Durian: Dual Reference Image-Guided Portrait Animation with Attribute Transfer

Hyunsoo Cha Byungjun Kim Hanbyul Joo
Seoul National University
{243stephen, byungjun.kim, hbjoo}@snu.ac.kr
https://hyunsoocha.github.io/durian

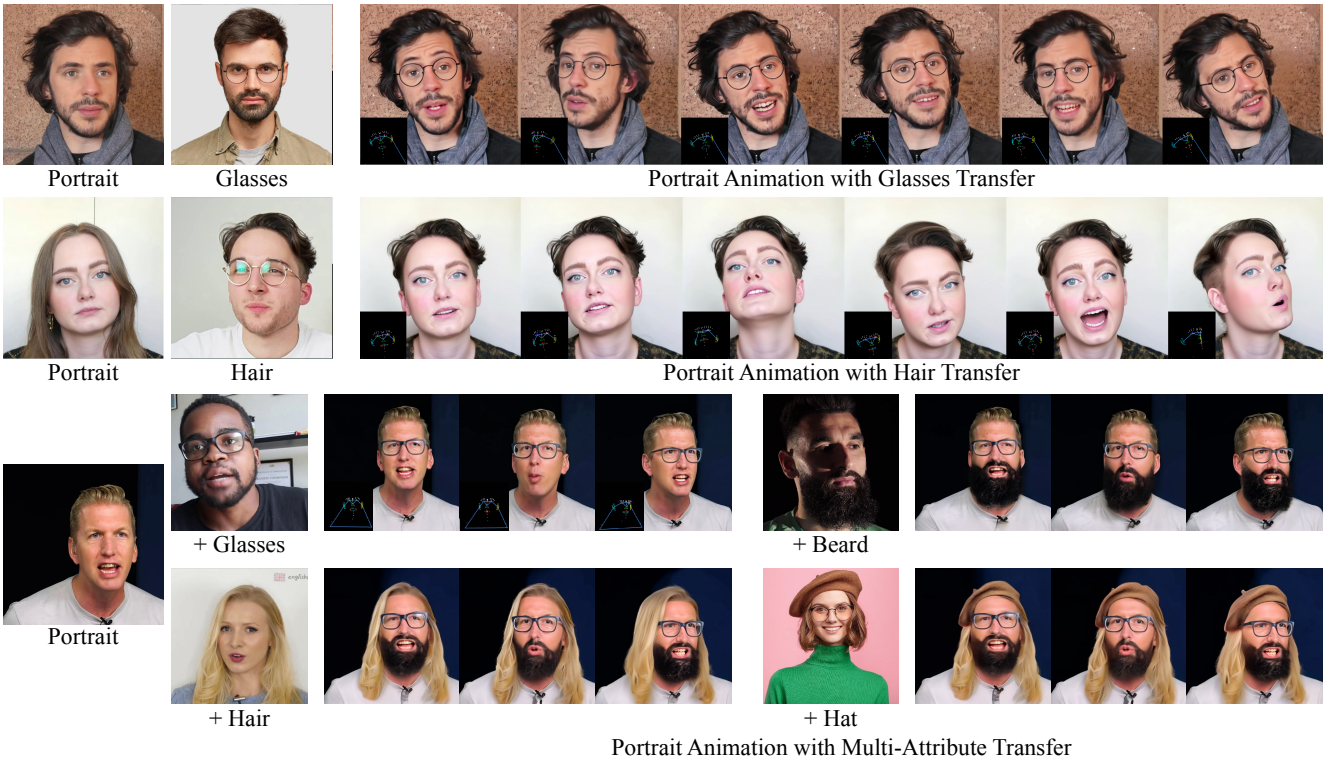

Figure 1. **Durian.** Given a portrait image and single or multiple reference images specifying target attributes (*e.g.*, hairstyle, eyeglasses), our method generates a portrait animation with facial attribute transfer conditioned on a keypoint sequence.

Abstract

We present Durian, the first method for generating portrait animation videos with cross-identity attribute transfer from one or more reference images to a target portrait. Training such models typically requires attribute pairs of the same individual, which are rarely available at scale. To address this challenge, we propose a self-reconstruction formulation that leverages ordinary portrait videos to learn attribute transfer without explicit paired data. Two frames from the same video act as a pseudo pair: one serves as an attribute reference and the other as an identity reference. To enable this self-reconstruction training, we introduce a Dual ReferenceNet that processes the two references separately and then fuses their features via spatial attention within a diffusion model. To make sure each reference functions as a specialized stream for either identity or attribute information, we apply complementary masking to the reference images. Together, these two components guide the model to reconstruct the original video, naturally learning cross-identity attribute transfer. To bridge the gap between self-reconstruction training and cross-identity inference, we introduce a mask expansion strategy and augmentation schemes, enabling robust transfer of attributes with varying spatial extent and misalignment.

Durian achieves state-of-the-art performance on portrait animation with attribute transfer. Moreover, its dual reference design uniquely supports multi-attribute composition and smooth attribute interpolation within a single generation pass, enabling highly flexible and controllable synthesis.

1. Introduction

Personalized appearance editing, such as virtually trying on glasses or experimenting with new hairstyles, is becoming a key feature of virtual styling applications. However, most existing solutions are highly specialized and limited in scope. Hairstyle preview apps typically rely on fixed templates, which may look realistic from a single view but fail to adapt to head pose or expression changes. Glasses try-on systems often depend on pre-scanned 3D product models, restricting users to a predefined catalog. Furthermore, these systems focus on a single attribute and cannot combine multiple elements, such as hair, glasses, or hats, within a unified experience.

A key challenge in building such a system is obtaining suitable training data. Disentangling identity from attributes ideally requires paired images of the same person with different attributes, which are rarely available and expensive to collect at scale. This difficulty grows exponentially for multiple attributes, as capturing all combinations quickly becomes infeasible. For example, Li et al. [33] collect multi-view images of subjects wearing different eyeglasses to model realistic glasses try-on, but the dataset remains too limited to generalize broadly. Zhang et al. [66] propose a synthetic pipeline that predicts a bald version of a portrait and generates reference hair images using a pretrained diffusion model. However, this approach is not easily scalable beyond hair.

This naturally raises the question: *can we train a model for portrait animation with attribute transfer without any explicit attribute-paired data?* Motivated by this question, we propose a **self-reconstruction framework** that learns this task directly from widely available in-the-wild portrait videos. During training, we randomly sample two frames from a single video: one as the attribute reference and the other as the identity reference. The remaining frames are treated as targets to be generated, conditioned on a keypoint sequence representing the motion of the video. To prevent identity leakage, we apply complementary masking to the two reference frames so that the network must disentangle and combine the attribute and identity information to reconstruct the original video.

To enable this framework, we design a **Dual ReferenceNet** architecture that explicitly encodes the attribute and portrait references through two separate branches and fuses their disentangled features for generation via spatial attention. This design enables the network to move beyond simple pose driving, generating keypoint-driven portrait animations that seamlessly combine the attribute from one image with the identity from the other. Surprisingly, although the model is trained with only a single attribute reference at a time, the spatial attention mechanism allows more advanced operations at inference time. Since different attributes (*e.g.*, hair, glasses, beard, hats) occupy distinct spatial regions, their features can be jointly injected without conflict, enabling seamless multi-attribute transfer. Furthermore, by interpolating the features of two attribute references, our model can achieve attribute interpolation, generating smooth transitions between the attributes. These emergent capabilities make our framework especially valuable for real-world styling scenarios, where users may want to explore diverse combinations and gradual transformations of facial attributes.

While self-reconstruction training is effective for learning to separate identity and attributes, it operates within a single video, leading to a domain gap when the model is applied to cross-identity inference, where the attribute and portrait come from different individuals. To mitigate this gap, we introduce a mask expansion strategy and lightweight augmentation schemes. These techniques expose the model to a broader range of attribute configurations during training, enabling robust transfer across spatial and structural variations of the attribute region. These designs form a unified framework capable of robust cross-identity attribute transfer. As a result, our method achieves a versatile system that generates portrait animations with diverse appearance edits in a zero-shot manner.

We summarize the key contributions of our work, as follows:

- We propose the first method to generate keypoint-driven portrait animations with transferred attributes directly from two images, generalized across diverse facial attributes beyond hair.
- We design a Dual ReferenceNet architecture that disentangles attribute and identity through two branches fused via spatial attention, enabling self-reconstruction training directly on uncurated in-the-wild videos without paired data.
- We propose a mask expansion strategy and lightweight augmentations to bridge the domain gap for cross-identity transfer, improving robustness to diverse spatial configurations.
- Our framework exhibits an emergent ability to support multi-attribute composition and interpolation in a single generation pass, without requiring any additional training.

2. Related Work

Face Editing. Generative models have advanced facial editing from unconditional synthesis to fine-grained manipulation of existing images [16, 22, 43]. Latent-space editing with StyleGAN [24] and GAN inversion [1, 44, 67] has been extended to video via latent trajectory modeling [49, 60] and 3D-aware editing [2, 57]. However, such approaches

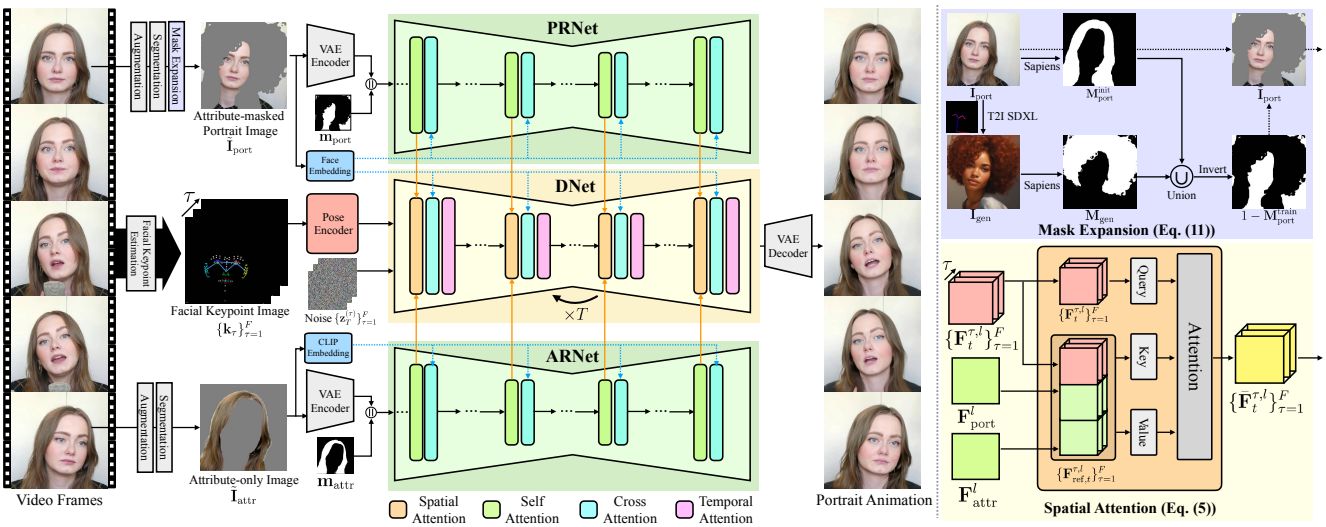

Figure 2. **Overview of Training Pipeline.** Given an attribute-masked portrait image $\tilde{\mathbf{I}}_{\text{port}}$ and an attribute-only image $\tilde{\mathbf{I}}_{\text{attr}}$, Durian synthesizes a portrait animation with the transferred attribute. These inputs are constructed by randomly sampling two frames from a training video and applying the estimated masks. A sequence of facial keypoints $\{\boldsymbol{k}_\tau\}_{\tau=1}^{F}$ is extracted from the video to guide the motion. During generation, spatial features from PRNet and ARNet are fused via spatial attention into the DNet, ensuring identity preservation and attribute consistency in the synthesized video.

often rely on attribute classifiers or fixed editing controls. Diffusion-based models have introduced more flexible editing through prompt-driven [3] or identity-preserving techniques [52, 61], with extensions to video improving temporal consistency [27, 31]. Still, these methods are limited to modifying existing content and cannot generate new motions or expressions.

Diffusion-based Attribute Transfer. Diffusion-based attribute transfer methods typically formulate editing as masked inpainting, where reference content is inserted into a target image using explicit masks [8, 9, 38, 46, 58]. These approaches have been adapted to domain-specific tasks such as hairstyle [12, 66], clothing [10, 28, 35], and makeup [65]. While effective for static images, they rely on category labels or mask annotations. Video extensions [15, 48] apply per-frame inpainting with post-hoc smoothing, but predefined masks are hard to specify for deformable facial attributes that vary over time. Recent works have also explored attribute transfer in 3D avatars [6, 7, 26, 29, 39, 51], but such approaches often require specialized capture setups or are not easily generalizable to in-the-wild scenarios. In contrast, our model performs attribute transfer and animation jointly in a single forward pass, conditioned only on a pair of reference images and a facial keypoint sequence. This eliminates the need for per-frame masks, text prompts, or category labels, enabling zero-shot transfer of diverse facial attributes.

Portrait Animation from a Single Image. Portrait animation aims to generate motion from a static image, typically guided by facial keypoints, audio, or motion trajectories. Early methods rely on GANs with implicit keypoint modeling [17, 54], while recent approaches use diffusion models [23, 59, 68] for improved realism and temporal stability. These methods primarily focus on reenactment and identity preservation. Others incorporate paired motion [56] or audio [59], but require multi-stage inference or fine-tuning. Our model jointly performs facial attribute transfer and motion generation, producing photorealistic, identity-preserving videos from diverse attribute references and keypoint-driven motion in a single pass.

3. Method

3.1. Overview: Learning Attribute Transfer from Self-Reconstruction

We propose a diffusion-based generative framework for portrait animation with cross-identity attribute transfer. At a high level, our model generates an F-frame animation sequence $\mathbf{V} = \{\mathbf{I}_\tau\}_{\tau=1}^{F}$ as:

$$\mathbf{V} = \text{Durian}(\mathbf{I}_{\text{attr}}, \mathbf{M}_{\text{attr}}, \mathbf{I}_{\text{port}}, \mathbf{M}_{\text{port}}, \mathbf{K}), \qquad (1)$$

conditioned on an attribute image $\mathbf{I}_{\text{attr}}$, a portrait image $\mathbf{I}_{\text{port}}$, and a sequence of driving facial keypoint images $\mathbf{K} = \{\boldsymbol{k}_\tau\}_{\tau=1}^{F}$. Each reference image has a binary mask: $\mathbf{M}_{\text{attr}}$ localizes the attribute region (*e.g.*, hair or glasses) in the reference image, while $\mathbf{M}_{\text{port}}$ specifies the candidate region in the portrait where the attribute will be transferred.

Using these masks, we construct two masked inputs: the *attribute-only image* $\tilde{\mathbf{I}}_{\text{attr}} = \mathbf{I}_{\text{attr}} \odot \mathbf{M}_{\text{attr}}$, where only the attribute region is preserved, and the *attribute-masked portrait image* $\tilde{\mathbf{I}}_{\text{port}} = \mathbf{I}_{\text{port}} \odot (1 - \mathbf{M}_{\text{port}})$, where the corresponding region is removed. These masked inputs are fed into the **Dual ReferenceNet**, consisting of the *Attribute ReferenceNet (ARNet)* and *Portrait ReferenceNet (PRNet)*, which extract multi-scale spatial features. These features are then injected into a diffusion-based generator, the *Denoising UNet (DNet)*, to synthesize the remaining frames of the video with keypoint guidance $\mathbf{K}$ (Sec. 3.2).

To enable training without requiring explicitly annotated triplets (*i.e.*, combinations of a target attribute image, an original portrait image, and an edited portrait image), we adopt a **self-reconstruction strategy** based on portrait videos [55, 62]. Specifically, we simulate attribute transfer by sampling two frames $\mathbf{I}_{\text{attr}}$ and $\mathbf{I}_{\text{port}}$ from the same video, treating one as the attribute reference and the other as the target portrait. We then construct the masked inputs $\tilde{\mathbf{I}}_{\text{attr}}$ and $\tilde{\mathbf{I}}_{\text{port}}$ using the same masking formulation as in inference, based on a segmentation mask of a randomly selected attribute. Although the two frames come from the same identity, the complementary masking enforces a clear separation between identity and attribute inputs, encouraging the model to learn meaningful mappings from these features to output frames without requiring cross-identity supervision. To enhance the model's ability to generalize beyond the self-attribute transfer setup, we introduce an augmentation scheme that improves robustness to spatial and appearance variations(Sec. 3.3).

At inference time, we estimate refined attribute masks by aligning the attribute image to the portrait through a lightweight alignment process, mitigating spatial misalignment between them. Conditioned on the two masked reference images and the driving keypoint sequence, our model then synthesizes portrait animations with attribute transfer. Notably, our design also supports multi-attribute composition and smooth interpolation within a single generation pass, without requiring additional training or post-processing (Sec. 3.4). Fig. 1 shows our generated portrait animations with attribute transfer.

3.2. Model Architecture: Dual ReferenceNet

Inspired by recent approaches [18, 23, 68] that leverage ReferenceNet to inject spatial features into diffusion models, we propose a **Dual ReferenceNet** architecture tailored for portrait animation with attribute transfer. Unlike previous work, our model includes two separate encoders: *Attribute ReferenceNet (ARNet)* and *Portrait ReferenceNet (PRNet)*, each sharing the same architecture as the *Denoising U-Net (DNet)* in the diffusion model, excluding the temporal layers. The networks follow the U-Net [36] architecture used in latent diffusion models [45], with each block containing convolutional layers followed by self- and cross-attention modules. The overall architecture is shown in Fig. 2.

Reference inputs. Given an attribute image $\mathbf{I}_{\text{attr}} \in \mathbb{R}^{3\times H\times W}$ and a portrait image $\mathbf{I}_{\text{port}} \in \mathbb{R}^{3\times H\times W}$, along with their binary masks $\mathbf{M}_{\text{attr}} \in \mathbb{R}^{1\times H\times W}$ and $\mathbf{M}_{\text{port}} \in \mathbb{R}^{1\times H\times W}$, which localize the attribute region and the candidate transfer region respectively, we construct two masked inputs: the attribute-only image $\tilde{\mathbf{I}}_{\text{attr}} = \mathbf{I}_{\text{attr}} \odot \mathbf{M}_{\text{attr}}$, where only the attribute region is preserved, and the attribute-masked portrait image $\tilde{\mathbf{I}}_{\text{port}} = \mathbf{I}_{\text{port}} \odot (1 - \mathbf{M}_{\text{port}})$, where the corresponding candidate region is removed. We then encode these masked images into latent representations using the pretrained VAE from the latent diffusion model [45], yielding $\boldsymbol{z}_{\text{attr}}, \boldsymbol{z}_{\text{port}} \in \mathbb{R}^{c\times h\times w}$. The corresponding masks $\mathbf{M}_{\text{attr}}, \mathbf{M}_{\text{port}}$ are downsampled to match the latent resolution, producing $\boldsymbol{m}_{\text{attr}}, \boldsymbol{m}_{\text{port}} \in \mathbb{R}^{1\times h\times w}$. These downsampled masks are concatenated with the latents along the channel dimension to form $(c+1)$-channel inputs $\tilde{\boldsymbol{z}}_{\text{attr}}, \tilde{\boldsymbol{z}}_{\text{port}} \in \mathbb{R}^{(c+1)\times h\times w}$ as follows:

$$\begin{aligned} \tilde{\boldsymbol{z}}_{\text{attr}} &= \text{concat}_{\text{c}}(\boldsymbol{z}_{\text{attr}}, \boldsymbol{m}_{\text{attr}}), \\ \tilde{\boldsymbol{z}}_{\text{port}} &= \text{concat}_{\text{c}}(\boldsymbol{z}_{\text{port}}, \boldsymbol{m}_{\text{port}}). \end{aligned} \tag{2}$$

Spatial attention. The augmented latents are passed to ARNet $\mathcal{E}_{\text{attr}}$ and PRNet $\mathcal{E}_{\text{port}}$ to extract multi-scale feature maps after convolutional layers of each block:

$$\begin{aligned} \mathcal{F}_{\text{attr}} &:= \{\mathbf{F}^l_{\text{attr}}\}_{l=1}^{L} = \mathcal{E}_{\text{attr}}(\tilde{z}_{\text{attr}}; \Theta_{\text{attr}}), \\ \mathcal{F}_{\text{port}} &:= \{\mathbf{F}^l_{\text{port}}\}_{l=1}^{L} = \mathcal{E}_{\text{port}}(\tilde{z}_{\text{port}}; \Theta_{\text{port}}). \end{aligned} \tag{3}$$

where $\Theta_{\{\text{attr},\text{port}\}}$ are the parameters of Dual ReferenceNet. Let $\mathbf{F}^{\tau,l}_t \in \mathbb{R}^{c_l\times h_l\times w_l}$ denote the feature map of the frame τ at the l-th block of the denoising U-Net. While the original denoising U-Net includes a self-attention layer at each resolution, we replace it with our spatial attention to integrate identity and attribute features in a spatially-aware manner. We denote width-wise concatenation as $\text{concat}_{\text{w}}(\cdot)$, and define our spatial attention $\text{SA}(\cdot,\cdot,\cdot)$ as:

$$\mathbf{F}^{\tau,l}_{\text{ref},t} := \text{concat}_{\text{w}}(\{\mathbf{F}^{\tau,l}_t, \mathbf{F}^l_{\text{port}}, \mathbf{F}^l_{\text{attr}}\}) \in \mathbb{R}^{c_l\times h_l\times 3w_l}, \tag{4}$$

$$\begin{aligned} \bar{\mathbf{F}}^{\tau,l}_t &= \text{SA}(\mathbf{F}^{\tau,l}_t, \mathbf{F}^l_{\text{port}}, \mathbf{F}^l_{\text{attr}}) \\ &= \text{Attention}(\mathbf{W}_Q\mathbf{F}^{\tau,l}_t, \mathbf{W}_K\mathbf{F}^{\tau,l}_{\text{ref},t}, \mathbf{W}_V\mathbf{F}^{\tau,l}_{\text{ref},t}), \end{aligned} \tag{5}$$

where $\bar{\mathbf{F}}^{\tau,l}_t \in \mathbb{R}^{c_l\times h_l\times w_l}$ is the feature map after the spatial attention, $\text{Attention}(Q,K,V) = \text{softmax}(QK^\top/\sqrt{d})V$ is the standard scaled dot-product attention [50], $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are linear projection layers. This width-wise concatenation preserves spatial resolution and allows the model to attend across all positions in the combined reference and target features. As a result, the model can leverage both attribute and portrait guidance at every step.

Cross-attention with semantic embeddings. After applying spatial attention, we further inject semantic guidance into both the Dual ReferenceNet and the denoising U-Net via cross-attention. For ARNet, we use the CLIP [42] embedding of the attribute-only image $\tilde{\mathbf{I}}_{\text{attr}}$ as the attribute embedding $\boldsymbol{\phi}_{\text{attr}}$, which is injected via cross-attention into each block of ARNet. For PRNet and DNet, we construct a portrait embedding $\boldsymbol{\phi}_{\text{port}}$ by combining ArcFace [14] and CLIP embeddings of the attribute-masked portrait image $\tilde{\mathbf{I}}_{\text{port}}$ following StableAnimator [47]. This embedding is injected into both PRNet and DNet to enhance identity preservation. We define the cross-attention operation $\text{CA}(\cdot,\cdot)$ as:

$$\text{CA}(\bar{\mathbf{F}},\, \boldsymbol{\phi}) = \text{Attention}(\mathbf{W}'_Q\bar{\mathbf{F}},\, \mathbf{W}'_K\boldsymbol{\phi},\, \mathbf{W}'_V\boldsymbol{\phi}), \quad (6)$$

where $\bar{\mathbf{F}}$ is the input feature map, $\boldsymbol{\phi}$ is the conditioning embedding, and $\mathbf{W}'_Q, \mathbf{W}'_K, \mathbf{W}'_V$ are learned linear projections. Let $\bar{\mathbf{F}}^l_{\text{attr}}$ and $\bar{\mathbf{F}}^l_{\text{port}}$ be the self-attended features of the l-th block in ARNet and PRNet, and $\bar{\mathbf{F}}^l_t$ the spatially attended feature of DNet. Then, the cross-attention updates are given by:

$$\begin{aligned} \tilde{\mathbf{F}}^l_{\{\text{attr,port}\}} &= \text{CA}(\bar{\mathbf{F}}^l_{\{\text{attr,port}\}},\, \boldsymbol{\phi}_{\{\text{attr,port}\}}), \\ \tilde{\mathbf{F}}^{\tau,l}_t &= \text{CA}(\bar{\mathbf{F}}^{\tau,l}_t,\, \boldsymbol{\phi}_{\text{port}}), \end{aligned} \quad (7)$$

where $\tilde{\mathbf{F}}^l_{\text{attr}}$, $\tilde{\mathbf{F}}^l_{\text{port}}$, and $\tilde{\mathbf{F}}^{\tau,l}_t$ are the feature maps after cross-attention in ARNet, PRNet, and DNet.

Temporal extension and keypoint guidance. Our model incorporates temporal awareness to generate coherent portrait animations by inserting temporal self-attention into each U-Net block, following Hu [23], Zhu et al. [68]. To control pose and expression, we use a sequence of facial keypoints $\mathbf{K} = \{\boldsymbol{k}_\tau\}_{\tau=1}^F$ extracted by Sapiens [25]. Each keypoint image $\boldsymbol{k}_\tau$ is encoded into a spatial feature map $\mathbf{F}^\tau_{\text{kpt}}$ via a pose encoder and combined with the noisy latent $\boldsymbol{z}^{(\tau)}_t$ following Zhu et al. [68]. For each frame τ, DNet ϵ_θ predicts the added noise $\hat{\epsilon}^{(\tau)}_t$ from the noisy latent $\boldsymbol{z}^{(\tau)}_t$ at timestep t, using the reference features, semantic embeddings, and keypoint features:

$$\hat{\epsilon}^{(\tau)}_t = \epsilon_\theta\left(\boldsymbol{z}^{(\tau)}_t,\, t,\, \mathcal{F}_{\text{attr}}, \mathcal{F}_{\text{port}}, \boldsymbol{\phi}_{\text{attr}}, \boldsymbol{\phi}_{\text{port}}, \mathbf{F}^\tau_{\text{kpt}}\right). \quad (8)$$

The predicted noise is used to recover the denoised latent $\boldsymbol{z}^{(\tau)}_0$, then decoded by the VAE decoder $\mathcal{D}$ to produce the final video frame as $\mathbf{I}_\tau = \mathcal{D}(\boldsymbol{z}^{(\tau)}_0)$ for $\tau = 1, \ldots, F$.

3.3. Training Strategy

Training loss. To effectively train our model, we adopt a two-stage training scheme following the previous approaches [23, 68]. In the first stage, we optimize the entire model except the temporal attention layers, treating each video frame as an independent training sample. We define the per-frame conditioning bundle as $\mathcal{C} := (\mathcal{F}_{\text{attr}}, \mathcal{F}_{\text{port}}, \boldsymbol{\phi}_{\text{attr}}, \boldsymbol{\phi}_{\text{port}})$, where $\mathcal{F}_{\text{port}}, \mathcal{F}_{\text{attr}}$ are the multi-scale spatial features from PRNet and ARNet and $\boldsymbol{\phi}_{\text{port}}, \boldsymbol{\phi}_{\text{attr}}$ are the semantic embeddings. Then, the training objective is the standard denoising diffusion loss:

$$\mathcal{L}^{(1)}_{\text{diff}} = \mathbb{E}_{\boldsymbol{z}_0,\,\epsilon,\,t}\left[\left\|\epsilon - \epsilon_\theta\left(\boldsymbol{z}_t,\, t,\, \mathcal{C},\, \mathbf{F}_{\text{kpt}}\right)\right\|^2\right], \quad (9)$$

where $\boldsymbol{z}_t$ is the noised latent at diffusion timestep t, ϵ is the sampled noise, and $\mathbf{F}_{\text{kpt}}$ is the feature map of the corresponding facial keypoint image. In the second stage, we freeze all modules except the temporal attention layers and train them using multi-frame inputs. The temporal objective considers a sequence of noised latents and corresponding keypoints:

$$\begin{aligned} \mathcal{L}^{(2)}_{\text{diff}} = \mathbb{E}_{\{\boldsymbol{z}^{(\tau)}_0\}_{\tau=1}^F,\, \boldsymbol{\epsilon}^{1:F},\, t}\Big[\Big\|\boldsymbol{\epsilon}^{1:F} - \epsilon_\theta\Big(\\ \{\boldsymbol{z}^{(\tau)}_t\}_{\tau=1}^F,\, t,\, \mathcal{C},\, \{\mathbf{F}^\tau_{\text{kpt}}\}_{\tau=1}^F\Big)\Big\|^2\Big] \end{aligned} \quad (10)$$

where $\boldsymbol{\epsilon}^{1:F} = \{\epsilon^{(\tau)}\}_{\tau=1}^F$ denotes the per-frame noise sequence. This staged training improves convergence and allows the temporal attention module to focus on modeling motion dynamics without disrupting the spatial fidelity learned in the first stage.

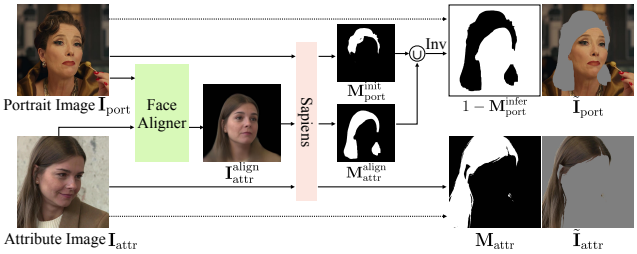

Figure 3. **Aligned Attribute Mask Estimation.** To improve attribute-portrait alignment, we estimate an aligned attribute mask via Face Aligner.

Attribute-aware mask expansion. To expose the model to diverse spatial extents of facial attributes during training, we introduce an attribute-aware mask expansion strategy, illustrated in the top right of Fig. 2. Given a training frame $\mathbf{I}$, we first select a target attribute (*e.g.*, hair, eyeglasses, beard) and obtain its binary mask $\mathbf{M}_{\text{attr}}$ using Sapiens [25]. To simulate variation in the shape and coverage of this attribute, we generate a modified image $\mathbf{I}_{\text{gen}}$ with SDXL [41] and ControlNet [64], conditioned on the facial keypoints of $\mathbf{I}$ and a text prompt describing an altered appearance (e.g., "long wavy hair"). To enable fully automated prompt generation without any human intervention, we construct a dictionary

of descriptive attribute modifiers (e.g., long, short, wavy, curly) and randomly sample their combinations to generate prompts for image generation. A new mask $\mathbf{M}_{\text{gen}}$ is then extracted from $\mathbf{I}_{\text{gen}}$ using Sapiens. The final training mask is computed as the union of the original and generated masks, and the two masked inputs are constructed as:

$$\begin{aligned} \mathbf{M}_{\text{port}}^{\text{train}} &= \mathbf{M}_{\text{attr}} \cup \mathbf{M}_{\text{gen}}, \\ \tilde{\mathbf{I}}_{\text{attr}} = \mathbf{I} \odot \mathbf{M}_{\text{attr}}, &\tilde{\mathbf{I}}_{\text{port}} = \mathbf{I} \odot (1 - \mathbf{M}_{\text{port}}^{\text{train}}), \end{aligned} \tag{11}$$

where $\odot$ denotes element-wise multiplication. Here, $\mathbf{M}_{\text{attr}}$ localizes the original attribute region, while $\mathbf{M}_{\text{port}}^{\text{train}}$ defines the expanded region into which the attribute will be inserted during generation. This expansion process is *attribute-aware* as it preserves the intended attribute category while diversifying its spatial extent. Unlike HairFusion [12], which expands masks using fixed heuristics specific to hair, our approach generalizes across multiple facial attributes and enables the model to learn spatially flexible yet semantically grounded transfer patterns.

Reference image augmentation. To address the limited diversity of self-reconstruction setups, we introduce an augmentation pipeline that improves robustness to pose, alignment, and appearance variations in attribute–portrait pairs. We perturb both the attribute-only and masked portrait images to simulate realistic spatial and photometric variations. We apply random affine transformations (translation, scaling, rotation) to induce spatial misalignment, and use the FLUX outpainting model [32] to inpaint newly exposed regions. Additionally, color jittering on tone, contrast, saturation, and hue accounts for appearance variations. This strategy exposes the model to diverse configurations, enabling more robust attribute transfer and animation under real-world variations.

3.4. Inference Framework and Extensions

Inference pipeline. At inference time, our system takes as input a portrait image, an attribute image, and a keypoint sequence. We first construct two masked reference images: the attribute-only image $\tilde{\mathbf{I}}_{\text{attr}}$ and the attribute-masked portrait image $\tilde{\mathbf{I}}_{\text{port}}$, by applying segmentation masks predicted by Sapiens [25] to the attribute image $\mathbf{I}_{\text{attr}}$ and the portrait image $\mathbf{I}_{\text{port}}$. To improve spatial alignment between the attribute and portrait inputs, we introduce a *Face Aligner* module, which repurposes a lightweight image-to-3D avatar model [11] solely for alignment. This module reconstructs a coarse 3D avatar from the attribute image and aligns its shape and pose to the portrait using FLAME [34] parameters $(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi})$ estimated by EMOCA [13]. From the resulting pose-aligned image $\mathbf{I}_{\text{attr}}^{\text{align}}$, we extract a refined attribute mask $\mathbf{M}_{\text{attr}}^{\text{align}}$ using Sapiens. This mask is then merged with the initial portrait mask $\mathbf{M}_{\text{port}}^{\text{init}}$ to define the final transferable region $\mathbf{M}_{\text{port}}^{\text{infer}} = \mathbf{M}_{\text{port}}^{\text{init}} \cup \mathbf{M}_{\text{attr}}^{\text{align}}$. The updated mask is applied to construct the final attribute-masked portrait image, $\tilde{\mathbf{I}}_{\text{port}} = \mathbf{I}_{\text{port}} \odot (1 - \mathbf{M}_{\text{port}}^{\text{infer}})$, as illustrated in Fig. 3. Finally, spatial features $\mathcal{F}_{\text{attr}}, \mathcal{F}_{\text{port}}$ and semantic embeddings $\boldsymbol{\phi}_{\text{attr}}, \boldsymbol{\phi}_{\text{port}}$ are extracted from the two masked reference images. Conditioned on these features and the keypoint sequence, DNet synthesizes a video of the target identity with the desired attribute through iterative denoising (Eq. (8)).

Multi-attribute transfer. Our model supports zero-shot composition of multiple attributes without additional training, by generalizing the spatial attention formulation in Eq. (5). Instead of using a single attribute feature, we concatenate multiple attribute feature maps along the width dimension:

$$\bar{\mathbf{F}}_t^l = \text{SA}\left(\mathbf{F}_t^l, \mathbf{F}_{\text{port}}^l, \text{concat}_{\text{w}}\left(\mathbf{F}_{\text{attr}}^{l,1}, \mathbf{F}_{\text{attr}}^{l,2}, \cdots, \mathbf{F}_{\text{attr}}^{l,N_{\text{attr}}}\right)\right), \tag{12}$$

where each $\mathbf{F}_{\text{attr}}^{l,k}$ denotes the feature map extracted from the k-th attribute-only image using the ARNet. To construct the final attribute-masked portrait in this setting, we also generalize the mask fusion process by taking the union of all aligned attribute masks:

$$\mathbf{M}_{\text{port}}^{\text{infer}} = \mathbf{M}_{\text{port}}^{\text{init}} \cup \bigcup_{k=1}^{N_{\text{attr}}} \mathbf{M}_{\text{attr}}^{\text{align},k}, \tag{13}$$

where each $\mathbf{M}_{\text{attr}}^{\text{align},k}$ is the aligned mask extracted from the k-th attribute image. This composite mask is then used to remove all attribute regions from the portrait image before generation. The rest of the attention computation remains unchanged, allowing the model to jointly attend to all attributes and synthesize coherent multi-attribute compositions without retraining.

Attribute interpolation. Our model enables zero-shot interpolation between two attributes of the same category (e.g., hairstyle A and B) without fine-tuning [7, 63]. Given two attribute-only images, we extract spatially attended features $\bar{\mathbf{F}}_t^{\tau,l,1}$ and $\bar{\mathbf{F}}_t^{\tau,l,2}$ using our spatial attention, and interpolate them as follows:

$$\bar{\mathbf{F}}_t^{\tau,l} = (1-\alpha)\,\bar{\mathbf{F}}_t^{\tau,l,1} + \alpha\,\bar{\mathbf{F}}_t^{\tau,l,2}, \tag{14}$$

where $\alpha \in [0, 1]$ controls the interpolation ratio. The interpolated feature $\bar{\mathbf{F}}_t^{\tau,l}$ is then passed to DNet for generation. This enables smooth and semantically consistent transitions between attributes.

4. Experiments

Experimental setup. To address the lack of ground-truth data for cross-identity attribute transfer, we design two evaluation settings: *self-attribute transfer* and *cross-attribute*

Table 1. **Quantitative Comparison.** We compare our method with recent approaches that (1) synthesize portraits with transferred hairstyles, and (2) animate the synthesized portrait image.

		Self-Attribute Transfer					Cross-Attribute Transfer				
Img.Gen.	Animation	$L_1 \downarrow$	PSNR↑	SSIM↑	LPIPS↓	FID↓	mCLIP-I↑	mDINO↑	ID-Sim.↑	$\text{VFID}_{\text{I3D}} \downarrow$	$\text{VFID}_{\text{ResNeXt}} \downarrow$
PbE	LivePortrait	0.1059	16.14	0.5641	0.2859	40.63	0.8499	0.6407	0.5630	37.6462	3.3868
	X-Portrait	0.1180	15.33	0.5270	0.2978	59.20	0.8393	0.5916	0.5458	36.7030	2.9008
	MegActor-$\sum$	0.1268	14.82	0.4840	0.3157	62.77	0.8535	0.6266	0.4863	38.2746	6.2743
HairFusion	LivePortrait	0.1438	13.76	0.4801	0.3792	46.24	0.8741	0.6843	0.6502	30.5632	2.6719
	X-Portrait	0.1511	13.30	0.4334	0.3733	59.02	0.8809	0.6914	0.6520	30.2570	4.9184
	MegActor-$\sum$	0.1650	12.75	0.4138	0.4015	65.59	0.8736	0.6708	0.6044	30.9702	5.3037
StableHair	LivePortrait	0.1122	15.84	0.5491	0.3041	43.74	0.8831	0.7051	0.6564	29.5014	3.9495
	X-Portrait	0.1229	15.04	0.5114	0.3117	53.36	0.8895	0.7239	0.6443	28.2627	1.5718
	MegActor-$\sum$	0.1301	14.62	0.4706	0.3347	63.47	0.8848	0.7271	0.6130	30.4087	**1.4672**
TriplaneEdit	LivePortrait	0.1023	16.52	0.5511	0.2924	57.86	0.8540	0.6163	0.2776	32.5660	8.9103
	X-Portrait	0.1051	16.05	0.5401	0.2760	60.25	0.8366	0.6216	0.2944	30.6319	2.9315
	MegActor-$\sum$	0.1248	15.10	0.4828	0.3293	70.41	0.8210	0.5674	0.2770	32.5679	2.8542
Ours		**0.0744**	**18.83**	**0.6527**	**0.1565**	**38.00**	**0.9043**	**0.7801**	**0.7098**	**27.1547**	2.4052

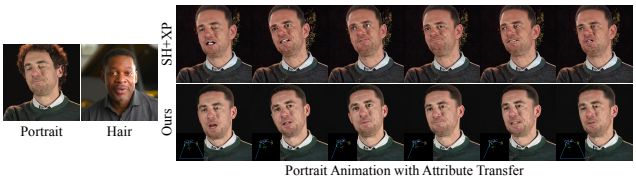

Figure 4. **Qualitative Comparison for Cross-Attribute Transfer.** We compare our method and the baselines that combine X-Portrait [56] with StableHair [66] in cross-identity transfer setup. We provide more results in our Supp. Mat.

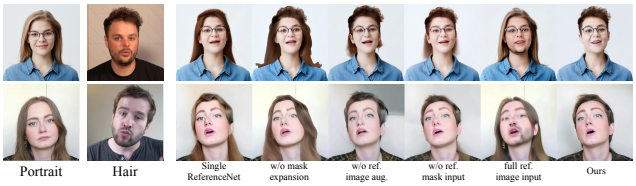

Figure 5. **Ablation Study.** Omitting components or altering training scheme degrades visual quality.

transfer. In **self-attribute transfer**, a single video is split into a portrait and an attribute image from different frames of the same identity, and the model reconstructs the original video. While useful for controlled evaluation, this provides only a pseudo ground-truth and mainly reflects reconstruction ability rather than the full complexity of cross-identity transfer. In **cross-attribute transfer**, the portrait and attribute images come from different individuals. Without exact ground-truth, this setting instead evaluates semantic consistency, identity preservation, and temporal realism. Together, the two settings offer a comprehensive evaluation of both low-level fidelity and high-level transfer quality.

Dataset. We train our model on CelebV-Text [62], VFHQ [55], and Nersemble [30], totaling 2,747 videos. For evaluation, we sample 200 videos for self-attribute transfer and 50 videos for cross-attribute transfer from CelebV-Text and VFHQ, ensuring diverse and unseen identities, head poses, and expressions. The masks for the portrait and attribute frames are generated following the procedure used in each compared method.

Metrics. For self-attribute transfer, we evaluate reconstruction fidelity using L_1, PSNR, SSIM, and LPIPS, and perceptual quality with FID [40]. For cross-attribute transfer, we measure attribute transfer quality with CLIP-I [20, 42] and DINO [4], identity preservation with ArcFace [14], and temporal realism with VFID [15] using I3D [5] and ResNeXt [19].

4.1. Comparison

Baselines. As no prior work directly tackles portrait animation with attribute transfer from in-the-wild reference images, we construct two-stage baselines by combining image-level attribute transfer with video animation methods, resulting in 12 model combinations. For attribute transfer (stage 1), we consider: Paint-by-Example (PbE) [58], a mask-conditioned diffusion method for reference image insertion; HairFusion [12] and StableHair [66], diffusion-based models for hairstyle transfer with and without masks; and TriplaneEdit [2], a 3D-aware GAN-based face editor. For portrait animation (stage 2), we use: LivePortrait [17], X-Portrait [56], and MegActor-$\sum$ [59].

Results. As shown in Tab. 1, our method consistently outperforms all baseline combinations across both fidelity and perceptual quality metrics in self-attribute transfer. Fig. 4 presents a qualitative comparison against baselines using LivePortrait [17] as the animation module (stage 2). Our method generates coherent and realistic hairstyle animations that preserve the identity and maintain consistency in spatial extent, shape, and fine details across frames. Please refer to

Table 2. **Ablation Study.** Bold indicates the best, underline the second.

Variant	$L_1 \downarrow$	PSNR↑	SSIM↑	LPIPS↓
single ReferenceNet	0.0813	17.95	0.6314	0.1973
w/o mask expansion	0.0881	17.16	0.5915	0.2073
w/o ref. image aug.	0.0900	16.97	0.5973	0.2248
w/o ref. mask input	0.0747	18.60	0.6511	0.1670
full ref. image input	**0.0670**	**19.47**	**0.6698**	**0.1310**
Ours	<u>0.0744</u>	<u>18.83</u>	<u>0.6527</u>	<u>0.1565</u>

our Supp. Mat. for additional qualitative comparisons with other baseline combinations.

4.2. Ablation Study

We evaluate the contributions of key components in our model and training strategy. Tab. 2 presents quantitative results, and Fig. 5 shows corresponding qualitative comparisons. **"single ReferenceNet"** replaces the dual-branch architecture with a shared encoder that receives the portrait and attribute images concatenated along the channel dimension, following CAT-VTON [10]. This setup fails to separate the roles of the two inputs, resulting in undesired blending of attribute and identity cues. **"w/o mask expansion"** omits the attribute-aware augmentation that simulates variations in spatial extent. Without this strategy, the model tends to rely on the default shape of the portrait's original attribute mask, making it less capable of handling diverse attribute shapes during inference. **"w/o ref. image aug."** disables spatial and photometric augmentations applied to the reference images during training. As a result, the model fails to accurately transfer the desired attribute with misaligned reference images. **"w/o ref. mask input"** removes the binary mask concatenation from the inputs to the ReferenceNets. This weakens spatial localization and often leads to artifacts or residual traces of the original attribute in the output. **"full ref. image input"** uses unmasked portrait and attribute images during training. Interestingly, this variant achieves the best quantitative scores in Tab. 2, which evaluates the self-attribute transfer setting, since full images simplify the task by allowing the model to copy content more easily. However, as shown in Fig. 5, this model fails to disentangle identity and attribute roles, leading to visible identity leakage during cross-identity transfer. **Ours** achieves spatially consistent, identity-preserving results, and quantitatively outperforms all other ablated variants except the full reference image variant.

4.3. Application

Multi-attribute transfer. Our model supports the composition of multiple attributes (*e.g.*, glasses, hat, hairstyle) in a single generation pass by extending the spatial attention mechanism as described in Eq. (12). Fig. 6 shows qualitative results where multiple attributes are simultaneously transferred from different reference images. Remarkably, our model not only combines multiple attributes seamlessly but also handles interactions between overlapping regions, such as between hair and a hat. Despite the reference images exhibiting diverse lighting conditions and spatial alignments, the model successfully integrates all attributes into the portrait image while maintaining a coherent and natural appearance.

Figure 6. **Multi-Attribute Transfer.** Our model supports composition of multiple attributes (*e.g.*, hair, eyeglasses, beard, hat) in a single forward pass without additional training.

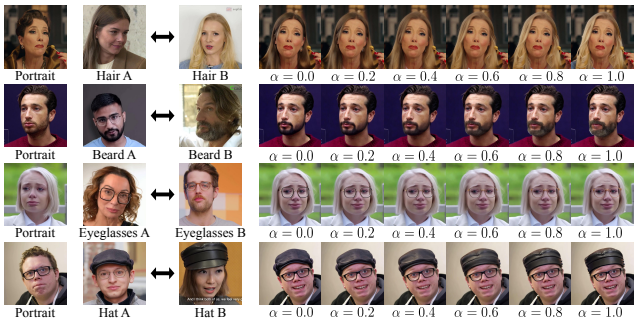

Figure 7. **Attribute Interpolation.** Our method generates zero-shot, single-stage portrait animations with interpolated attributes, even for rigid objects such as hats and eyeglasses. The animations interpolate naturally according to the α values.

Attribute interpolation. Our model enables attribute interpolation by linearly blending the reference features of two attributes, as in Eq. (14). Fig. 7 shows the results with smooth transitions in shape and appearance. The interpolations exhibit smooth changes in visual attributes, demonstrating that our model effectively captures semantically meaningful directions in the attribute feature space.

5. Conclusion

We present Durian, a zero-shot framework for portrait animation with cross-identity attribute transfer, given a portrait image and one or more reference images specifying the target attributes. Our diffusion model, equipped with a Dual ReferenceNet, learns attribute transfer directly from uncurated portrait videos through a self-reconstruction training strategy, eliminating the need for triplet supervision. This is further enhanced by our attribute-aware mask expansion and augmentation scheme. Moreover, Durian naturally extends to multi-attribute composition and attribute interpolation within a single generation pass, without requiring any additional training.

References

[1] R. Abdal, Y. Qin, and P. Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2

[2] B. B. Bilecen, Y. Yalin, N. Yu, and A. Dundar. Reference-based 3d-aware image editing with triplanes. *arXiv preprint arXiv:2404.03632*, 2024. 2, 7, 13

[3] T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[4] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 7, 12

[5] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 7

[6] H. Cha, B. Kim, and H. Joo. Pegasus: Personalized generative 3d avatars with composable attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[7] H. Cha, I. Lee, and H. Joo. Perse: Personalized 3d generative avatars from a single portrait. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3, 6

[8] L. Chen, Q. Mao, Y. Gu, and M. Z. Shou. Edit transfer: Learning image editing via vision in-context relations. *arXiv preprint arXiv:2503.13327*, 2025. 3

[9] X. Chen, L. Huang, Y. Liu, Y. Shen, D. Zhao, and H. Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[10] Z. Chong, X. Dong, H. Li, S. Zhang, W. Zhang, X. Zhang, H. Zhao, D. Jiang, and X. Liang. Catvton: Concatenation is all you need for virtual try-on with diffusion models. *arXiv preprint arXiv:2407.15886*, 2024. 3, 8

[11] X. Chu and T. Harada. Generalizable and animatable gaussian head avatar. *Advances in Neural Information Processing Systems*, 2024. 6

[12] C. Chung, S. Park, J. Kim, and J. Choo. What to preserve and what to transfer: Faithful, identity-preserving diffusion-based hairstyle transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 3, 6, 7, 13, 15

[13] R. Daněček, M. J. Black, and T. Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 6

[14] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 5, 7

[15] Z. Fang, W. Zhai, A. Su, H. Song, K. Zhu, M. Wang, Y. Chen, Z. Liu, Y. Cao, and Z.-J. Zha. Vivid: Video virtual try-on using diffusion models. *arXiv preprint arXiv:2405.11794*, 2024. 3, 7, 12

[16] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 2014. 2

[17] J. Guo, D. Zhang, X. Liu, Z. Zhong, Y. Zhang, P. Wan, and D. Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024. 3, 7, 12, 13

[18] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 4, 12

[19] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 7

[20] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 7, 12

[21] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 2017. 12

[22] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020. 2

[23] L. Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3, 4, 5

[24] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[25] R. Khirodkar, T. Bagautdinov, J. Martinez, S. Zhaoen, A. James, P. Selednik, S. Anderson, and S. Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, 2024. 5, 6, 12

[26] B. Kim, S. Saito, G. Nam, T. Simon, J. Saragih, H. Joo, and J. Li. Haircup: Hair compositional universal prior for 3d gaussian avatars. *arXiv preprint arXiv:2507.19481*, 2025. 3

[27] G. Kim, H. Shim, H. Kim, Y. Choi, J. Kim, and E. Yang. Diffusion video autoencoders: Toward temporally consistent face video editing via disentangled video encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[28] J. Kim, G. Gu, M. Park, S. Park, and J. Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3, 12, 15

[29] T. Kim, B. Kim, S. Saito, and H. Joo. Gala: Generating animatable layered assets from a single scan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[30] T. Kirschstein, S. Qian, S. Giebenhain, T. Walter, and

M. Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics*, 2023. 7, 12

[31] M. Ku, C. Wei, W. Ren, H. Yang, and W. Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024. 3

[32] B. F. Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 6, 14

[33] J. Li, S. Saito, T. Simon, S. Lombardi, H. Li, and J. Saragih. Megane: Morphable eyeglass and avatar network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[34] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics*, 2017. 6

[35] Y. Li, H. Zhou, W. Shang, R. Lin, X. Chen, and B. Ni. Anyfit: Controllable virtual try-on for any combination of attire across any scenario. *arXiv preprint arXiv:2405.18172*, 2024. 3

[36] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015. 4

[37] D. Morelli, M. Fincato, M. Cornia, F. Landi, F. Cesari, and R. Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 12

[38] C. Mou, Y. Wu, W. Wu, Z. Guo, P. Zhang, Y. Cheng, Y. Luo, F. Ding, S. Zhang, X. Li, et al. Dreamo: A unified framework for image customization. *arXiv preprint arXiv:2504.16915*, 2025. 3

[39] H. Nam, D. Kim, J. Oh, and K. M. Lee. Decloth: Decomposable 3d cloth and human body reconstruction from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5636–5645, 2025. 3

[40] G. Parmar, R. Zhang, and J.-Y. Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 7

[41] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 5

[42] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021. 5, 7, 12

[43] D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *Proceedings of the International Conference on Machine Learning*, 2015. 2

[44] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[45] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 4

[46] W. Song, H. Jiang, Z. Yang, R. Quan, and Y. Yang. Insert anything: Image insertion via in-context editing in dit. *arXiv preprint arXiv:2504.15009*, 2025. 3

[47] S. Tu, Z. Xing, X. Han, Z.-Q. Cheng, Q. Dai, C. Luo, and Z. Wu. Stableanimator: High-quality identity-preserving human image animation. *arXiv preprint arXiv:2411.17697*, 2024. 5

[48] Y. Tu, H. Luo, X. Chen, S. Ji, X. Bai, and H. Zhao. Videoanydoor: High-fidelity video object insertion with precise motion control. *arXiv preprint arXiv:2501.01427*, 2025. 3

[49] R. Tzaban, R. Mokady, R. Gal, A. Bermano, and D. Cohen-Or. Stitch it in time: Gan-based facial editing of real videos. In *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 2022. 2

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 4

[51] C. Wang, D. Kang, H. Sun, S. Qian, Z. Wang, L. Bao, and S.-H. Zhang. Mega: Hybrid mesh-gaussian head avatar for high-fidelity rendering and head editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3

[52] Q. Wang, X. Bai, H. Wang, Z. Qin, A. Chen, H. Li, X. Tang, and Y. Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 3

[53] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018. 12

[54] T.-C. Wang, A. Mallya, and M.-Y. Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3

[55] L. Xie, X. Wang, H. Zhang, C. Dong, and Y. Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 4, 7, 12

[56] Y. Xie, H. Xu, G. Song, C. Wang, Y. Shi, and L. Luo. X-portrait: Expressive portrait animation with hierarchical motion attention. In *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 2024. 3, 7, 13

[57] Y. Xu, Z. Shu, C. Smith, S. W. Oh, and J.-B. Huang. In-n-out: Faithful 3d gan inversion with volumetric decomposition for face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[58] B. Yang, S. Gu, B. Zhang, T. Zhang, X. Chen, X. Sun, D. Chen, and F. Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3, 7, 12, 13

[59] S. Yang, H. Li, J. Wu, M. Jing, L. Li, R. Ji, J. Liang, H. Fan, and J. Wang. Megactor-sigma: Unlocking flexible mixed-modal control in portrait animation with diffusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 3, 7, 13

[60] X. Yao, A. Newson, Y. Gousseau, and P. Hellier. A latent transformer for disentangled face editing in images and videos.

In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 2

[61] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3

[62] J. Yu, H. Zhu, L. Jiang, C. C. Loy, W. Cai, and W. Wu. Celebv-text: A large-scale facial text-video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 4, 7, 12

[63] K. Zhang, Y. Zhou, X. Xu, B. Dai, and X. Pan. Diffmorpher: Unleashing the capability of diffusion models for image morphing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 6

[64] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 5

[65] Y. Zhang, L. Wei, Q. Zhang, Y. Song, J. Liu, H. Li, X. Tang, Y. Hu, and H. Zhao. Stable-makeup: When real-world makeup transfer meets diffusion model. *arXiv preprint arXiv:2403.07764*, 2024. 3

[66] Y. Zhang, Q. Zhang, Y. Song, J. Zhang, H. Tang, and J. Liu. Stable-hair: Real-world hair transfer via diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 2, 3, 7, 13, 15

[67] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*, 2016. 2

[68] S. Zhu, J. L. Chen, Z. Dai, Z. Dong, Y. Xu, X. Cao, Y. Yao, H. Zhu, and S. Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, 2024. 3, 4, 5, 12