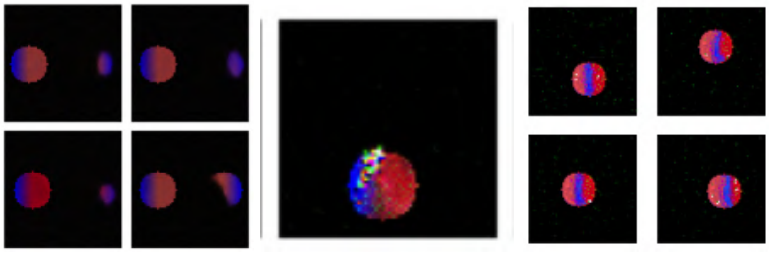


Figure 17: Qualitative interpolation between the two endpoint classes on the synthetic dataset. Left: Classifier Guidance (CG). Middle: Classifier-Free Guidance (CFG). Right: CFG with Spectral Normalization (CFG+SN).

7.5.1 Classifier Guidance (CG)

When using traditional classifier guidance, we observe that the transition between the two modes remains essentially bimodal: at intermediate guidance strengths the sampler still produces two distinct discs corresponding to the endpoints rather than a single blended object. Concretely:

- **No smooth interpolation:** samples at intermediate steps show two separate discs, indicating a failure to merge the modes.
- **No gradual blending of colors:** there is no perceptible gradient between the pure "left" and "right" hues.

These artifacts suggest that CG overly biases generation toward the two learned class prototypes.

7.5.2 Classifier-Free Guidance (CFG)

Classifier-free guidance sometimes remedies the bimodal failure by producing a single disc at intermediate latents, effectively interpolating between the endpoint classes. However, this comes at the cost of:

- **Color consistency issues:** the hue across the disc can fluctuate, leading to patchy or blotchy appearance.
- **Edge artifacts:** small speckle noise and ringing around the disc boundary become visible.

7.5.3 Classifier-Free Guidance with Spectral Normalization (CFG+SN)

Augmenting CFG with spectral normalization on the score network further improves interpolation quality. As shown in Figure 17, CFG+SN yields:

- **Smooth single-ball blends:** the sampler consistently produces a single disc whose color transitions smoothly from one endpoint hue to the other.
- **Minor remaining artifacts:** slight edge noise and minor texture imperfections persist, and color saturation across the disc can be marginally uneven.

8 Conclusion, Discussion and Future Work

We test how well a diffusion model edits faces without any extra training or labels. We confirm that zero-shot multi-guidance handles small edits like adding a smile or slight aging. When we move to larger changes, making a face look much older or rotating it from full left to full right, direct multi-guidance fails. For age progression, small gaps (e.g., "20s" $\rightarrow$ "30s") produce some wrinkles, but larger age jumps cause odd artifacts. For pose rotation, guiding the model with a simple yaw classifier over 4000 steps still yields blurry and distorted outputs without a clear rotation. We then test three auxiliary techniques to improve pose control: (1) training TransFusion from scratch with LLaMA 3 weights in its encoder (which does not converge under our compute limits), (2) adding spectral normalization to the U-Net during inference (which does not improve pose interpolation), and (3) using classifier-free guidance to blend frontal and rotated embeddings (which gives modest gains but

still needs unlabeled front faces). We also evaluated our guidance methods on a controlled synthetic dataset of 15 000 colored-disc images interpolating between left and right endpoints, finding that classifier guidance fails to blend the modes (producing two separate discs), classifier-free guidance can yield a single disc but with color-consistency and edge artifacts, and CFG+SN produces the smoothest interpolation with only minor noise. Overall, multi-guidance diffusion excels at small, local edits (e.g. smiles) but fails on large, global changes like full rotations; CFG and spectral normalization smooth interpolation and stabilize gradients yet still exhibit artifacts, and TransFusion+LLaMA-3 via LoRA does not converge under current compute/data. Next, we'll combine CFG+SN for stronger global control and scale up data/token coverage or adapter capacity to make TransFusion viable.

Discussion Our experiments highlight several key points:

- **Limits of Pure Zero-Shot.** Direct gradient-based guidance produces realistic small edits but cannot bridge large attribute gaps. When the model has never seen intermediate poses, it cannot invent them.
- **Training TransFusion with LLaMA 3.2 Weights.** We train TransFusion from scratch by initializing its encoder with LLaMA 3 weights, hoping this provides better pose control. However, the strong language priors and our limited data mean the model only learns to denoise, not rotate faces.
- **Spectral Normalization.** Although SN keeps gradients stable, it does not add any new 3D information. In practice, it makes little difference for pose interpolation and sometimes worsens artifacts.
- **Classifier-Free Guidance.** CFG mixes internal pose representations and yields smoother transitions.
- **Other Attempts.** We also try using a continuous classifier (based on the yaw detector from yakhyo (2025). Unfortunately, we struggled to make it converge as we needed to retrain it on noisy examples. It was landing reasonable results for slightly noisy images but struggled but higher noise level.

Future Work We see two main avenues to build on this work:

- **Incorporate 3D Priors.** Embedding explicit 3D geometry, such as a neural radiance field or a 3D morphable model, into the diffusion guidance process may allow more reliable pose control without full supervision. For example, a pretrained 3DMM encoder could supply intermediate pose embeddings that bridge extreme yaw angles.
- **Develop Inference-Time Controllers.** Designing novel guidance mechanisms that manipulate pretrained latent spaces (e.g., CLIP, latent diffusion, or a separate pose encoder) during inference could steer pose or age attributes more robustly. A dynamic, iterative controller that adjusts guidance weights based on intermediate outputs may reduce reliance on labeled intermediate samples.

References

Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16123–16133, 2022.

Xiangyi Chen and Stéphane Lathuilière. Face aging via diffusion-based editing. *arXiv preprint arXiv:2309.11321*, 2023.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

Justin Deschenaux, Igor Krawczuk, Grigorios Chrysos, and Volkan Cevher. Going beyond compositions, ddpms can produce zero-shot interpolations. *arXiv preprint arXiv:2405.19201*, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Omar Elharrouss, Noor Almaadeed, and Somaya Al-Maadeed. Lfr face dataset:left-front-right dataset for pose-invariant face recognition in the wild. In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, pp. 124–130, 2020. doi: 10.1109/ICIoT48696.2020.9089530.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yuming Gu, Hongyi Xu, You Xie, Guoxian Song, Yichun Shi, Di Chang, Jing Yang, and Linjie Luo. Diffportrait3d: Controllable diffusion for zero-shot portrait view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10456–10465, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Tianyang Hu, Fei Chen, Haonan Wang, Jiawei Li, Wenjia Wang, Jiacheng Sun, and Zhenguo Li. Complexity matters: Rethinking the latent space for generative modeling. *Advances in Neural Information Processing Systems*, 36:29558–29579, 2023.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6007–6017, 2023.

Hanjae Kim, Jiyoung Lee, Seongheon Park, and Kwanghoon Sohn. Hierarchical visual primitive experts for compositional zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5675–5685, 2023.

Yeonkyeong Lee, Taeho Choi, Hyunsung Go, Hyunjoon Lee, Sunghyun Cho, and Junho Kim. Exp-gan: 3d-aware facial image generation with expression control. In *Computer Vision – ACCV 2022: 16th Asian Conference on Computer Vision, Macao, China, December 4–8, 2022, Proceedings, Part VII*, pp. 151–167, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-26292-0. doi: 10.1007/978-3-031-26293-7_10. URL `https://doi.org/10.1007/978-3-031-26293-7_10`.

Gen Li and Yuling Yan. Adapting to unknown low-dimensional structures in score-based diffusion models. *arXiv preprint arXiv:2405.14861*, 2024.

Xiyao Liu, Yang Liu, Yuhao Zheng, Ting Yang, Jian Zhang, Victoria Wang, and Hui Fang. Semantics-guided Generative Diffusion Model with a 3DMM Model Condition for Face Swapping. *Computer Graphics Forum*, 2023. ISSN 1467-8659. doi: 10.1111/cgf.14949.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Safa C Medin, Bernhard Egger, Anoop Cherian, Ye Wang, Joshua B Tenenbaum, Xiaoming Liu, and Tim K Marks. Most-gan: 3d morphable stylegan for disentangled face image manipulation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 1962–1971, 2022.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Brian B Moser, Arundhati S Shanbhag, Federico Raue, Stanislav Frolov, Sebastian Palacio, and Andreas Dengel. Diffusion models, image super-resolution, and everything: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, 29, 2016.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Yichun Shi, Divyansh Aggarwal, and Anil K Jain. Lifting 2d stylegan for 3d-aware face generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6258–6266, 2021.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Keqiang Sun, Shangzhe Wu, Ning Zhang, Zhaoyang Huang, Quan Wang, and Hongsheng Li. Cgof++: Controllable 3d face synthesis with conditional generative occupancy fields. *IEEE transactions on pattern analysis and machine intelligence*, 46(2):913–926, 2023.

Anwaar Ulhaq and Naveed Akhtar. Efficient diffusion models for vision: A survey. *arXiv preprint arXiv:2210.09292*, 2022.

yakhyo. head-pose-estimation. `https://github.com/yakhyo/head-pose-estimation/tree/main`, 2025.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.