Towards High-fidelity Head Blending with Chroma Keying for Industrial Applications

Hah Min Lew[1]*, Sahng-Min Yoo[2]*‡, Hyunwoo Kang[3]*‡, and Gyeong-Moon Park[4]†
[1]Klleon AI Research, [2]Samsung Research, [3]Hyperconnect, [4]Kyung Hee University
hahmin.lew@klleon.io {yoosahngmin, khw7147}@gmail.com gmpark@khu.ac.kr

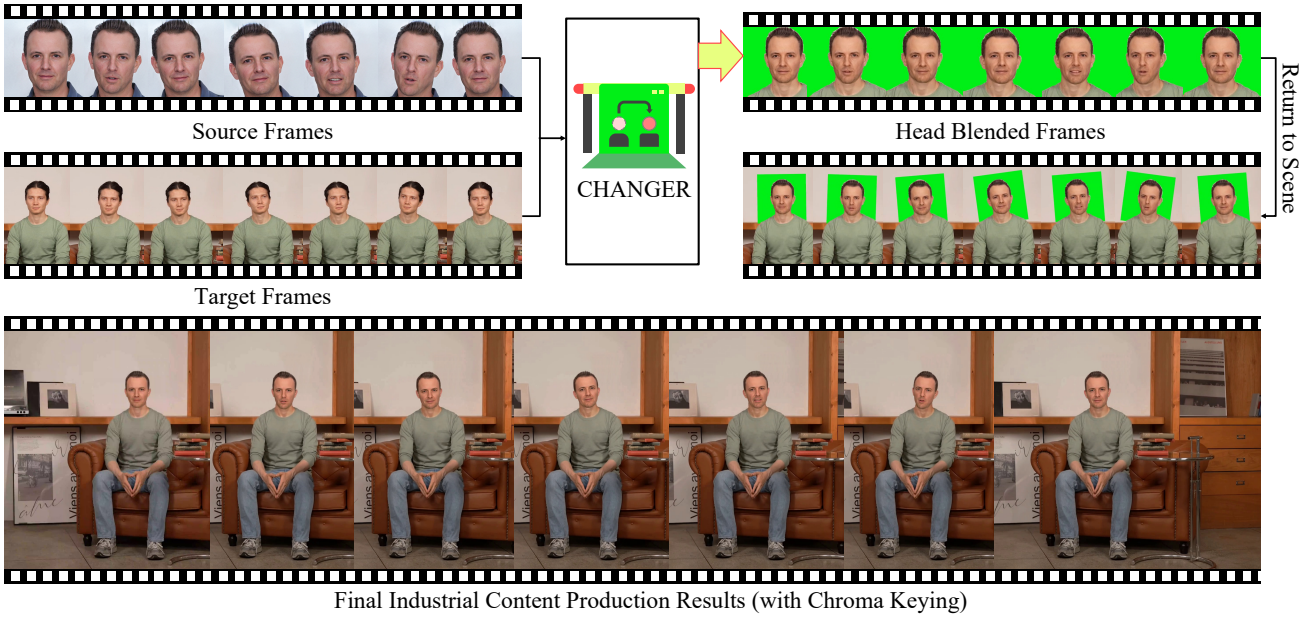

Figure 1. **Illustration of our CHANGER pipeline.** After acquiring the actor's frames (source), we can seamlessly blend acting scenes into the desired scenes with our CHANGER. Chroma keying ensures high-fidelity backgrounds. Here, both of the source and the target actors are virtual humans.

Abstract

We introduce an industrial Head Blending pipeline for the task of seamlessly integrating an actor's head onto a target body in digital content creation. The key challenge stems from discrepancies in head shape and hair structure, which lead to unnatural boundaries and blending artifacts. Existing methods treat foreground and background as a single task, resulting in suboptimal blending quality. To address this problem, we propose ***CHANGER****, a novel pipeline that decouples background integration from foreground blending. By utilizing chroma keying for artifact-free background generation and introducing* ***Head shape and long Hair augmentation*** *(****H²*** *augmentation) to simulate a wide range of head shapes and hair styles, CHANGER improves generalization on innumerable various real-world cases. Furthermore, our* ***Foreground Predictive Attention Transformer*** *(****FPAT****) module enhances foreground blending by predicting and focusing on key head and body regions. Quantitative and qualitative evaluations on benchmark datasets demonstrate that our CHANGER outperforms state-of-the-art methods, delivering high-fidelity, industrial-grade results.*

1. Introduction

In the realm of modern digital content creation, Head Blending, the seamless integration of an actor's head onto

Project page: https://hahminlew.github.io/changer
*Equal contribution
†Corresponding author
‡Work done at Klleon AI Research

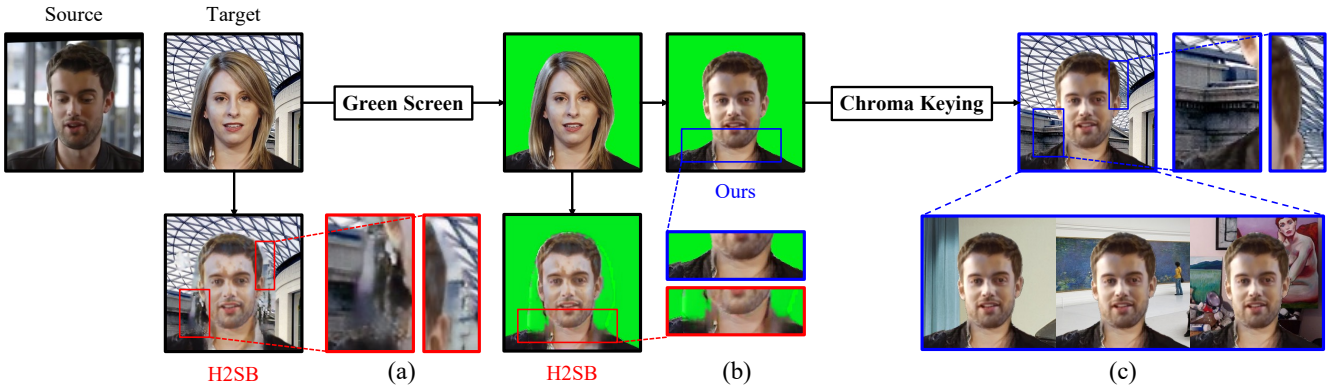

Figure 2. **Motivations of our work**. We propose CHANGER to consider the real-world application. As shown in (a), the existing work (H2SB [16]) shows severe artifacts on inpainting regions. To inpaint the background flawlessly, we propose to introduce chroma keying in the head blending framework. However, it still shows low-fidelity results to inpaint the body, which is hidden due to the head shape and hair difference described in a red box of (b). CHANGER generates the high-fidelity foreground with H^2 augmentation and Foreground Predictive Attention Transformer (FPAT), which is explained in Section 3.2 and 3.3, respectively. CHANGER removes artifacts as shown in the blue boxes of (b) and (c), and easily changes various high-fidelity real-world backgrounds. All backgrounds in the figure are from the benchmark dataset [12].

a body filmed in separate takes or contexts is a critical yet under-explored task. We focus on a such process, which is essential for various applications such as visual effects (VFX) post-production, digital human creation, and virtual avatar generation. In these scenarios, integrating an actor's head with footage where the body and surrounding environment may differ significantly is often necessary.

The main challenge of Head Blending arises from the discrepancies between the actor's head and the target body, including differences in head shape and hair structure. These discrepancies often lead to unnatural boundaries or blending artifacts, which can be particularly problematic in professional applications where high fidelity and visual coherence are ultimate. The existing method, Head2Scene Blender (H2SB) [16], approaches this task by treating the foreground and background generation as a single process. H2SB shows unsatisfactory results (Figure 2(a), (b)), especially around the boundary regions. Although the generation region has two distinct background and foreground parts, H2SB considers the region at once, which results in an unclear border of a human and artifacts. Moreover, H2SB lacks in mimicking the cross-identity head blending and fails to cover large inpainting regions.

To this end, we propose ***CHANGER***, a novel pipeline for ***C***onsistent ***H***ead blending with predictive ***A***tte***N***tion ***G***uided foreground ***E***stimation under chroma key setting for indust***R***ial applications. We decompose the problem into two distinct sub-tasks: background integration and foreground blending. This decomposition allows for a more focused treatment of each aspect of the task, ensuring higher fidelity in both the background and foreground.

The background integration challenge is addressed by incorporating chroma keying [14], a widely used technique in content production where a uniformly colored background (e.g., a green screen) is replaced with the desired scene. This allows for flawless background generation, eliminating the artifacts that arise when the foreground and background are blended simultaneously. By decoupling the foreground blending from the background, we ensure that the visual integrity of the scene is preserved, even in complex environments.

For the foreground blending, we tackle the problem of seamlessly integrating the actor's head onto the body of the target, particularly in cases where significant differences exist in head shape and hair structure. To generate the high-fidelity foreground, we devise two contributions, one from a data-centric and the other from a model-centric perspective.

First, we propose a novel data augmentation technique called ***H****ead shape and long* ***H****air* ***augmentation*** (***H^2 augmentation***), which simulates a wide range of head shapes and hair styles in the self-supervised training. This enables our model to better generalize to real-world variations and handle the significant visual discrepancies that often arise in professional content production.

Second, we introduce the ***F***oreground ***P***redictive ***A***ttention ***T***ransformer (***FPAT***), a novel architecture designed for foreground blending. FPAT predicts the exact regions of the head and body that require attention and apply targeted attention to these areas during the blending process. By explicitly restricting the attention to these key regions, FPAT enhances the blending quality, particularly in areas where head shape and hair differences pose a challenge.

To summarize, we propose the first comprehensive solution for the Head Blending task in industrial content produc-

tion. Unlike the previous approach that treats this process as part of general head or face swapping, our method focuses explicitly on the seamless blending of an actor's head with the target body, ensuring realistic and high-quality results. Our method, CHANGER, significantly outperforms state-of-the-art techniques, as demonstrated through both quantitative metrics and qualitative evaluations on benchmark datasets.

In summary, our main contributions are as follows:

- We propose CHANGER, a novel pipeline that utilizes chroma keying for the first time to decouple background integration from the head blending process, addressing the common artifacts seen in prior methods.
- We introduce H^2 augmentation, a data-centric approach designed to handle significant variations in head shape and hair structure, enhancing the robustness of the Head Blending process.
- We present the FPAT module, which uses predictive attention to focus on key regions of the head and body, resulting in high-fidelity blending with minimized artifacts.
- CHANGER significantly outperforms existing methods on benchmark datasets, both quantitatively and qualitatively, showcasing its effectiveness in industrial content production scenarios.

2. Related Work

Head Blending. Head Blending aims to replace the head in a target image with a source head, ensuring that the result is seamless and maintains the consistency of the skin color of the target image. To address this problem, H2SB [16] proposes a semantic-guided color reference creation module based on [23] for re-coloring the head and filling the neck and the background at once. However, we empirically find that H2SB results in inadequate generation results which is unsuitable for real-world application. H2SB [16] relies on a single feature correspondence matrix, which is proposed for image translation [23] and has a simple U-Net structure. In contrast, we propose a novel approach that manages the background with chroma keying for high-quality and efficient background changes.

Mask-Aware Transformers. Transformer [18] is a model that processes a sequence of tokens with an attention mechanism. The original attention in the transformer computes the similarity between the query and the key, where all tokens have participated. On the other hand, there are lines of work that restrict the region of computing attention using the pre-defined masks that reflect the prior knowledge of the task. Transformer decoder block [18] uses causal attention in a transformer decoder to block that later tokens affect the previous tokens to be generated. Swin Transformers [9] and ConViT [5] restrict or prioritize the attention region in spatially closed patches to inject the spatial inductive bias to the model, and OAMixer [6] reweighs the attention with a soft mask to strengthen the relationship between semantically related tokens to improve generalization and mitigates with background bias. In this work, we design a novel foreground predictive attention transformer (FPAT) by leveraging a neck and body region prediction module. FPAT differs from previous mask-aware transformers in that the mask for attention is not given, but we predict the foreground mask implicitly within training our head blending pipeline.

3. Method

In this section, we present the detailed methodology behind CHANGER, our novel framework for Head Blending. CHANGER addresses the primary challenges of blending an actor's (source) head onto a target body by decomposing the task into two key sub-tasks: background integration and foreground blending. CHANGER handles the background integration via combining chroma keying. We detail the network input and output preparation for chroma keying in Section 3.1. To generate high-fidelity foregrounds, we propose new data augmentation and model design. To address the constraints of self-supervised training that uses the same images for both source and target, where cross-identity settings lack ground truth, we propose a novel augmentation method called H^2 augmentation. This technique broadens the diversity of the input X, enhancing the ability of the model to adapt to various identities. Further details on this approach can be found in Section 3.2. In Section 3.3, we detail our Foreground Predictive Attention Transformer (FPAT) which enhances foreground blending. The overall network of CHANGER is shown in Figure 3.

3.1. Chroma Keying for Head Blending

We propose a chroma key setting for a head blending task to divide the labor of generating the background region to chroma keying. To this end, we modify the input of the network X to have a green background, ensuring the output of the network Y also maintains this green background as shown in Figure 3(a). To do so, we first paint the background of the target image I_T as green and acquire I_T^{green} by finding the foreground with the state-of-the-art face parsing network [22]. Then, we extract the head mask of the source, $\mathtt{M}_S^{head}$, and the target head mask $\mathtt{M}_T^{head}$. We obtain a union mask of the source head mask $\mathtt{M}_S^{head}$ and the target head mask $\mathtt{M}_T^{head}$ as follows:

$$\mathtt{M}_{union}^{head} = \begin{cases} \mathtt{M}_S^{head} \oplus \mathtt{M}_{h^2}^{head}, & \text{(train)}, \\ \mathtt{M}_S^{head} \oplus \mathtt{M}_T^{head}, & \text{(test)}, \end{cases} \tag{1}$$

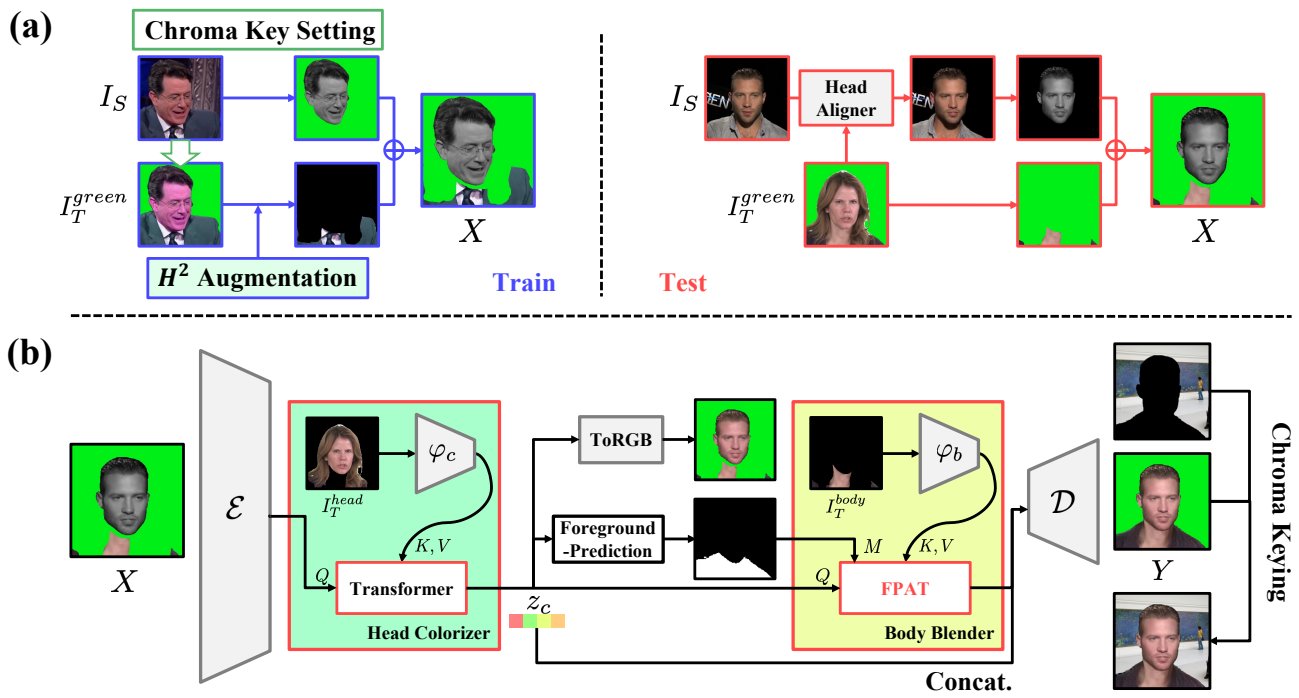

Figure 3. **Network overview of CHANGER.** (a) We visualize how we conduct the input of the network (X) at the train (blue) and the test (red). We apply H^2 augmentation during the training to improve the fidelity of the generated image by improving the diversity of the input. (b) We visualize the network of CHANGER. The head colorizer colorizes the gray head of X, and the body blender inpaints the hidden body with a foreground mask-aware attention mechanism. Please refer to the detailed explanations of H^2 augmentation and FPAT in Section 3.2 and 3.3, respectively.

where $\oplus$ is a union operation. Note that we use $\mathtt{M}_{h^2}^{head}$, which is the output of H^2 augmentation during training to obtain $\mathtt{M}_{union}^{head}$ since the M_S^{head} and M_T^{head} is identical in the self-supervised setting. Finally, we formulate input X as the following equation:

$$X = I_S^{gray} + I_T^{green} \otimes (1 - \mathtt{M}_{union}^{head}) + I^{green} \otimes \mathtt{M}^{ip}, \quad (2)$$

where the gray-scale source head, $I_S^{gray} = \mathtt{g}(I_S \otimes \mathtt{M}_S^{head})$, where $\mathtt{g}(x)$ indicates a gray-scaling function and $\otimes$ is the pixel-wise multiplication, and an inpainting mask $\mathtt{M}^{ip} = \mathtt{M}_{union}^{head} - \mathtt{M}_S^{head}$. During test environments, any targets are acceptable in our pipeline. If the target is filmed on a green screen, we directly apply our CHANGER.

3.2. H^2 Augmentation

Since we train the model in a self-driven manner during training, the target image is generated from the source image. We propose a simple but powerful H^2 augmentation that manipulates the input X during the self-identity head blending training to simulate the various cross-identity blending scenarios, especially the settings where the foreground blending region is large. Existing methods lack variation in inpainting regions, critical for self-driven training. To tackle the issue, we carefully designed promising computer vision techniques and the stochastic sampling method. Since the head shape and the hair difference between the source and the target generates a large mismatch region, H^2 augmentation includes a head shape and a long hair augmentation.

Head Shape Augmentation. Since an outline of a source head differs from a target head in the cross-identity blending scenarios, an empty region between a source head and a target body is quite diverse. To mimic possible mismatches appearing in the blending procedure under a self-supervised manner, we randomly augment the region by transforming the source head mask with a head shape augmentation $\mathcal{T}_{head}$, which includes an affine transformation, squeezing, expanding, and varying dilation widths. Therefore, from the source head mask $\mathtt{M}_S^{head}$, we obtain the augmented head mask as following:

$$\mathtt{M}_{h^1}^{head} = \mathcal{T}_{head}(\mathtt{M}_S^{head}). \quad (3)$$

Long Hair Augmentation. To mimic the hair differences in the cross-identity setting, we randomly sample an identity whose hair is long enough to cover its clothing and body. With the hair mask of the sampled identities $\mathtt{M}_{long}^{hair}$, we apply the long hair augmentation $\mathcal{T}_{hair}$ to the augmented head

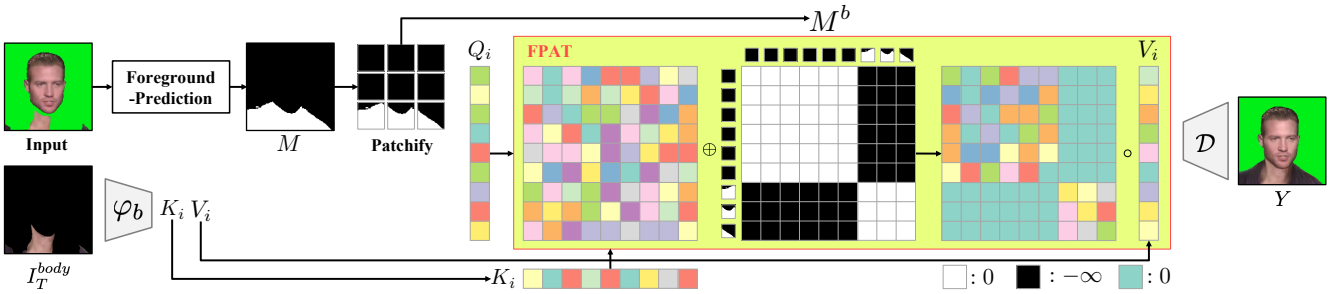

Figure 4. **Visualization of attention mechanism of our Foreground Predictive Attention Transformer (FPAT) block.** The Foreground-Prediction module predicts the foreground mask M of the body and the neck region, and the attention is reweighted according to M.

mask $\mathtt{M}_{h^1}^{head}$, and get the final head mask as follows:

$$\mathtt{M}_{h^2}^{head} = \mathcal{T}_{hair}(\mathtt{M}_{h^1}^{head}) = \begin{cases} \mathtt{M}_{h^1}^{head} \oplus \mathtt{M}_{long}^{hair}, & \text{if } p \geq \epsilon, \\ \mathtt{M}_{h^1}^{head}, & \text{otherwise,} \end{cases} \tag{4}$$

where p is sampled from uniform distribution and $\epsilon \in (0, 1)$ is a fixed threshold. The visualizations of Eq. (3) and (4) are shown in our supplementary material. With the augmented head mask $\mathtt{M}_{h^2}^{head}$, we obtain the union mask with the source head mask $\mathtt{M}_{union}^{head} = \mathtt{M}_{h^2}^{head} \oplus \mathtt{M}_{S}^{head}$ then produce the input X with Eq. (2).

Our proposed H^2 augmentation creates various union masks from the source head and enables diverse neck and body completion regions in a self-supervised manner. H^2 augmentation is a delicate solution that addresses unique characteristics during head blending training. The ablation study in Table 4 and Figure 6 demonstrates the novelty of H^2 augmentation.

3.3. Foreground Predictive Attention Transformer

The network architecture for the foreground blending is divided into two components: (1) Head Colorizer that transfers the color of the target to the gray-scaled source head, and (2) Body Blender that generates the body for a seamless connection between the source head and the target body via our novel Foreground Predictive Attention Transformer (FPAT).

Head Colorizer in Figure 3(b) transfers the color from the head of the target image into the head of the source image. Since the input of the network X has a gray head, the model should colorize the head of the input by referring to the conditioned target image. Head colorizer is composed of cross-attention transformer blocks with a query from the embedded features of the input X using the encoder $\mathcal{E}$ and the key and the value from the target head I_T^{head} using φ_c, which is a conditional projection embedder [3, 4, 13, 18]. The head colorizer outputs the intermediate hidden representation $z_c \in \mathbb{R}^{C \times h \times w}$, where C is the number of channels and (h, w) is the resolution of the feature.

The Body Blender in Figure 3(b) generates the occluded body region by incorporating our novel FPAT. Here, the body blender aims to ensure coherent edge continuation and seamless head-body connection, while avoiding inappropriate influences from background or mismatched regions. Therefore, the body blender requires a sophisticated attention mechanism that computes region-selective attention: for occluded clothing, exclusively from other clothing regions; for the head-body junction, solely from the facial area. To address these distinctive requirements, we introduce the Foreground Predictive Attention Transformer (FPAT), a novel architecture that redefines masked attention in the context of head blending.

FPAT computes the masked attention between tokens from the intermediate feature of the colorized head (z_c) to the feature of the target body I_T^{body}. Here, the masked attention is applied within respective regions, i.e., foreground to foreground and background to background, respectively. FPAT starts with the output of the head colorizer z_c, and predicts a foreground region as a binary mask $M \in \mathbb{R}^{h \times w}$, as shown in Figure 4. Then, FPAT patchifies the mask M with N patches; M_i for $i \in \{1, ..., N\}$.

With the patchified mask, FPAT computes the binary attention mask M^b as following:

$$M_{ij}^b = \begin{cases} 0, & \text{if } M_i \text{ and } M_j \text{ are the same type of patches,} \\ -\infty, & \text{otherwise,} \end{cases} \tag{5}$$

where 'the same type of patches' refers to pairs of patches that are either both classified as foreground or both as background, and M_{ij}^b is the (i, j)-th element of M^b.

FPAT masks the attention between a query from the latent representation z_c^p, key and value from the target head feature z_{body}^p as follows:

$$\text{FPAT}(Q_i,\ K_i,\ V_i) = \text{softmax}(\frac{Q_i K_i^T}{\sqrt{d_k}} + M^b) \cdot V_i, \tag{6}$$

where $Q_i = z_c^p \cdot W_i^Q$, $K_i = z_{body}^p \cdot W_i^K$, and $V_i = z_{body}^p \cdot W_i^V$. The dimension of learnable projection parameter matrices [18] is $W_i^Q \in \mathbb{R}^{N \times D_K}$, $W_i^K \in \mathbb{R}^{N \times D_K}$, and $W_i^V \in \mathbb{R}^{N \times D_V}$, respectively.

Method	PSNR ↑	LPIPS ↓	L_1 ↓	SSIM ↑	FPS ↑	MACs ↓	Param. ↓
H2SB [16]	12.397	0.134	0.125	0.743	28.10	122.07G	24.37M
H2SB [16] + CK	12.974	0.086	0.119	0.737			
Ours	**27.845**	**0.011**	**0.014**	**0.950**	**60.57**	**81.73G**	**8.92M**

Table 1. **Quantitative comparison with H2SB [16] and our CHANGER.** "CK" is an abbreviation of chroma keying.

Method	BG ↑	ID ↑	Natural ↑	Holistic ↑
H2SB [16]	0.696	1.234	1.035	0.725
H2SB [16] + CK	0.720	1.188	0.847	0.642
Ours	**1.110**	**1.300**	**1.226**	**1.091**

Table 2. **Quantitative comparison from the user study.** "CK" is an abbreviation of chroma keying.

Note that FPAT updates the body and the neck parts of the hidden representation z_c^p by only referring to the body and the neck parts of z_{body}^p. Our proposed FPAT generates high-fidelity foregrounds by restricting the attention region with the predicted regions.

3.4. Training Objectives

We formulate z by channel-wisely concatenating the output of the head colorizer z_c and the output of the body blender z_b. z is the input for the decoder $\mathcal{D}$:

$$\hat{Y} = \mathcal{D}(z), \tag{7}$$

where $\hat{Y}$ contains spatial information of the colorized head from I_S^{gray} and the the body completion.

We define the final output Y by utilizing I_T as follows:

$$Y = \hat{Y} \otimes \mathtt{M}_{union}^{head} + I_T \otimes (1 - \mathtt{M}_{union}^{head}). \tag{8}$$

We use five loss functions, **(1)** $\mathcal{L}_{rec} = ||Y \otimes \mathtt{M}_S^{head} - I_T \otimes \mathtt{M}_S^{head}||_1$, the reconstruction loss for the final output head and the ground truth, **(2)** $\mathcal{L}_{hc} = ||Y - I_T^{hc}||_1$, the reconstruction loss for the output of `ToRGB` block, **(3)** $\mathcal{L}_{mask} = ||M_{gt} - M||_1$ [10], the loss for the output of the Foreground-Prediction module where M_{gt} is the ground-truth full body mask, **(4)** perceptual loss $\mathcal{L}_{per} = \sum_{i=1}^{L} ||\Phi_i(Y) - \Phi_i(I_T)||_1$, and **(5)** adversarial loss $\mathcal{L}_{adv}$. I_T^{hc} is a target image without neck, and body completion and Φ is a pre-trained VGG19 network [17].

The final objective function is as follows:

$$\begin{aligned}\mathcal{L}_{total} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{hc}\mathcal{L}_{hc} + \lambda_{mask}\mathcal{L}_{mask} \\ + \lambda_{per}\mathcal{L}_{per} + \lambda_{adv}\mathcal{L}_{adv},\end{aligned} \tag{9}$$

where λ_{rec}, λ_{hc}, λ_{mask}, λ_{per}, and λ_{adv} are weights for the loss $\mathcal{L}_{rec}$, $\mathcal{L}_{hc}$, $\mathcal{L}_{mask}$, $\mathcal{L}_{per}$, and $\mathcal{L}_{adv}$, respectively.

4. Experiments

Implementation Details. We combined three different benchmark datasets with training and testing our CHANGER and the state-of-the-art model H2SB [16]: **(1)** VoxCeleb1 [11], **(2)** VoxCeleb2 [2], and **(3)** HDTF [24]. An Adam optimizer [7] with hyperparameters of $\beta_1 = 0.9$ and $\beta_2 = 0.999$ was used for every models. We usedlearning rates 1e-4 and 4e-4 for the generator and discriminator, respectively. We used $\epsilon = 0.5$ in Eq. (4). We used 4 NVIDIA RTX 3090 (24 GB) GPUs to train our CHANGER. Due to the limitations of GPU resources, the experiments were conducted on 256-resolution images. However, the core components of CHANGER, H^2 augmentation and FPAT, are inherently resolution-agnostic. This design ensures that our CHANGER can seamlessly scale to higher resolutions for future applications.

4.1. Comparison with the State-of-the-Art

Quantitative Comparison. In order to provide a clear benchmark for advancement in the specific domain of head blending, we quantitatively compared our method with H2SB [16], the only existing method designed for the head blending task. Moreover, we generated the self-blending results and prepared a ground truth to evaluate the performance quantitatively.

Table 1 shows that our CHANGER outperforms both H2SB and H2SB + CK quantitatively. In addition, we analyzed the computational efficiency of our method and the baseline method in terms of inference time. Our method achieved comparable speed performance (2.2 times faster *FPS* than H2SB) to the baseline methods while requiring only a fraction of the computational cost (33% less *MACs* than H2SB) and 64% fewer parameters (*Param.*). These findings demonstrate the practical feasibility of our method for real-world applications.

To assess perceptual quality of the results, we conducted a user study referring to the state-of-the-art network [16], which rates (1) the fidelity of background regions (*BG*), (2) the identity similarity according to the head skin colorization fidelity (*ID*), (3) the naturalness of generated neck and body (*Natural*), and (4) overall perceptual qualities of head blending images (*Holistic*). We attached the user study material in supplementary.

The user study results from 21 human evaluators rating of 0 to 2 for various criteria are shown in Table 2. In the

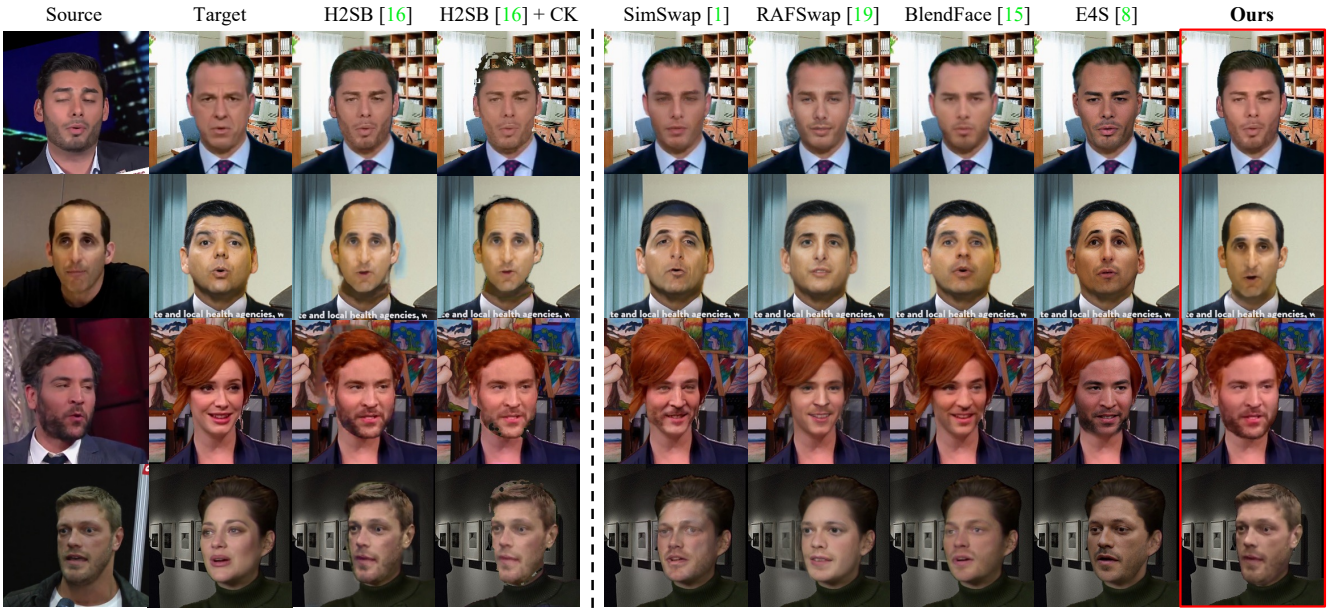

Figure 5. We compared qualitative results between CHANGER with the state-of-the-art head blending and face swapping models: H2SB [16], SimSwap [1], RAFSwap [19], BlendFace [15], and E4S [8]. "CK" represents inference on the chromakey configuration.

blind test between our CHANGER with H2SB and H2SB $+ CK$, most of the users ranked our head blending results highest as depicted in the table, even though duplication pick was allowed. Ours outperformed the baselines regarding identity preservation and the fidelity of the head and background region with overall user satisfaction.

Qualitative Comparison. Qualitative comparisons were conducted more extensively, encompassing a wider range of comparative face swapping methods.

Figure 5 shows the head blending results of various source and target images against the state-of-the-art head blending and face swapping frameworks. As shown in Figure 5, our CHANGER shows high-fidelity head blending results with high-quality backgrounds. H2SB shows poor background results. Although H2SB with our chroma keying (H2SB $+ CK$) shows slightly improved background outcomes, it still suffers from low-fidelity results on body blending. The body blending results are improved significantly in Ours due to the proposed H^2 augmentation and FPAT.

4.2. Ablation Study

In our ablation study, we examined the respective efficacy of the CHANGER components by removing or altering the key components. Table 3 and Table 4 show the performance of the possible variants of CHANGER based on our main proposal and the details on H^2 augmentation, respectively.

Quantitative Results. We explored the effectiveness of our key components by removing H^2 augmentation (model "A") and FPAT (model "B") as shown in Table 3. We replaced FPAT with a conventional cross-attention transformer for model "B". Quantitative results show that our full model achieves the best performance in overall metrics.

Type	FPAT	H^2	PSNR ↑	LPIPS ↓	L_1 ↓	SSIM ↑
A	✓		16.965	0.122	0.067	0.863
B		✓	27.199	0.012	0.015	0.949
Ours	✓	✓	**27.845**	**0.011**	**0.014**	**0.950**

Table 3. **Ablation study on our proposal.** Quantitative results when excluding the proposed H^2 augmentation and FPAT, respectively.

We evaluated the components in our proposed H^2 augmentation: (1) the head shape augmentation ("Head") and (2) the long hair augmentation ("Hair") in Table 4. Model "C" (without the long hair augmentation) showed better performance than model "D" (without the head shape augmentation) in terms of SSIM and PSNR. For LPIPS, model "D" shows better results. These results imply that head shape augmentation leads the better accuracy in the reconstruction of the image, while long hair augmentation allows the better perceptual quality.

Qualitative Results. We also qualitatively analyzed the effect of each component by conducting cross-head blending to examine the large differences in head shape and hair, i.e., v-chin to u-chin and short-hair to long-hair as depicted in Figure 6. The head shape augmentation allowed the generation of a more acceptable neckline (see top results of models "B" and "C" compared to models "A" and "D"). On the

Type	Head	Hair	PSNR ↑	LPIPS ↓	L_1 ↓	SSIM ↑
C	✓		26.437	0.023	0.020	0.948
D		✓	25.037	0.018	0.020	0.932
Ours	✓	✓	**27.845**	**0.011**	**0.014**	**0.950**

Table 4. **Ablation study on H^2 augmentation.** All combinations made by the two components of H^2 augmentation were re-trained and evaluated. "Head" is the head shape augmentation. "Hair" is the long hair augmentation.

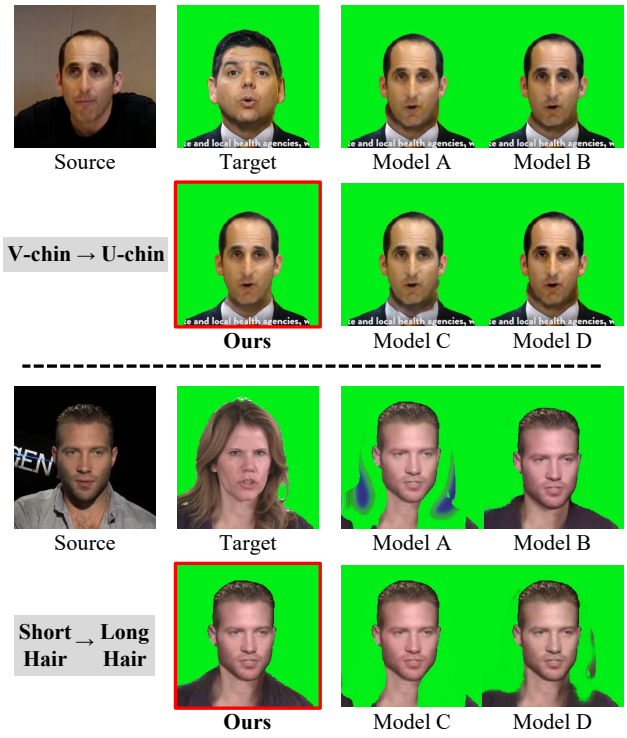

Figure 6. **Ablation study on model designs.** Both the head shape augmentation and the long hair augmentation are crucial components for high-fidelity head blended images. FPAT allows fine-grained foreground generation.

other hand, the long-hair augmentation tended to the better generation of hidden regions due to the long hair of the target (see bottom results of models "B" and ours compared to models "A" and "C"). Without H^2 augmentation or employing either one, both head blending quality and fidelity deteriorate (see models "A", "C", and "D"). With FPAT, the model showed better foreground blending compared to ours and model "B". The proposed H^2 augmentation and FPAT contributed significantly to performing head blending in the chroma key setting.

4.3. Discussions

Limitations. As shown in Figure 7, CHANGER encounters challenges under certain extreme conditions where the target image has too rich hair so that the body region is

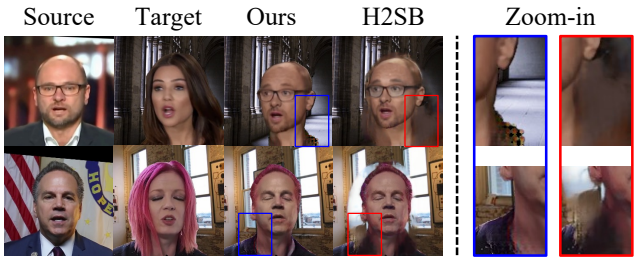

Figure 7. **Limitations.** When the target has certain extreme attributes such as too rich hair or pinkish hair, CHANGER suffers artifacts on generated body regions.

Figure 8. **Various industrial application examples.** By leveraging chroma key technique with our proposed CHANGER pipeline, we can obtain various high-fidelity head blended videos in the wild environments. The red boxes represent the source images.

largely hidden. Also, CHANGER sometimes fails to generate high-fidelity head blending results when the target has extreme attributes, e.g., pinkish hair. Despite these imperfections, we emphasize that across all rows in Figure 5, the results of CHANGER consistently surpass the achievements of H2SB, the state-of-the-art of head blending.

Social Impacts. As shown in Figure 8, our CHANGER pipeline can achieve various high-fidelity industrial content productions in the wild. However, due to its high performance, our technology might cause cultural, political, and ethical social problems, such as indistinguishable deep fake videos, invasion of privacy, or even defamation.

5. Conclusion

In this work, we presented CHANGER, a novel head blending pipeline for high-fidelity industrial content production within chroma key settings for the first time. Our approach, proposing H^2 Augmentation and Foreground Predictive Attention Transformer (FPAT) led to realistic and seamless foreground blending. Through various experiments and comparative analysis, we demonstrated that CHANGER offers significant qualitative and quantitative improvements over the state-of-the-art model. We believe that the superiority and cost-effectiveness of CHANGER pave the way for its adoption in real-world applications, offering robust solutions for generating high-quality head blending contents.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2710018251), and Korea Planning & Evaluation Institute of Industrial Technology (KEIT) grant funded by the Korea government (MOTIE) (RS-2024-00444344), and in part by the IITP grant funded by the Korea Government (MSIT) (Artificial Intelligence Innovation Hub) under Grant 2021-0-02068, and by the IITP grant funded by the Korea government (MSIT) (No.RS-2022-00155911, Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University)).

References

[1] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2003–2011, 2020. 7

[2] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 6

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 5

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 5

[5] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021. 3

[6] Hyunwoo Kang, Sangwoo Mo, and Jinwoo Shin. Oamixer: Object-aware mixing layer for vision transformers. *arXiv preprint arXiv:2212.06595*, 2022. 3

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[8] Zhian Liu, Maomao Li, Yong Zhang, Cairong Wang, Qi Zhang, Jue Wang, and Yongwei Nie. Fine-grained face swapping via regional gan inversion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8578–8587, 2023. 7

[9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[10] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 6, 2

[11] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017. 6

[12] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE conference on computer vision and pattern recognition*, pages 413–420. IEEE, 2009. 2

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 5, 1, 2

[14] Shigeru Shimoda, Masaki Hayashi, and Yasuaki Kanatsugu. New chroma-key imagining technique with hi-vision background. *IEEE Transactions on broadcasting*, 35(4):357–361, 1989. 2

[15] Kaede Shiohara, Xingchao Yang, and Takafumi Taketomi. Blendface: Re-designing identity encoders for face-swapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7634–7644, 2023. 7

[16] Changyong Shu, Hemao Wu, Hang Zhou, Jiaming Liu, Zhibin Hong, Changxing Ding, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Few-shot head swapping in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10789–10798, 2022. 2, 3, 6, 7

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6, 3

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 5

[19] Chao Xu, Jiangning Zhang, Miao Hua, Qian He, Zili Yi, and Yong Liu. Region-aware face swapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7632–7641, 2022. 7

[20] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 1, 2

[21] Sahng-Min Yoo, Tae-Min Choi, Jae-Woo Choi, and Jong-Hwan Kim. Fastswap: A lightweight one-stage framework for real-time face swapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3558–3567, 2023. 3

[22] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 3

[23] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 3

[24] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 6

[25] Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10477–10486, 2023. 1, 2

Supplementary Material

This supplementary document provides an intensive insight into our work presented in the main paper, consisting of qualitative comparisons with the state-of-the-art inpainting methods, the notation and visualization of our Head shape and long Hair (H^2) augmentation, more detailed analysis and descriptions of the proposed Foreground Predictive Attention Transformer (FPAT), implementation details on the training objectives, experimental details on the user study, and a head blending video results in our project page.

A. Qualitative Comparisons with Recent Inpainting Models

In this section, we investigate the performance of our CHANGER compared to the state-of-the-art inpainting models through qualitative comparisons.

Baselines. We establish the state-of-the-art inpainting models as follows: (1) Stable Diffusion Inpainting (SDI) [13], (2) Paint-by-Example (PBE) [20], (3) ProPainter [25].

Figure 1 shows the results from head blending video compared with the recent diffusion-based or video-based inpainting models. *SDI* and *PBE* mainly suffered from background generation (green boxes) and artifacts of the foreground region. *ProPainter* showed blurry foreground generation (orange boxes). Our results show not only the highest fidelity in the background inpainting region as well as the foreground but also stability in a time-consistency perspective, which ensures the quality of the video output.

B. Notation and Visualization Summary of H^2 Augmentation

We provide a detailed explanation of the various notations used in our method, especially for the proposed Head shape and long Hair (H^2) augmentation, in Table 1. We also visualize the process of H^2 augmentation in Figure 2. Please refer to the descriptions in the table and figure, to ensure clarity in interpreting our work.

C. More Details in FPAT

C.1. Attention Map of FPAT

The Foreground Predictive Attention Transformer (FPAT) stands at the forefront of our model structure, primarily focusing on the enhancement of the fidelity of foreground blending. In the diverse situations created by various head shape and hairstyle differences between the source and target images, FPAT aims to predict the foreground region and then attend to the predicted foreground region. We demonstrate the effectiveness of the FPAT qualitatively.

Figure 3 presents predicted masks (M), attention maps (*Attention*), and head blending results (Y) obtained by FPAT on various source and target pairs. For each input, two distinct attention maps are depicted: one for the neck (upper row) and another for the cloth (lower row). The small red boxes inside the images in the X column represent the patches used to generate queries for our proposed FPAT transformer layer. The images in the *Attention* column depict the calculated attention derived from these queries and keys, where higher values are represented closer to yellow and lower values closer to blue.

The predicted mask results show that FPAT effectively reconstructs obscured foreground areas caused by long hair. Furthermore, meaningful attention is trained within the predicted region, as depicted in the attention maps. Specifically, during the generation of the neck region (upper row), the model focuses explicitly on the neck area of the target image. In contrast, when generating the occluded attire region (lower row), the model focuses on relevant clothing areas, indicating its ability to create images with attention to pertinent regions.

C.2. Detailed Explanation of FPAT Mechanism

Our FPAT starts with the input z_c, and predicts a foreground region, including the body and the neck, as a binary mask $M \in \mathbb{R}^{h \times w}$ with Foreground-Prediction module. The FPAT block refers to the target body information I_T^{body} and updates z_c using the information of M to generate the neck and body via the Foreground-Aware Transformer block. FPAT patchifies the feature output of the head colorizer $z_c \in \mathbb{R}^{C \times h \times w}$ and get $z_c^p \in \mathbb{R}^{N \times P^2 C}$, where (P, P) is the resolution of the patches and $N = hw/P^2$ is the number of patches. FPAT also patchifies the embedded feature of the target body I_T^{body} as $z_{body}^p \in \mathbb{R}^{N \times P^2 C}$, and the predicted body and neck mask M as $M^p \in \mathbb{R}^{N \times P^2}$. Then, FPAT averages M^p along the channel axis to acquire $M_{\text{avg}}^p \in \mathbb{R}^N$ as following:

$$[M_{\text{avg}}^p]_n = \frac{1}{P^2} \Sigma_{m=1}^{P^2} M_{nm}^p, \tag{10}$$

where $[M_{\text{avg}}^p]_n$ is the n-th patch of M_{avg}^p and M_{nm}^p is the (n, m)-th element of M^p. Next, we divide patches into two groups: (1) a set of patches S_b that includes the predicted body and neck parts and (2) a set of patches S_{nb} that does not include them by thresholding M_{avg}^p by following:

$$\begin{aligned} S_b &= \{i \in 1, ..., N \mid [M_{\text{avg}}^p]_i \geq \tau\} \\ S_{nb} &= \{i \in 1, ..., N \mid [M_{\text{avg}}^p]_i < \tau\}, \end{aligned} \tag{11}$$

where τ is the hyperparameter. Then, FPAT computes the binary mask $M^b \in \mathbb{R}^{N \times N}$ as following:

$$M_{ij}^b = \begin{cases} 0, & \text{if } i, j \in S_b \ , \ i, j \in S_{nb}, \\ -\infty, & \text{otherwise}, \end{cases} \tag{12}$$

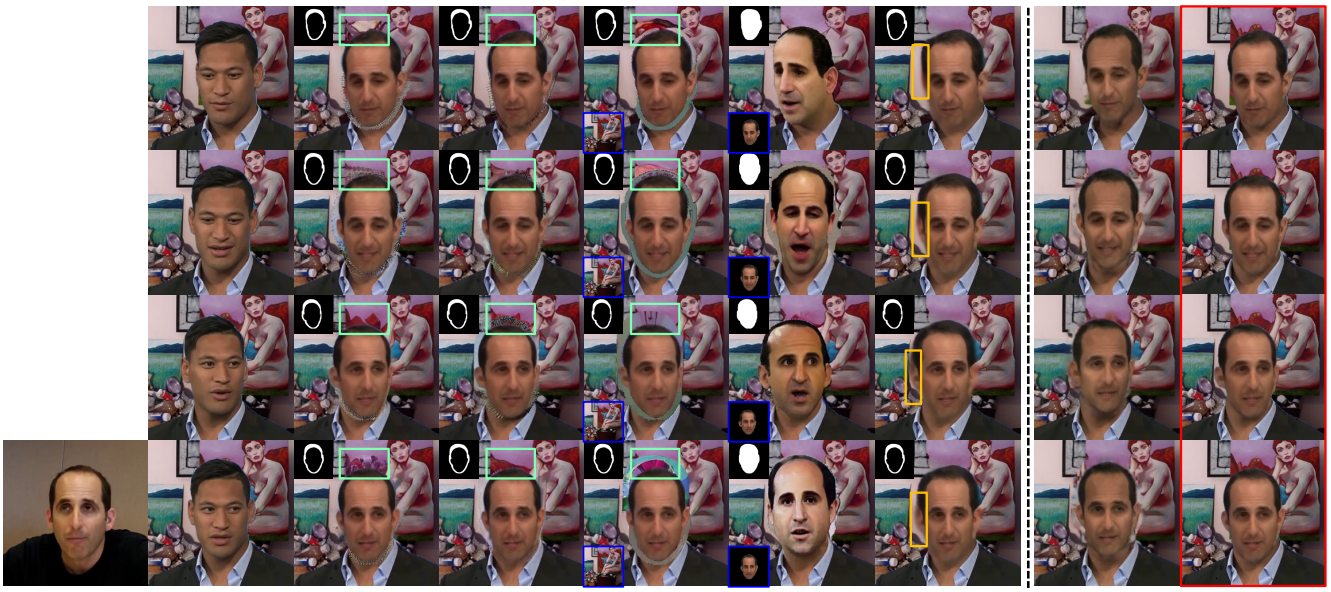

Figure 1. Qualitative comparisons of using recent inpainting baselines [13, 20, 25] and the head blending model [16] on sequential frames of a target video. We tested both scenarios with and without text prompting (Prompt) for *SDI*. For *PBE*, we separated scenarios; the background (BG) and the foreground (FG) references (bottom-left blue boxes of each column).

Notation	Dimension	Description
I_S	$\mathbb{R}^{3\times H\times W}$	Source image.
I_T	$\mathbb{R}^{3\times H\times W}$	Target image.
I_S^{gray}	$\mathbb{R}^{1\times H\times W}$	Gray-scale image from source.
I_T^{green}	$\mathbb{R}^{3\times H\times W}$	Target image with a green screen background.
I_T^{head}	$\mathbb{R}^{3\times H\times W}$	Target head image, used in Head Colorizer, made by only leaving the head region from the target image.
I_T^{body}	$\mathbb{R}^{3\times H\times W}$	Target body image, used in Body Blender, made by excluding head, neck, and background.
$\mathtt{M}_S^{head}$	$\mathbb{R}^{1\times H\times W}$	Head mask from source.
$\mathtt{M}_T^{head}$	$\mathbb{R}^{1\times H\times W}$	Head mask from target.
$\mathtt{M}_{h^1}^{head}$	$\mathbb{R}^{1\times H\times W}$	Augmented head mask made by transformation $\mathcal{T}_{head}$.
$\mathtt{M}_{h^2}^{head}$	$\mathbb{R}^{1\times H\times W}$	Augmented head mask from $\mathtt{M}_{h^1}^{head}$, made by transformation $\mathcal{T}_{hair}$.
$\mathtt{M}_{union}^{head}$	$\mathbb{R}^{1\times H\times W}$	Union mask of $\mathtt{M}_{h^2}^{head}$ and $\mathtt{M}_S^{head}$ during training, union mask of $\mathtt{M}_T^{head}$ and $\mathtt{M}_S^{head}$ during testing.
$\mathtt{M}^{ip}$	$\mathbb{R}^{1\times H\times W}$	Inpainting region subtracting $\mathtt{M}_S^{head}$ from $\mathtt{M}_{union}^{head}$.
M	$\mathbb{R}^{1\times H\times W}$	Predicted foreground mask which is further used as an input of the FPAT blocks.
X	$\mathbb{R}^{3\times H\times W}$	Input for our CHANGER.
Y	$\mathbb{R}^{3\times H\times W}$	Head blended outputs of our CHANGER.

Table 1. **Notations and corresponding descriptions in our CHANGER.**

where M_{ij}^b is the (i,j)-th element of M^b.

Finally, FPAT masks the attention between a query from the latent representation z_c^p, key and value from the target head feature z_{body}^p.

D. Training Objectives Details

We train the model with $\mathcal{L}_{total}$, which is a summation of **(1)** $\mathcal{L}_{rec}$, the reconstruction loss for the final output head and the ground truth, **(2)** $\mathcal{L}_{hc}$, the reconstruction loss for the output of `ToRGB` block, **(3)** $\mathcal{L}_{mask}$ [10], the loss for the

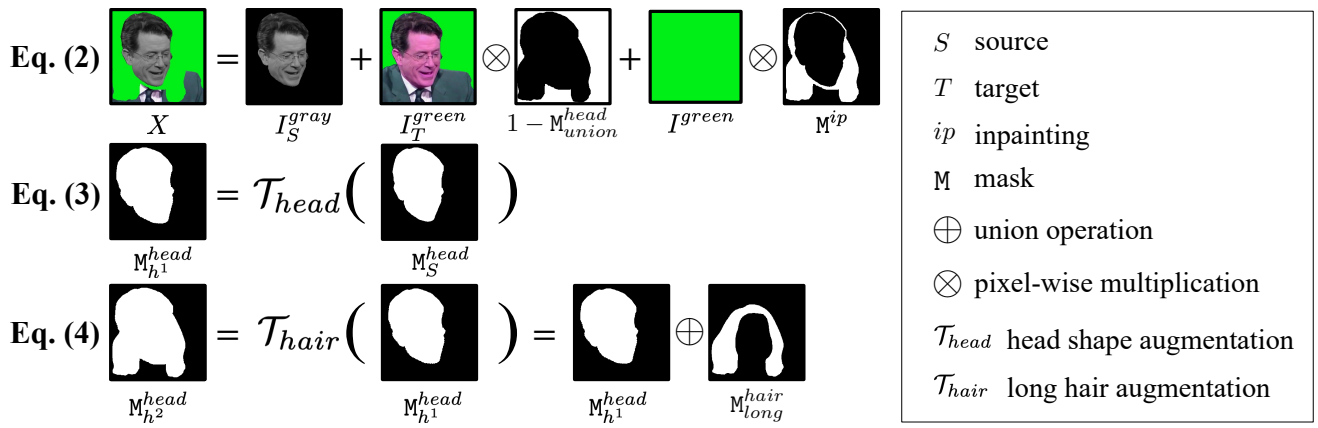

Figure 2. **Visualization of H^2 Augmentation.** Eq. (2) is the input X formulation during training. Inspired by [21], we apply the same color jitter to both I_T^{green} and the ground truth during the training phase. Eq. (3) shows the head shape augmentation. Eq. (4) shows the long hair augmentation.

Figure 3. The foreground mask predicted by FPAT (M), the attention map used in the transformer layer (*Attention*), and the head blending result (Y) when input source image I_S and target image I_T are used. We visualize the similarity between the query patch (red box) and each key patch in the depicted image as an attention map. Blue represents low values and yellow represents high values.

output of the Foreground-Prediction module, **(4)** perceptual loss $\mathcal{L}_{per}$, and **(5)** adversarial loss $\mathcal{L}_{adv}$ for our objective functions.

Corresponding objective functions are as follows:

$$\begin{aligned}\mathcal{L}_{total} &= \lambda_{rec}\mathcal{L}_{rec} + \lambda_{hc}\mathcal{L}_{hc} + \lambda_{mask}\mathcal{L}_{mask} \\ &\quad + \lambda_{per}\mathcal{L}_{per} + \lambda_{adv}\mathcal{L}_{adv}, \end{aligned} \tag{13}$$

$$\mathcal{L}_{rec} = ||Y \otimes \mathtt{M}_S^{head} - I_T \otimes \mathtt{M}_S^{head}||_1, \tag{14}$$

$$\mathcal{L}_{hc} = ||Y - I_T^{hc}||_1, \tag{15}$$

$$\mathcal{L}_{mask} = ||M_{gt} - M||_1, \tag{16}$$

$$\mathcal{L}_{per} = \sum_{i=1}^{L} ||\Phi_i(Y) - \Phi_i(I_T)||_1, \tag{17}$$

$$\begin{aligned}\mathcal{L}_{adv}^{D_I} &= -\mathbb{E}_{I_T \sim p_{data}}[\log(D_I(I_T))] \\ &\quad - \mathbb{E}_{Y \sim p_Y}[\log(1 - D_I(Y))], \end{aligned} \tag{18}$$

$$\mathcal{L}_{adv}^{\mathcal{D}(z)} = -\mathbb{E}_{Y \sim p_Y}[D_I(Y)], \tag{19}$$

where λ_{rec}, λ_{hc}, λ_{mask}, λ_{per}, and λ_{adv} are weights for the loss $\mathcal{L}_{rec}$, $\mathcal{L}_{hc}$, $\mathcal{L}_{mask}$, $\mathcal{L}_{per}$, and $\mathcal{L}_{adv}$, respectively. I_T^{hc} is a target image without neck, and body completion and Φ is a pre-trained VGG19 network [17], and D_I is a discriminator. We used λ_{rec} = 10, λ_{hc} = 10, λ_{mask} = 10, λ_{per} = 1, and λ_{adv} = 1.

E. User Study

We elucidate the details of our user study in the attached *User Study.pdf*. The material includes the questionnaire design, participant demographics, and methodology. We also present the comprehensive results in Excel format.

F. Project Page

The head blending video results are shown on our project page linked in the footnote of the main paper. The video results demonstrate the effectiveness and robustness of CHANGER in various industrial scenarios and suggest its potential for adoption in the industrial field. We submit a project page created in HTML to firmly prove the finality and completeness of the results displayed on our project page as of the submission date.