H-Adapter: Pose-Robust Hairstyle Transfer via Attention-Derived, Source-Aligned Hair Masks

Seulgi Jeong*, Yunseong Cho, and Sanghun Park✉

SNOW Corp., Republic of Korea
wjdtmfrl3682@gmail.com, {yunseong.cho,sanghun.park}@snowcorp.com

Abstract. Hairstyle transfer has practical applications such as virtual try-on, yet remains challenging when the source and reference exhibit large head-pose discrepancies. We propose H-Adapter, which improves pose robustness by training with a region-specific loss that disentangles hair and non-hair objectives and thereby induces spatially disentangled cross-attention, from which a source-aligned hair edit mask is derived to guide diffusion-based inpainting. Experiments on pose-agnostic and pose-different subsets demonstrate strong quantitative results, including the best FID, FID_{CLIP}, and CLIP-I under pose differences, while maintaining competitive non-hair preservation and improving qualitative fidelity to fine-grained reference hairstyle details. Beyond source-conditioned transfer, H-Adapter supports practical extensions including reference-guided text-to-image generation, auxiliary prompt-based hair color control, and compatibility with an identity-preserving IP-Adapter variant. We also introduce a VLM-as-a-judge protocol and observe consistent gains in hairstyle faithfulness, non-hair preservation, and artifact quality.

Keywords: Hairstyle transfer · Diffusion inpainting · Cross-attention

1 Introduction

Recent progress in image synthesis and editing has led to higher-fidelity generation and more precise local image manipulations. Hairstyle transfer has emerged as an important image editing task with practical applications such as virtual hairstyle try-on. Given a source and a reference image, the goal is to transfer the reference hair color and shape while preserving non-hair content (*e.g.*, identity, background, and clothing). This task imposes region-dependent objectives: the model must reflect reference hairstyle features in hair regions while preserving the source content in non-hair regions. This coupling of conflicting objectives makes it difficult to localize edits without degrading non-hair content.

A key challenge arises from the non-rigid, pose-dependent nature of hair: changes in head pose can substantially shift its spatial placement and silhouette, so successful transfer requires determining where hair should appear on the

* Work done during an internship at SNOW Corp.
✉ Corresponding author.

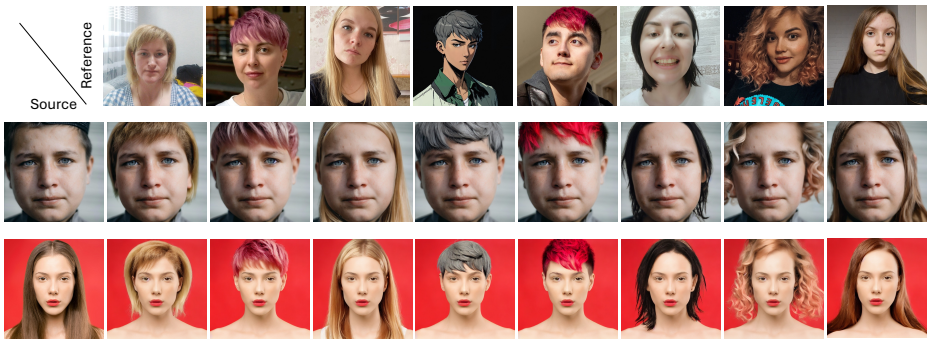


Fig. 1: Qualitative results of H-Adapter (Ours) on EasyPortrait [15] and web-crawled images. H-Adapter preserves image quality and reference hairstyle features under diverse, unconstrained conditions. Source and license information for the Pexels, Pixabay, and Unsplash images used in this figure is provided in the supplementary material.

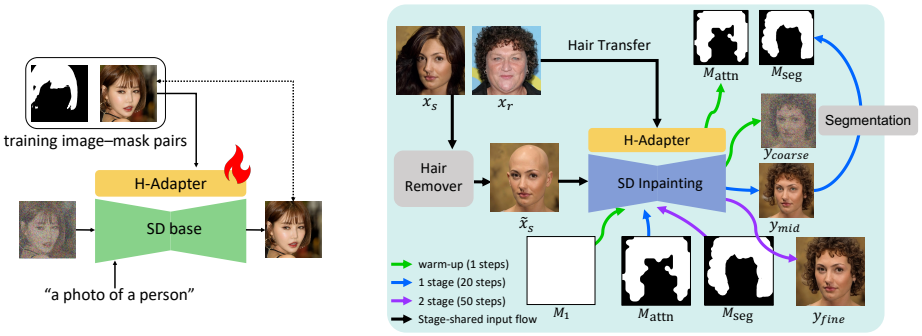
source head geometry and in what approximate shape, especially under large source–reference pose discrepancies.

GAN-based approaches [4, 13, 19, 27, 31, 34, 36, 44] have driven much of the recent progress in hairstyle transfer. These methods typically rely on aligned face datasets, which can be effective in controlled settings but limit robustness in unconstrained imagery. To improve generalization under unconstrained conditions, diffusion-based approaches [5, 39, 41] have also been explored. However, under pose mismatch, two failure modes remain common: (1) inaccurate hair-region localization, leading to misplaced or over-extended edits; and (2) limited reproduction of fine-grained reference structure and texture, even when global placement is reasonable.

Existing methods often fail to localize the editable hair region under large pose mismatch. We therefore seek *source-aligned spatial guidance* that specifies where hair should be synthesized on the source head geometry, and use it to constrain diffusion-based inpainting [26] to the hair region.

We propose H-Adapter, a hair-aware adapter trained with a novel *region-specific loss* to spatially disentangle hair and non-hair objectives, built upon IP-Adapter [37]. Concretely, the loss encourages the model to learn reference hairstyle appearance only within hair regions, while suppressing reference conditioning in non-hair regions by regularizing them toward the pretrained diffusion model's predictions. This training yields cross-attention maps with increased spatial separation between hair and non-hair content, from which we derive a coarse, source-aligned hair mask that guides diffusion-based inpainting. Figure 1 provides representative in-the-wild examples, illustrating the practical applicability of H-Adapter under source–reference pose discrepancies and other unconstrained image variations.

Beyond improving localization under pose mismatch, the adapter-based design also enables flexible inference-time use: H-Adapter supports auxiliary prompt-based hair-color control in hairstyle transfer, general prompt-driven generation,



(a) Training pipeline of H-Adapter. The adapter is optimized with a region-specific objective using training image-mask pairs to encourage hair-region-focused conditioning.

(b) Two-stage inference pipeline. HairRemover [1] converts the source image x_s into a hair-removed base image $\tilde{x}_s$. Green arrows denote coarse mask extraction (M_{attn}), blue arrows denote intermediate generation (y_{mid}) and segmentation (M_{seg}), and purple arrows denote final inpainting to produce y_{fine}.

Fig. 2: Overview of our method. We train an H-Adapter with a region-specific objective and apply a source-aligned coarse attention mask to localize diffusion inpainting for reference-guided hairstyle transfer.

and compatibility with identity-preserving IP-Adapter variants. Our contributions are summarized as follows:

- We introduce a region-specific loss that disentangles hair and non-hair objectives for faithful reference hairstyle transfer.
- We derive a source-aligned coarse attention mask from H-Adapter cross-attention to guide diffusion inpainting with pose and shape consistency.
- We demonstrate plug-and-play inference, including source-free reference-guided text-to-image generation, optional text-guided hair-color control, and compatibility with other IP-Adapter variants.

We achieve strong results under pose mismatch across quantitative, qualitative, VLM-as-a-judge, and human preference evaluations.

2 Related Work

2.1 Hairstyle Transfer

Prior work on hairstyle transfer has largely focused on structured control signals to improve editability and fidelity. Several methods explicitly disentangle hair attributes to enable more controllable editing [19,27,31,36,44]. Text-driven control is further explored in HairCLIP [34].

More recent approaches expand the conditioning space for practical use, ranging from diffusion-based text conditioning for multi-color hair editing [39] to HairCLIPv2 [36], which supports multi-modal, user-interactive controls via

proxy feature blending; a follow-up work [35] generalizes this proxy-based editing paradigm, including 3D-aware hair editing.

However, large head-pose differences between the source and reference images remain challenging. To address this, prior works propose various alignment and pose-conditioning strategies. Barbershop [44] performs GAN-based image compositing with segmentation masks. Style-Your-Hair [13] aligns hairstyles by optimizing a target-hair latent code toward the source pose. HairFiT [4] leverages multi-view datasets and flow-based alignment to better handle viewpoint changes, while HairNet [45] further addresses hairstyle transfer with pose changes by removing the source hair and transferring a pose-aligned reference hairstyle. HairFastGAN [19], a later GAN-based method designed to mitigate head-pose mismatch, introduces a Rotate Encoder that transforms the face latent to enable pose-aware transfer. Stable-Hair [41] exploits video data by transferring hairstyles across frames of the same identity to increase pose robustness. HairFusion [5] injects pose cues (*e.g.*, DensePose [9]) into attention to encourage pose-aligned hair generation.

Despite these efforts, handling large pose gaps between the source and reference images remains challenging. Several existing approaches mitigate pose variation through pose-aware alignment, latent transformations, or auxiliary pose cues [4, 5, 13, 19]. In contrast, our method derives a source-aligned coarse edit region from cross-attention maps induced by our region-specific loss and uses it to guide diffusion inpainting for pose-consistent hairstyle transfer.

2.2 Image Editing

Controlling Attention for Editing Recent advances in text-to-image diffusion models [18, 23, 25, 26, 28] have improved generation fidelity and enabled more precise local control in image editing. In particular, several works manipulate attention to localize edits and preserve structure. Prompt-to-Prompt [10] performs text-driven editing by manipulating cross-attention maps to localize changes across denoising steps, without relying on explicit masks. Plug-and-Play methods [32] inject self-attention from a guidance image into the generation process to preserve structure in image-to-image translation.

Mask Generation from Attention Several works derive spatial masks from attention maps to localize edits. MasaCtrl [2] separates foreground and background using cross-attention and constrains each region to attend only to the corresponding source region. DiffEdit [7] derives edit masks from differences in diffusion noise predictions, and InstDiffEdit [46] selects representative tokens and aggregates token relevance to generate masks in real time.

Our approach also leverages attention-derived masks, but targets hair transfer under pose mismatch. Unlike prior work on general edit localization, we derive a source-aligned hair mask from cross-attention maps that separate hair and non-hair content.

3 Method

3.1 Overview

An overview of our framework is shown in Fig. 2. Given a source image x_s and a reference image x_r, our goal is to transfer the reference hairstyle to the source while preserving non-hair regions. Our method consists of three parts: (i) training IP-Adapter as H-Adapter with a region-specific objective that decouples hair-region reference learning from non-hair consistency regularization (Sec. 3.2), (ii) an inference pipeline that leverages a source-aligned coarse attention mask to restrict diffusion inpainting [26] to the target hair region (Sec. 3.3), and (iii) extensions enabled by the plug-and-play compatibility of H-Adapter with text prompts and other adapters (Sec. 3.4).

3.2 Training the H-Adapter

We adopt a pretrained IP-Adapter [37] to inject reference hairstyle features through cross-attention during diffusion denoising. However, directly applying a standard IP-Adapter to hairstyle transfer is non-trivial, as it is trained to inject conditioning features globally and thus tends to affect non-hair regions as well. As shown in Fig. 5(g) (rows 2–3), directly applying IP-Adapter often fails to transfer hairstyles reliably, leading to incomplete or incorrect hair edits. To confine the conditioning effect to hair, we train the adapter with a region-specific objective, as illustrated in Fig. 2a.

We train the adapter using triplets (x_i, x_j, M_i), where x_i is the target image used for diffusion denoising, x_j is the reference image used as the H-Adapter condition, and M_i is the binary hair mask of x_i. Since paired images of the same identity with controlled hairstyle changes are difficult to obtain in practice, we train with a mixture of self-supervised and video-based supervision. Specifically, for a subset of samples, we adopt a self-supervised setup where the same image serves as both the reference and the target (*i.e.*, $x_j = x_i$) providing direct reconstruction supervision. For the remaining samples, we construct reference–target pairs from videos such that x_j and x_i depict the same identity but correspond to different frames (*i.e.*, $x_j \neq x_i$), exposing the adapter to natural hairstyle variations while preserving identity consistency.

Region-Specific Training Objective The region-specific objective is designed to impose region-dependent behavior on the H-Adapter across hair and non-hair regions. In hair regions, the masked denoising loss encourages the adapter to incorporate reference hairstyle cues into the denoising prediction. In non-hair regions, its reference-conditioning effect should be suppressed so that the prediction remains consistent with the pretrained diffusion model. By separating these two objectives, the loss encourages the reference branch to become hair-focused and to produce spatially disentangled attention maps between hair and non-hair content, which can later be used as a source-aligned localization cue during inference. We formalize this design with two complementary loss terms.

First, we apply the standard diffusion denoising loss only within the hair region:

$$L_{\text{hair}} = \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,I),t}\left[\|(\epsilon - \epsilon_{\theta_{\text{H}}}(z_t, t, c_t, c_i)) \cdot M_i\|_2^2\right], \tag{1}$$

where z_t is the noisy latent at timestep t, $\epsilon \sim \mathcal{N}(0, I)$ is Gaussian noise, c_t and c_i denote the text and reference-image conditions, and $\epsilon_{\theta_{\text{H}}}$ is the noise predictor augmented with the H-Adapter.

Second, to suppress the H-Adapter's influence on non-hair regions, we regularize the model to match the pretrained diffusion model's noise prediction outside the hair mask:

$$L_{\text{non-hair}} = \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,I),t}\left[\|(\epsilon_{\theta_{\text{H}}}(z_t, t, c_t, c_i) - \epsilon_{\theta_0}(z_t, t, c_t)) \cdot (1 - M_i)\|_2^2\right], \tag{2}$$

where ϵ_{θ_0} is the baseline noise predictor without the H-Adapter.

The final training objective is

$$L = L_{\text{hair}} + \lambda_{\text{non-hair}}\, L_{\text{non-hair}}, \tag{3}$$

where $\lambda_{\text{non-hair}}$ controls the strength of the non-hair consistency term and is set to 0.1 in all experiments.

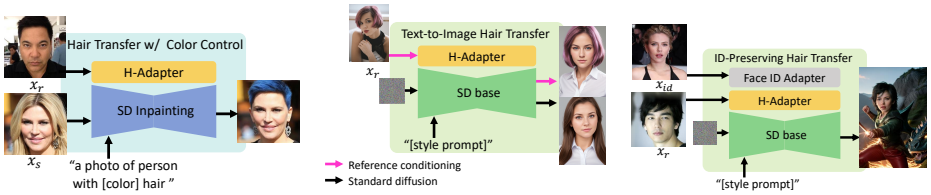
3.3 Inference-Time Hairstyle Transfer

Since the proposed training objective is independent of the inpainting process, we train the H-Adapter on the base text-to-image backbone. The learned adapter is then directly transferred to the inpainting backbone used by our coarse-to-fine inference pipeline. We obtain the hair-removed base image $\tilde{x}_s$ from the source image x_s by prompting a pretrained instruction-based image editing model [1]. We then perform a two-stage coarse-to-fine inpainting process, as shown in Fig. 2b, to obtain the final hairstyle transfer result.

Source-Aligned Coarse Attention Mask A key observation is that the proposed region-specific objective yields cross-attention maps with spatial separation between hair and non-hair regions. Figure 5(b) visualizes the cross-attention maps for all K tokens $(t_0, \ldots, t_{K-1})$. We use the token that consistently attends to non-hair regions as a separator token t_s, and aggregate the remaining token maps to obtain an attention-derived coarse inpainting mask:

$$M_{\text{attn}} = \text{Binarize}\left(\sum_{k\neq s} CA(t_k)\right), \tag{4}$$

where $CA(t_k)$ denotes the cross-attention map of token t_k. In our implementation, $K = 16$ and we set $s = 8$ based on the validation protocol described in Appendix F of the Supplementary Material. The resulting attention-derived mask is used in two ways. First, it serves as the inpainting mask M_{attn} in the two-stage inference pipeline. Second, the same attention-based localization is recomputed at each denoising timestep for reference gating, as described next.



(a) Auxiliary prompt-based control **(b)** Source-free reference-guided text-to-image generation **(c)** Identity-preserving hair transfer

Fig. 3: Flexible extensions of H-Adapter. H-Adapter supports auxiliary prompt control (λ; Eq. (5)), source-free reference-guided text-to-image generation, and composition with identity-preserving adapters. Input images in panels (b) and (c) are adapted from Pexels stock images under the Pexels License and from © JCS, Wikimedia Commons, licensed under CC BY 3.0, respectively.

Per-timestep Reference Gating During inference, we recompute an attention-derived spatial mask M_t using the same separator-token aggregation described above and use it to gate the H-Adapter reference branch in cross-attention. Following an existing masked reference-injection cross-attention implementation, [1] we compute

$$Z = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V + \lambda\Big(M_t \odot \mathrm{softmax}\left(\frac{QK'^\top}{\sqrt{d}}\right)V'\Big), \tag{5}$$

where Q is the query projected from the current spatial features, (K, V) are the key and value projected from the textual context used in standard cross-attention, and (K', V') are those from the H-Adapter conditioning features. The output combines the standard cross-attention output with a masked H-Adapter conditioning branch, which is spatially gated by M_t and scaled by λ. The operator $\odot$ denotes element-wise multiplication.

Two-Stage Inpainting with a Warm-Up Step A two-stage inpainting pipeline with a warm-up step is employed. Figure 2b illustrates the overall procedure. Since a source-aligned coarse mask is not available at the start of inference, we initialize inpainting with an all-one mask and run a single warm-up denoising step to extract cross-attention maps, from which the attention-based mask M_{attn} is derived. Using M_{attn}, denoising is run for a small number of steps (*e.g.*, $\sim$20) to obtain an intermediate image y_{mid}. A pretrained segmentation model [38] then predicts a pixel-level hair mask M_{seg} from the intermediate image y_{mid}. Denoising is then restarted from the initial noise latent z_T with M_{seg} as the inpainting mask, yielding the final output y_{fine} with a pixel-level editable region. Overall, this coarse-to-fine strategy progressively estimates the hair mask at increasing spatial precision, enabling robust source-pose-aligned hairstyle transfer.

[1] `https://github.com/huggingface/diffusers`

3.4 Extensions for Reference-Guided Hairstyle Transfer

As shown in Fig. 3, H-Adapter supports flexible compositions with existing diffusion conditioning mechanisms. We describe three practical usages built on the same interface.

Auxiliary Prompt-Based Hair Color Control In the source-conditioned hairstyle transfer setting, we optionally incorporate an auxiliary text prompt (*e.g.*, hair color) as an additional condition. A relative weighting factor λ (Eq. (5)) controls the trade-off between auxiliary prompt controllability and reference faithfulness.

Reference-guided text-to-image generation H-Adapter supports reference-guided text-to-image generation by injecting reference hairstyle features via cross-attention, without requiring a source-image condition.

Compatibility with Identity-Preserving Adapters H-Adapter is composable with identity-preserving adapters by jointly conditioning the diffusion model on (i) a source identity embedding and (ii) the reference hairstyle embedding. This enables identity-consistent hairstyle transfer while retaining prompt-driven content generation.

4 Experiments

4.1 Experimental setup

H-Adapter is fine-tuned from IP-Adapter-Plus [37] on Stable Diffusion v1.5 [26] using our region-specific loss. We construct self-paired samples from FFHQ [12] images and same-identity video pairs from CelebV-HQ [43] frames. At inference time, H-Adapter is applied with the Stable Diffusion v1.5 inpainting model, using hair-removed base images synthesized by FLUX.2 [1]. Full implementation details are provided in Appendix B of the Supplementary Material.

4.2 Evaluation Protocol

Baselines Comparisons are conducted against state-of-the-art hairstyle transfer methods, including HairCLIPv2 [36], Style-Your-Hair (SYH) [13], HairFastGAN [19], Stable-Hair [41], and HairFusion [5]. For each baseline, images are generated by following the official code release. We additionally report results obtained by replacing H-Adapter with the standard IP-Adapter [37], where the IP-Adapter is trained on the same dataset as ours for a fair comparison. As an additional qualitative comparison, we further include recent general-purpose image editors—FLUX.2 [1], Nano Banana Pro [8], and Qwen-Image-2.0-Pro [42]—as well as the GAN-inversion-based method Barbershop [44].

Metrics To evaluate overall distributional fidelity, we compute FID between the source images and the generated images. FID_{CLIP} [16] is computed similarly to FID, but uses a CLIP image encoder [24] instead of Inception V3 [30]. To

Fig. 4: Qualitative comparisons under head-pose differences. The first two columns show the source and reference images; the remaining columns show outputs from each method. These examples span source–reference yaw gaps from 18.65° to 53.21°.

evaluate non-hair preservation, reconstruction quality is measured on the intersection of the non-hair regions of the source and generated images using PSNR and SSIM [33]. Hairstyle transfer faithfulness is measured by the CLIP-I [16] similarity between the reference image and the generated image.

Evaluation subsets CelebA-HQ [11] contains 30,000 unpaired images. We build two pair-disjoint evaluation subsets of 3,000 source–reference pairs by randomly pairing distinct images from the dataset, excluding hat images. The pose-agnostic subset is sampled without conditioning on head pose, while the pose-different subset includes only pairs with an absolute yaw difference $> 15°$ to test robustness to pose variation. For HairFusion [5], facial landmark extraction fails for some samples; to ensure fair comparison, we remove the affected pairs (62 pose-agnostic, 101 pose-different) for all methods and report results on the remaining pairs.

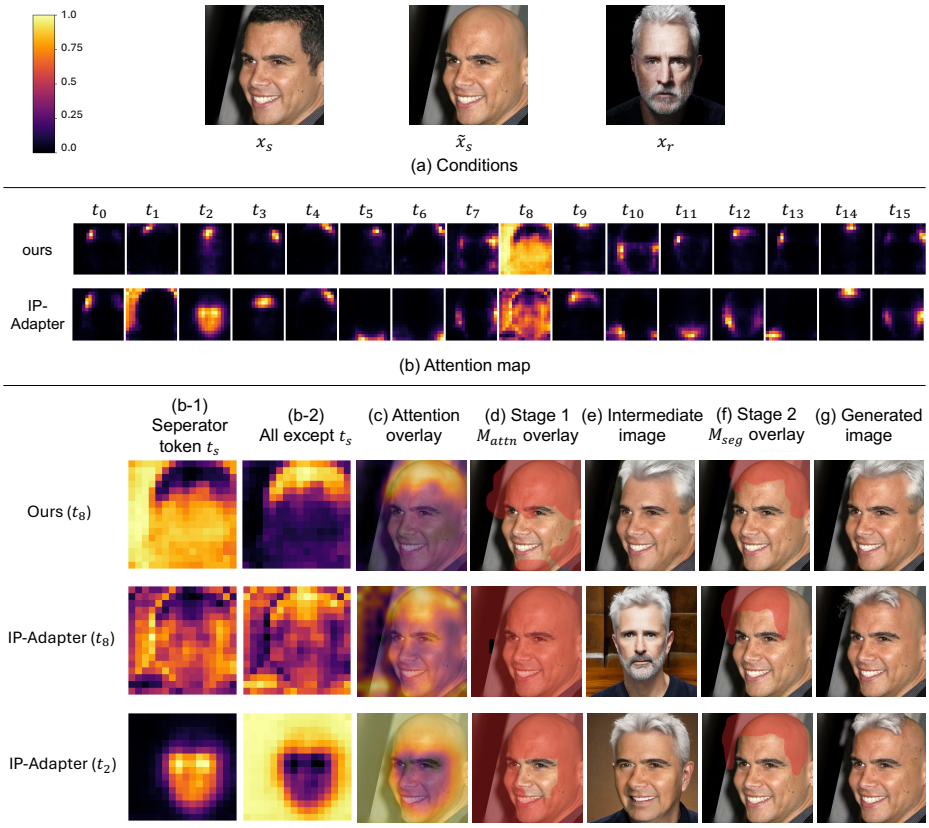


Fig. 5: Stage-wise visualization of editing guidance in the proposed inference pipeline, comparing IP-Adapter and H-Adapter (ours). (a) Conditioning inputs: source x_s, hair-removed base $\tilde{x}_s$, and reference x_r. (b) Token-wise cross-attention for 16 tokens $t_0, \ldots, t_{15}$: (b-1) the attention map of the separator token t_s, with its token index shown in parentheses; and (b-2) the aggregated attention map obtained by summing all tokens except t_s. (c) Overlay of (b-2) on $\tilde{x}_s$. (d) Stage-1 inpainting mask M_{attn} overlaid on $\tilde{x}_s$. (e) Stage-1 intermediate output. (f) Stage-2 inpainting mask M_{seg} overlaid on $\tilde{x}_s$. (g) Stage-2 final output. The figure shows that region-specific loss yields more source-aligned guidance, reducing mask misalignment and improving pose- and shape-consistent hairstyle transfer.

4.3 Qualitative Evaluation on Pose-Different Pairs

Figure 4 shows qualitative results on source—reference pairs with different head poses, following the baseline setup described in Sec. 4.2. Our method transfers the reference hairstyle while aligning its shape and spatial placement to the source head geometry and pose, yielding coherent hair boundaries and natural integration with the source face. In contrast, existing methods often fail to generate hair at an appropriate location or with a plausible shape aligned

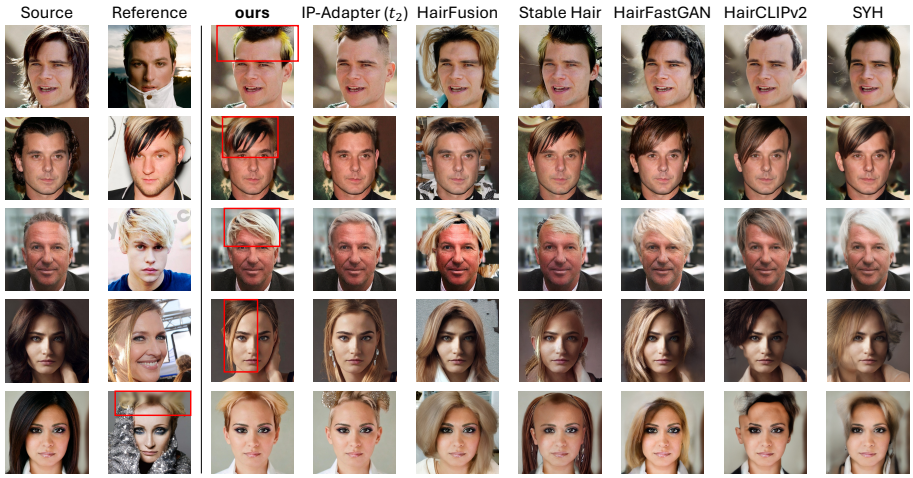


Fig. 6: Qualitative analysis of reference-feature faithfulness. The first two columns show the source and reference conditions, while the remaining columns compare method outputs.

to the source head geometry, particularly with large differences in head shape or pose. For general-purpose editors, the outputs can appear visually realistic, but the qualitative examples indicate that jointly preserving the source content and faithfully reflecting the reference hairstyle remains challenging under pose mismatch.

4.4 Analysis of Attention-Derived Localization

We analyze the contribution of the region-specific loss to source-aligned mask generation in our two-stage inference pipeline. To isolate its effect, we compare H-Adapter with an IP-Adapter baseline trained under the same setting but without the region-specific loss. For the IP-Adapter baseline, we report both the same t_8-based extraction used in our pipeline and an alternative t_2-based variant to illustrate the instability of attention-based localization without the proposed objective. As shown in Fig. 5, H-Adapter, trained with the region-specific loss, produces a coarse but source-aligned attention mask in Stage 1, yielding an intermediate image from which Stage 2 extracts a more accurate pixel-level hair mask for localized inpainting. In contrast, the IP-Adapter baselines exhibit less reliable attention localization, causing overly broad edits and boundary drift that degrade pose and shape consistency. Detailed per-stage analyses are provided in Appendix A of the Supplementary Material.

4.5 Qualitative Analysis of Reference Feature Faithfulness

We compare reference hairstyle transfer fidelity across our method, existing methods, and an IP-Adapter baseline obtained by replacing our H-Adapter us-

Table 1: Quantitative comparison on the pose-different subset.

Method	FID ↓	FID_{CLIP} ↓	SSIM ↑	PSNR ↑	CLIP-I ↑
Ours	**12.47**	**3.98**	**0.831**	23.06	**0.659**
IP-Adapter (t_8)	15.27	8.83	0.803	21.56	0.639
IP-Adapter (t_2)	12.53	4.26	0.825	22.70	0.651
HairFusion	28.03	8.80	0.756	17.26	0.626
Stable-Hair	25.79	8.70	0.798	22.39	0.640
HairFastGAN	12.78	4.53	0.817	24.40	0.649
HairCLIPv2	13.44	7.91	0.824	23.63	0.623
Style-Your-Hair	15.95	8.54	0.816	22.80	0.649

ing the t_2-based mask. As shown in Fig. 6, our method faithfully transfers both fine-grained details and global hairstyle structure from the reference while remaining consistent with the source head geometry. In the first two rows, it captures the color accents of the reference hairstyle. The third row demonstrates improved transfer of fine-grained parting details. The fourth row reproduces a thin highlight strand. The last row highlights stable faithfulness under imperfect references: even when the reference hair region is blurred, our method preserves the intended hairstyle attributes. Overall, these examples demonstrate faithful transfer of reference hairstyle attributes, preserving both fine-grained details and global structure.

4.6 Quantitative Evaluation

Table 1 reports quantitative results on the pose-different subset, while results on the pose-agnostic subset are provided in Appendix E of the Supplementary Material. On the pose-different subset, our method outperforms prior approaches as well as an IP-Adapter baseline obtained by replacing our H-Adapter, achieving the best FID and FID_{CLIP} scores, indicating improved visual fidelity. Notably, it attains the highest CLIP-I, suggesting the most faithful transfer of reference hairstyle attributes under pose mismatch. Moreover, it maintains high SSIM with competitive PSNR, indicating that structural similarity outside the edited hair region is largely preserved.

4.7 VLM-as-a-Judge Evaluation

While the above metrics quantify overall quality and pixel-level non-hair consistency, they cannot isolate hair-specific transfer quality, semantic non-hair preservation, or localized artifacts. We therefore adopt a VLM-as-a-judge framework [40] to score each result along three axes: (i) Hair Fidelity Score (`HFS`)—how faithfully the reference hairstyle is reproduced; (ii) Non-hair Preservation Score (`NPS`)—how well identity, background, and other non-hair regions are preserved; and (iii) Artifact Quality Score (`AQS`)—the absence of artifacts such as blending seams, color bleeding, or unnatural boundaries.

Table 2: VLM-as-a-judge scores (GPT-4o) with 95% bootstrap percentile CIs ($n_{\text{boot}} = 1{,}000$ over per-triplet template-mean scores).

Method	HFS↑	NPS↑	AQS↑
Ours	**3.11** [2.99, 3.23]	**4.23** [4.15, 4.32]	**3.73** [3.63, 3.83]
HairFusion	2.55 [2.44, 2.65]	3.55 [3.46, 3.64]	3.09 [2.99, 3.18]
Stable-Hair	2.90 [2.77, 3.02]	3.42 [3.33, 3.51]	2.75 [2.67, 2.84]
HairFastGAN	2.83 [2.72, 2.95]	3.91 [3.82, 4.00]	3.29 [3.19, 3.38]
HairCLIPv2	2.10 [1.97, 2.23]	4.05 [3.95, 4.13]	3.57 [3.46, 3.67]
Style-Your-Hair	2.87 [2.74, 3.02]	3.98 [3.90, 4.07]	3.61 [3.52, 3.70]

Table 3: Human preference study results, including valid pairwise votes, H-Adapter preference rates, 95% confidence intervals, and two-sided exact binomial test p-values against a 50% chance level.

Baseline	Votes	Ours (%)	95% CI	p-value
HairCLIPv2	624	81.1	[77.8, 84.0]	< 0.001
HairFastGAN	595	55.3	[51.3, 59.2]	0.011
HairFusion	678	80.4	[77.2, 83.2]	< 0.001
Stable-Hair	656	72.9	[69.3, 76.1]	< 0.001
Style-Your-Hair	640	72.3	[68.8, 75.7]	< 0.001
Overall	3193	72.7	[71.1, 74.2]	< 0.001

Evaluation Protocol. For each model, we evaluate 200 triplets (source, reference, generated output). The source–reference pairs are randomly sampled from CelebA-HQ and shared across all models. We adopt a *pointwise* scoring protocol [3] in which each VLM judge assigns a 1–5 Likert score to each axis via separate prompts to avoid inter-axis interference [14]. To mitigate prompt-sensitivity bias [29], we construct three lexically distinct prompt variants per axis and average across templates in each axis. Each prompt incorporates rubric anchors [17] and chain-of-thought rationale elicitation [14]. To reduce single-judge bias, the pipeline is executed with three VLM judges—GPT-4o [20], GPT-5.2 [21,22], and Gemini-2.5-Flash [6]—with prompt templates, consistency and pairwise validation reported in Appendix D of the Supplementary Material.

Evaluation Results. We present results with GPT-4o as the representative judge, as it achieves the highest consistency (Krippendorff's $\alpha \geq 0.90$ on all axes; mean per-triplet run-to-run $\sigma \leq 0.07$).

As shown in Tab. 2, our method achieves the highest mean on all three axes. While the AQS gap over Style-Your-Hair falls within overlapping confidence intervals, Style-Your-Hair ranks third on HFS. Other baselines show similar trade-offs (*e.g.*, HairCLIPv2 is second on NPS but last on HFS), whereas ours is the only method that maintains top-tier performance across all criteria without such compromises. Note that HFS scores are generally low across all methods due to the strict rubric requiring simultaneous match in color, texture, length, silhou-

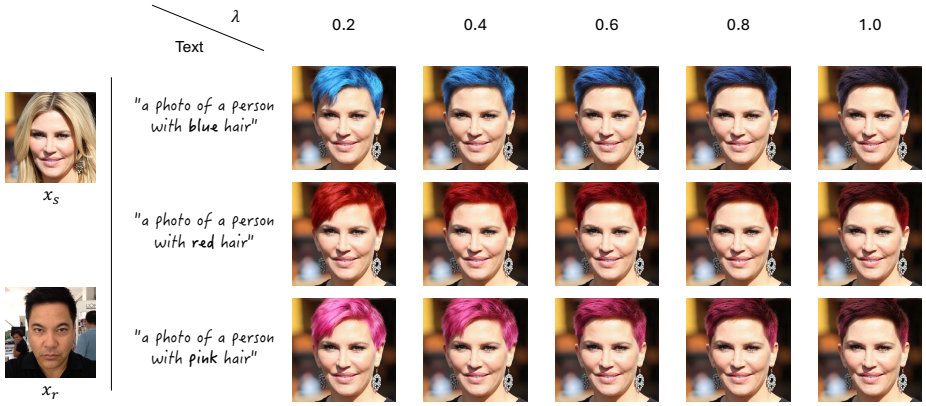


Fig. 7: Auxiliary hair-color control with H-Adapter. Inputs and the random seed are fixed. Rows vary hair-color prompts and columns vary λ (0.2–1.0). Smaller λ improves prompt responsiveness, while larger λ strengthens reference faithfulness.

ette, and parting; ours is the only method exceeding 3.0. Pairwise validation and per-judge breakdowns are provided in Appendix D of the Supplementary Material.

4.8 Human Preference Study

To complement automatic and VLM-based evaluations, we conduct a two-alternative forced-choice (2AFC) human preference study. Given the same source and reference images, participants select the result they preferred overall between two anonymized outputs, one from H-Adapter and one from a baseline. Participants may answer an arbitrary number of questions. As shown in Tab. 3, we report 3,193 valid pairwise votes from 53 respondents. Our method is preferred in 72.7% of comparisons (binomial test, $p < 0.001$) and is favored over every baseline individually.

4.9 Flexible Usage of H-Adapter

Beyond the default hairstyle transfer setting, H-Adapter also supports auxiliary prompt-based hair-color control, reference-guided text-to-image generation, and composition with IP-Adapter FaceID Plus for identity-preserving generation. As shown in Figs. 7 and 8, these usages demonstrate that the learned hair-conditioning branch can be flexibly combined with text prompts and identity conditions without retraining. Additional qualitative results, including in-the-wild examples, stylized inputs, reference-guided text-to-image generation, and identity-preserving adapter analysis, are provided in Appendix C of the Supplementary Material.

Fig. 8: Qualitative results of H-Adapter combined with IP-Adapter FaceID Plus [37]. The source image conditions FaceID Plus for identity preservation. Rows vary text prompts, and columns vary reference images for H-Adapter, demonstrating reference-guided hair transfer across diverse contexts while preserving identity. The source image is adapted from © JCS, Wikimedia Commons, licensed under CC BY 3.0.

5 Conclusion

We present H-Adapter, a pose-robust hairstyle transfer framework that trains an IP-Adapter with a region-specific objective to disentangle hair and non-hair regions. The resulting cross-attention provides a source-aligned coarse mask, which guides a two-stage diffusion inpainting pipeline to align the transferred hairstyle with the source head geometry. Across quantitative, qualitative, VLM-as-a-judge, and human preference evaluations, H-Adapter improves reference faithfulness and non-hair preservation under pose mismatch, while supporting flexible uses such as prompt-based hair-color control and identity-preserving generation. More broadly, our results suggest that region-specific training can turn cross-attention into a useful localization signal in reference-guided local editing.

References

1. Black Forest Labs: Flux.2 [klein]: Towards interactive visual intelligence, `https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence`, accessed: 2026-03-05
2. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22560–22570 (2023)

3. Chen, D., Chen, R., Zhang, S., Wang, Y., Liu, Y., Zhou, H., Zhang, Q., Wan, Y., Zhou, P., Sun, L.: Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In: Forty-first International Conference on Machine Learning (2024)
4. Chung, C., Kim, T., Nam, H., Choi, S., Gu, G., Park, S., Choo, J.: Hairfit: pose-invariant hairstyle transfer via flow-based hair alignment and semantic-region-aware inpainting. arXiv preprint arXiv:2206.08585 (2022)
5. Chung, C., Park, S., Kim, J., Choo, J.: What to preserve and what to transfer: Faithful, identity-preserving diffusion-based hairstyle transfer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2582–2590 (2025)
6. Comanici, G., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
7. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427 (2022)
8. Google DeepMind: Gemini 3 pro image – nano banana pro, `https://deepmind.google/models/gemini-image/pro/`, accessed: 2026-06-22
9. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7297–7306 (2018)
10. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
11. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
12. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
13. Kim, T., Chung, C., Kim, Y., Park, S., Kim, K., Choo, J.: Style your hair: Latent optimization for pose-invariant hairstyle transfer via local-style-aware hair alignment. In: European conference on computer vision. pp. 188–203. Springer (2022)
14. Ku, M., Jiang, D., Wei, C., Yue, X., Chen, W.: Viescore: Towards explainable metrics for conditional image synthesis evaluation. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12268–12290 (2024)
15. Kvanchiani, K., Petrova, E., Efremyan, K., Sautin, A., Kapitanov, A.: Easyportrait–face parsing and portrait segmentation dataset. arXiv preprint arXiv:2304.13509 (2023)
16. Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T., Lehtinen, J.: The role of imagenet classes in fr\'echet inception distance. arXiv preprint arXiv:2203.06026 (2022)
17. Lee, S., Kim, S., Park, S., Kim, G., Seo, M.: Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 11286–11315 (2024)
18. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
19. Nikolaev, M., Kuznetsov, M., Vetrov, D., Alanov, A.: Hairfastgan: Realistic and robust hair transfer with a fast encoder-based approach. Advances in Neural Information Processing Systems **37**, 45600–45635 (2024)

20. OpenAI: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
21. OpenAI: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
22. OpenAI: Update to gpt-5 system card: Gpt-5.2. `https://openai.com/index/gpt-5-system-card-update-gpt-5-2/` (2025), acessed: 2026-06-30
23. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
25. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
27. Saha, R., Duke, B., Shkurti, F., Taylor, G.W., Aarabi, P.: Loho: Latent optimization of hairstyles via orthogonalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1984–1993 (2021)
28. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
29. Slyman, E., Tanjim, M., Kafle, K., Lee, S.: Calibrating mllm-as-a-judge via multimodal bayesian prompt ensembles. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17224–17234 (2025)
30. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
31. Tan, Z., Chai, M., Chen, D., Liao, J., Chu, Q., Yuan, L., Tulyakov, S., Yu, N.: Michigan: multi-input-conditioned hair image generation for portrait editing. arXiv preprint arXiv:2010.16417 (2020)
32. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1921–1930 (2023)
33. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
34. Wei, T., Chen, D., Zhou, W., Liao, J., Tan, Z., Yuan, L., Zhang, W., Yu, N.: Hairclip: Design your hair by text and reference image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18072–18081 (2022)
35. Wei, T., Chen, D., Zhou, W., Liao, J., Wang, C., Zhang, W., Hua, G., Yu, N.: Unifying multi-modal hair editing via proxy feature blending. IEEE Transactions on Pattern Analysis and Machine Intelligence (2026)
36. Wei, T., Chen, D., Zhou, W., Liao, J., Zhang, W., Hua, G., Yu, N.: Hairclipv2: Unifying hair editing via proxy feature blending. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23589–23599 (2023)

37. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
38. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 325–341 (2018)
39. Zeng, Y., Zhang, Y., Jiachen, L., Shen, L., Deng, K., He, W., Wang, J.: Hairdiffusion: Vivid multi-colored hair editing via latent diffusion. Advances in Neural Information Processing Systems **37**, 5048–5073 (2024)
40. Zhang, X., Lu, Y., Wang, W., Yan, A., Yan, J., Qin, L., Wang, H., Yan, X., Wang, W.Y., Petzold, L.R.: Gpt-4v (ision) as a generalist evaluator for vision-language tasks. arXiv preprint arXiv:2311.01361 (2023)
41. Zhang, Y., Zhang, Q., Song, Y., Zhang, J., Tang, H., Liu, J.: Stable-hair: Real-world hair transfer via diffusion model. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10348–10356 (2025)
42. Zhao, B., Wu, C., Li, D., Meng, H., Li, J., Zhang, J., Zhou, J., Lin, J., Gao, K., Cao, K., et al.: Qwen-image-2.0 technical report. arXiv preprint arXiv:2605.10730 (2026)
43. Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., Zhang, L., Liu, Z., Loy, C.C.: Celebv-hq: A large-scale video facial attributes dataset. In: European conference on computer vision. pp. 650–667. Springer (2022)
44. Zhu, P., Abdal, R., Femiani, J., Wonka, P.: Barbershop: Gan-based image compositing using segmentation masks. arXiv preprint arXiv:2106.01505 (2021)
45. Zhu, P., Abdal, R., Femiani, J., Wonka, P.: Hairnet: Hairstyle transfer with pose changes. In: European Conference on Computer Vision. pp. 651–667. Springer (2022)
46. Zou, S., Tang, J., Zhou, Y., He, J., Zhao, C., Zhang, R., Hu, Z., Sun, X.: Towards efficient diffusion-based image editing with instant attention masks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7864–7872 (2024)