

HairShifter: Consistent and High-Fidelity Video Hair Transfer via Anchor-Guided Animation

Wangzheng Shi*

Xiamen University

School of Informatics

Xiamen, China

shiwangzheng@stu.xmu.edu.cn

Yinglin Zheng*

Xiamen University

School of Informatics

Xiamen, China

zhengyinglin@stu.xmu.edu.cn

Yuxin Lin

Xiamen University

School of Informatics

Xiamen, China

linyx@stu.xmu.edu.cn

Jianmin Bao

Microsoft Research Asia

Beijing, China

jianbao@microsoft.com

Ming Zeng[†]

Xiamen University

School of Informatics

Xiamen, China

zengming@xmu.edu.cn

Dong Chen

Microsoft Research Asia

Beijing, China

doch@microsoft.com



Figure 1: High-Fidelity Video Hairstyle Transfer with HairShifter. We successfully transfers the reference hairstyle while precisely preserving the subject’s identity, expression, and background details across dynamic motion.

ABSTRACT

Hair transfer is increasingly valuable across domains such as social media, gaming, advertising, and entertainment. While significant progress has been made in single-image hair transfer, video-based hair transfer remains challenging due to the need for temporal consistency, spatial fidelity, and dynamic adaptability. In this work, we propose HairShifter, a novel "Anchor Frame + Animation" framework that unifies high-quality image hair transfer with smooth and coherent video animation. At its core, HairShifter integrates a Image Hair Transfer (IHT) module for precise per-frame transformation and a Multi-Scale Gated SPADE Decoder to ensure seamless spatial blending and temporal coherence. Our method maintains hairstyle fidelity across frames while preserving non-hair regions. Extensive experiments demonstrate that HairShifter achieves state-of-the-art performance in video hairstyle transfer, combining superior visual quality, temporal consistency, and scalability. **The code will be publicly available.** We believe this work will open new avenues for video-based hairstyle transfer and establish a robust baseline in this field.

CCS CONCEPTS

- Computing methodologies → Machine learning; Computer vision.

*Equal contribution.

[†]Corresponding authors.

KEYWORDS

Video Hair Transfer, Image Animation, Video Editing

1 INTRODUCTION

Hair Transfer (also called Hair Replacement) is an important task in the area of Intelligent Multimedia Processing. It has numerous applications in industries such as social media, gaming, advertising, film, and television.

Currently, a serial of works on single-image hair transfer[27, 29, 44, 45, 56, 59, 63] achieve satisfactory results. For example, the methods based on Generative Adversarial Networks (GANs)[22, 25–27, 29, 38, 44, 45, 63], and the methods based on Latent Diffusion Models(LDMs)[17, 33, 50, 56, 59, 64]. Despite these progress, extending these successes to video remains a significant challenge. We analyze potential approaches for video hair transfer and their limitations.

A basic approach for video hair transfer involves applying single-image methods frame by frame. However, this naive frame-wise transfer lacks temporal coherence, resulting in visible flickering, structural jittering, and instability in non-hair areas. Alternatively, while modern text-to-video (T2V) generation methods [3, 5, 10, 23, 31, 46] can model temporal dependencies, their open-domain nature and limited fine-grained control hinder precise hair transfer and identity preservation. Furthermore, their applicability is often restricted by sequence length limitations.

Method	Temporal Consistency	Non-Hair Region Fidelity	Hair Transfer Quality	Frame Scalability
Image Hair Transfer[29, 56, 59]	✗	✓	✓	✓
T2V Generation[3, 10, 31, 46]	✓	✗	✗	✗
Video Editing[1, 4, 53]	✓	✗	✗	✗
Portrait Animation[12, 43, 48, 51]	✓	✗	✗	✓
HairShifter (Ours)	✓	✓	✓	✓

Table 1: Comparison of Video Hair Transfer Paradigms. Only our method satisfies all core requirements.

Video editing models [1, 4, 53] offer some promise for preserving background regions by using masks for inpainting. However, their effectiveness is limited by the practical challenge of obtaining accurate auxiliary masks. Furthermore, portrait animation techniques [12, 43, 48, 51], while capable of generating temporally smooth sequences through motion transfer, inherently deform the entire image. This global deformation makes them ill-suited for the selective replacement required for hair transfer, as it inevitably leads to unwanted modifications in non-hair areas.

We summarize these methods and their properties in Table 1, revealing none of these existing paradigms fully satisfy the core requirements for ideal video hair transfer: *Temporal Consistency*, *Non-Hair Region Fidelity*, *Hair Transfer Quality*, and *Frame Scalability*.

To address these challenges, we propose an "**Anchor Frame + Animation**" pipeline. This approach **combines high-fidelity image hair transfer with animation's temporal smoothness** in two key stages. First, we create a **high-quality "anchor frame"** using image hair transfer. This static image embodies the desired hairstyle and the identity of the driving video subject. The anchor frame then serves as the consistent hair source. Second, we **animate this anchor frame using motion cues** from the driving video. This process ensures temporally coherent hair transfer, dynamically applying the hairstyle while preserving identity and expressions throughout the video.

To enhance anchor frame quality, we utilize an **Image Hair Transfer (IHT) module**. IHT is based on the state-of-the-art image hair transfer model (**Stable-Hair** [59]) but with better robustness to pose variations, thereby generating high-quality hair-swapped anchor frame.

Even with high-quality anchor frames, the '**Anchor + Animation**' strategy faces challenges in preserving non-hair regions and achieving harmonious blending. To tackle this, we propose a **de-coupled training strategy and a Multi-Scale Gated SPADE Decoder**. The decoupled training strategy uses IHT to generate pseudo driving frames, compelling the network to **learn hair appearance from the source frame** and **non-hair features from the pseudo driving frame**. The MSG-SPADE Decoder employs a dual-pathway architecture and multi-scale gated fusion to achieve seamless blending and enhance spatio-temporal consistency.

Our framework, named *HairShifter*, balances high-quality hair generation, dynamic adaptability, temporal continuity, non-hair region preservation, and seamless blending. It is efficient, has no length limitations, and achieves state-of-the-art performance, as demonstrated in Figure 1 and our experiments.

In summary, our contributions are outlined as follows:

- To the best of our knowledge, we propose the first framework specifically designed for video hairstyle transfer, effectively balancing generation quality, temporal consistency,

and preservation/blending fidelity, achieving state-of-the-art results.

- We design a novel training strategy that repurposes an image hair transfer module to decouple appearance and motion, resolving the core conflict between maintaining temporal consistency and preserving non-hair region details.
- We develop the **MSG-SPADE Decoder**, whose **dual-path**, reference-guided architecture with gated fusion enhances identity retention and achieves seamless spatial integration for temporally stable video synthesis.

2 RELATED WORK

Hair Transfer. Early GAN-based hair transfer focused on editable synthesis via latent space manipulation, such as attribute decomposition[38], preservation[63], and style-space modification[26, 27, 44, 47]. Subsequent works enhanced control and efficiency, with HairCLIPV2[45] enabling text-driven editing and HairFastGAN[29] achieving fast, pose-robust generation. Recently, diffusion models[7, 17, 33], often guided by spatial constraints like ControlNet[57], have gained prominence. Methods like Stable-Hair and HairDiffusion[56] improved attribute alignment and fine-grained control. However, these techniques are fundamentally designed for static images and lack the temporal modeling.

Portrait Animation. Existing portrait animation methods animate source images using motion priors, such as keypoints or flow fields[8, 14, 18, 19, 28, 36, 43, 55, 60]. While advanced techniques have achieved photorealistic rendering[41], enhanced spatiotemporal coherence[51], performed cross-identity synthesis[48], and captured subtle facial dynamics[12], they typically operate holistically on the entire image. This lack of part-specific control renders them unsuitable for tasks like hairstyle transfer that demand region-aware identity preservation.

Video Editing. Text-to-video (T2V) generation[3, 10, 13, 31, 46, 49, 54] and editing frameworks[5, 11, 23, 52, 58] struggle with fine-grained spatial control, often hampered by a reliance on accurate masks. Even advanced methods that enhance spatiotemporal locality, such as UniEdit[1], StableVideo[4], and VideoGrain[53], do not offer robust mechanisms for targeted manipulation. This architectural deficiency limits their utility for applications like video hairstyle transfer, which mandate precise, spatially-aware and identity-preserving editing.

3 METHODOLOGY

3.1 Framework Overview

We introduce *HairShifter*, a novel framework conceived for high-fidelity, temporally consistent video hair transfer while meticulously preserving non-hair attributes. At its core, HairShifter employs an "**Anchor Frame + Animation**" pipeline.

The inference process, illustrated in Figure 2(b), involves two main stages. First, a high-quality static anchor image I_s containing the target hairstyle is generated using an **Image Hair Transfer (IHT) module**. Second, our **animation network G** synthesizes the output video frame-by-frame by animating this anchor I_s , guided by the motion and non-hair context derived from the driving video $V_d = \{I_d^t\}, t \in \{1, \dots, T\}$.



To enable this targeted animation during inference, the network G is trained using the specific strategy depicted in Figure 2(a). This training regime addresses the core challenges: (1) **Forced Decoupling** is achieved via a novel **training strategy** using **pseudo driving frames** (Sec 3.2), forcing the network to **separate hair appearance** (from source I_s) **from non-hair dynamics** (from driving frame I_d^t). (2) **Information Preservation** relies on **Disentangled Feature Encoding** (Sec 3.3) to separately capture hair features and essential non-hair context from the driving frame. (3) **Seamless Blending** is handled by our innovative **Multi-Scale Gated Fusion** (Sec 3.4), which intelligently fuses these decoupled feature streams at multiple scales for artifact-free synthesis.

The interplay between the high-quality anchor generated by the IHT module (used for anchor generation at inference and pseudo-frame generation during training) and the animation network G , trained with these specialized decoupling, preservation, and blending mechanisms, allows HairShifter to effectively navigate the complexities of video hair transfer. Subsequent sections detail these components and strategies.

3.2 Pseudo Driving Frame Generation

A core challenge in designing the animation network G is enabling **controllable video hair transfer during inference**. Specifically, the network must learn to **synthesize an output frame** that **combines the desired hair appearance** from a static source image I_s with the **non-hair dynamics** (pose, expression, identity, background) derived from a driving frame I_d . Standard image animation training protocols, which typically train the network to **reconstruct I_d using I_s and I_d as inputs** (often where I_s is derived from I_d or a related frame), are unsuitable for this task. In such setups, the driving frame I_d inherently contains the correct target hair information, providing a "shortcut" that prevents the network from learning to rely exclusively on I_s for hair appearance and I_d solely for non-hair dynamics.

To enforce this crucial decoupling, we introduce a novel training strategy centered around the generation of a *pseudo driving frame*, denoted as I_d' . This I_d' is constructed by taking the original **driving frame I_d** and **replacing its hair** with an **arbitrary, incorrect hairstyle**, represented by a **random reference image R_{random}** . Generating this I_d' requires a capable image hair transfer mechanism. Crucially, this mechanism must be able to: (1) generate high-fidelity, realistic hairstyles, (2) accurately transfer the style from R_{random} onto the identity and pose present in I_d , and (3) be robust to potential variations in pose between I_d and R_{random} , a common scenario.

For this purpose, we utilize a high-fidelity **Image Hair Transfer(IHT)** module, adapted from **existing state-of-the-art diffusion-based methods [59]** and fine-tuned on video data to enhance **robustness to pose variations** (details of IHT can be found in the supplementary material). Using this module, we generate the pseudo driving frame as $I_d' = \text{IHT}(I_d, R_{random})$. Thus, I_d' retains the **essential non-hair dynamics** (pose, expression, identity, background) of the original driving frame I_d but **presents deliberately incorrect hair information**.

The animation network G is then trained to reconstruct the original ground truth frame I_d using the source image I_s (which contains

the *correct* target hair) and the pseudo driving frame I_d' as inputs: $I_p = G(I_s, I_d') \approx I_d$. This training objective fundamentally changes the learning dynamics. Because the hair information in I_d' is explicitly incorrect, the network is forced to disregard it and learn to extract the **target hair appearance** solely from the source image I_s . Simultaneously, since I_d' contains the **correct non-hair information** (pose, expression, identity alignment, background), the network learns to **extract these dynamic elements** from the driving input. This constraint effectively compels G to master the essential decoupling required for controllable hair transfer: Hair Appearance from I_s , Non-Hair Dynamics from the driving input (I_d' during training, I_d during inference).

It is worth noting that the same robust image hair transfer module used here to generate I_d' during training is also employed during the inference stage to create the high-quality static anchor frame I_s that provides the consistent target hair appearance throughout the generated video.

3.3 Disentangled Feature Encoding

Our animation network G leverages the robust motion estimation and warping capabilities of the **LivePortrait [12] framework**. Specifically, we utilize its core components: the **Motion Estimator (E_M)**, adapt its appearance encoder as our **Hair Feature Encoder (E_H)**, and employ its **motion-driven feature Warping module (W)**.

The motion estimator E_M extracts **motion representations** (capturing pose, expression, etc.) from both the source image I_s and the corresponding driving frame I_d (representing a **target time step**). Concurrently, the Hair Feature Encoder E_H processes the source image I_s to **extract features f_h** representing its **hair appearance**. The warping module W then utilizes the estimated motion difference between the source and driving frame to dynamically transform the hair features f_h , producing warped hair features $f_w = W(E_M(I_s), E_M(I_d), E_H(I_s))$. These features f_w effectively carry the target hair appearance information, spatially aligned with the posture and expression depicted in the driving frame.

While f_w provides the hair information, accurately reconstructing the non-hair regions (identity, background, etc.) requires context directly from the driving frame. To ensure meticulous **Information Preservation** (a core design goal from Sec 3.1), we introduce a dedicated **Non-hair Context Encoder E_C** .

Structurally identical to E_H , the encoder E_C processes the driving frame I_d to **extract context features f_c** . These features f_c capture the **essential spatial layout, identity cues, and structural detail** pertaining specifically to the non-hair regions of the target frame I_d . This provides the necessary reference for reconstructing everything *except* the hair according to the driving video's content.

Thus, we obtain two primary disentangled feature streams: f_w encoding the motion-aligned hair appearance derived from I_s , and f_c encoding the non-hair context derived from I_d . Alongside these, the **binary hair mask m_c** of the driving frame I_d (derived using a face parsing network[61]), is also utilized as guidance(*hair mask guidance*). This mask provides explicit spatial guidance regarding the hair region location in the driving frame. All three inputs – the warped hair features f_w , the non-hair context features f_c , and the hair mask m_c – are then fed into our proposed **Multi-Scale Gated SPADE Decoder D** (detailed in Section 3.4). This decoder, replacing



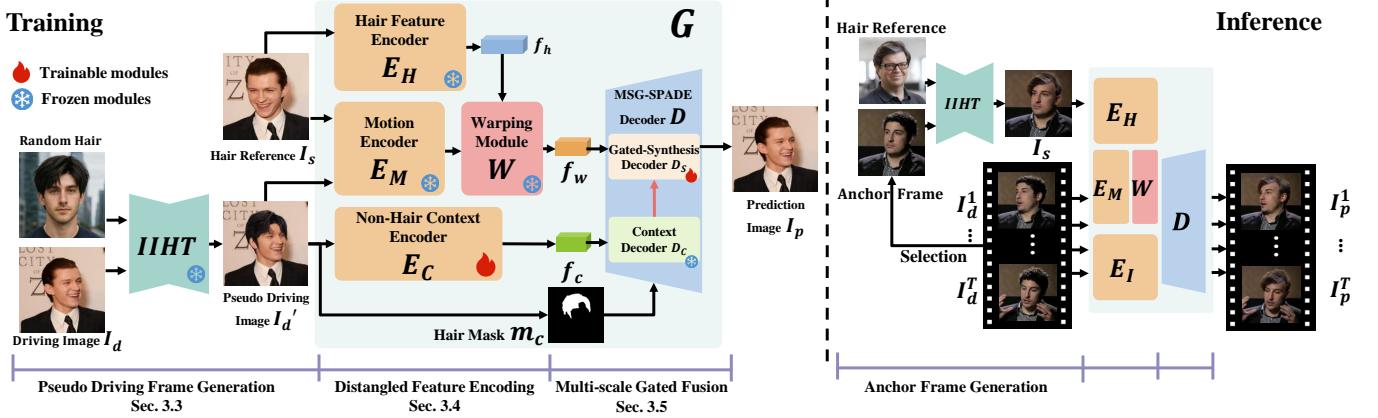


Figure 2: Overview of the HairShifter (a) During training, we enforce hair/non-hair decoupling using Pseudo Driving Frame Generation and Disentangled Feature Encoding, with the Multi-Scale Gated Fusion handling seamless feature fusion for synthesis. (b) During inference, the Image Hair Transfer(IHT) creates the static anchor frame I_s , which is subsequently animated by the animation network G guided by the driving video.

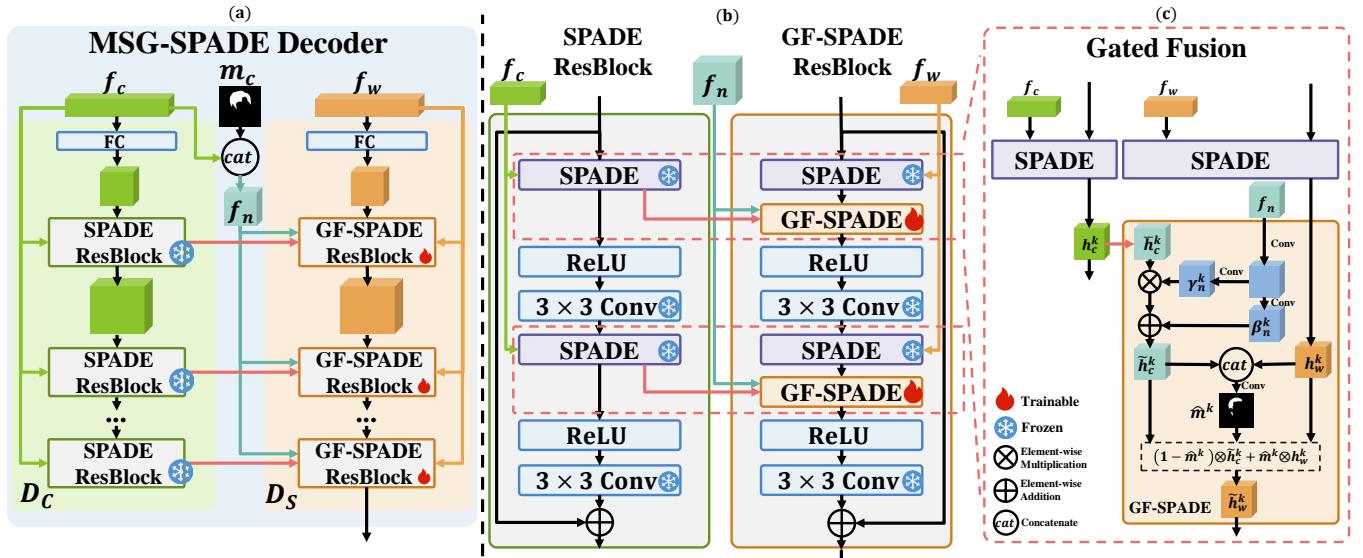


Figure 3: The Multi-Scale Gated SPADE (MSG-SPADE) Decoder. (a) Features dual pathways for synthesis (D_S , hair f_w) and context (D_C , non-hair f_c). (c) The key Gated Fusion (GF-SPADE) blocks within D_S employ hair mask(m_c) guided modulation and learned gating to fuse features across multiple scales, achieving seamless hair integration while preserving context details.

the standard generator, leverages all inputs to synthesize the final prediction: $I_p = D(f_w, f_c, m_c)$.

3.4 Multi-Scale Gated SPADE Decoder

A critical challenge in video hair transfer is seamlessly integrating the synthesized target hairstyle while meticulously preserving the identity, expression, and background details present in the driving video's non-hair regions. To address this, we propose the **Multi-Scale Gated SPADE(MSG-SPADE) Decoder**, a novel architecture designed to replace standard SPADE-based decoders[30] commonly found in animation frameworks like LivePortrait[12].

As illustrated in Figure 3(a), the MSG-SPADE Decoder employs a dual-pathway structure to process distinct feature streams. The

Gated Synthesis Decoder (D_S) primarily operates on the warped feature volume f_w , which encodes the target hair appearance dynamically aligned with the driving video's motion. Concurrently, the **Context Decoder (D_C)** processes the context feature volume f_c , derived from the driving frame (I_d' during training, I_d during inference), capturing essential non-hair information such as identity cues, pose, and structural details.

Both pathways are fundamentally constructed using cascaded Residual Blocks, detailed in Figure 3(b), incorporating spatially-adaptive normalization via standard SPADE blocks. The context pathway D_C functions as a standard SPADE decoder, directly utilizing these blocks conditioned on f_c . The synthesis pathway D_S also employs standard SPADE blocks, conditioned on f_w , but introduces

our core innovation: following each standard SPADE block within D_S , a **Gated Fusion SPADE (GF-SPADE)** block is inserted. This GF-SPADE block is specifically designed to **fuse the output features** from the corresponding standard SPADE block in the context pathway (D_C) with the features generated by the preceding standard SPADE block in the synthesis pathway (D_S), thereby selectively integrating contextual information at multiple scales.

The core innovation lies in the multi-scale gated fusion mechanism integrated within the GF-SPADE blocks of the synthesis pathway D_S . As shown in Figure 3(c), the **output activation** for the **k -th SPADE block** in the context pathway and the synthesis pathway are denoted as \tilde{h}_c^k and \tilde{h}_w^k respectively. Before fusion, the context activation h_c^k is further modulated with the guidance of the context features f_c and the hair mask m_c of the driving frame (both concatenated into f_n), yielding \tilde{h}_c^k :

$$\tilde{h}_c^k = \gamma_n^k(f_n) \otimes \tilde{h}_c^k + \beta_n^k(f_n), \quad f_n = \text{Concat}(f_c, m_c) \quad (1)$$

where γ_n^k and β_n^k are modulation parameters derived from f_n , and \otimes denotes element-wise multiplication. This modulation step helps align the context information precisely before it's blended with the synthesis activation, and incorporating the **hair mask guidance** (m_c) provides explicit spatial localization, help improve blending accuracy, particularly at the hair boundaries.

Next, a **spatial gating mask** \hat{m}^k is computed at each block k . This gate adaptively determines the contribution of each pathway at every spatial location. It is derived from the concatenation of the modulated context activation \tilde{h}_c^k and the synthesis activation \tilde{h}_w^k :

$$\hat{m}^k = \sigma(\text{Conv}(\text{Concat}(\tilde{h}_w^k, \tilde{h}_c^k))) \quad (2)$$

where Conv represents a convolutional layer and σ is the sigmoid activation function, constraining the gate values to the range $[0, 1]$.

Finally, the modulated **context activation** \tilde{h}_c^k and the **synthesis activation** \tilde{h}_w^k are fused using the learned gate \hat{m}^k to produce the output activation \tilde{h}_w^k for the k -th GF-SPADE block:

$$\tilde{h}_w^k = (1 - \hat{m}^k) \otimes \tilde{h}_c^k + \hat{m}^k \otimes h_w^k \quad (3)$$

This output \tilde{h}_w^k then proceeds through the remainder of the GF-SPADE ResBlock (e.g., another convolution, ReLU).

By performing this **gated fusion** at multiple scales **throughout** the decoder, the MSG-SPADE architecture effectively integrates the target hairstyle features (via f_w and h_w^k) with the **identity-preserving** structural details from the driving frame's **context** (via f_c and h_c^k). The learned gating mechanism ensures that non-hair regions are faithfully reconstructed based on the context pathway, while the hair region smoothly incorporates the synthesized appearance. This approach significantly mitigates structural drift and identity inconsistencies often observed in video editing tasks, leading to spatially coherent synthesis with sharp hair boundaries and well-preserved facial and background details.

3.5 Training Objectives and Inference Pipeline

3.5.1 Optimization Objectives. We optimize the animation network G using a composite loss function. Following standard practices in portrait animation and generative modeling, similar to frameworks like LivePortrait [12], our objective includes conventional loss terms. These typically comprise an adversarial loss (\mathcal{L}_{adv}) to encourage photorealistic outputs, a perceptual loss (\mathcal{L}_p) utilizing features

from a pre-trained VGG network [37] to enforce high-level feature similarity, and a global L1 loss (\mathcal{L}_{rec}) promoting overall **structural fidelity** between the generated frame I_p and the **ground truth frame** I_d .

To further enhance the preservation and detail within specific regions of interest, we incorporate **localized L1 losses** that specifically **target the hair and non-hair face regions**. These losses leverage binary masks, m_{hair} for the hair region and m_{face} for the non-hair face region (excluding hair), which are derived from ground truth facial parsing annotations [61]. By focusing optimization on these critical areas, we aim to achieve more precise results. The localized losses \mathcal{L}_{hair} and \mathcal{L}_{face} are defined as:

$$\mathcal{L}_{hair} = \|m_{hair} \otimes (I_d - I_p)\|_1, \quad \mathcal{L}_{face} = \|m_{face} \otimes (I_d - I_p)\|_1. \quad (4)$$

Here, $\|\cdot\|_1$ signifies the L1 norm, ensuring that errors within these masked areas are directly minimized.

The final objective function guiding the training is a weighted summation of both the standard and these localized refinement loss components:

$$\mathcal{L}_{total} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_p\mathcal{L}_p + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{hair}\mathcal{L}_{hair} + \lambda_{face}\mathcal{L}_{face} \quad (5)$$

where the λ coefficients ($\lambda_{adv}, \lambda_p, \lambda_{rec}, \lambda_{hair}, \lambda_{face}$) serve as **hyperparameters** to balance the relative influence of each loss term during optimization.

3.5.2 Inference Pipeline. During inference, given a driving video sequence composed of frames I_d^t where $t \in \{1, \dots, T\}$, and a **target hairstyle** reference image R_{target} , the goal is to generate a **corresponding output video** sequence frame by frame, denoted as I_p^t . This process involves two main stages, leveraging the pre-trained IHT module and the disentanglement-trained network G . First, a **static source image** I_s is prepared. An **anchor frame** I_{anchor} is selected¹ from V_d , and the IHT module **synthesizes** $I_s = \text{IHT}(I_{anchor}, R_{target})$. This image I_s provides the **target hair appearance** while retaining the **identity from I_{anchor}** . Second, the network G performs frame-by-frame animation. For each driving frame I_d^t , G synthesizes the output frame I_p^t . It uses I_s as the **source for hair appearance** and I_d^t to provide the **motion cues** and the **non-hair context** (identity, pose, background). Inside, G generates motion-aligned hair features derived from I_s and context features capturing the non-hair regions of I_d^t . The core synthesis relies on the Multi-Scale Gated SPADE Decoder D , which effectively fuses these two sets of features using its learned gating mechanism (Eq. 3).

This inference strategy capitalizes on the disentanglement learned during training. By processing the static source I_s and the dynamic driving frames I_d^t through complementary pathways within G and fusing them intelligently via the MSG-SPADE Decoder, the framework **generates a temporally coherent video** with the desired hairstyle, while robustly preserving the subject's identity and non-hair region details.

¹In our practice, we automatically select the source frame with the facial pose most similar to the reference hair image.

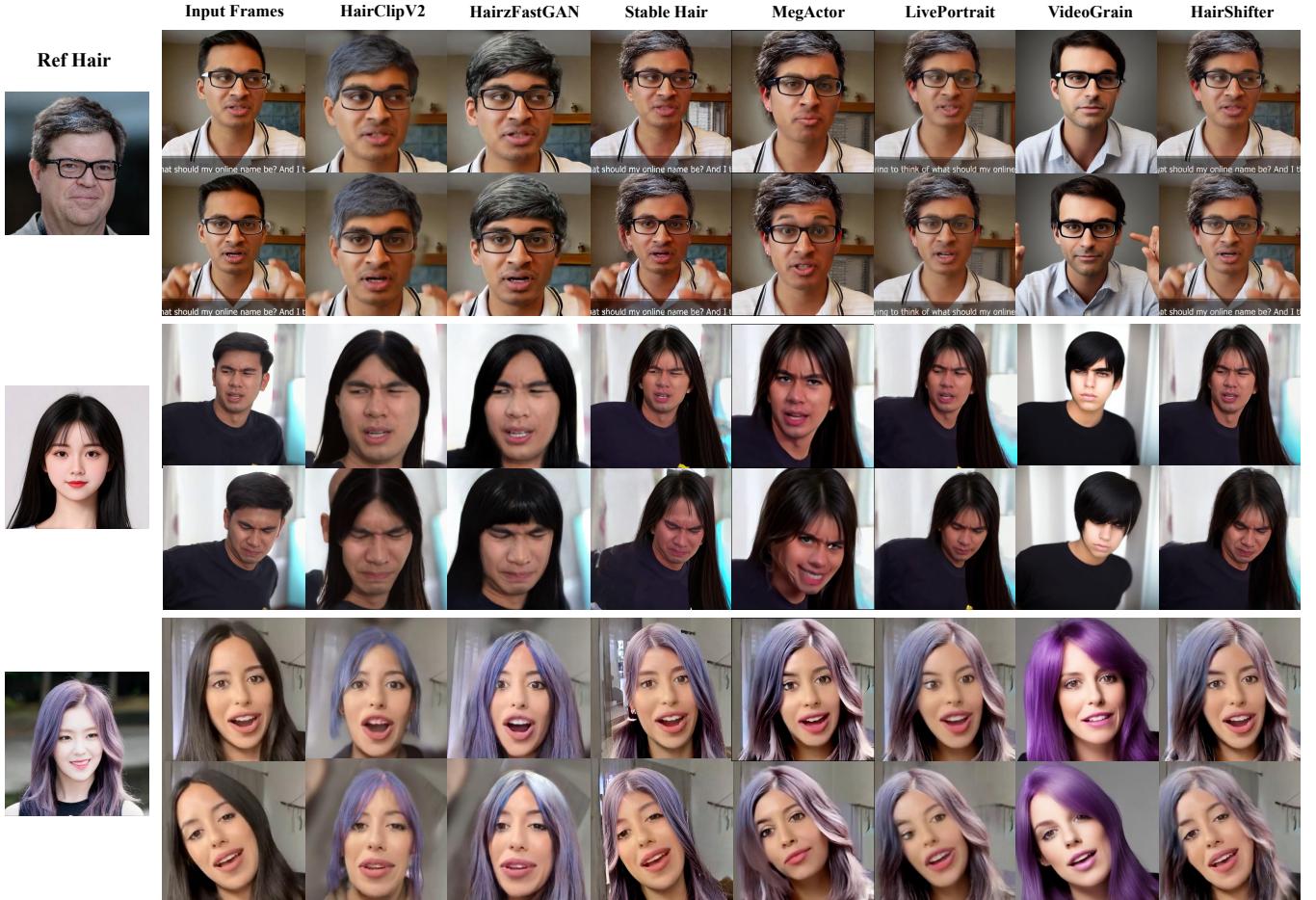


Figure 4: Qualitative Comparison with other methods. HairShifter achieves more refined and stable hairstyle transfer while consistently preserving the details of irrelevant regions between input frames

Method	Video Quality			Non-Hair Region Fidelity			Temporal Consistency			
	FID-VID \downarrow	FVD \downarrow	PSNR \uparrow	SSIM \uparrow	L1 \downarrow	IDS \uparrow	Vbench-TF (%) \uparrow	Vbench-BC (%) \uparrow	Vbench-MS (%) \uparrow	CLIP-score \uparrow
HairClipV2[45]	62.826	1097.600	33.745	0.953	0.028	0.735	96.431	94.729	96.979	97.582
HairFastGAN[29]	29.819	409.640	32.166	0.952	0.030	0.718	97.732	95.714	98.196	98.889
Stable-Hair[59]	90.099	1161.048	31.781	0.909	0.031	0.690	94.658	92.321	95.565	95.338
Megactor[51]	36.993	367.061	30.918	0.766	0.071	0.504	98.450	95.178	99.129	99.117
LivePortrait[12]	<u>21.007</u>	<u>242.575</u>	31.062	0.821	0.055	0.568	99.191	96.235	<u>99.406</u>	<u>99.637</u>
VideoGrain[53]	48.286	566.635	30.076	0.824	0.071	0.363	98.506	<u>96.481</u>	98.989	98.946
HairShifter	11.686	140.261	<u>32.577</u>	0.959	0.025	0.778	<u>99.032</u>	96.501	99.457	99.717

Table 2: Quantitative Comparison of Baselines. Bold and underlined scores indicate the 1st- and 2nd-ranked methods.

4 EXPERIMENTS

4.1 Implementation Detail

We employ the CelebV-HQ[62] for model training. Following our decoupled strategy, we use anchor frames I_s and IHT-generated pseudo driving frames I_d' to reconstruct the target frame I_d . We adapt the LivePortrait [12] framework by loading and freezing its pre-trained weights for the motion components (E_M, E_H, W), the Context Decoder D_C , and the standard SPADE blocks within the

Synthesis Decoder D_S . Training exclusively focuses on the newly introduced Non-hair Context Encoder E_C and the GF-SPADE blocks integrated into D_S . This targeted training runs for 100 epochs on three RTX 3090 GPUs with a batch size of 4 and a learning rate of 2.0×10^{-5} . In all of our experiments, the loss weights were set as follows: $\lambda_{adv} = 1, \lambda_p = 1, \lambda_{rec} = 1, \lambda_{hair} = 1$, and $\lambda_{face} = 1$. In practice, these weights achieve a well-balanced trade-off across the multiple objectives.



4.2 Evaluation Protocol

4.2.1 Baselines. We compare our method against several state-of-the-art video generation and editing frameworks, grouped by their underlying paradigm. First, we evaluate against single-image hairstyle transfer models applied frame-by-frame, including **HairClipV2**[45], **HairFastGAN**[29], and **Stable-Hair**[59]. Second, we include portrait animation approaches such as **LivePortrait**[12] and **MegActor**[51]. For fair comparison, these animation models utilize the same high-quality anchor frame generated by our IHT module as input to drive the video synthesis. Finally, we benchmark against recent diffusion-based video editing models **VideoGrain**[53]. For the text-driven VideoGrain [53], prompts were generated from reference images using *Google Gemini 2.0 Flash*[39], and the editing region was set to the entire input image. Note that text inherently provides less precise visual control than image references, which may affect fine-grained style replication.

4.2.2 Metrics. We evaluate the generated videos using a diverse set of metrics assessing three core aspects: video quality and fidelity, non-hair region preservation, and temporal consistency. For overall video quality and fidelity, we compute **FID-VID**[16] to measure the Fréchet distance between Inception feature distributions of generated and source frames, indicating visual quality. We also use **FVD**[40], which leverages I3D features[2] to assess both appearance and temporal dynamics. To evaluate non-hair region fidelity and identity preservation, we calculate several metrics specifically on the non-hair areas identified using a face parsing network[61]. These include **PSNR**[35] and **SSIM**[42] for structural and pixel-level similarity, and **L1 Loss** for average pixel difference. Additionally, we employ ArcFace[6] to compute identity similarity (**IDS**) based on facial embeddings, quantifying identity preservation. For temporal consistency, we adopt relevant metrics from the VBench[21] benchmark: **Temporal Flicker (TF)**, **Background Consistency (BC)**, and **Motion Smoothness (MS)**, where higher percentages indicate better performance for all three. Furthermore, following previous work[46], we compute the average **CLIP-score**[32] measuring cosine similarity between CLIP embeddings of consecutive frames, where higher scores indicate better frame-to-frame visual coherence.

4.3 Comparison with Baselines

Qualitative Comparison. Figure 4 provides a qualitative comparison using test set videos and social media portraits with distinctive hairstyles as references. Frame-wise methods (HairCLIPV2, HairFastGAN, Stable-Hair) suffer from severe flickering due to poor spatio-temporal consistency. HairCLIPV2 and HairFastGAN also exhibit background inconsistencies due to the face alignment pre-processing required by their official implementations. While Stable-Hair achieves good single-frame quality, its inability to consistently adapt the hairstyle under pose variation (e.g., rows 4, 8) highlights the difficulty of achieving dynamic adaptability with static models. Portrait animation methods (MegActor, LivePortrait), despite starting from high-quality IHT anchors, struggle with precise non-hair region preservation: both exhibit progressive identity drift and background inconsistencies (e.g., hand motion, row 1), revealing the limitation of holistic deformation for targeted replacement tasks, as discussed in the Introduction. VideoGrain performs poorly in

both non-hair region preservation (altering identity and background details) and achieving fine-grained, reference-driven high-quality hair transfer. In contrast, our HairShifter achieves strong dynamic adaptability, robust spatio-temporal consistency, and accurate non-hair region preservation, even under complex motion and occlusion. These results demonstrate the effectiveness of our "Anchor Frame + Animation" approach, particularly the synergy between the IHT anchor generation and the MSG-SPADE Decoder's ability to seamlessly integrate the hair while preserving context.

Quantitative Comparison. We perform quantitative evaluation on 200 videos from the CelebV-HQ dataset [62] (100 identities, 100 frames each, outside the training set), using random reference hairs from CelebA-HQ [20]. Table 2 presents the results. HairShifter achieves state-of-the-art performance across the board, validating its effectiveness in balancing the conflicting demands of video hair transfer. In terms of overall video quality, our method obtains the best FID-VID and FVD scores, indicating superior visual realism and addressing the challenge of high-quality generation. For non-hair region fidelity, HairShifter leads significantly in SSIM, L1 Loss, and IDS, while also achieving a highly competitive PSNR. This quantitatively confirms its success in precise non-hair region preservation, directly validating the effectiveness of our decoupled training strategy and the MSG-SPADE Decoder's reference-guided fusion in maintaining identity and background details. Regarding temporal consistency, our method achieves the highest scores for VBench BC, VBench MS, and CLIP-score, along with a near-best VBench TF. This demonstrates its capability to maintain spatio-temporal consistency, mitigating flickering and ensuring smooth transitions, thereby validating the robustness of our "Anchor Frame + Animation" approach. These quantitative results strongly support our qualitative observations and the claims made regarding HairShifter's advancements.

Method	Accuracy	Preservation	Naturalness	Temporal Consistency
HairClipV2	1.340	1.769	1.184	1.211
HairFastGAN	1.646	2.463	1.490	1.803
Stable-Hair	2.313	2.680	1.680	1.551
Megactor	3.769	3.279	3.605	3.864
LivePortrait	4.136	3.442	3.864	4.224
VideoGrain	1.490	1.224	1.714	2.109
HairShifter	4.361	4.599	4.333	4.463

Table 3: User Study on Video Hair Transfer. HairShifter outperforms all others across these four metrics.

User Study. To incorporate human perceptual judgment, we conducted a user study with 20 participants evaluating 30 video triplets (original, reference, generated) across 7 methods. Participants rated results based on four criteria directly related to our core challenges: **Accuracy** (quality of hair transfer), **Preservation** (fidelity of non-hair regions), **Naturalness** (overall realism), and **Temporal Consistency** (smoothness). As shown in Table 3, HairShifter was overwhelmingly preferred across all criteria, significantly outperforming competing methods. This confirms that our technical contributions translate into perceptually superior results, successfully balancing generation quality, preservation, and temporal coherence in a way users find most compelling.

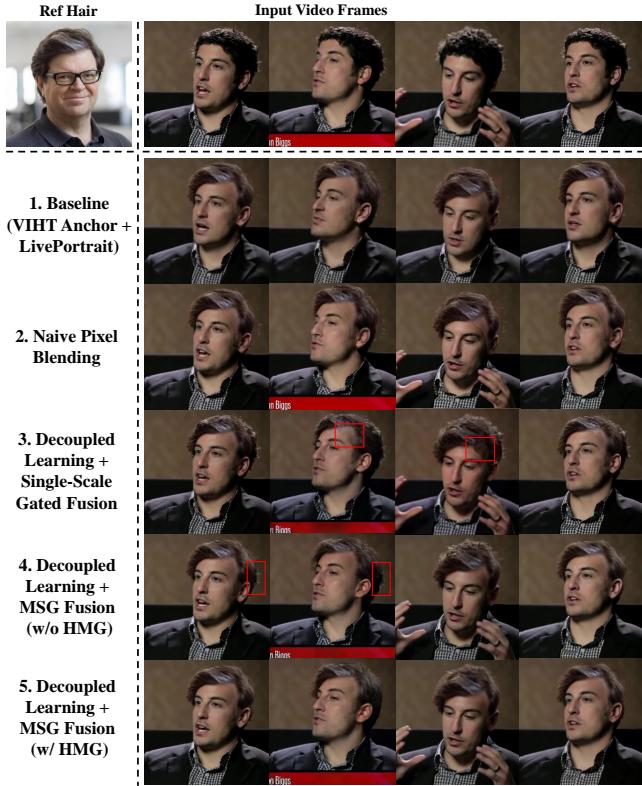


Figure 5: Visual comparison for the Ablation Study. Comparing settings from Baseline (top) to HairShifter (bottom), showcasing progressive gains in blending quality and non-hair region preservation.

#	Gated Fusion	Pixel Blending	FID-VID↓	FVD↓	SSIM↑	IDS↑	Vbench-MS↑
1	N/A	✗	21.007	242.575	0.821	0.568	99.406
2	N/A	✓	15.011	183.667	0.957	0.740	99.023
3	Single-scale	✓	14.801	146.743	0.934	0.769	99.131
4	Multi-scale w/o HMG	✓	12.299	140.534	0.940	0.770	99.337
5	Multi-scale w/ HMG	✓	11.686	140.261	0.959	0.778	99.457

Table 4: Quantitative Ablation Study. Results demonstrate the progressive contribution of each component, with HairShifter(Setting 5) achieving the best overall performance.

4.4 Ablation Study

To validate the effectiveness of our key contributions – the decoupled training strategy and the Multi-Scale Gated SPADE (MSG-SPADE) Decoder with Hair Mask Guidance (HMG) – we conduct an ablation study. We evaluate five settings, progressively incorporating our proposed components. The **Baseline (Setting 1)** employs the standard LivePortrait [12] framework driven by our IHT anchor. **Setting 2 (+ Naive Pixel Blending)** then applies post-processing: it composites the synthesized hair from the Setting 1’s predicted frame onto the original driving frame, using a hair mask softened at the edges with Gaussian blur. **Setting 3 (+ Decoupled Learning & Single-Scale Fusion)** then introduces our decoupled training strategy and applies learned gated fusion (Eq. 3) at a single decoder scale. Building on this, **Setting 4 (+ Multi-Scale Fusion w/o HMG)** extends the gated fusion across multiple scales but omits Hair Mask

Guidance (HMG) during feature modulation in Eq. 1 (use $f'_n = f_c$ instead of f_n). Finally, **Setting 5 (+ HMG)** represents our complete model, utilizing multi-scale gated fusion with HMG (using f_n in Eq. 1).

The quantitative results in Table 4 and qualitative comparisons in Figure 5 clearly demonstrate the benefits of each added component. The Baseline (Setting 1) suffers from poor non-hair region fidelity (low SSIM/IDS). While Naive Blending (Setting 2) enforces pixel preservation, it leads to unnatural artifacts and degrades overall quality (worse FID/FVD), highlighting the necessity for learned integration. Introducing single-scale gated fusion with decoupled learning (Setting 3) improves integration and fidelity over naive blending. Extending fusion to multiple scales (Setting 4) significantly enhances spatial coherence and temporal stability (better FID/FVD, Vbench-MS vs. Setting 3). Crucially, the comparison between Setting 4 and Setting 5 isolates the impact of HMG. Adding HMG yields substantial improvements in non-hair fidelity (highest SSIM/IDS) and overall realism (lowest FID/FVD), this component fulfills its specific purpose of providing explicit spatial cues for boundary refinement, leading to visibly sharper results. This systematic improvement validates our approach, proving the effectiveness of combining decoupled learning with mask-aware multi-scale gated fusion for achieving identity-preserving, spatially coherent, and temporally stable video hair transfer.

4.5 Limitations and Discussions

As shown in Fig 6, although our method can inherently handle various hairstyles, the **hair transfer from long hair into short hair still remains challenging**. This relies on the consistent video inpainting for the hair regions. Besides, our current framework primarily bases hair motion on head movement, **limiting its ability to simulate secondary dynamics like inertia during rapid rotations**. Additionally, complex **hand-hair interactions in the driving video can sometimes degrade blending quality near the interaction points**. Future work could explore incorporating physics-based modeling or specific interaction handling to address these limitations and enhance realism.



Figure 6: Example limitation showing artifacts when transferring long hair onto a short-haired subject.

5 CONCLUSION

We introduced *HairShifter*, the first dedicated framework for video hairstyle transfer. Our “Anchor Frame + Animation” approach integrates high-fidelity anchor generation via Image Hair Transfer(IHT) module with a specialized animation network. Trained using a novel decoupled strategy for effective hair/non-hair disentangle, the network employs our Multi-Scale Gated SPADE (MSG-SPADE) Decoder for seamless blending and precise identity/background preservation. Demonstrating state-of-the-art results, HairShifter successfully balances quality, temporal consistency, and fidelity, establishing a robust baseline for video hairstyle transfer.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable feedback and constructive suggestions. This work was supported by National Natural Science Foundation (Grant No. 62072382) and Yangtze Charitable Foundation.

REFERENCES

- [1] Jianhong Bai, Tianyu He, Yuchi Wang, Junliang Guo, Haoji Hu, Zuozhu Liu, and Jiang Bian. 2024. Unedit: A unified tuning-free framework for video motion and appearance editing. *arXiv preprint arXiv:2402.13185* (2024).
- [2] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.
- [3] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. 2023. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 23206–23217.
- [4] Wenhao Chai, Xun Guo, Gaopang Wang, and Yan Lu. 2023. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 23040–23050.
- [5] Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. 2023. Flatten: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922* (2023).
- [6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.
- [7] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.
- [8] Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. 2024. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8498–8507.
- [9] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lemitsky, and Egor Zakharov. 2022. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*. 2663–2671.
- [10] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. 2023. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373* (2023).
- [11] Yuchao Gu, Yipin Zhou, Bichen Wu, Licheng Yu, Jia-Wei Liu, Rui Zhao, Jay Zhangjie Wu, David Junhao Zhang, Mike Zheng Shou, and Kevin Tang. 2024. Videoswap: Customized video subject swapping with interactive semantic point correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7621–7630.
- [12] Jianzhu Guo, Dingyung Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. 2024. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168* (2024).
- [13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahu Lin, and Bo Dai. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725* (2023).
- [14] Yue Han, Junwei Zhu, Keke He, Xu Chen, Yanhao Ge, Wei Li, Xiangtai Li, Jiangning Zhang, Chengjie Wang, and Yong Liu. 2024. Face-Adapter for Pre-trained Diffusion Models with Fine-Grained ID and Attribute Control. In *European Conference on Computer Vision*. Springer, 20–36.
- [15] Thorsten Hempel, Ahmed A Abdelrahman, and Ayoub Al-Hamadi. 2022. 6d rotation representation for unconstrained head pose estimation. In *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2496–2500.
- [16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).
- [17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
- [18] Fa-Ting Hong and Dan Xu. 2023. Implicit identity representation conditioned memory compensation network for talking head video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 23062–23072.
- [19] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. 2022. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3397–3406.
- [20] Huaibo Huang, Ran He, Zhenan Sun, Tieniu Tan, et al. 2018. Introvae: Introspective variational autoencoders for photographic image synthesis. *Advances in neural information processing systems* 31 (2018).
- [21] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Champaisit, et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21807–21818.
- [22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.
- [23] Hyeonho Jeong and Jong Chul Ye. 2023. Ground-a-video: Zero-shot grounded video editing using text-to-image diffusion models. *arXiv preprint arXiv:2310.01107* (2023).
- [24] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II* 14. Springer, 694–711.
- [25] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.
- [26] Sasikarn Kwanmuang, Pakkapon Phongthawee, Patsorn Sangkloy, and Supasorn Suwajanakorn. 2023. StyleGAN Salon: multi-view latent optimization for pose-invariant hairstyle transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8609–8618.
- [27] Taewoo Kim, Chaeyeon Chung, Yoonseo Kim, Sunghyun Park, Kangyeol Kim, and Jaegul Choo. 2022. Style your hair: Latent optimization for pose-invariant hairstyle transfer via local-style-aware hair alignment. In *European Conference on Computer Vision*. Springer, 188–203.
- [28] Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. 2022. Implicit warping for animation with image sets. *Advances in Neural Information Processing Systems* 35 (2022), 22438–22450.
- [29] Maxim Nikolaev, Mikhail Kuznetsov, Dmitry P Petrov, and Aibek Alanov. 2024. Hairfastgan: Realistic and robust hair transfer with a fast encoder-based approach. *Advances in Neural Information Processing Systems* 37 (2024), 45600–45635.
- [30] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2337–2346.
- [31] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. 2023. Faterzero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15932–15942.
- [32] Alex Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
- [34] Rohit Saha, Brendan Duke, Florian Shkurti, Graham W Taylor, and Parham Arabi. 2021. Loho: Latent optimization of hairstyles via orthogonalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1984–1993.
- [35] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. 2006. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing* 15, 11 (2006), 3440–3451.
- [36] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First order motion model for image animation. *Advances in neural information processing systems* 32 (2019).
- [37] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [38] Zhentao Tan, Menglei Chai, Dongdong Chen, Jing Liao, Qi Chu, Li Yuan, Sergey Tulyakov, and Nenghai Yu. 2020. Michigan: multi-input-conditioned hair image generation for portrait editing. *arXiv preprint arXiv:2010.16417* (2020).
- [39] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricu, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).
- [40] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717* (2018).
- [41] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10039–10049.
- [42] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.
- [43] Huawei Wei, Zejun Yang, and Zhisheng Wang. 2024. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694* (2024).

- [44] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Zhentao Tan, Lu Yuan, Weiming Zhang, and Nenghai Yu. 2022. Hairclip: Design your hair by text and reference image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18072–18081.
- [45] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Weiming Zhang, Gang Hua, and Nenghai Yu. 2023. Hairclipv2: Unifying hair editing via proxy feature blending. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 23589–23599.
- [46] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7623–7633.
- [47] Yiqian Wu, Yong-Liang Yang, and Xiaogang Jin. 2022. Hairmapper: Removing hair from portraits using gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4227–4236.
- [48] You Xie, Hongyi Xu, Guoxian Song, Chao Wang, Yichun Shi, and Linjie Luo. 2024. X-portrait: Expressive portrait animation with hierarchical motion attention. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.
- [49] Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. 2024. Easyanimate: A high-performance long video generation method based on transformer architecture. *arXiv preprint arXiv:2405.18991* (2024).
- [50] Shiyuan Yang, Xiaodong Chen, and Jing Liao. 2023. Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*. 3190–3199.
- [51] Shurong Yang, Huadong Li, Juhao Wu, Minhao Jing, Linze Li, Renhe Ji, Jiajun Liang, and Haoqiang Fan. 2024. Megactor: Harness the power of raw video for vivid portrait animation. *arXiv preprint arXiv:2405.20851* (2024).
- [52] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. 2023. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*. 1–11.
- [53] Xiangpeng Yang, Linchao Zhu, Hehe Fan, and Yi Yang. 2025. VideoGrain: Modulating Space-Time Attention for Multi-grained Video Editing. *arXiv preprint arXiv:2502.17258* (2025).
- [54] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. 2023. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10459–10469.
- [55] Bohan Zeng, Xuhui Liu, Sicheng Gao, Boyu Liu, Hong Li, Jianzhuang Liu, and Baochang Zhang. 2023. Face animation with an attribute-guided diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 628–637.
- [56] Yu Zeng, Yang Zhang, Liu Jiachen, Linlin Shen, Kaijun Deng, Weizhao He, and Jinbao Wang. 2024. HairDiffusion: Vivid Multi-Colored Hair Editing via Latent Diffusion. *Advances in Neural Information Processing Systems* 37 (2024), 5048–5073.
- [57] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3836–3847.
- [58] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. 2023. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077* (2023).
- [59] Yuxuan Zhang, Qing Zhang, Yiren Song, Jichao Zhang, Hao Tang, and Jiaming Liu. 2024. Stable-hair: Real-world hair transfer via diffusion model. *arXiv preprint arXiv:2407.14078*.
- [60] Jian Zhao and Hui Zhang. 2022. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3657–3666.
- [61] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. 2022. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18697–18709.
- [62] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. 2022. CelebV-HQ: A large-scale video facial attributes dataset. In *European conference on computer vision*. Springer, 650–667.
- [63] Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. 2021. Barber-shop: Gan-based image compositing using segmentation masks. *arXiv preprint arXiv:2106.01505* (2021).
- [64] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. 2024. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *European Conference on Computer Vision*. Springer, 195–211.

A APPENDIX

A.1 Adaptation of Image Hair Transfer(IHT)

The foundation of our video hair transfer pipeline relies on generating a high-fidelity initial frame incorporating the target hairstyle onto the subject’s identity. For this purpose, we develop the Image Hair Transfer (IHT) module, which builds upon the robust Stable-Hair [59] framework, known for its strong diffusion-based image synthesis. A key challenge with the original Stable-Hair lies in its **training on single-image derivatives**, which inadvertently **couples the input face pose with the reference hair pose**. This limits its effectiveness when transferring hairstyles between images with differing poses, a common scenario in video. To address this, we **improve Stable-Hair by fine-tuning it with video data**, specifically using input and reference frames sampled from *different timestamps* within the same sequence. This forces the model to learn pose-invariant hair representations, significantly enhancing its robustness to the pose variations encountered in video applications.

To **improve the IHT module**, we **fine-tune the Hair Extractor and Latent IdentityNet of Stable-Hair**, with a **combined image-video dataset**. For the image portion, we sample 2000 images from **FFHQ**. Following the methodology of Stable-Hair, we generate training triplets – comprising a **GT image, a Reference image, and a Bald proxy image – all derived from one source image**. For the video portion, we sample approximately **18,000 frames from CelebV-HQ**. For these, the **GT image and Bald proxy image originate from the same frame** while the **Reference image is another frame** from the same video, often chosen for pose variation via **SixDRepNet[15]**. This process yields a combined training dataset of over 60,000 triplets. **Fine-tuning was conducted for 100,000 steps on three A6000 GPUs**, using a **batch size of 8 and a learning rate of 2×10^{-5}** .

As shown in Figure 7, the resulting IHT module demonstrates improved capability in generating spatially coherent and structurally accurate hairstyles, despite substantial pose discrepancies between the anchor frame and the reference hair image. This establishes a stable, high-fidelity foundation for subsequent animation. Within our framework, IHT serves two distinct functions: supporting training by generating pseudo driving frames, and enabling inference by synthesizing the high-quality anchor frame that defines the target hair appearance.

A.2 Text Prompt Generation for VideoGrain Baseline

To evaluate the text-driven VideoGrain baseline [53] comparably against reference-image-driven methods, we generated descriptive text prompts from the target hair reference images (R_{target}). We utilized the Multimodal Large Language Model (MLLM) *Google Gemini 2.0 Flash* [39] for this task. Each reference image was input to the MLLM along with the following instruction prompt:

Analyze the given facial image and generate a descriptive text embedding. The text should include the following details:
 1. Gender - Clearly specify whether the person appears to be male, female, or another gender identity.
 2. Hair Color - Describe the hair color, such as black, blonde, brown, red, etc.
 3. Hair Shape - Specify the shape or style of the



Figure 7: Comparison of our adapted Image Hair Transfer (IHT) module with the baseline Stable-Hair. Our fine-tuned IHT demonstrates improved robustness, generating a hairstyle more consistent with the input posV Ce compared to the baseline.

hair, such as straight, curly, wavy, short, long, etc.

4. Hair Structure – Describe the hair texture and structure, including volume, layering, smoothness, or other defining characteristics.

Avoid mentioning the background, environment, or any elements unrelated to the person’s facial features. Focus solely on the person’s facial and hair attributes.

The text description generated by the MLLM for a given reference image, focusing on the hair attributes as requested, was then directly used as the input prompt for VideoGrain when processing the corresponding driving video (V_d). For these experiments, the editing region was set to the entire image frame, omitting a specific mask. This approach ensured the textual guidance for VideoGrain was derived from the same visual targets used across all compared methods.

A.3 Computational Cost and Real-Time Capability

HairShifter contains approximately 3.45B parameters and **sustains 20+ FPS on 512×512 resolution videos using a single RTX 3090 GPU (23 GB VRAM)** without any acceleration, significantly surpassing existing methods or alternative solutions that typically operate near **1 FPS**. Inference requires a one-time invocation of the high-fidelity *IHT module* for the anchor frame, incurring **74.23 TFLOPs**. All remaining frames are handled by the animation module with only **1.57 TFLOPs per frame**. The average computational cost per frame over a video of length N is:

$$C(N) = \frac{74.23 + 1.57 \times N}{N} \text{ TFLOPs},$$

which quickly approaches 1.57 as N grows. This amortized formulation ensures that per-frame overhead becomes negligible in longer sequences, enabling **real-time, high-resolution video editing** with consistent performance.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009