## 6. Additional Related Work

### 6.1. Head Swap

Many existing methods, such as those proposed in [2, 20, 40], optimize their approaches based on these cropped datasets, leading to inherent limitations in handling cases where the full head or surrounding region should be harmonized with the body. Consequently, these methods struggle with occlusions, head orientations beyond a narrow frontal distribution, and varying hair structures. While FaceX [20] and REFace [2] leverage diffusion models for head swapping, they still rely on face-centered training data, inheriting the same dataset-induced weaknesses. HSDiffusion [53], although diffusion-based, assumes a simple alignment mechanism where the center points of the head and body images are matched before compositing. However, without explicit modeling of head orientation differences, this approach often results in unnatural compositions when the source and target images have misaligned orientations. In contrast, HeSer [50] attempts to address these limitations by incorporating more varied head orientations. However, it operates under a few-shot learning paradigm, making it less flexible and scalable compared to zero-shot approaches. Additionally, the recent method GHOST 2.0 [18] adopts HeSer's blending technique, requiring precise image alignment similar to HeSer. This introduces more complex data preprocessing steps. Furthermore, due to dataset limitations, it struggles to handle cases where the subject has extremely long hair.

## 7. Implementation Details

Our model is composed of three key components: the H-Net, which utilizes an SDXL inpainting model [52]; the S-Net, which employs the UNet from the original SDXL [41]; and a pretrained IP-Adapter [57] and a pretrained face encoder from PhotoMaker [35]. We train our model on the SHHQ dataset [16], adopting the data handling procedures from HID [25] with modified captions as detailed in Sec. 7.1. For data augmentation, we apply GAGAvatar [11] to 70% of the images. The corresponding masks are augmented through dilation (with a 90% probability), concatenation (50%), and conversion to bounding boxes (50%). The model is trained for 70 epochs using the AdamW optimizer [37] with a learning rate of $1 \times 10^{-5}$ and a batch size of 6 per GPU. For reproducibility, we fix the random seed to 42 for both training and inference. During inference, we use a classifier-free guidance (CFG) [22] scale of 2.0 with 30 denoising steps and generate images at a resolution of $1024 \times 1024$. In addition, for simplicity and efficiency, we employ the DeepXception model [10] to generate the segmentation mask during inference.

### 7.1. Datasets

We leverage the SHHQ dataset [16] following the approach in HID [25], but modify the captions. By replacing the original text embedding of 'hairstyle' with a fused embedding from the hair image and text 'hairstyle', HID eliminates the need for hairstyle descriptions in the prompt.Instead, we generate image captions about the hairstyle using the multimodal large language model, GPT-4o [1] and add the generated captions to each original caption used in HID after removing "with hairstyle". To explicitly indicate the hair region, we provide both the input image and the cropped hair portion of the input image to the model.

### 7.2. Loss

To train our model, we use a composite loss function that balances overall image fidelity with accuracy in the specific head region. Our final loss, $\mathcal{L}_{\text{total}}$, is formulated as a weighted sum of two Mean Squared Error terms:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{global}} + \lambda_2 \mathcal{L}_{\text{head}} \tag{4}$$

Here, $\mathcal{L}_{\text{global}}$ is the standard MSE loss between the prediction and the ground truth over the entire image. $\mathcal{L}_{\text{head}}$ is an MSE loss computed exclusively on the head region, isolated using a mask $M$. For all experiments, we set the balancing hyperparameters $\lambda_1$ and $\lambda_2$ to 1.0.

### 7.3. GAGAvatar Augmentation

For GAGAvatar [11] augmentation, we employ a balanced sampling strategy to ensure robustness. Head pose differences are distributed as $5^\circ$ (37%), 5–$10^\circ$ (31%), and $> 15^\circ$ (32%). For expression variations, we sample across a wide range of $[-0.52, 0.99]$, achieving a mean cosine similarity of $0.67$ ($\sigma = 0.18$). This diverse distribution allows the model to generalize across various motion scales. Our pipeline is fully automatic and does not require manual alignment between the source and target. Since the model generates the head within the target bounding box while centering the head using a normal map, it remains robust to large pose disparities.

## 8. Additional Experiments

### 8.1. Comparisons with Additional Baselines

We further compare our AHS with four additional baselines: Nano Banana [13], Qwen-Image-Edit [55], HeSer [50], and Ghost 2.0 [18]. As shown in Fig. 10 and Tab. 3, while Nano Banana and Qwen-Image-Edit prioritize consistency, they often produce images identical to the input or suffer from severe copy-and-paste artifacts, which paradoxically leads to a high FID score. Similar to REFace [2], both HeSer and Ghost 2.0 rely on a conventional face-swap paradigm based on a crop-and-align pipeline. This approach is inherently


Figure 10. **Qualitative comparison with additional baselines.**

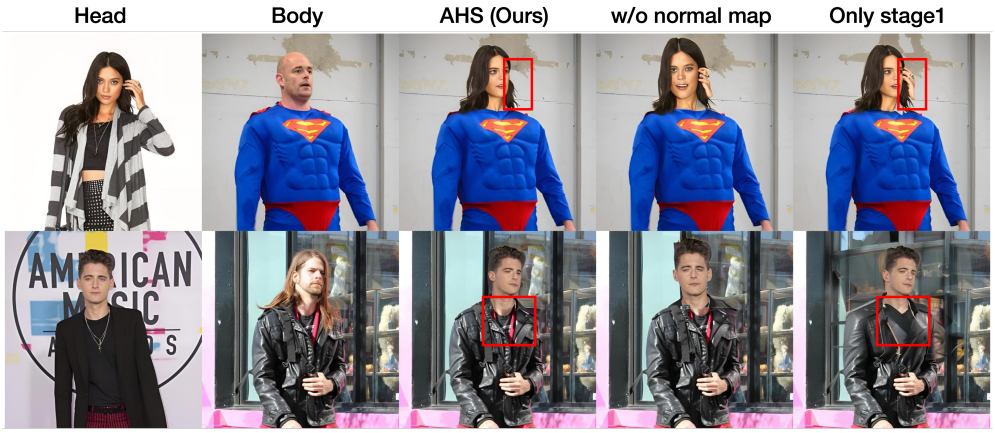


Figure 11. **Qualitative results of additional ablation study.**

| Method | CLIP-I(Head) ↑ | FID ↓ | FID(Crop) ↓ | ID sim↑ | Head orientation↓ | Expression ↓ |
|---|---|---|---|---|---|---|
| Ghost 2.0 | 0.8487 | 7.968 | 19.749 | 0.483 | **4.831** | 10.626 |
| HeSer [50] | 0.8507 | 6.444 | 15.221 | 0.503 | 8.734 | 11.022 |
| Nano Banana [13] | 0.8634 | **4.241** | **2.809** | 0.474 | 10.725 | 7.449 |
| Qwen-Image-Edit [55] | 0.8789 | 4.377 | 4.422 | 0.536 | 15.504 | 7.522 |
| Ours | 0.9139 | 9.613 | 6.719 | 0.625 | 8.427 | 6.887 |
| Ours w/o normal | **0.9238** | 10.055 | 5.944 | **0.729** | 17.472 | 10.799 |
| Ours stage1 | 0.9157 | 9.073 | 6.221 | 0.631 | 9.000 | **6.761** |

Table 3. **Quantitative Results.**

unsuitable for head swapping as it is unable to handle regions outside the fixed facial crop, such as long hair. As a result, all three methods suffer from prominent bounding box artifacts and degradation in structural completeness compared to our method.

## 8.2. Additional Ablation Study

As shown in Fig. 11 and Tab. 3, omitting surface normals degrades head orientation and expression accuracy, as they provide essential geometric guidance for motion-identity decoupling. Furthermore, while bounding-box-only in-

ference lacks boundary constraints leading to background flickering and deformation.

### 8.3. Lighting Condition Augmentation

Regarding lighting variations and complex occlusions, we realize that GAGAvatar alone is insufficient to handle these factors. To address this, we apply a data augmentation strategy using relighting models such as IC-Light [59]. Specifically, during training, we further augment 70% of the images augmented by GAGAvatar by applying IC-Light. As qualitatively shown in Fig. 12, this approach mitigates sensitivity to lighting changes and further enhances overall lighting consistency.


Figure 12. **Results of IC-light Augmentation.**

### 8.4. Inference Mask

As detailed in Section 3.3, this section elaborates on our inference methodology. Generating realistic hair presents a unique challenge, as the target area requires significant creative flexibility. To accommodate this, we employ a two-step inference process. Initially, we perform inference using a simple bounding box as the mask. While this approach provides ample space for hair synthesis, its broad nature can lead to undesirable artifacts, such as the deformation of clothing or the background outside the primary head region. To address this, we refine the mask in a second step. First, we extract a precise head region mask from the intermediate output. We then create a new, more accurate mask by computing the union of this mask and a mask from the body image. This refined mask is used to perform a second round of inference. This strategy ensures that the inpainting process is precisely focused on the desired areas, preventing modifications to irrelevant regions. As illustrated in Figure 13, the intermediate result, while plausible, exhibits clothing distortion. In contrast, the final output is cleanly reconstructed because the refined mask correctly excludes the clothing area from the inpainting process.

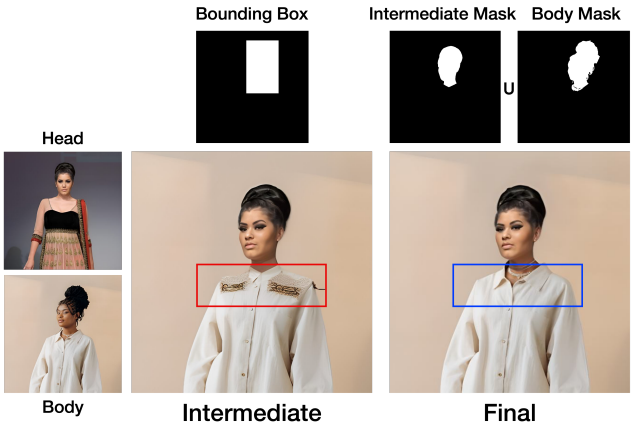


Figure 13. **Inference mask results.**

## 9. Failure Cases

Despite robust normal estimation in profile views, Fig. 14, our method faces three main challenges: (1) identity preservation under extreme poses, (2) restoration of masked-out facial occlusions, and (3) maintaining consistent facial scales when aligning with the body geometry. These cases arise from the inherent difficulty of hallucinating out-of-distribution spatial and structural information.

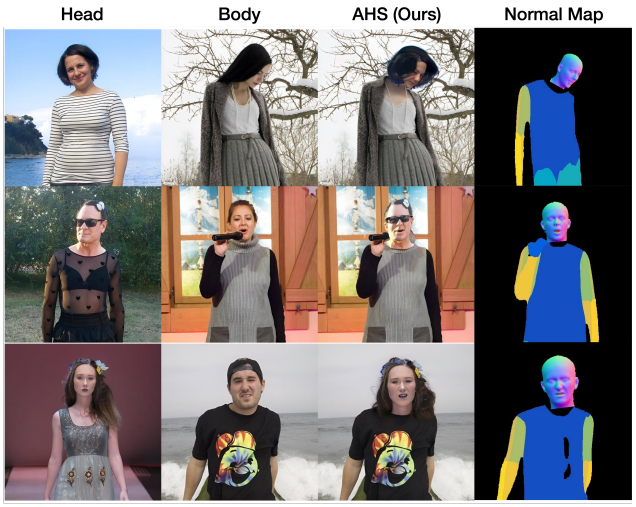


Figure 14. **Failure Cases.**

## 10. Additional Qualitative Results

We provide additional qualitative results in Fig. 15 and Fig. 16 generated by our proposed approach.

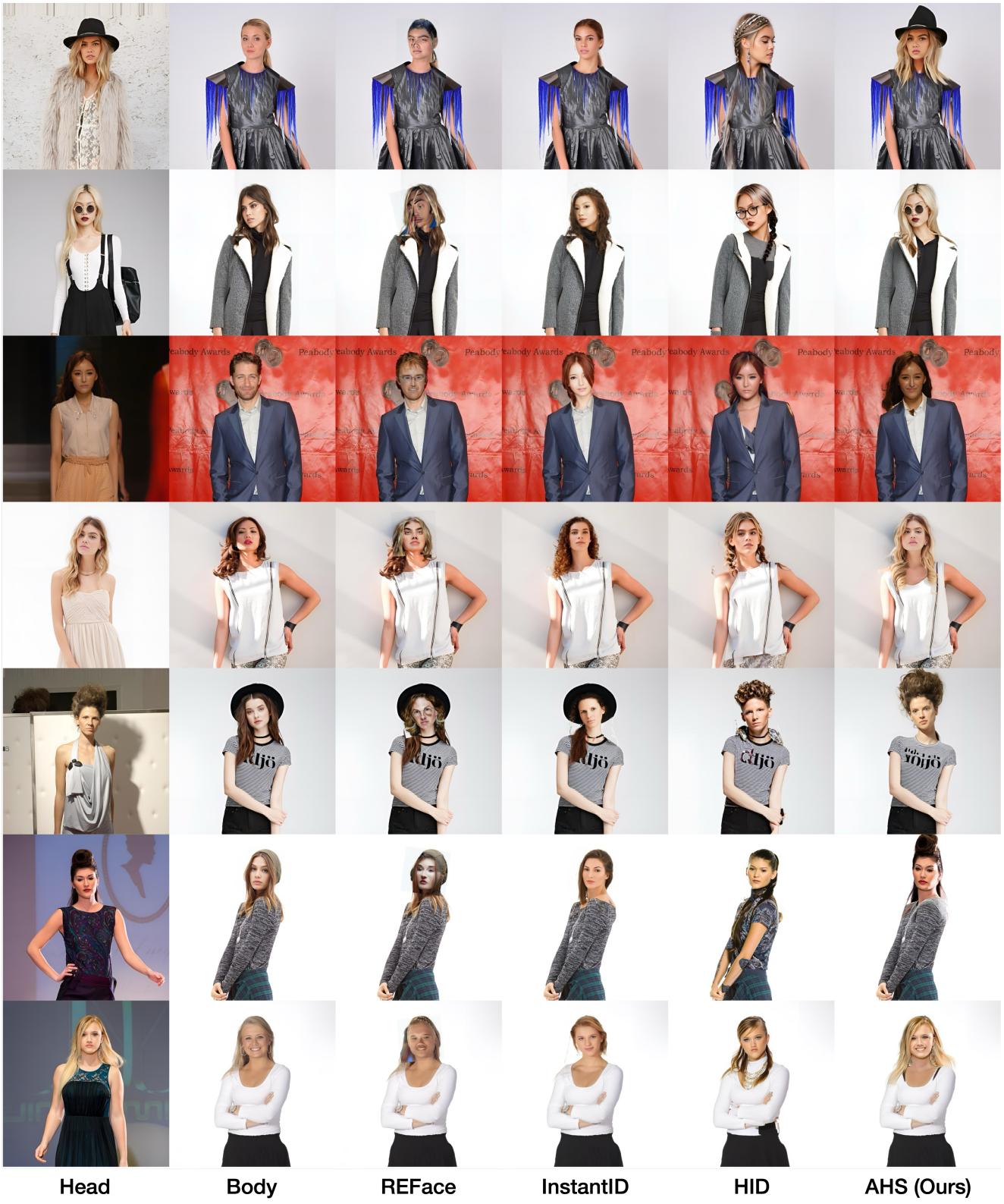


Figure 15. **Qualitative comparison.** The images in the *Head* column are combined with those in the *Body* column. The last four columns are the head-swapped results produced by each method.


Figure 16. **Qualitative comparison.** The images in the *Head* column are combined with those in the *Body* column. The last four columns are the head-swapped results produced by each method.