UNIBIOTRANSFER: A UNIFIED FRAMEWORK FOR MULTIPLE BIOMETRICS TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

DeepFace generation has traditionally followed a task-driven paradigm. However, such a single-task setting severely limits model generalization and scalability. A unified model capable of solving multiple deepface generation tasks in a single pass represents a promising and practical future direction, but remains extremely challenging due to data scarcity and serious cross-task conflicts arising from heterogeneous attributes transformations. To this end, we propose **UniBioTransfer**, the first unified framework capable of handling both conventional deepface tasks, e.g., face transfer and face reenactment, and tasks with huge shape-varying transformations, e.g., hair transfer and head transfer. Furthermore, *UniBioTransfer* can generalize to unseen tasks, like lip, eye, and glasses transfer, with minimal model fine-tuning. Specifically, *UniBioTransfer* achieves the goal by addressing data insufficiency challenge in multi-task generation through a unified data construction strategy that can handle diverse and distinct attributes, and mitigates cross-task interference via an innovative BioMoE, a Mixture-of-Experts based model coupled with a novel two-stage training strategy that effectively disentangles task-specific knowledge. Extensive experiments demonstrate the effectiveness, generalization, and scalability of *UniBioTransfer*, outperforming both existing multi-task models and task-specific methods across a wide range of tasks, while maintaining a comparable model size and training time. **We will release all codes upon acceptance.**

1 INTRODUCTION

DeepFace generation plays a pivotal role in generative AI, enabling real world applications in digital content creation, film and entertainment production, privacy-preserving face anonymity, etc. Typically, it refers to the deep-learning-based synthesis and manipulation of biometric attributes within the head region in different levels as show in Figure 1. However, the current landscape of DeepFace research remains largely task-specific, such as head swapping (Shu et al., 2022), face swapping (Zhao et al., 2023b), face reenactment (Chen et al., 2025), and hair transfer (Zhang et al., 2025). Multi-task frameworks (Baliah et al., 2025) also require extensive retraining and fine-tuning for each new task, resulting in limited generalization and impractical for real-world deployment.

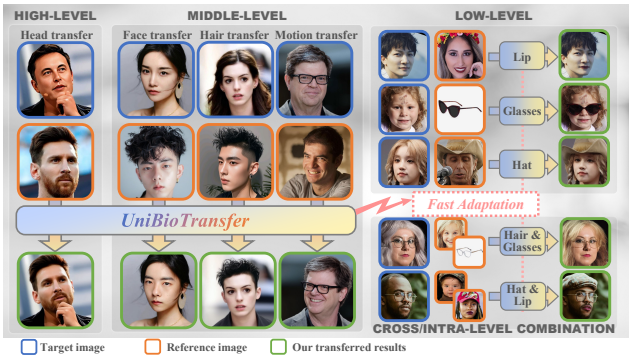


Figure 1: **UniBioTransfer**, the first unified framework for different deepface generation tasks.

These limitations highlight the need for a unified framework to efficiently handle multiple generation tasks within a single model. We observe that these tasks share the **same domain and paradigm**: (1) they all involving the transfer or preservation of head-region biometric attributes (e.g., face identity, hair, expression), and (2) they follow an identical paradigm of transferring reference attributes onto a target image. The two core commonalities motivate the design of a framework that can lever-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

age inter-task knowledge to jointly address and benefit multiple tasks. Such an approach improves practicality and scalability by eliminating the need to maintain separate task-specific models.

To achieve this, we need to address two critical challenges. ***1). Data Dilemma.*** The scarcity of paired data for deepface generation tasks necessitates a self reconstruction training paradigm, where the target, reference, and ground-truth images are derived from the same identity. Typically, a target is constructed by masking out specific regions of the ground truth, a strategy proven effective in previous works (Baliah et al., 2025). However, this naive mask-based strategy (Baliah et al., 2025) fails for large structural changes because mask silhouette leaks geometry, causing models to simply inpaint within the mask rather than perform true shape transfer (Figure 2). Other methods attempted to tackle the issue by compromising model design (Zhang et al., 2025), and artifacts/error-prone (Chung et al., 2025; Zhang et al., 2025), but inevitably introduced substantial artifacts (Figure 4c). ***2). Cross-Task Conflict Dilemma.*** Although deepface tasks all focus on manipulation of the human head region, they also exhibit distinct task-specific objectives and feature distributions. For instance, face transfer emphasizes identity preservation, while face reenactment focuses on pose&expression transfer, and hair or head transfer requires structural adaptation (Figure 1). These divergent objectives create inherent conflicts when jointly optimized within a single model. Naively training a shared network across multiple tasks often leads to gradient interference, where updates beneficial for one task degrade the performance of others.



Figure 2: Limitations of traditional mask-based strategy for attributes with significant structural changes.

Hence, we propose **UniBioTransfer**, the first **Uni**fied Deepface Generation Framework for Multiple **Bio**metrics **Transfer** within the same model. To bridge the data gap, we propose a **unified data construction strategy**. We first define all tasks as transferring a set of attributes (e.g., hair, identity) from a reference image I_{ref} to a target image I_{tgt} while preserving the remaining attributes of the target (see Section 3). Based on this, we formulate a unified data construction paradigm termed **attribute corruption**: starting from a ground-truth image I_{gt}, we divide its attributes into those to be transferred and those to be preserved, then corrupt each set to create I_{tgt} and I_{ref}. For relatively-static attributes like facial identity, simple masking suffices to effectively corrupt the identity information in the ground-truth image. However, for spatially-dynamic attributes where masks often leak geometry, we propose a novel **Swapping-based Corruption** strategy to prevent trivial solutions in tasks with large shape variations. Specifically, we leverage an off-the-shelf generative model to transfer these regions with novel geometries, compelling the network to learn genuine shape transfer from the reference (Figure 2b) rather than falling into a trivial inpainting-like solution (Figure 2a).

For the cross-task conflict dilemma, we propose **BioMoE**, an MoE-based architecture tailored for our task, and a **Two-Stage Training Strategy**. Unlike a conventional "black-box" router that relies solely on input tokens, our **BioMoE** additionally incorporates spatial structural cues by providing the router with each token's relative position to a set of facial landmarks, enabling the routed experts to specialize in distinct anatomical regions (e.g., identity-rich centers vs. geometry-heavy boundaries). To further mitigate gradient interference in multi-task learning and to facilitate adaptive capacity allocation across tasks, we further propose a ***Two-Stage Training Strategy***. Stage I (Task-Isolated Initialization) trains each task independently to bypass early gradient conflicts and establish task weight profiles. Then a ***Result-Based Adaptive LoRA Rank Allocation*** mechanism dynamically allocates LoRA capacity for task-specific experts before Stage II joint training of the entire framework. We summarize our contribution as follows:

1). We present the first framework that effectively unifies diverse biometrics transfer tasks.

2). We propose a **Unified Data Construction Strategy**, featuring a *Swapping-based Corruption Strategy* for spatial varying attributes, and a novel **BioMoE** via *Two-Stage Training Strategy*, to reduced inter-task interference and improved cross-task collaborative generation.

3). We achieve state-of-the-art performance across a wide range of representative deepface generation tasks, showing strong generalization and scalability, at a comparable computational costs to those of task-specific models.

2 Related Works

DeepFace Generation. *Task-Specialized Methods* Early task-specialized methods largely relied on GANs and warping, offering efficiency but struggling with fidelity. Recent diffusion approaches substantially improve identity preservation, texture realism, and compositional control across tasks such as face transfer (Liu et al., 2023; Zhao et al., 2023b; Xu et al., 2024a; He et al., 2025; Baliah et al., 2025; Lee et al., 2024; Ye et al., 2025; Luo et al., 2025; Wang et al., 2025; Li et al., 2023; Cui et al., 2024), motion transfer (face reenactment) (Guo et al., 2024; Wei et al., 2025a; Yang et al., 2025; Xu et al., 2025b; Jiang et al., 2024; Ma et al., 2024; Chen et al., 2025; Shurong et al., 2024; Qiu et al., 2025; Shurong et al., 2025), hair transfer (Wei et al., 2022; 2023; Nikolaev et al., 2024; Chung et al., 2025; Zhang et al., 2025), and head transfer (Shu et al., 2022; Groshev et al., 2025; Wang et al., 2022; Kang et al., 2025). Overall, such methods suffer from limited generalization—models trained for one task transfer poorly to others. This confines their applicability to single-task scenarios, which limits their versatility as the range of deepface applications expands. *Multi-Task Methods* Recognizing the limitations of task-specific models, several works have aimed to unify different DeepFace tasks. Early efforts (Xu et al., 2022; Han et al., 2024; Xu et al., 2025a; Wei et al., 2025b), such as FaceAdapter (Han et al., 2024) and RigFace (Wei et al., 2025b), unified face swapping and face reenactment by exploiting their shared underlying pattern: the recombination of identity and motion (pose and expression) from target and reference faces. Nevertheless, their heavy dependence on 3DMMs restricts their applicability to tasks involving significant shape deformations beyond facial regions, a constraint that our approach effectively addresses.

Reference-Based Image Editing. Beyond the DeepFace domain, general image editing models have made significant strides. Instruction-driven editors offer broad, promptable control without task-specific tuning, but the visual quality can be inconsistent. Even with stronger preservation in recent large vision-language models like Kontext (Labs et al., 2025), Qwen-Image-Edit (Wu et al., 2025), reference-based image editing (Brooks et al., 2023; He et al., 2023; Yu et al., 2025; Cao et al., 2023; Hertz et al.; Khachatryan et al., 2023) usually fails to precisely transfer the reference attribute to the target (Xia et al., 2025). Consequently, these general editing methods are rarely utilized in practical deepface generation because they are unreliable under multi-image (reference-based) inputs (see Sec. B.1). However, they can be effectively leveraged for our training data construction, as detailed in Sec. 3.1.

Mixture-of-Experts (MoE) and Multi-task Learning. Mixture-of-Experts (MoE) is widely used across a wide range of fields (e.g., recommendation system (Ma et al., 2018; Tang et al., 2020), and Large Language Model (Fedus et al., 2022; Lepikhin et al., 2020; Dai et al., 2024; Zhou et al., 2022)). In vision field, recent works (Zhao et al., 2023a; Sun et al., 2024; Mou et al., 2024) introduce MoE to handle multi-control image generation, but they are not inherently for multi-task learning as they focus on the fusion of conditions which are complementary and additive, rather than resolving fundamental cross-task conflicts. Many works (Chen et al., 2023; Zhu et al., 2022b; Chen et al., 2024; Ci et al., 2023; Hu et al., 2024; Qin et al., 2024) successfully introduce MoE to address cross-task challenges using dynamic routing. Despite their success, these models are not directly suitable for our objectives. Adapting "all-routed" MoE architectures like AdaMV-MoE (Chen et al., 2023) to our backbone is impractical: since state-of-the-art generation models are predominantly dense, retraining a sparse MoE from scratch requires prohibitive resources, and upcycling often incurs significant performance loss. M3DT (Kong et al., 2025) employs full-scale experts identical to the backbone FFN for each task group, yet this would lead to a prohibitive memory requirement in our case, as our backbone model (Rombach et al., 2022) is already massive. While our *BioMoE* can maintain compatibility with pre-trained dense models like Stable Diffusion while enabling precise specialization and parameter efficiency.

3 Method

A unified model capable of handling multiple deepface transfer tasks is promising but highly challenging. At the data level, constructing effective training pairs for shape-varying attributes (e.g., hair, glasses, hats) is difficult. While at the model level, naively sharing all parameters across heterogeneous tasks leads to severe task interference. To address these challenges, we introduce **UniBioTransfer**, a unified framework for multi-level deepface generation. It features a ***Unified Data Construction Strategy*** to alleviate data scarcity and a customized MoE-based architecture, ***BioMoE***,

which mitigates inter-task interference through a ***Two-Stage Training Strategy*** *with Result-Based Adaptive LoRA Rank Allocation*, enabling strong performance across diverse generation tasks.

Problem Definition and Pipeline Overview. We formulate various deepface tasks as swapping a set of attributes X (e.g., face identity, hair, pose, expression, skin tone) from a reference image I_{ref} onto a target image I_{tgt}, while preserving the remaining attributes Y from I_{tgt}. The desired output is $I_{\text{out}} = X_{\text{ref}} \cup Y_{\text{tgt}}$. We adopt Stable Diffusion v1.5 as the backbone for a fair comparison with previous methods. As shown in Figure 3, given a target image I_{tgt} and a reference image I_{ref}, we first mask out irrelevant attributes in both. We extract semantic features of masked I_{ref} via the CLIP encoder (Radford et al., 2021) and identity features (for tasks where I_{ref} provides identity attributes like face/head transfer) via the ArcFace model (Deng et al., 2019). An MLP then projects these raw features into tokens for cross-attention conditioning in UNets. We form blended landmarks to guide motion, taking pose&expression from I_{ref} (for motion transfer) or I_{tgt} (for other tasks) and identity from the counterpart. The masked I_{tgt} is encoded to a latent and concatenated with noise along the channel dimension as input to the main UNet. The masked I_{ref} flows through the refNet (Reference-UNet), and its tokens are injected via cross-attention into the main UNet. We use only the first half of the refNet to improve efficiency. After iterative denoising, the latent is decoded back into pixel space by VAE decoder to produce the final result I_{out}. Refer to Appendix A.1 for details.

3.1 A Unified Data Construction Strategy

The lack of paired data in deepface generation necessitates self-reconstruction training, meaning the I_{tgt}, I_{ref}, and I_{gt} come from the same identity. However, this setup risks trivial learning for shape-varying tasks, such as hair or head transfer, with models preserving spatial attributes in I_{tgt} rather than transferring them from I_{ref} (Figure 2). Task-specific solutions for shape-varying tasks remain insufficient: StableHair (Zhang et al., 2025) relies on a bald converter, introducing error accumulation; HS-Diffusion (Wang et al., 2022) adds a mask predictor, breaking model unification; and HairFusion (Chung et al., 2025) uses overly aggressive masking, causing background distortion and boundary artifacts (Figure 4c).

To address these limitations, we introduce our unified data construction strategy. We formulate the data construction process for various deepface tasks as ***Attribute Corruption***: beginning with a ground-truth image $I_{\text{gt}} = X_{\text{gt}} \cup Y_{\text{gt}}$, where X_{gt} and Y_{gt} are the ground-truth attributes corresponding to the sets X and Y, the goal of data construction is to synthesize the training pair $(I'_{\text{tgt}}, I'_{\text{ref}})$ by strategically corrupting attributes in I_{gt}. Specifically, the target image I'_{tgt} is constructed by corrupting the attributes X_{gt} to $X_{\text{corrupted}}$ while preserving Y_{gt}, resulting in $I'_{\text{tgt}} = X_{\text{corrupted}} \cup Y_{\text{gt}}$. Conversely, the reference image I'_{ref} is formed by corrupting Y_{gt} to $Y_{\text{corrupted}}$ while preserving X_{gt}, yielding $I'_{\text{ref}} = X_{\text{gt}} \cup Y_{\text{corrupted}}$. As a result, the constructed data pair $(I'_{\text{tgt}}, I'_{\text{ref}}) \rightarrow I_{\text{gt}}$ compels the model to learn the intended transfer behavior: to extract attribute set X from I'_{ref} and merge it with attribute set Y from I'_{tgt}. **The key idea is to ensure that the corrupted attributes $X_{\text{corrupted}}$ do not leak the original geometry of X_{gt}.**

Spatially-Dynamic Attributes Corruption Spatially-dynamic attributes include those where the shape is a critical component to be transferred, such as hair or accessories (e.g., hats, glasses). Simple masking is troublesome here, as the mask boundary itself leaks the ground-truth silhouette, allowing the model to find a trivial inpainting-like solution instead of learning shape transfer (Baliah et al., 2025). To address this, we propose a Swapping-Based Corruption Strategy.

Swapping-Based Corruption Strategy. We observe that if the shape information of spatially dynamic attributes in the target image, e.g. hair, is sufficiently corrupted, the model can no longer recover the target geometry through a trivial inpainting-like process and is thus forced to learn the desired geometry from the reference image. Motivated by this observation, we can take advantage of existing generative models to corrupt spatially dynamic attributes, to generate a target image with a completely different geometry, ensuring the model cannot rely on the target's geometry and must learn to extract it entirely from the reference (Figure 2). However, we emphasize that the external generative model used in our Swapping-based Corruption strategy serves ***purely*** as a data construction tool to introduce structural variations and prevent shape leakage. Table 6 shows that we achieve similar performance regardless of the choice of a stronger or a weaker external generator, while consistently outperforming both external generators themselves. More discussion is in Appendix B.1.

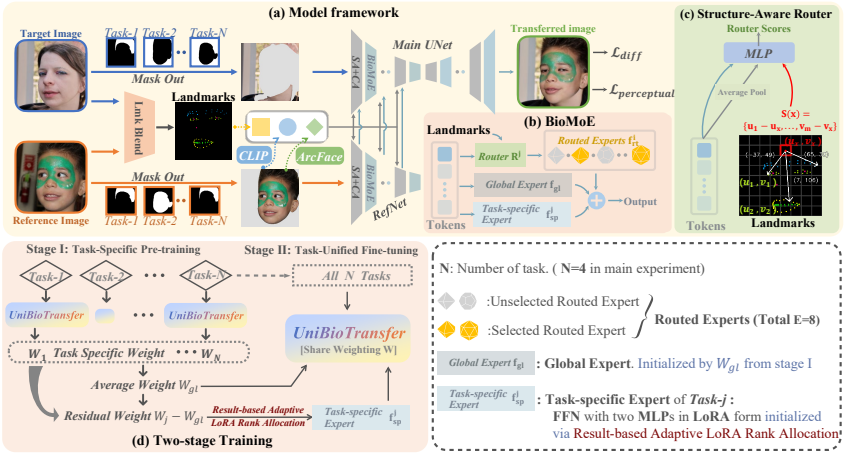


Figure 3: **UniBioTransfer** architecture overview. (a) Overall framework. (b) We introduce an MoE-enhanced Feed Forward Network (FFN). (c) Expert selection is guided by a Structure-Aware Router. (d) The entire system is optimized using a two-stage training strategy designed to stabilize routing and promote expert specialization.

Relative-Static Attributes Corruption Relative-static attributes, including both structural and non-structural elements whose spatial locations and overall shapes remain largely consistent during transfer, are processed using the conventional masking strategy (Baliah et al., 2025). Please refer to Appendix A.3 for details.

3.2 BioMoE

A unified model capable of handling multiple deepface tasks is highly desirable for efficiency but faces a key challenge: task interference due to disparate task requirements. For example, face transfer necessitates high-precision identity transfer within the central facial region, whereas hair/head transfers focus on geometric deformations at the peripheral boundaries. An alternative is to directly adopt the standard Mixture of Experts (MoE) architecture in LLM (Dai et al., 2024) and multi-task vision (Chen et al., 2023) fields. However, this typically requires a pre-trained MoE image generation base model, whereas state-of-the-art image generation models are predominantly dense. Other approaches (Kong et al., 2025) copy full-scale experts from the backbone FFN for each task group. Yet, this would lead to a prohibitive memory requirement, as the base model (Rombach et al., 2022) for our tasks is already heavy.

To resolve these conflicting constraints, we introduce our **BioMoE**. We achieve an effective design with acceptable overhead by retaining only one identical FFN from the pretrained dense model (Rombach et al., 2022) for knowledge sharing, while delegating dynamic and specialized knowledge to lightweight routed experts and task-specific experts. Specifically, the proposed BioMoE module consists of three types of experts, balancing shared knowledge, task-specific knowledge, and dynamic specialization:

1). Global Expert. The original FFN structure from the pretrained model is preserved to serve as a common knowledge base that processes all input tokens, capturing the foundational paradigm of all tasks: extracting reference attributes and blending them into the target image.

2). Routed Experts. A pool of $E = 8$ lightweight experts is shared across all tasks for dynamic and structure-aware specialization. For each input token, a task-specific router selects the top-K experts from this pool to process the token. Each expert follows the FFN's two-layer MLP structure, but with a reduced inner dimension (reduced by a factor of $F = 16$ compared to original FFN's dimension) for efficiency.

3). Task-specific Experts. To encode highly specialized information unique to each task's distinct requirements (e.g., precise identity process for face/motion transfer or geometric deformation for hair/head transfer), we append a task-specific expert for each task, efficiently implemented using Low-Rank Adaptation (LoRA). Formally, the forward process of BioMoE is as follows: $f(x) = f_{\text{gl}}(x) + \sum_{i=1}^{E} \mathcal{G}_i^{(j)}(x) \cdot f_{\text{rt}}^{(i)}(x) + f_{\text{sp}}^{(j)}(x)$, where f_{gl} is the global shared FFN,

$\mathcal{G}^{(j)}(x) \in \mathbb{R}^E$ contains the routing weights (with only top-K non-zero values), $f_{\text{rt}}^{(i)}$ denotes the i-th dynamically routed expert, and $f_{\text{sp}}^{(j)}$ represents the task-specific expert of task-j. Meanwhile, $\mathcal{G}^{(j)}(x) = \text{top}_K\left(\text{softmax}\left(R^{(j)}(\text{concat}(x, \text{pool}(X), \mathcal{S}(x))) + \epsilon\right)\right)$, in which the top_K operator sets all values to be zero except the largest K values, $\mathcal{G} \in \mathbb{R}^E$ denotes the routing weights, $R^{(j)}$ is the task-specific router for task j, $\text{concat}(x, \text{pool}(X), \mathcal{S}(x))$ concatenates the token feature x, the pooled sequence features, and structure information $\mathcal{S}(x)$ of the token, E is the number of experts, and j is the task id. ϵ is a noise term added during training to encourage exploration and improve expert load balancing.

Structure-Aware Routing Different deepface generation tasks emphasize distinct semantic regions—for instance, face transfer and reenactment focus on facial identity, while hair transfer targets the scalp area. To exploit this spatial locality, we introduce a structure-aware router. Unlike conventional MoE routers that rely solely on token features, our router is an MLP that also incorporates structure information (relative positions to each facial landmarks). This design enables the routing mechanism to dynamically select experts based on spatial context. For example, tokens corresponding to facial regions can be routed toward experts specialized in identity preservation, while tokens from the scalp region activate experts focusing on hair transfer. The structure information $\mathcal{S}(x)$ for each token is computed as: $\mathcal{S}(x) = \text{concat}\left(u_1 - u_x, v_1 - v_x, \ldots, u_M - u_x, v_M - v_x\right)$, where (u_m, v_m) denotes the coordinates of the m-th landmark, (u_x, v_x) is the spatial position of token x in the feature map, and M is the number of key landmarks.

3.3 Two-Stage Training Strategy

A straightforward approach to training our model would be to initialize all components and jointly train on all tasks from the outset. However, this is suboptimal as: (1) the global expert contains the majority of the parameters, thus joint training from the start would induce significant gradient conflicts within this large shared module, slowing convergence; (2) the optimal rank for the task-specific experts can vary significantly depending on the task and layer complexity. Hence, we propose an efficient two-stage training strategy.

Stage I: Task-Isolated Initialization We first train a separate version of the model for each individual task using the same model structure. Specifically, a standard FFN is trained for each task separately, with gradients isolated between them and without routed experts. This stage serves a dual purpose: 1). allowing task-specific modules to rapidly achieve relatively optimal performance on each task, providing a conflict-free initialization for stage II, and 2). establishing a comprehensive weight profile for each task capability, enabling our *Result-Based Adaptive LoRA Rank Allocation*.

Result-Based Adaptive LoRA Rank Allocation. A key mechanism in our two-stage training is the Result-Based Adaptive LoRA Rank Allocation, which utilizes the weight from 1_{st} stage to allocate parameter budgets in a dynamic and data-driven manner, ensuring each task receives its optimal capacity before the final joint optimization. Unlike conventional "gradient-oriented" adaptive LoRA methods (Zhang et al., 2023), which are inherently "greedy" as they depend on short-term gradient information that can be noisy and misleading, we propose to utilize the pretrained weights from stage I, which reflect long-term task requirements. Specifically, we perform Singular Value Decomposition (SVD) on the residual matrix $(W_j - W_{\text{gl}})$ for each task-specific expert. We then select the rank r such that the top r singular values capture a predefined percentage (e.g., 20%) of the total matrix energy. This adaptive approach automatically allocates more parameters (a higher rank) to tasks and layers that require more complex, specialized adaptation, while using a lower rank for simpler ones, resulting in a highly efficient and optimized architecture.

Stage II: Task-Unified Fine-tuning We then construct our final MoE-based unified model. In detail, we initialize the weights of the global expert in each MoE layer by averaging the weights of the corresponding FFNs from the model trained in stage one ($W_{\text{gl}} = \frac{1}{N}\sum_{j=1}^{N} W_j$). Next, the task-specific experts need to be initialized to capture the residual information for each task. For each task j, the task-specific expert's weights are initialized by fitting the difference between the weights of the task-specific FFN (W_j) and the averaged global expert (W_{gl}), where the LoRA rank is determined by the proposed Adaptive LoRA Rank Allocation in Appendix 3.3.

Training Loss can be found at Appendix A.2.

Table 1: Comparisons on **High-Level** and **Mid-Level** tasks with **multi-task and unified approaches**. Note only ours can perform all four tasks simultaneously.For all metrics, ↑ indicates higher is better and ↓ indicates lower is better. Best results are highlighted in **bold**.

Method	Medium Level Tasks									High Level Tasks		
	Face Transfer			Hair Transfer			Motion Transfer			Head Transfer		
	ID sim↑	pose dist.↓	expr. dist.↓	CLIP dist.↓	ID sim↑	non-hair SSIM↑	ID sim↑	pose dist.↓	expr. dist.↓	CLIP dist.↓	ID sim↑	pose dist.↓
Multi-Task Methods (Training separately for different tasks)												
REFace (WACV 25)	0.631	3.75	1.04	-	-	-	-	-	-	0.639	0.540	5.43
Unified Method (Train once for all tasks)												
Face-Adapter (ECCV 24)	0.519	4.91	1.33	-	-	-	0.503	**6.89**	2.27	-	-	-
RigFace (arXiv 25)	0.357	5.46	2.11	-	-	-	0.413	9.33	2.32	-	-	-
UniBioTransfer(Ours)	**0.637**	**3.63**	**1.03**	**0.421**	**0.887**	**0.91**	**0.602**	7.09	**2.09**	**0.460**	**0.545**	**5.28**

Table 2: Comparisons on **High-Level** and **Mid-Level** deepface tasks with **task-specific methods**. **Note only ours performs all four tasks simultaneously in a single framework.**

Method	Medium Level Tasks									High Level Tasks		
	Face Transfer			Hair Transfer			Motion Transfer			Head Transfer		
	ID sim↑	pose dist.↓	expr. dist.↓	CLIP dist.↓	ID sim↑	non-hair SSIM↑	ID sim↑	pose dist.↓	expr. dist.↓	CLIP dist.↓	ID sim↑	pose dist.↓
HunyuanPortrait (CVPR 25)	-	-	-	-	-	-	0.580	9.22	2.18	-	-	-
CanonSwap (ICCV 25)	0.551	**2.92**	**0.87**	-	-	-	-	-	-	-	-	-
Stable-Hair (AAAI 25)	-	-	-	0.468	0.855	0.87	-	-	-	-	-	-
HairFusion (AAAI 25)	-	-	-	0.478	0.848	0.80	-	-	-	-	-	-
GHOST2.0 (arXiv 25)	-	-	-	-	-	-	-	-	-	0.618	0.461	5.78
UniBioTransfer(Ours)	**0.637**	3.63	1.03	**0.421**	**0.887**	**0.91**	**0.602**	**7.09**	**2.09**	**0.460**	**0.545**	**5.28**

4 Experiments and Discussions

Baselines. We compare UniBioTransfer with three categories of methods: 1) **task-specific methods**, including CanonSwap (Luo et al., 2025) for face transfer, HunyuanPortrait (Xu et al., 2025b) for motion transfer, StableHair and HairFusion for hair transfer, and GHOST2.0 (Groshev et al., 2025) for head transfer; 2) **multi-task methods**, represented by REFace (Baliah et al., 2025), which requires separate retraining for different tasks; and 3) **unified methods**, including Face-Adapter and RigFace. Additional baseline details are provided in Appendix A.5.

Metrics. Following prior works, we adopt face identity similarity (ID Sim.), pose distance (Pose Dist.) (Ruiz et al., 2018), expression distance (Expr. Dist.), CLIP distance (CLIP Dist.), and structural similarity (SSIM) for quantitative evaluation.

More Details for implementation and dataset can be found in Appendix A.

4.1 Evaluations and Discussions

UniBioTransfer supports a broad range of biometric transfer tasks spanning three hierarchically organized levels. Qualitative comparisons are presented in Figure 4, with additional results provided in the supplementary material.

High-Level Head Transfer. As shown in Table 1, existing unified baselines cannot handle high-level head transfer because of its considerable structural complexity. In contrast, UniBioTransfer provides a unified solution and substantially outperforms both state-of-the-art task-specific head-transfer methods in Table 2 and the retraining-based multi-task method REFace in Table 1.

Medium-Level Facial and Hair Transfer. UniBioTransfer is the only unified model that jointly supports face transfer, face reenactment, hair transfer, and high-level head transfer within a single framework, as evaluated in Tables 1 and 2.

Hair Transfer. Despite substantial variations in hair geometry, texture, and color, UniBioTransfer effectively handles hair transfer, unlike existing multi-task and unified methods. To the best of our knowledge, it is the first unified deepface framework to support this task and outperforms the specialized StableHair and HairFusion methods.

Face and Motion Transfer. For face and motion transfer, UniBioTransfer surpasses multi-task and unified baselines on most metrics while achieving competitive or superior performance compared with task-specific methods.

Low-Level Fine-Grained Feature Transfer. Starting from the unified pretrained model, UniBioTransfer adapts to low-level tasks, such as eye and lip transfer, using less than 20% of the training data and 10% of the training cost required by face transfer. We

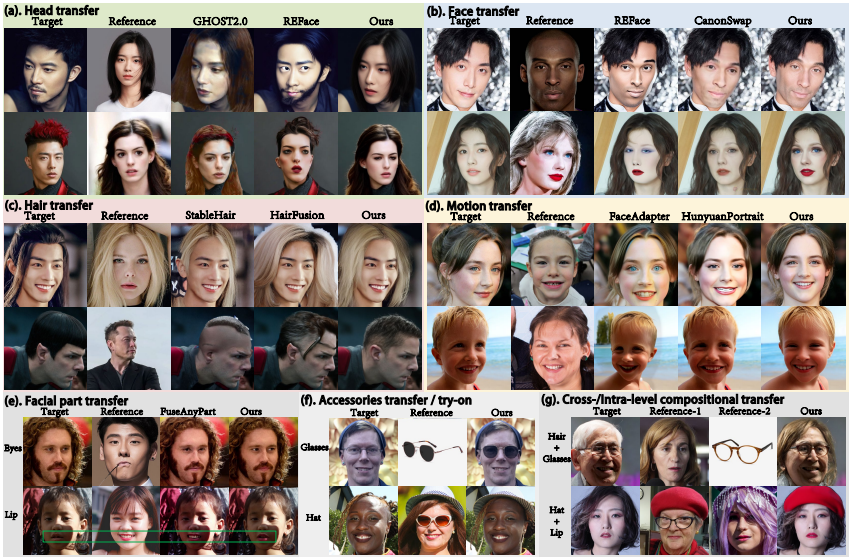


Figure 4: Visual comparisons on in-the-wild web images.

initialize each new task by inheriting the corresponding experts and router. As shown in Table 3, our method outperforms the specialized FuseAnyPart (Cui et al., 2024). Figure 4(e) further shows that FuseAnyPart struggles to preserve expressions, consistent with the limitation reported in its original paper. Its lower lip CLIP distance partly arises from directly copying the source mouth region at the expense of the target expression and skin tone. We further evaluate *Ours (w/o Uni.)*, which is trained from the base model without unified pretraining, incurs approximately $5\times$ higher training cost while still underperforming the full model. These results demonstrate that unified pretraining provides transferable representations for efficient adaptation to new low-level tasks.

Table 3: Quantitative comparisons on **Low-Level** Swap tasks.

Method	Facial-Part Transfer (Low Level Tasks)								
	Eye			Nose			Lip		
	CLIP-L2↓	Expr.-L2↓	SSIM↑	CLIP-L2↓	Expr.-L2↓	SSIM↑	CLIP-L2↓	Expr.-L2↓	SSIM↑
FuseAnyPart	0.449	0.798	0.898	0.332	0.802	0.898	**0.372**	0.794	0.902
Ours (w/o Uni.)	0.471	0.697	0.889	0.337	0.672	0.859	0.384	1.009	0.867
Ours	**0.427**	**0.599**	**0.927**	**0.327**	**0.532**	**0.937**	0.378	**0.579**	**0.937**

Cross-Level and Intra-Level Compositional Transfer. As illustrated in Figure 1, UniBioTransfer efficiently adapts to multi-attribute composition by inheriting two task-specific experts from the parent tasks and introducing a dedicated router to balance their token-wise contributions. This design enables conflict-free composition within a shared knowledge base. Further explanations of the attribute-composition strategy are provided in the supplementary material.

4.2 Ablation Studies

We conduct comprehensive ablation studies to evaluate the effectiveness of the model architecture, training strategy, routing mechanism, and data-construction approach.

Table 4: Ablations on model design and training strategy.

Model Design				
Setting	**Face ↑**	**Hair ↓**	**Motion ↓**	**Head ↓**
w/o BioMoE (Naive Parameter Sharing)	0.478	0.474	9.74	0.574
w/o BioMoE (Task-Specific Model)	0.622	0.428	7.14	0.465
w/o Structure-Aware Routing	0.634	0.423	7.13	0.469
Ours	**0.637**	**0.421**	**7.09**	**0.460**
Training Strategy				
Setting	**Face ↑**	**Hair ↓**	**Motion ↓**	**Head ↓**
w/o Two-Stage Training	0.507	0.443	8.33	0.491
w/o Result-based Rank Allocation	0.634	0.433	7.35	0.464
Ours	**0.637**	**0.421**	**7.09**	**0.460**

Table 5: Ablation on the data-corruption strategy using CLIP distance.

Method	Hair Transfer ↓	Head Transfer ↓
w/o swapping-based corruption	0.492	0.555
Ours	**0.421**	**0.460**

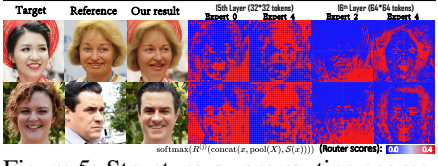


Figure 5: Structure-aware routing scores.

Table 6: Ablation on low reliance of external generators for training-data construction.

Method	Hair CLIP↓	BG SSIM↑	FID↓
HairCLIPv2 (2023)	0.485	0.82	8.18
StableHair (2025)	0.468	0.87	7.55
Ours (via HairCLIPv2)	0.424	**0.91**	**4.82**
Ours (via StableHair)	**0.421**	**0.91**	4.83

Table 7: Runtime Analysis. We achieve training and inference efficiency comparable to task-specific baselines.

Method	Task	VRAM (GB)	Time (s/img)	Train (GPU-days)
REFace	Face	11.8	12	18 (A100)
StableHair	Hair	19.3	16	Unknown
HunyuanPortrait	Motion	18.4	103	384 (A100)
GHOST2.0	Head	8.0	< 1	232 (A100)
Ours	All	14.5	15	34 (A40)
Ours (Adapt.)	Unseen	15.4	16	3 (A40)

Ablation on Model Design. To evaluate BioMoE, Table 4 compares the full model with two variants: *Single-Task Models*, which train an independent model for each task, and *Naive Sharing*, which jointly trains all tasks with fully shared parameters. The inferior performance of Naive Sharing reveals substantial cross-task interference, motivating the proposed BioMoE architecture. Unified training achieves comparable or superior performance to task-specific models while consolidating all tasks into a single checkpoint. It also provides an effective initialization for adapting to low-level and compositional tasks with limited additional data and computation, as further evidenced by the advantage over *Ours (w/o Uni.)* in Table 3.

Structure-Aware Routing vs. Vanilla Routing. Unlike a vanilla router that relies solely on input-token features, our structure-aware router additionally incorporates structural cues. Figure 5 visualizes the average expert probabilities (after softmax and before top-K selection) during head-transfer inference, revealing semantically meaningful expert specialization. For instance, Expert 4 in Layer 15 primarily attends to the forehead, while Expert 2 in Layer 16 focuses on the neck region, potentially facilitating seamless blending among the transferred head, target neck, and background.

Ablation on Training Strategy. Table 4 further compares our two-stage training strategy with a one-stage baseline that jointly optimizes the MoE model from scratch, as well as our result-based adaptive LoRA rank allocation with conventional gradient-based allocation. The full model consistently achieves superior performance, demonstrating the effectiveness of both designs. Further analyses are provided in Appendix B.2.

Ablation on Data Construction. We first evaluate the proposed *swapping-based corruption* strategy for spatially dynamic tasks, including hair and head transfer. In Table 5, we compare it against a simple masking-based construction strategy. The clear performance degradation of the masking-based variant confirms that preventing the model from exploiting trivial shortcuts is essential for learning effective shape and appearance transfer.

Low Reliance on External Generative Models. Table 6 then compares hair-transfer models trained using data constructed using either the stronger StableHair (Zhang et al., 2025) or the weaker HairCLIPv2 (Wei et al., 2023). The two variants achieve similar performance and consistently outperform the external generators themselves, indicating that our swapping-based corruption strategy is robust to the choice and quality of the external generative model.

4.3 Runtime and Efficiency Analysis

As shown in Table 7, UniBioTransfer achieves training and inference efficiency comparable to task-specific/multi-task baselines, despite being trained using less powerful GPUs. Additional inference-cost comparisons are provided in Appendix Table 12.

5 Conclusions

To conclude, we present **UniBioTransfer**, a unified, multi-task framework for diverse DeepFace generation tasks. Leveraging our unified data construction strategy and a novel BioMoE component with a two-stage training strategy, the framework effectively handles shape variations, mitigates task interference, and exploits task synergy, outperforming SOTA unified models while maintaining competitive or superior performance compared to task-specific models. Importantly, it can be efficiently adapted to new tasks with minimal fine-tuning, showing strong generalization capability and scalability.

AI USE STATEMENT

In this work, we used generative AI tools for execution. This covers method implementation. We have not used generative AI tools to formulate or prove mathematical claims, translate text, clean or reformat datasets, or conduct qualitative or thematic analysis. External generative models used to construct training pairs are part of the method and are described in the paper. Additionally, we used generative AI tools to aid or polish the writing, and for retrieval and discovery, including searching for and identifying related work. We reviewed the AI-assisted text, literature suggestions, and research suggestions. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Training and evaluation use public portrait datasets. The model can be misused for non-consensual identity manipulation.

REPRODUCIBILITY STATEMENT

Implementation, data construction, and evaluation protocols are in Sec. A. We will release all source code and model weights upon acceptance.

REFERENCES

Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Trans. Graph.*, 40(3), May 2021. ISSN 0730-0301. doi: 10.1145/3447648. URL https://doi.org/10.1145/3447648.

Sanoojan Baliah, Qunxing Lin, Shengcai Liao, Xiaodan Liang, and Muhammad Haris Khan. Realistic and efficient face swapping: A unified approach with diffusion models. pp. 1062–1071, February 2025. doi: 10.1109/wacv61041.2025.00112. URL http://dx.doi.org/10.1109/wacv61041.2025.00112.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402. IEEE, June 2023. doi: 10.1109/cvpr52729.2023.01764. URL http://dx.doi.org/10.1109/cvpr52729.2023.01764.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22503–22513. IEEE, October 2023. doi: 10.1109/iccv51070.2023.02062. URL http://dx.doi.org/10.1109/iccv51070.2023.02062.

Jiaxin Chen, Di Huang, Yunhong Wang, Yutong Zhang, and Hanwen Zhong. Transforming vision transformer: Towards efficient multi-task asynchronous learner. In *Advances in Neural Information Processing Systems 37*, NeurIPS 2024, pp. 81130–81156. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-2579. URL http://dx.doi.org/10.52202/079017-2579.

Tianlong Chen, Xuxi Chen, Xianzhi Du, Abdullah Rashwan, Fan Yang, Huizhong Chen, Zhangyang Wang, and Yeqing Li. Adamv-moe: Adaptive multi-task vision mixture-of-experts. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17300–17311. IEEE, October 2023. doi: 10.1109/iccv51070.2023.01591. URL http://dx.doi.org/10.1109/iccv51070.2023.01591.

Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(3):2403–2410, April 2025. ISSN 2159-5399. doi: 10.1609/aaai.v39i3.32241. URL https://ojs.aaai.org/index.php/AAAI/article/download/32241/34396.

Chaeyeon Chung, Sunghyun Park, Jeongho Kim, and Jaegul Choo. What to preserve and what to transfer: Faithful, identity-preserving diffusion-based hairstyle transfer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(3):2582–2590, April 2025. ISSN 2159-5399. doi: 10.1609/aaai.v39i3.32261. URL http://dx.doi.org/10.1609/aaai.v39i3.32261.

Yuanzheng Ci, Yizhou Wang, Meilin Chen, Shixiang Tang, Lei Bai, Feng Zhu, Rui Zhao, Fengwei Yu, Donglian Qi, and Wanli Ouyang. Unihcp: A unified model for human-centric perceptions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17840–17852. IEEE, June 2023. doi: 10.1109/cvpr52729.2023.01711. URL http://dx.doi.org/10.1109/cvpr52729.2023.01711.

Siying Cui, Senzhang Wang, Yaohua Wang, Yu Zheng, Aixi Zhang, and Wei-Long Zheng. Fuseanypart: Diffusion-driven facial parts swapping via multiple reference images. In *Advances in Neural Information Processing Systems 37*, NeurIPS 2024, pp. 80864–80884. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-2569. URL http://dx.doi.org/10.52202/079017-2569.

Damai Dai, Chengqi Deng, Chenggang Zhao, Ruijian Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yongle Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. pp. 1280–1297, 2024. doi: 10.18653/v1/2024.acl-long.70. URL http://dx.doi.org/10.18653/v1/2024.acl-long.70.

Trung Tuan Dao, Duc Hong Vu, Cuong Pham, and Anh Tran. Efhq: Multi-purpose extremepose-face-hq dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Egorovlvan. Glasses dataset. kaggle-datasets egorovlvan glasses-dataset, url is removed to avoid being desk rejected, 2021.

Mustafa Ekrem Erakın, Uğur Demir, and Hazım Kemal Ekenel. On recognizing occluded faces in the wild. In *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–5. IEEE, 2021.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Alexander Groshev, Pavel Paramonov, and Andrey Kuznetsov. Ghost 2.0: Generative high-fidelity one shot transfer of heads. 2025. doi: 10.2139/ssrn.5221493. URL http://dx.doi.org/10.2139/ssrn.5221493.

Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *CoRR*, abs/2407.03168, 2024. doi: 10.48550/ARXIV.2407.03168. URL https://doi.org/10.48550/arXiv.2407.03168.

Yue Han, Junwei Zhu, Keke He, Xu Chen, Yanhao Ge, Wei Li, Xiangtai Li, Jiangning Zhang, Chengjie Wang, and Yong Liu. Face-adapter for pre-trained diffusion models with fine-grained id and attribute control. In *Computer Vision – ECCV 2024*, pp. 20–36. Springer Nature Switzerland, November 2024. ISBN 9783031729737. doi: 10.1007/978-3-031-72973-7_2. URL http://dx.doi.org/10.1007/978-3-031-72973-7_2.

Dailan He, Xiahong Wang, Shulun Wang, Guanglu Song, Bingqi Ma, Hao Shao, Yu Liu, and Hongsheng Li. High-fidelity diffusion face swapping with id-constrained facial conditioning. *arXiv preprint arXiv:2503.22179*, 2025.

Tianyu He, Junliang Guo, Runyi Yu, Yuchi Wang, Jialiang Zhu, Kaikai An, Leyi Li, Xu Tan, Chunyu Wang, Hu Han, HsiangTao Wu, Sheng Zhao, and Jiang Bian. Gaia: Zero-shot talking avatar generation. *arXiv preprint arXiv:2311.15230*, 2023. doi: 10.48550/arxiv.2311.15230. URL https://arxiv.org/pdf/2311.15230.

Amir Hertz, Ron Mokady, Jay M. Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8153–8163, 2024.

Ying Hu, Miao Li, Fanbin Mo, Ji Wu, and Xun Zhu. Uni-med: A unified medical generalist foundation model for multi-task learning via connector-moe. In *Advances in Neural Information Processing Systems 37*, NeurIPS 2024, pp. 81225–81256. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-2582. URL http://dx.doi.org/10.52202/079017-2582.

Liangwei Jiang, Ruida Li, Zhifeng Zhang, Shuo Fang, and Chenguang Ma. Emojidiff: Advanced facial expression control with high identity preservation in portrait generation. *arXiv preprint arXiv:2412.01254*, 2024.

Taewoong Kang, Sohyun Jeong, H. R. Jang, and Jaegul Choo. Zero-shot head swapping in real-world scenarios. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10805–10814. IEEE, June 2025. doi: 10.1109/cvpr52734.2025.01009. URL http://dx.doi.org/10.1109/cvpr52734.2025.01009.

Tero Karras, Samuli Laine, and Timo Aila. 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.

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15908–15918. IEEE, October 2023. doi: 10.1109/iccv51070.2023.01462. URL http://dx.doi.org/10.1109/iccv51070.2023.01462.

Yilun Kong, Guozheng Ma, Qi Zhao, Haoyu Wang, Li Shen, Xueqian Wang, and Dacheng Tao. Mastering massive multi-task reinforcement learning. 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, J. English, Zion English, Peter D. Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pp. arXiv–2506, 2025. doi: 10.48550/arxiv.2506.15742. URL https://arxiv.org/pdf/2506.15742.

Jaeseong Lee, Junha Hyung, Sohyun Jung, and Jaegul Choo. Selfswapper: Self-supervised face swapping via shape agnostic masked autoencoder. In *Computer Vision – ECCV 2024*, pp. 383–400. Springer Nature Switzerland, November 2024. ISBN 9783031730016. doi: 10.1007/978-3-031-73001-6_22. URL http://dx.doi.org/10.1007/978-3-031-73001-6_22.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Fırat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020. doi: 10.48550/arxiv.2006.16668. URL https://arxiv.org/pdf/2006.16668.

Yixuan Li, Chao Ma, Yichao Yan, Wenhan Zhu, and Shuicheng Yan. 3d-aware face swapping. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12705–12714. IEEE, June 2023. doi: 10.1109/cvpr52729.2023.01222. URL http://dx.doi.org/10.1109/cvpr52729.2023.01222.

Zhian Liu. Segnext-faceparser, 2023. URL A-github-repo.url-is-removed-to-avoid-being-desk-rejected.

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*, pp. 8578–8587, 2023.

Xiangyang Luo, Ye Zhu, Yunfei Liu, Lijian Lin, Cong Wan, Zijian Cai, Yu Li, and Shao-Lun Huang. Canonswap: High-fidelity and consistent video face swapping via canonical space modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10064–10074, 2025.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pp. 1930–1939. ACM, July 2018. doi: 10.1145/3219819.3220007. URL https://dl.acm.org/doi/pdf/10.1145/3219819.3220007.

Yue Ma, Hongyu Liu, Hongfa Wang, Hongjian Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, and Qifeng Chen. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–12. ACM, December 2024. doi: 10.1145/3680528.3687587. URL https://dl.acm.org/doi/pdf/10.1145/3680528.3687587.

Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019. doi: 10.1109/TAFFC.2017.2740923.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4296–4304, 2024.

Maxim Nikolaev, Mikhail Kuznetsov, Dmitry Vetrov, and Aibek Alanov. Hairfastgan: Realistic and robust hair transfer with a fast encoder-based approach. In *Advances in Neural Information Processing Systems 37*, volume 37 of *NeurIPS 2024*, pp. 45600–45635. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-1450. URL http://dx.doi.org/10.52202/079017-1450.

Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model for id-consistent human faces. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 241–261, 2024. doi: 10.1007/978-3-031-72913-3_14.

Lixiong Qin, Mei Wang, Xuannan Liu, Yuhang Zhang, Wei Deng, Xiaoshuai Song, Weiran Xu, and Weihong Deng. Faceptor: A generalist model for face perception. In *Computer Vision – ECCV 2024*, pp. 240–260. Springer Nature Switzerland, October 2024. ISBN 9783031727542. doi: 10.1007/978-3-031-72754-2_14. URL http://dx.doi.org/10.1007/978-3-031-72754-2_14.

Di Qiu, Zhengcong Fei, Rui Wang, Jianjun Bai, Changqian Yu, Mingyuan Fan, Guibin Chen, and Xiang Wen. Skyreels-a1: Expressive portrait animation in video diffusion transformers. *arXiv preprint arXiv:2502.10841*, 2025. doi: 10.48550/arxiv.2502.10841. URL https://arxiv.org/pdf/2502.10841.

Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 2021.

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*, pp. 10684–10695, 2022.

Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10779–10788. IEEE, June 2022. doi: 10.1109/cvpr52688.2022.01052. URL http://dx.doi.org/10.1109/cvpr52688.2022.01052.

Yang Shurong, Huadong Li, Juhao Wu, Minhao Jing, Linze Li, Renhe Ji, Jiajun Liang, and Haoqiang Fan. Megactor: Harness the power of raw video for vivid portrait animation. *arXiv preprint arXiv:2405.20851*, 2024. doi: 10.48550/arxiv.2405.20851. URL https://arxiv.org/pdf/2405.20851.

Yang Shurong, Huadong Li, Juhao Wu, Minhao Jing, Linze Li, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Jinfeng Wang. Megactor-sigma: Unlocking flexible mixed-modal control in portrait animation with diffusion transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(9):9256–9264, April 2025. ISSN 2159-5399. doi: 10.1609/aaai.v39i9.33002. URL https://ojs.aaai.org/index.php/AAAI/article/download/33002/35157.

Yanan Sun, Yanchen Liu, Yinhao Tang, Wenjie Pei, and Kai Chen. Anycontrol: create your artwork with versatile control on text-to-image generation. In *European Conference on Computer Vision*, pp. 92–109. Springer, 2024.

Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *Fourteenth ACM Conference on Recommender Systems*, RecSys '20, pp. 269–278. ACM, September 2020. doi: 10.1145/3383313.3412236. URL http://dx.doi.org/10.1145/3383313.3412236.

Qinghe Wang, Lijie Liu, Hua Miao, Qian He, Pengfei Zhu, Bing Cao, and Qinghua Hu. Hs-diffusion: Semantic-mixing diffusion for head swapping, 2022. URL https://arxiv.org/pdf/2212.06458.

Runqi Wang, Sijie Xu, Tianyao He, Yang Chen, Wei Zhu, Dejia Song, Nemo Chen, Xu Tang, and Yao Hu. Dynamicface: High-quality and consistent video face swapping using composable 3d facial priors. *arXiv preprint arXiv:2501.08553*, 2025. doi: 10.48550/arxiv.2501.08553. URL https://arxiv.org/pdf/2501.08553.

Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024.

Mengting Wei, Yante Li, Tuomas Varanka, Yan Jiang, and Guoying Zhao. Magicportrait: Temporally consistent face reenactment with 3d geometric guidance. *arXiv preprint arXiv:2504.21497*, 2025a.

Mengting Wei, Tuomas Varanka, Yante Li, Xingxun Jiang, Huai-Qian Khor, and Guoying Zhao. Towards consistent and controllable image synthesis for face editing. *arXiv preprint arXiv:2502.02465*, 2025b. doi: 10.48550/arxiv.2502.02465. URL https://arxiv.org/pdf/2502.02465.

Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Zhentao Tan, Lu Yuan, Weiming Zhang, and Nenghai Yu. Hairclip: Design your hair by text and reference image. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18051–18060. IEEE, June 2022. doi: 10.1109/cvpr52688.2022.01754. URL http://dx.doi.org/10.1109/cvpr52688.2022.01754.

Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Weiming Zhang, Gang Hua, and Nenghai Yu. Hairclipv2: Unifying hair editing via proxy feature blending. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23532–23542. IEEE, October 2023. doi: 10.1109/iccv51070.2023.02156. URL http://dx.doi.org/10.1109/iccv51070.2023.02156.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12863–12872, 2021.

Bin Xia, Bohao Peng, Yuechen Zhang, Junjia Huang, Jiyang Liu, Jingyao Li, Haoru Tan, Sitong Wu, Chengyao Wang, Yitong Wang, et al. Dreamomni2: Multimodal instruction-based editing and generation. *arXiv preprint arXiv:2510.06679*, 2025.

Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 656–665, 2022. doi: 10.1109/cvprw56347.2022.00081.

Chao Xu, Jiangning Zhang, Yue Han, Guanzhong Tian, Xianfang Zeng, Ying Tai, Yabiao Wang, Chengjie Wang, and Yong Liu. Designing one unified framework for high-fidelity face reenactment and swapping. In *Computer Vision – ECCV 2022*, pp. 54–71. Springer Nature Switzerland, 2022. ISBN 9783031197840. doi: 10.1007/978-3-031-19784-0_4. URL http://dx.doi.org/10.1007/978-3-031-19784-0_4.

Chao Xu, Yijie Qian, Shaoting Zhu, Baigui Sun, Jian Zhao, Yong Liu, and Xuelong Li. Uniface++: Revisiting a unified framework for face reenactment and swapping via 3d priors. *International Journal of Computer Vision*, 133(7):4538–4554, March 2025a. ISSN 1573-1405. doi: 10.1007/s11263-025-02395-6. URL http://dx.doi.org/10.1007/s11263-025-02395-6.

Chen Xu, Keke He, Junwei Zhu, Yanhao Ge, Wei Li, and Chengjie Wang. Hifivfs: High fidelity video face swapping. *arXiv preprint arXiv:2411.18293*, 2024a. doi: 10.48550/arxiv.2411.18293. URL https://arxiv.org/pdf/2411.18293.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1481–1490, 2024b.

Zunnan Xu, Zhen Yu, ZIQING zhou, Jun Zhou, Xiaoyu Jin, Fa-Ting Hong, Xiaozhong Ji, Junwei Zhu, Chengfei Cai, Shiyu Tang, Qin Lin, Xiu Li, and Qinglin Lu. Hunyuanportrait: Implicit condition control for enhanced portrait animation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15909–15919. IEEE, June 2025b. doi: 10.1109/cvpr52734.2025.01483. URL http://dx.doi.org/10.1109/cvpr52734.2025.01483.

Zejun Yang, Huawei Wei, and Zhisheng Wang. Realportrait: Realistic portrait animation with diffusion transformers. volume 39, pp. 9345–9353. Association for the Advancement of Artificial Intelligence (AAAI), April 2025. doi: 10.1609/aaai.v39i9.33012. URL https://ojs.aaai.org/index.php/AAAI/article/download/33012/35167.

Fulong Ye, Miao Hua, Pengze Zhang, Xinghui Li, Qichao Sun, Songtao Zhao, Qian He, and Xinglong Wu. Dreamid: High-fidelity and fast diffusion-based face swapping via triplet id group learning, 2025. URL https://arxiv.org/abs/2504.14509.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoxia Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26125–26135. IEEE, June 2025. doi: 10.1109/cvpr52734.2025.02433. URL http://dx.doi.org/10.1109/cvpr52734.2025.02433.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lq62uWRJjiY.

Yuxuan Zhang, Qing Zhang, Yiren Song, Jichao Zhang, Hao Tang, and Jiaming Liu. Stable-hair: Real-world hair transfer via diffusion model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(10):10348–10356, April 2025. ISSN 2159-5399. doi: 10.1609/aaai.v39i10.33123. URL https://ojs.aaai.org/index.php/AAAI/article/download/33123/35278.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:11127–11150, 2023a.

Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8568–8577. IEEE, June 2023b. doi: 10.1109/cvpr52729.2023.00828. URL http://dx.doi.org/10.1109/cvpr52729.2023.00828.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M. Dai, Zhifeng Chen, Quoc V. Le, and James Laudon. Mixture-of-experts with expert choice routing. volume 35, pp. 7103–7114, 2022.

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*, pp. 650–667, 2022a. doi: 10.1007/978-3-031-20071-7_38.

Jinguo Zhu, Xizhou Zhu, Wenhai Wang, Xiaohua Wang, Hongsheng Li, Xiaogang Wang, and Jifeng Dai. Uni-perceiver-moe: Learning sparse generalist models with conditional moes. *Advances in Neural Information Processing Systems*, 35:2664–2678, 2022b.

Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. Barbershop: Gan-based image compositing using segmentation masks. *ACM Trans. Graph.*, 40(6), December 2021. ISSN 0730-0301. doi: 10.1145/3478513.3480537. URL https://doi.org/10.1145/3478513.3480537.

The supplementary material is organized as follows:

- Section A provides comprehensive implementation details:
  - Sec. A.1 & A.2: The model pipeline, model training process, and algorithms for initialization and forwarding.
  - Sec. A.3: Data construction strategies.
  - Sec. A.4 & A.5: Evaluation protocols and baseline settings.
  - Sec. A.6: Adaptation to compositional transfer.
- Section B presents extensive analysis and ablation studies:
  - Sec. B.1 discusses the generative models used in data corruption.
  - Sec. B.2: Quantitative ablations on adaptive LoRA rank allocation.
  - Sec. B.3: Visual examples of our data filtering.
  - Sec. B.4 discusses StyleGAN-based methods.
- Section C provides extended quantitative comparisons against state-of-the-art methods:
  - Sec. C.1: Additional metrics.
  - Sec. C.2: A computational efficiency analysis.
  - Sec. C.3: Evaluations under challenging conditions like extreme poses, exaggerated expressions, and occlusions.
  - Sec. C.4: Human evaluation.
- Section D provides additional visual comparisons across various tasks.
- Section E discusses the limitations of our work.

A Implementation Details

A.1 Model and training process

Overall Pipeline Our framework is built upon the Stable Diffusion v1.5(Rombach et al., 2022) backbone. As shown in Fig. 3, given a target image I_{tgt} and a reference image I_{ref}, we first mask out irrelevant attributes in both. Following REFace (Baliah et al., 2025), we extract semantic features of masked I_{ref} via the CLIP vision encoder (Radford et al., 2021) and identity features (for tasks that I_{ref} provides identity attribute like face transfer and head transfer) via the Arcface face recognizer (Deng et al., 2019). Task-specific single-layer MLPs project these raw features into tokens for cross attention conditioning in Unets. We form blended landmarks to guide motion, taking pose/expression from the original driving image (I_{ref} for reenactment, I_{tgt} otherwise) and identity from the counterpart. The masked target is encoded by the VAE to a latent and concatenated with noise along the channel dimension as input to the main UNet. The masked reference flows through the refNet (Reference-UNet), and its tokens are injected via cross-attention into the main UNet (the context tokens in main Unet are concatenated with the tokens from refNet). Different from common refNet used in other works (Hu, 2024; Xu et al., 2024b; Wei et al., 2024; Zhang et al., 2025; Chung et al., 2025; Baliah et al., 2025; Han et al., 2024), we use only the first half of the refNet as we found it is enough to achieve a good performance and yields better efficiency. Finally, the main UNet performs iterative denoising on the concatenated latents, which are then decoded back into pixel space by the VAE decoder to produce the final swapped result I_{out}.

Training process

In the main paper for simplicity, we only present the MoE mechanism for FFN. Here we show a more detailed MoE mechanism.

In stage one, we isolate gradients across tasks. After training for 100K steps, we proceed to stage-II. Initialization of stage-II:

1) Weight initialization of stage II. As in Section 3.3, we average Stage-I FFNs to initialize the shared/global expert: $W_{\text{gl}} = \frac{1}{N}\sum_j W_j$. For task j, its task-specific expert is initialized on the residual $W_j - W_{\text{gl}}$. We perform SVD on the residual and select the smallest rank r such that the cumulative energy exceeds a global threshold $\tau = 0.2$, yielding an adaptive yet unified criterion across tasks/layers. Note that each specialized expert is actually an FFN with 2 MLPs in LoRA form (but we depicted it as a single LoRA structure for simplicity in the algorithm part and figure). And W represents each constituent linear projection matrix of the FFN (i.e., the up- and down-projections). The weight residual decomposition SVD$(W_j - W_{\text{gl}})$ and subsequent LoRA parameterization are executed independently for each linear layer.

Pseudocode: Weight initialization of stage II

1. Inputs: Stage-I FFNs $\{W_j\}_{j=1}^{N}$, global threshold τ.
2. Global: $W_{\text{gl}} \leftarrow \frac{1}{N}\sum_j W_j$.
3. For each task j: SVD $(W_j - W_{\text{gl}}) = U\Sigma V^\top$.
4. Pick r s.t. $\sum_{i\le r}\sigma_i^2 / \sum_i \sigma_i^2 \ge \tau$ (same τ for all).
5. Initialize task-j LoRA by rank-r reconstruction of the residual.

2) In Stage II, we introduce E=8 routed experts and a per-task router $R^{(j)}$ (single-layer MLP). For each token, the router activates the top K=2 experts. All routed experts share the FFN topology but reduce the hidden width by factor F=16, balancing capacity and cost. Parameters are randomly initialized before joint training.

After initialization of stage II, we jointly train the N=4 tasks (including face transfer, hair transfer, motion transfer, and head transfer). We train for 120K steps.

The model is trained on 4 NVIDIA A40 GPUs for around 8.5 days. Each GPU receives data from one task and the batch size is set to one. In Stage I, each task is optimized separately without gradient synchronization. In Stage II, we start from the Weight initialization of stage II and perform joint training with synchronous updates at each iteration. We utilize the AdamW optimizer across

Algorithm 1 Weight initialization of stage II

Require: set of N trained task-specific FFN weights $\{W_1, \ldots, W_N\}$, energy threshold τ.
Ensure: Initialized Global Expert W_{gl}, set of task-specific experts $\{(A_j, B_j)\}_{j=1}^{N}$.
Declarations:
$j \in \{1, \ldots, N\}$ ▷ Task ID
$W_j, \bar{W}, W_{\text{gl}} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ ▷ FFN weight matrices
$S \in \mathbb{R}^{\min(d_{\text{out}}, d_{\text{in}})}$ ▷ Singular values
Step 1: Consensus Initialization
$\bar{W} \leftarrow \frac{1}{N} \sum_{j=1}^{N} W_j$ ▷ Compute consensus (average) weights
$W_{\text{gl}} \leftarrow \bar{W}$ ▷ Initialize Global Expert
Step 2: Residual-based Adaptive LoRA Initialization
for $j = 1$ to N **do**
  $U, S, V^\top \leftarrow \text{SVD}(W_j - W_{\text{gl}})$ ▷ Singular Value Decomposition of residual
  $E_{\text{total}} \leftarrow \sum_k S_k^2$ ▷ Calculate total energy
  $r_j \leftarrow 0, \quad E_{\text{cum}} \leftarrow 0$
  while $E_{\text{cum}}/E_{\text{total}} < \tau$ **do** ▷ Determine adaptive rank
    $r_j \leftarrow r_j + 1$
    $E_{\text{cum}} \leftarrow E_{\text{cum}} + S_{r_j}^2$
  end while
  Initialize LoRA parameters A_j, B_j with rank r_j to approximate the residual.
end for

Algorithm 2 BioMoE Forward Process

Require: Input token feature $x \in \mathbb{R}^D$, Global context feature $\text{pool}(X) \in \mathbb{R}^D$, Structure information $\mathcal{S}(x) \in \mathbb{R}^{2M}$, Task ID j.
Ensure: Output feature $f(x) \in \mathbb{R}^D$.
Declarations:
$f_{\text{gl}} : \mathbb{R}^D \to \mathbb{R}^D$ ▷ Global Shared Expert
$\{f_{\text{rt}}^{(i)}\}_{i=1}^{E}, f_{\text{rt}}^{(i)} : \mathbb{R}^D \to \mathbb{R}^D$ ▷ Set of E Routed Experts
$f_{\text{sp}}^{(j)} : \mathbb{R}^D \to \mathbb{R}^D$ ▷ Task-specific Expert of task-j.
$R^{(j)} : \mathbb{R}^{D_{\text{gate}}} \to \mathbb{R}^E$ ▷ Task-j Structure-Aware Router
$K = 2$ ▷ Number of active experts
$\epsilon \sim \mathcal{N}(0, 1)$ ▷ Gaussian noise for exploration
Step 1: Structure-Aware Routing
$g \leftarrow \text{concat}(x, \text{pool}(X), \mathcal{S}(x))$ ▷ Assemble gating input
$s \leftarrow R^{(j)}(g)$ ▷ Predict expert scores
$\mathcal{G}^{(j)}(x) \leftarrow \text{top}_K(\text{softmax}(s + \epsilon))$ ▷ Select top-K experts
Step 2: Aggregate Experts and Final Fusion
$f(x) \leftarrow f_{\text{gl}}(x) + f_{\text{sp}}^{(j)}(x) + \sum_{i=1}^{E} \mathcal{G}_i^{(j)}(x) \cdot f_{\text{rt}}^{(i)}(x)$

all training stages. In Stage I (Task-Isolated Initialization), the base learning rate is set to 1.0×10^{-5}, with 10000 warm-up steps starting from a factor of 1.0×10^{-6} of the base learning rate. In Stage II (Task-Unified Fine-tuning), we increase the base learning rate to 4.0×10^{-4}, along with 10000 warm-up steps starting from a factor of 0.1.

We will release the source code and model weight which contains more details for better reproducibility.

A.2 Training loss

Following REFace (Baliah et al., 2025), our training objective consists of a standard DDPM loss $\mathcal{L}_{\text{diff}}$ and a multi-step DDIM-based perceptual loss $\mathcal{L}_{\text{perceptual}}$:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{perceptual}} \mathcal{L}_{\text{perceptual}}(I_{\text{gt}}, \hat{I}) + \lambda_{\text{id}} \mathcal{L}_{\text{id}}(I_{\text{gt}}, \hat{I}) \tag{1}$$

where $\mathcal{L}_{\text{diff}}$ is the standard noise prediction loss at latent space, and $\mathcal{L}_{\text{perceptual}}, \mathcal{L}_{\text{id}}$ are image-space enhancement losses computed between DDIM-sampled image $\hat{I}$ and the ground-truth I_{gt} at pixel space.

Latent-space Loss. We utilize the standard noise prediction loss $\mathcal{L}_{\text{diff}}$ to train the diffusion model:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,1),t}[\|\epsilon - \epsilon_\theta(z_t, t, \mathbf{c})\|_2^2], \tag{2}$$

where z_t is the noisy latent at timestep t and $\mathbf{c}$ represents the task-specific conditions, including the reference features and structure guidance. This loss allows the model to efficiently learn the mapping for various biometric transfer tasks in the compressed latent space.

Image-space Loss. To further refine the visual details and enhance identity preservation, we incorporate image-space losses computed on the reconstructed image $\hat{I}$:

$$\mathcal{L}_{\text{img}} = \lambda_{\text{perceptual}}\mathcal{L}_{\text{perceptual}}(I_{\text{gt}}, \hat{I}) + \lambda_{\text{id}}\mathcal{L}_{\text{id}}(I_{\text{gt}}, \hat{I}). \tag{3}$$

The reconstructed image $\hat{I}$ is obtained by applying $N_{\text{step}} = 4$ DDIM denoising steps starting from a noisy latent. Each DDIM step requires a complete feed-forward pass of the network. Consequently, the memory requirement for computing $\mathcal{L}_{\text{img}}$ is significantly higher than that for $\mathcal{L}_{\text{diff}}$, as it involves N_{step} times the computation and gradient storage. This leads to a reduced batch size and slower training speed when these losses are active.

Training Schedule. Given the computational overhead of the image-space losses, we adopt a two-phase weighting strategy (referred to as sub-stages to distinguish them from the two stages in our main pipeline). In sub-stage (i), we set $\lambda_{\text{perceptual}} = 0$ and $\lambda_{\text{id}} = 0$, using $\mathcal{L}_{\text{diff}}$ to achieve fast convergence in structure and basic attributes. The batch size is 4 per GPU. In sub-stage (ii), we activate the image-space losses to boost perceptual realism and identity fidelity, setting $\lambda_{\text{perceptual}} = 0.05$ and $\lambda_{\text{id}} = 0.3$. The batch size is 1 per GPU. We allocate the GPU training time between sub-stage (i) and sub-stage (ii) in a ratio of approximately 6:1.

A.3 Training Data

A.3.1 Datasets

To acquire real images and videos for training data construction, for the face transfer task, we utilize the entire VFHQ (Xie et al., 2022) dataset, which consists of 15K videos and a subset of the Arc2Face dataset (Papantoniou et al., 2024), comprising 60K images across 30K distinct identities. For glasses tasks, we use dataset from (Egorovlvan, 2021) containing 7K real images with glasses after filter. For hat tasks we use images with hats in FFHQ (Karras et al., 2019) (separate from testset) and VFHQ (Xie et al., 2022) dataset, resulting in 5K real images. For other tasks, celebV-HQ (Zhu et al., 2022a) and VFHQ (Xie et al., 2022) are used. After filter, we get 80K real images from the two datasets. For each real image, we will generate one to two training pairs.

A.3.2 Training data construction

Our unified data corruption strategy handles different attribute types, (a) Relative-static attributes. Relative-static attributes include both structural and non-structural elements whose spatial locations and overall shapes remain largely consistent during transfer. For structural attributes, e.g., face, eyes, lips, nose, which have well-defined geometric boundaries, we apply mask-based corruption, a widely used and effective strategy that removes texture and color information, forcing the model to reconstruct these details from the reference. For non-structural attributes, e.g., skin tone, we employ standard data augmentation techniques to introduce variation while preserving structure. Taking face transfer as an example, the target image I'_{tgt} is constructed by masking out the face in I_{gt}, preserving only the background, skin tone (from the neck region), and pose-expression (from landmarks). The reference image I'_{ref} is created by corrupting all attributes except the face: background, hair, skin tone (via lighting augmentation), and pose-expression (via warp augmentation or selecting a different frame from the video). (b) Spatially-dynamic attributes. Taking hair transfer as an example, the target image I'_{tgt} is constructed by corrupting the hair in I_{gt} using the swapping-based corruption strategy.

In Fig. 6, we show four specific examples of data construction.

A.3.3 Preprocess

All inputs are resized to 512×512, and irrelevant regions are masked out for each image (e.g., for head transfer we preserve only the head region in the reference image and mask out the head region

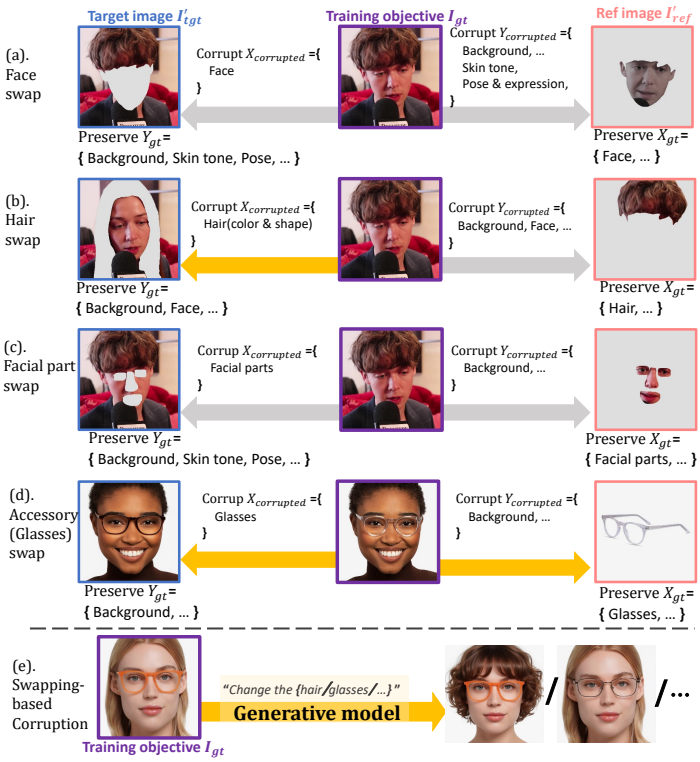


Figure 6: We show the specific examples of data construction on 4 tasks in (a)-(d). Grey arrow means the corruption process only involves Relative-Static Attributes Corruption. Yellow arrow means it involves Spatially-Dynamic Attributes Corruption, thus need the Swapping-Based Corruption, as shown in (e).

in the target image) to ensure consistency with the training process. We use the semantic mask extracted from SegNeXt-FaceParser (Liu, 2023; Liu et al., 2023) following previous work (Baliah et al., 2025; Han et al., 2024; Wei et al., 2025b). The network outputs at 512×512 resolution.

For landmark detection, we utilize the MediaPipe model, specifically using the facial part landmarks including pupil points while discarding all face contour points. For face-related tasks, when the facial structures (measured by nose-to-mouth distance, nose-to-eye distance, and inter-pupil distance) of the two input images differ substantially, we employ a rule-based landmark blending strategy to combine the identity from one landmark set with the motion from the other, so that the blended landmarks can better guide both identity preservation and motion transfer. Given two MediaPipe 3D landmark sets, A (expression source) and B (identity target), we compute blended landmarks by preserving B's facial proportions and injecting A's expression. We estimate a canonical vertical facial direction by fitting a line to the eye midpoint, nose tip, and mouth center. Then we translate the mouth region and the eye/eyebrow region of A along this vertical direction such that the projected nose-to-mouth and nose-to-eye distances match those measured on B. Finally, we enforce the inter-pupil distance of B through symmetric horizontal translations of the left/right eye regions, while leaving other landmarks unchanged. The blended 3D landmarks are finally projected into 2D landmarks as output.

A.4 Evaluation Protocol

Following the evaluation protocol established in previous work (Baliah et al., 2025), we construct our test set using 1,000 image pairs sampled from the FFHQ dataset. To further ensure a fair comparison with other SoTA methods (e.g., HairFusion) that rely on some pre-trained detection models (e.g., dlib), we exclude 13 pairs where these detectors completely fail. Consequently, our final evaluation set consists of 987 image pairs. Following REFace, we adopt face identity similarity (ID Sim.), pose distance (Pose Dist.) (Ruiz et al., 2018), expression distance (Expr. Dist.), CLIP distance (CLIP Dist.), and structural similarity (SSIM) for quantitative evaluation. For the Expression Distance (Expr. Dist.), we compute the L2 distance between the expression coefficients extracted by the Deep3DFaceRecon model.

A.5 Evaluation of Baselines

For most state-of-the-art (SOTA) methods, we utilize their official codebases to perform inference on the test set.

RigFace. Due to its reliance on multiple pre-trained detection models, RigFace is sensitive to failure cases. During metric calculation, we exclude failed pairs and compute metrics only on successful samples.

REFace for head transfer. REFace supports both face and head transfer via mask shuffling as mentioned in its paper. As the official release only includes training code for face transfer, we modify the training code to support head transfer by expanding the mask region to encompass the hair and accessories, following the methodology described in its paper. We then retrain the model using the official hyperparameters.

GHOST 2.0. For the head transfer task, we follow the setting of previous work (Kang et al., 2025), where the goal is to transfer the head along with the skin tone of the reference. GHOST 2.0 (Groshev et al., 2025) adopts a different setting that preserves the target's skin tone. Therefore, in our visual comparisons, we should only focus on the structural alignment and identity transfer quality rather than skin tone consistency for this specific baseline.

FaceAdapter. For the face transfer task, we follow the setting of previous works (Baliah et al., 2025; Luo et al., 2025), where the face shape in the target image should be preserved. FaceAdapter (Han et al., 2024) adopts a different setting that transfers the reference's face shape. Therefore, in our visual comparisons, we should ignore the face shape preservation for this specific baseline.

A.6 Adaptation to Compositional Transfer

Handling Multiple References. To address compositional tasks that involve multiple reference images (e.g., combining hair from one image and glasses from another), we extract the needed

Table 8: Comparison with Qwen-Image-Edit (general-purpose image editing model). For each task, we report the most representative metric, that is identity similarity, hair clip distance, pose distance, and head clip distance. For all metrics, ↑ indicates higher is better and ↓ indicates lower is better. Best results are in **bold**.

Method	Face transfer↑	Hair transfer↓	Reenact↓	Head transfer↓
Qwen-Image-Edit (arXiv 25)	0.442	0.486	17.13	0.582
Ours	**0.637**	**0.421**	**7.09**	**0.460**

attribute from each reference image using its corresponding mask. These attribute regions are then normalized via cropping and spatially concatenated to form a single composite reference image of size 512×512.

B Further Analysis and Ablation Study

B.1 Discussions on the generative model used in our Swapping-based Corruption.

Our swapping-based corruption strategy consists of two steps: 1. Using a generative model to synthesize a new image from a real image, thereby corrupting the target attribute. 2. Applying a post-synthesis filter to discard synthesized pairs that exhibit poor preservation of non-target attributes.

B.1.1 (a) Discussion on Qwen-Image-Edit.

For the generative model, we typically utilize a general-purpose image editing model, Qwen-Image-Edit.

Limitations in reference-based transfer. As illustrated in Fig. 7-(a) and Section D (Fig. 11-14), Qwen-Image-Edit exhibits unstable, high-variance behavior across samples under the multiple-image mode (i.e., our reference-based transfer setting). Specifically, it may under-follow the reference, over-transfer and damage the target, or hallucinate in the edited region. This high variability, together with its consistently low quantitative results in Tab. 8, makes general-purpose multi-image editors unsuitable for our task.

For Tab. 8, we use the following prompts (image 1: target, image 2: reference): *(1). Face transfer:* "Replace the face in image 1 with the face of image 2. Ensure 100% fidelity to Image 2's facial identity. Only change face region. Strictly preserve the expression, head pose, hair, background, lighting, and skin tone of image 1." *(2). Hair transfer:* "Replace the hair (style, color, and texture) in image 1 with the hair from image 2. Ensure 100% fidelity to Image 2's hair. Only change hair region. Strictly preserve the face, head pose, and background of image 1." *(3). Motion transfer:* "Use the head pose and facial expression of the subject in image 2 to drive the identity of the subject in image 1. Ensure 100% fidelity to Image 2's pose and expression. Only alter head pose and facial expression in image 1. Preserve the face identity, hair, and background of image 1." *(4). Head transfer:* "Replace the entire head (including face and hair) of the subject in image 1 with the head of the subject in image 2. Ensure 100% fidelity to Image 2's head. Only change head region. Preserve the body and background of image 1."

Sufficiency for attribute corruption. However, for our data corruption purposes, reference transfer is not required. As shown in Fig. 7-(b), the model under the single-image mode (instruction-based image editing) is capable of generating plausible and diverse attribute variations.

B.1.2 (b) Discussion on StableHair and self-evolution.

For specific tasks such as hair and head transfer, where spatially-dynamic attributes primarily involve hair, we employ the task-specific model StableHair (Zhang et al., 2025) for efficiency. As demonstrated in the visual comparisons for hair transfer (Fig. 4(c) and Fig. 11), although StableHair displays suboptimal preservation of face lighting and background, our model, trained on data synthesized by StableHair, consistently achieves better preservation fidelity.

Table 9: Quantitative results of the self-evolution experiment on hair transfer task. By leveraging our initial model to construct higher-quality training data, we can obtain an evolved model with improved performance.

Method	CLIP dist.↓	ID sim↑	non-hair SSIM↑
Stable-Hair (AAAI 25)	0.468	0.855	0.87
Our initial model	0.421	0.887	0.91
Our evolved model	**0.419**	**0.888**	**0.94**

Table 10: We compare our Result-based Adaptive LoRA rank against the conventional gradient-based and the uniform rank. For each task, we report the most representative metric, that is identity similarity, hair clip distance, pose distance, and head clip distance.

LoRA Rank Allocation Method	Face transfer↑	Hair transfer↓	Reenact↓	Head transfer↓
Uniform	0.636	0.419	10.13	0.460
Gradient-based	0.634	0.433	7.35	0.464
Result-based (**Ours**)	0.637	0.421	7.09	0.460

Remarks. The proposed data construction strategy can be seen as a way to **transform imperfect generative models into stronger ones**. Specifically, the main strict requirement for the external generative model used in our swapping-based corruption is the ability to introduce diverse attribute variations. It does not need to transfer attributes from a reference image, as we only aim to corrupt specific attributes in the GT image to get target image. Nor does it need to perfectly preserve non-target regions on its own, since low-quality pairs can be easily identified and filtered out using preservation metrics, including background SSIM, identity similarity, pose and expression distance (see Sec. B.3 for visual examples of discarded and retained pairs). Consequently, our trained model can surpass the capabilities of the generator used to create its training data.

This also suggests the possibility of 'self-evolution': by bootstrapping from a suboptimal generative model to synthesize data, we can obtain an improved model, which can then be used to generate even better training data, leading to further model enhancements.

We perform a further experiment on the hair transfer task: we first obtain an initial model (the model presented in our main experiments) trained on data synthesized by the suboptimal generative model StableHair. We then leverage this initial model to construct higher-quality training data, which exhibits better background consistency between the ground truth and target images compared to the original StableHair data. Consequently, this evolved model achieves even greater preservation performance as shown in Tab. 9.

Note. Our Swapping-based Corruption is NOT distillation. Distillation would use an external model in the forward direction to synthesize the ground-truth image that the student is trained to imitate. Ours is the reverse: the external model only **corrupts** selected attributes of a real image, and the supervision target remains that original real image I_{gt}.

B.2 Quantitative ablation of Result-based Adaptive LoRA Rank Allocation.

In Tab. 10, we compare our Result-based Adaptive LoRA Rank Allocation with uniform rank across tasks per layer, and gradient-based adaptive rank, while maintaining a consistent total parameter count. While the other three tasks maintain high performance, motion transfer suffers a **significant decline**. Without enough task-specific capacity for the motion transfer task, the model is heavily influenced by the other three tasks, which are all about transfering some visual elements from reference to target. As a result, it now struggles to drive the target with reference motion during motion transfer. This finding is corroborated by the specific rank distribution: taking the first MoE-FFN layer in the first middle block of the main UNet as an example, the assigned ranks for the face transfer, hair transfer, motion transfer and head transfer are 49, 37, 112 and 48 respectively. That is to say, our strategy assigns the highest rank (i.e., 112) to the motion transfer task, whereas the ranks for the other three tasks remain below 50. This indicates that motion transfer weights diverge significantly from other tasks. Our result-based approach effectively addresses this cross-task difference by automatically identifying and assigning higher ranks to the most conflict-prone tasks.

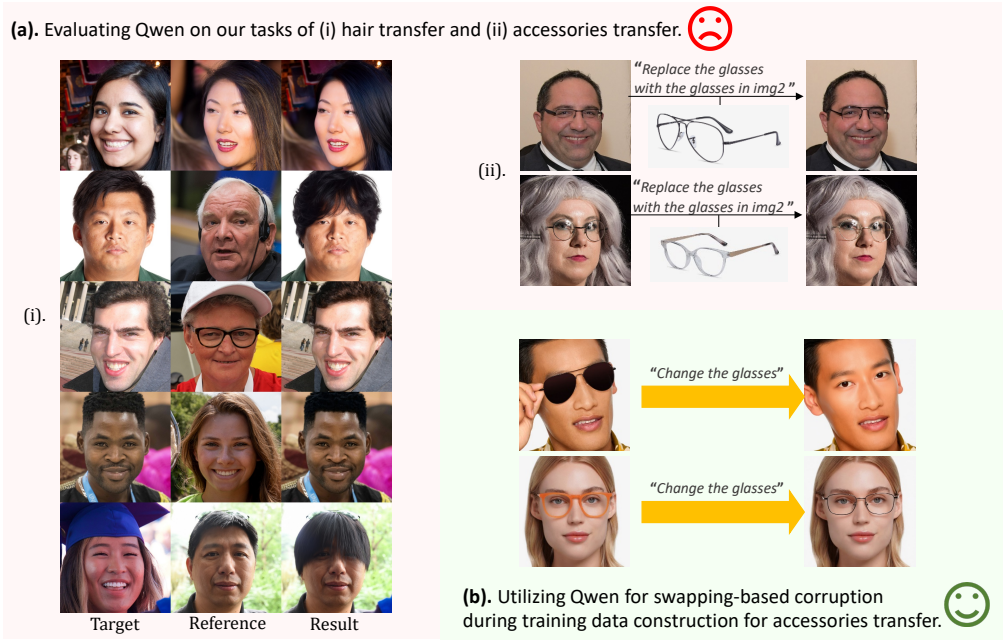


Figure 7: Analysis of the general image editing model Qwen-Image-Edit (Wu et al., 2025). **(a)** Under the multiple-image mode (reference-based editing), Qwen-Image-Edit exhibits many unstable behavior like under-follow the reference, over-transfer and damage the target, or hallucinate in the edited region. (More examples are shown in Section D: Fig. 11- 14). The consistently poor quantitative results shown in Tab. 8 further confirm its **unsuitability for the transfer task**. **(b)** However, under the single-image mode (instruction-based editing), it can still generate plausible and diverse attribute variations, which is **sufficient for swapping-based corruption**.

B.3 Examples of discarded and retained pairs in data construction

Fig. 8 shows examples of retained and discarded image pairs after the filtering step for the hair transfer task. Specifically, pairs are discarded when the ground-truth image I_{gt} and the constructed target image I_{tgt} exhibit low background SSIM, low face identity similarity, large pose L2 distance, or large expression L2 distance. We use percentile-based thresholds rather than absolute values. For each metric, we sort samples from best to worst (for similarity style metrics, higher is better; for distance style metrics, lower is better) and keep the top $Fraction_{\text{metric}}$ fraction. Any sample that fails any metric is discarded (union of rejects), so the final yield is the proportion remaining after all filters are applied. For example, for hair transfer, we use $Fraction_{\text{face_sim}} = 0.8$ (face identity similarity), $Fraction_{\text{pose_dist}} = 0.8$ (pose L2 distance), $Fraction_{\text{exp_dist}} = 0.8$ (expression L2 distance), and $Fraction_{\text{bg_sim}} = 0.5$ (background SSIM).

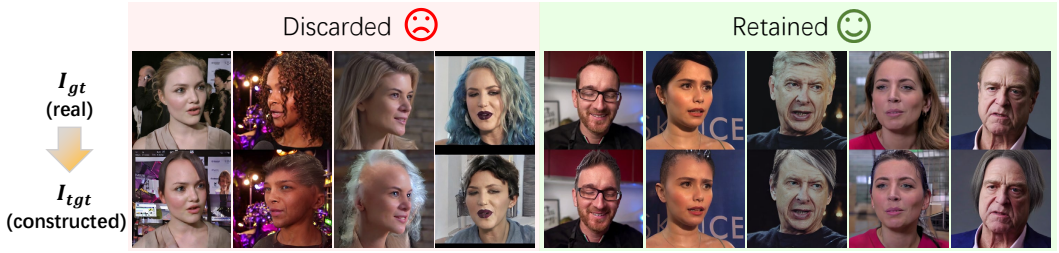


Figure 8: Examples of discarded and retained image pairs after filtering for hair transfer. (a) Discarded pairs due to low background SSIM, low face identity similarity, large pose L2 distance, and large expression L2 distance, respectively; (b) Retained pairs after filtering.

B.4 Discussion on StyleGAN-based Methods

Prior to the proliferation of diffusion models, StyleGAN-based latent manipulation (Abdal et al., 2021; Wu et al., 2021; Zhu et al., 2021) was the dominant paradigm for high-resolution portrait editing. Methods like StyleFlow (Abdal et al., 2021), StyleSpace (Wu et al., 2021), and Barbershop (Zhu et al., 2021) achieve impressive results by exploring disentangled latent spaces for attribute editing and image compositing.

Despite these significant achievements, they suffer from severe limitations in real-world scenarios compared to more recent feed-forward methods. While these methods exhibit excellent performance when both the reference and target images share strictly aligned, frontal poses, they struggle significantly with pose discrepancies. In practical applications, reference and target images are rarely strictly aligned and typically exhibit different head poses. Under such unconstrained conditions, methods like Barbershop (see Fig. 9) fail to geometrically adapt the spatial structure; they either rigidly copy and paste the reference attributes, or generate glaring artifacts and distorted boundaries, as shown in Fig. 9. Furthermore, to achieve an acceptable blending outcome, these methods rely on computationally heavy test-time latent optimization, which takes ≥ 4 minutes per image pair. Such extreme inefficiency makes them highly unscalable and impractical compared to the single feed-forward pass of our proposed method.

C More Quantitative Comparison

C.1 More metrics for head transfer

In the main paper, we only show some key metrics for head transfer to save space. Here we show more metrics in Tab. 11.

C.2 Efficiency comparison

To complement the quality metrics, we also report computational efficiency in Tab. 12, including peak CUDA memory and inference time per image. Our method achieves competitive efficiency among unified methods while maintaining strong transfer performance across multiple tasks.



Figure 9: **Qualitative comparison against the StyleGAN-based method, Barbershop (Zhu et al., 2021).** As indicated by the red arrows, Barbershop encounters obvious visual artifacts and distortions when there is a pose discrepancy between the reference and target images. Moreover, due to its heavy reliance on latent optimization, it requires ≥ 4 minutes to infer a single image pair, rendering it highly unscalable and unsuitable for our unified, efficient multi-task generation task.

Table 11: Quantitative comparison with more metrics for head transfer. For all metrics, ↑ indicates higher is better and ↓ indicates lower is better. Best result from Unified Methods are marked Red. Overall best results are highlighted in **bold**, while overall second-best results are underlined.

Method	CLIP dist.↓	ID sim↑	pose dist.↓	expr. dist.↓	CLIP dist. (hair)↓	FID↓
Task-Specific Methods						
GHOST2.0 (arXiv 25)	0.618	0.461	5.78	1.867	0.497	31.38
Multi-Task Methods (training separately for different tasks)						
REFace (WACV 25)	0.639	0.540	5.43	1.776	0.549	8.33
Unified Methods (train once for all tasks)						
Ours	**0.460**	**0.545**	**5.28**	**1.772**	**0.399**	**8.28**

C.3 Comparisons under complex scenes: extreme poses, exaggerated expressions, and occlusions

We show quantitative comparisons under complex scenes in Tab.13 and visual comparisons in Fig.15, Fig.16, Fig.17, and Fig.18. To construct the visual comparison set, we manually select images with occlusions from the FFHQ dataset (Karras et al., 2019). Because exaggerated expressions and extreme poses are rare in FFHQ, we source representative samples from the AffectNet dataset (Mollahosseini et al., 2019) and the ExtremePose-Face-HQ (EFHQ) dataset (Dao et al., 2024), respectively.

For the quantitative analysis, we also evaluate performance across three distinct scenarios: exaggerated expressions using AffectNet, extreme poses using EFHQ, and occlusions using the Real World Occluded Faces (ROF) dataset (Erakın et al., 2021). For each scenario, we sample 333 images and pair each with a standard image from the FFHQ test set, yielding 333 test pairs per setting. When computing metrics, as some approaches fail to process certain images due to brittle preprocessing steps (e.g., failed face detection), we omit these failure cases from the affected methods' evaluations to ensure fair metric computation.

Table 12: Efficiency comparison: Runtime and Peak CUDA memory. For all metrics, ↓ indicates lower is better. For multi-task methods, we report the maximum runtime and peak CUDA memory across all tasks. We achieve comparable inference efficiency to all other diffusion-based methods.

Method	Peak CUDA memory (GB)↓	Infer time (second/img)↓
Task-Specific Methods		
CanonSwap (ICCV 25)	11.7	< 1
Stable-Hair (AAAI 25)	19.3	16
HairFusion (AAAI 25)	10.9	22
HunyuanPortrait (CVPR 25)	18.4	103
GHOST2.0 (arXiv 25)	8.0	< 1
Multi-Task Methods (training separately for different tasks)		
REFace (WACV 25)	11.8	12
Unified Methods (train once for all tasks)		
Face-Adapter (ECCV 24)	6.8	3
RigFace (arXiv 25)	21.9	18
Ours	14.5	15

Table 13: Comparisons on the most competitive methods for each task under complex scenes.

Task	**Method**	**Exaggerated Expressions**			**Extreme Pose**			**Occlusions**		
		ID sim↑	pose dist.↓	expr. dist.↓	ID sim↑	pose dist.↓	expr. dist.↓	ID sim↑	pose dist.↓	expr. dist.↓
Face Transfer	REFace	0.509	5.10	1.28	0.293	10.09	1.10	0.311	5.93	1.07
	CanonSwap	0.446	**2.27**	**0.91**	0.141	**3.73**	0.75	0.152	**2.53**	**0.76**
	Ours	**0.518**	3.54	1.04	**0.302**	4.25	**0.73**	**0.316**	4.80	1.06
		CLIP dist.↓	ID sim↑	non-hair SSIM↑	CLIP dist.↓	ID sim↑	non-hair SSIM↑	CLIP dist.↓	ID sim↑	non-hair SSIM↑
Hair Transfer	Stable-Hair	0.460	0.869	0.871	**0.477**	0.831	0.880	0.472	0.826	0.848
	Ours	**0.436**	**0.904**	**0.931**	**0.477**	**0.899**	**0.941**	**0.452**	**0.883**	**0.930**
		ID sim↑	pose dist.↓	expr. dist.↓	ID sim↑	pose dist.↓	expr. dist.↓	ID sim↑	pose dist.↓	expr. dist.↓
Motion Transfer	Face-Adapter	0.492	**7.75**	2.31	0.337	**11.93**	2.60	0.472	10.61	2.46
	HunyuanPortrait	0.357	8.39	2.42	0.237	13.48	2.65	0.484	11.84	2.56
	Ours	**0.497**	8.16	**2.27**	**0.338**	12.63	**2.21**	**0.509**	**10.15**	**2.32**
		CLIP dist.↓	ID sim↑	pose dist.↓	CLIP dist.↓	ID sim↑	pose dist.↓	CLIP dist.↓	ID sim↑	pose dist.↓
Head Transfer	GHOST2.0	0.495	0.369	6.07	0.525	0.257	8.32	0.512	0.332	8.15
	REFace	0.552	0.568	5.87	0.559	0.465	14.56	0.555	0.458	11.54
	Ours	**0.416**	**0.573**	**5.20**	**0.479**	**0.484**	**8.02**	**0.424**	**0.490**	**7.43**

As shown in Tab. 13, Fig. 15, Fig. 16, Fig. 17 and Fig. 18, all methods naturally suffer performance degradation under these highly challenging conditions. Nevertheless, our method still achieves the best overall performance.

C.4 Human Evaluation

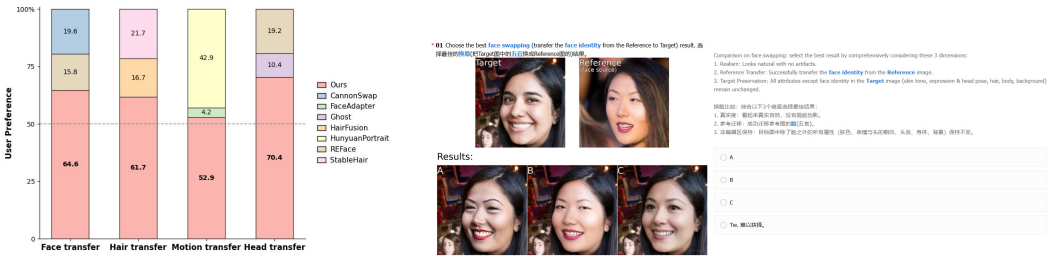
We recruited 16 participants for a user study, voting on the best result among 3 methods (ours + two top baselines) across the four main tasks based on realism, transfer consistency, and preservation. Each task contains 5 questions, resulting in 20 questions in total. As shown in Fig. 10(a), our method achieves preference rates of 64.6%, 61.7%, 52.9%, and 70.4% on face transfer, hair transfer, motion transfer, and head transfer, respectively, significantly outperforming the baseline methods. An example of the questionnaire presented to participants is shown in Fig. 10(b).

D More Visual Comparison

We provide more per-task visual results in Fig. 11, Fig. 12, Fig. 13, and Fig. 14. And we provide complex-scene visual comparisons (extreme poses, exaggerated expressions, occlusions) in Fig. 15 to Fig. 18.

E Limitations

Despite the improvements achieved by **UniBioTransfer**, some limitations remain:



(a) User preference results (b) Example questionnaire question

Figure 10: **Human evaluation study.** (a) User preference voting results across the four main tasks comparing our method against two top baselines. (b) An example question from the user study questionnaire.

- **Performance on Asian faces.** Our model sometimes exhibits suboptimal performance on certain ethnic groups, particularly Asian faces. This may be mostly attributed to the data distribution of the training datasets, which is also faced by many works in the field.
- Due to computational resource constraints, we only conducted experiments using Stable Diffusion v1.5 as the backbone (And for fair comparison with previous works like StableHair, REFacce). Theoretically, our entire framework, including the unified data construction, BioMoE and two-stage training strategy, can be applied to other backbones like FLUX. We leave these explorations for future work.

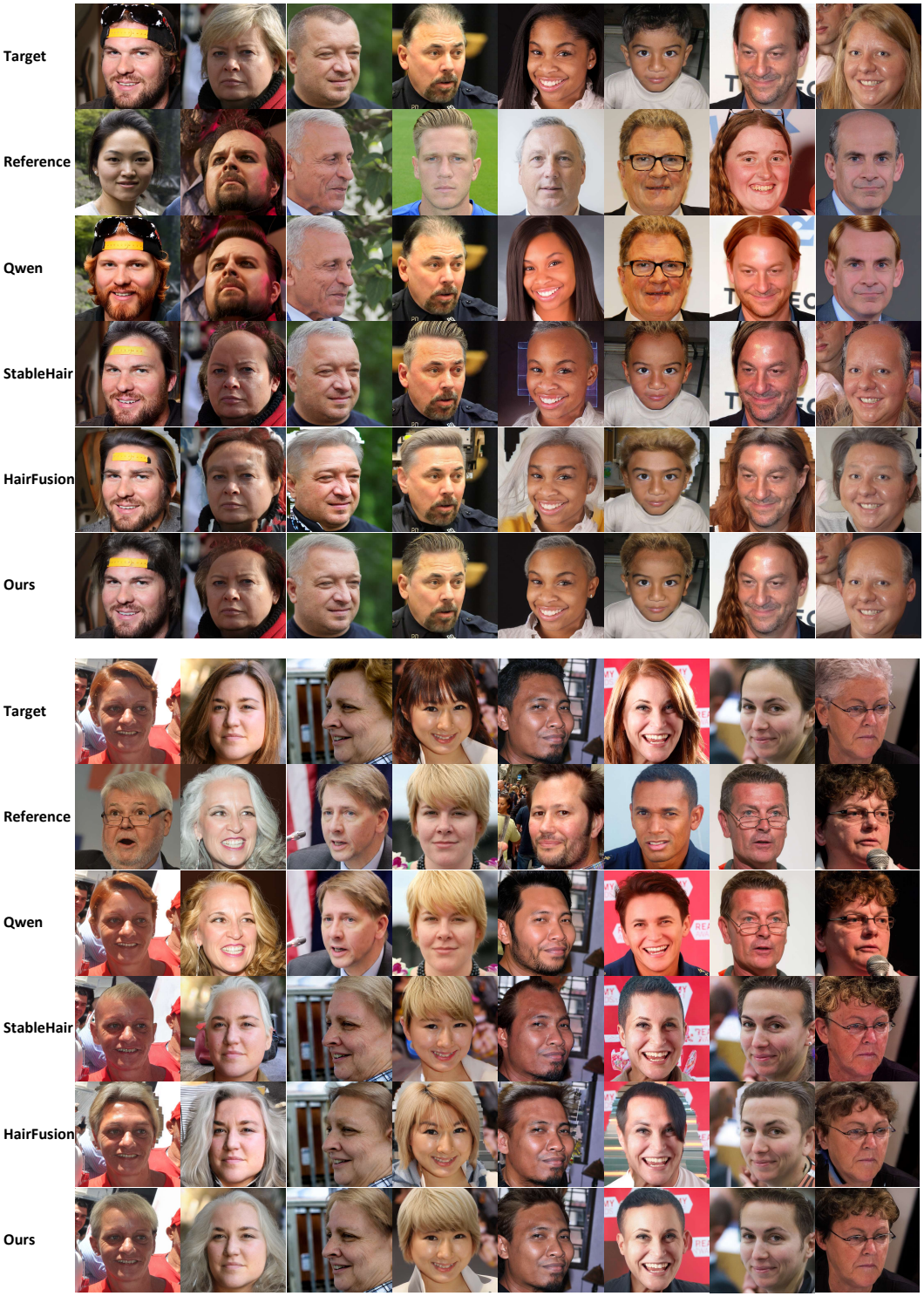


Figure 11: Hair transfer visual comparison.

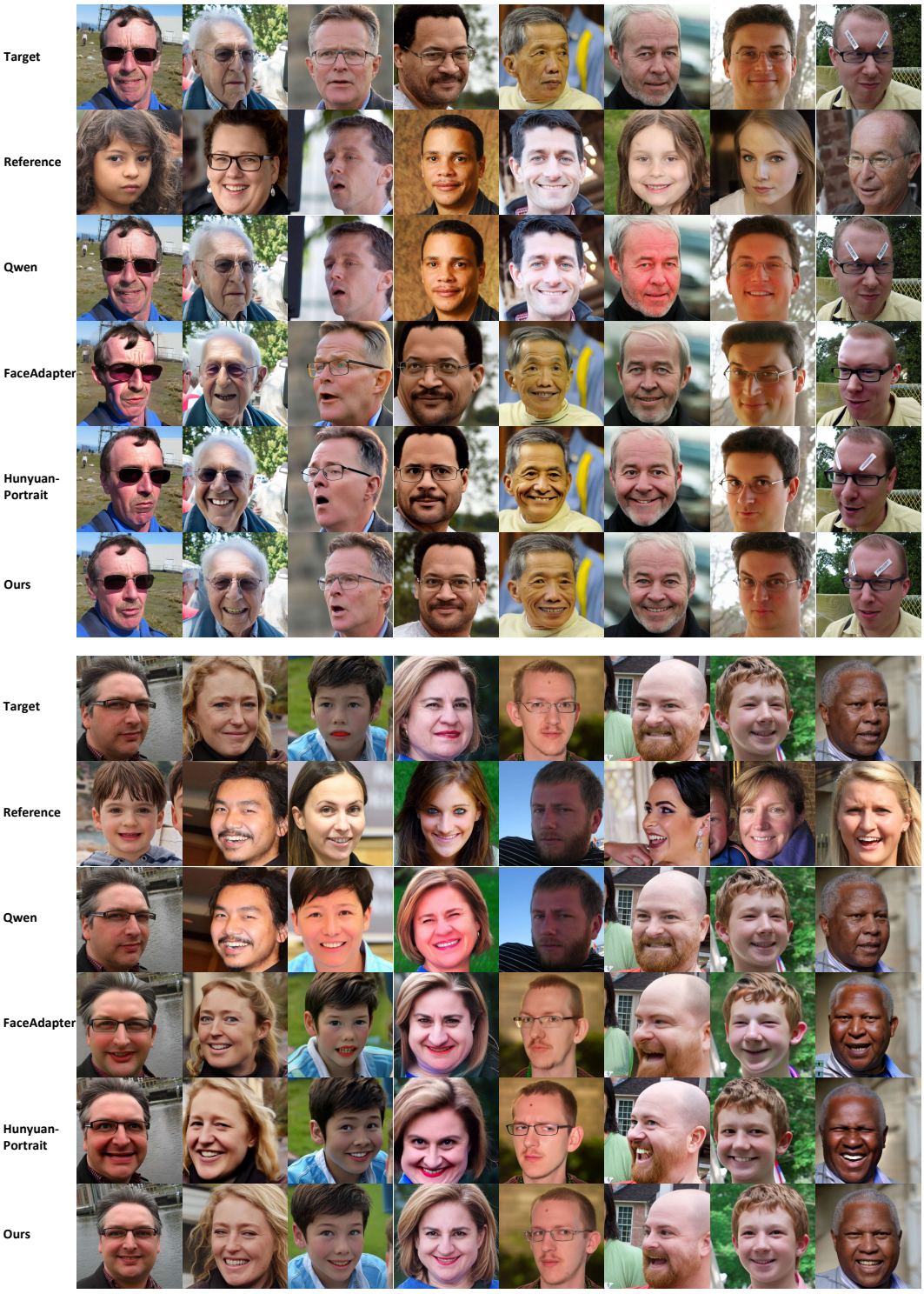


Figure 12: Motion transfer visual comparison. HunyuanPortrait tends to change the face skin, or lighting. The faceAdapter shows suboptimal face fidelity.



Figure 13: Head transfer visual comparison.

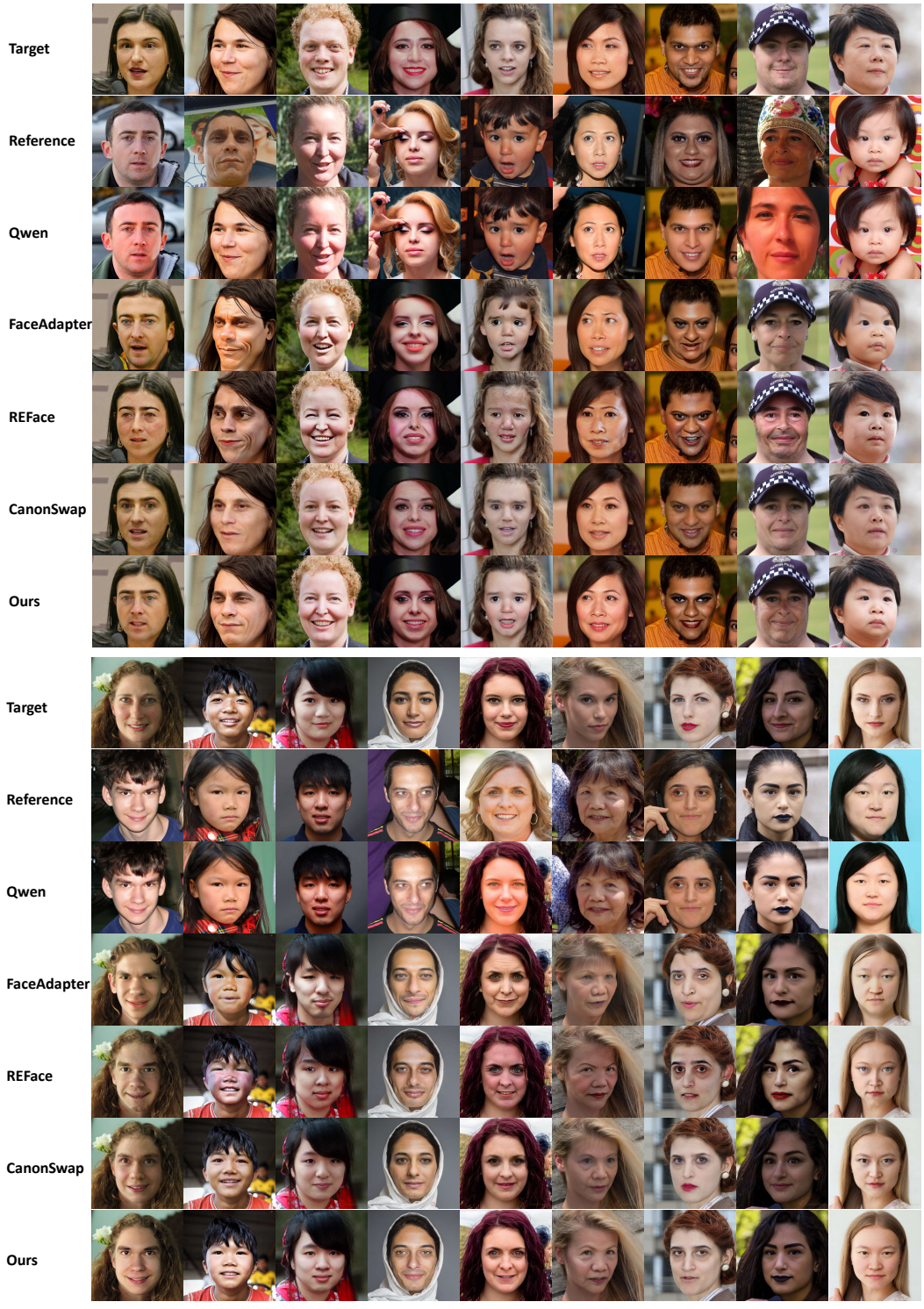


Figure 14: Face transfer visual comparison.

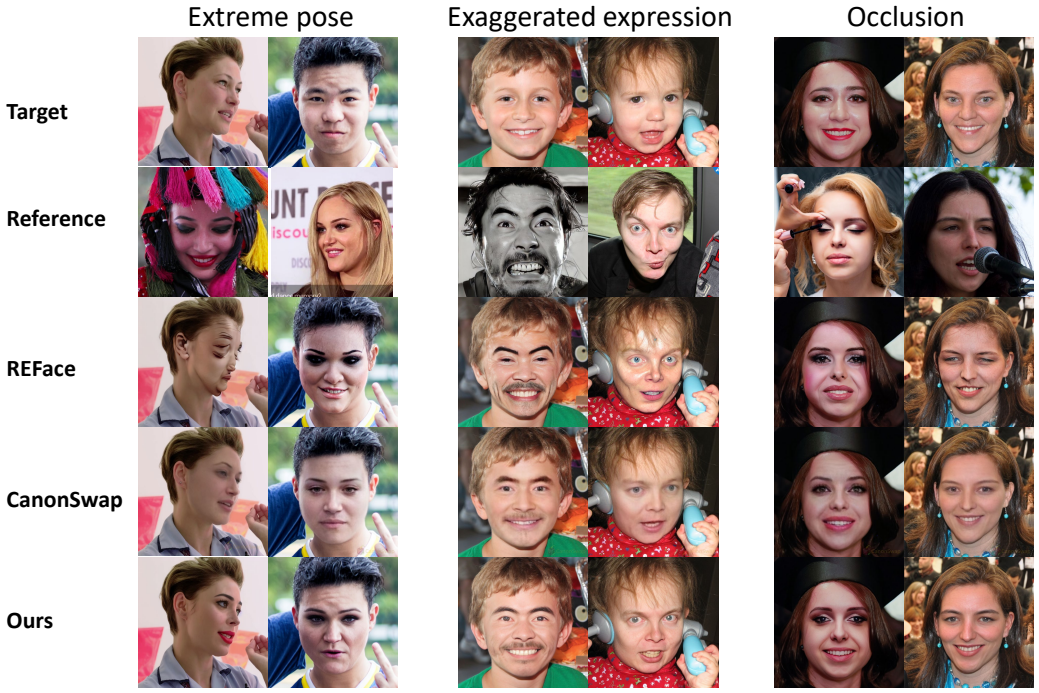


Figure 15: Visual comparison of **face transfer** under complex scenes.



Figure 16: Visual comparison of **hair transfer** under complex scenes.



Figure 17: Visual comparison of **motion transfer (face reenactment)** under complex scenes.



Figure 18: Visual comparison of **head transfer** under complex scenes.