Swap the Right Identity: Spatio-Temporal Diffusion Preference Optimization for Character Swapping

Hongdeng Shen[1,2], Jikang Cheng[1,2], Duo Li[1,3], Xiongzheng Li[1✉], Yunlong Feng[1†], and Jing Li[1]

[1] HUJING Digital Media & Entertainment Group, Beijing, China
[2] Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China
[3] Nanjing University, Nanjing, China

Abstract. Character swapping aims to faithfully transfer a target identity while preserving the source scene layout and pose, yielding visually seamless composites. However, existing general image editing methods often struggle to simultaneously achieve identity consistency, background preservation, and pose or structural stability. This challenge is further exacerbated by the lack of high-quality character swapping datasets. To address these challenges, we propose Spatio-Temporal Identity Preference Alignment (**SIPA**) framework, that systematizes diffusion-based preference optimization from both spatial and temporal perspectives. Spatially, we introduce Region-Weighted Preference (RWP), which increases the weighting of critical regions (e.g., the face or body) during DPO training, focusing preference signals on the desired local differences and thereby substantially reducing data dependence while improving training stability. Temporally, we propose Time-Injected Identity Preference (TIP): during low-noise steps, we explicitly inject face-difference cues into the DPO preference term to further enhance facial identity consistency. In addition, we build and plan to release a two-stage dataset tailored for character swapping. Extensive quantitative and qualitative experiments demonstrate that SIPA consistently outperforms prior methods across identity fidelity, structure preservation, and visual quality, achieving state-of-the-art performance.

Keywords: Image Editing · Character Swapping · Diffusion Preference Optimization

1 Introduction

Character swapping aims to accurately transfer a reference identity to a target image while strictly preserving the layout, geometric structure, and pose of the original scene [11, 30]. Crucially, this task is fundamentally distinct from

✉ Corresponding author: lxz@tju.edu.cn
† Project lead.

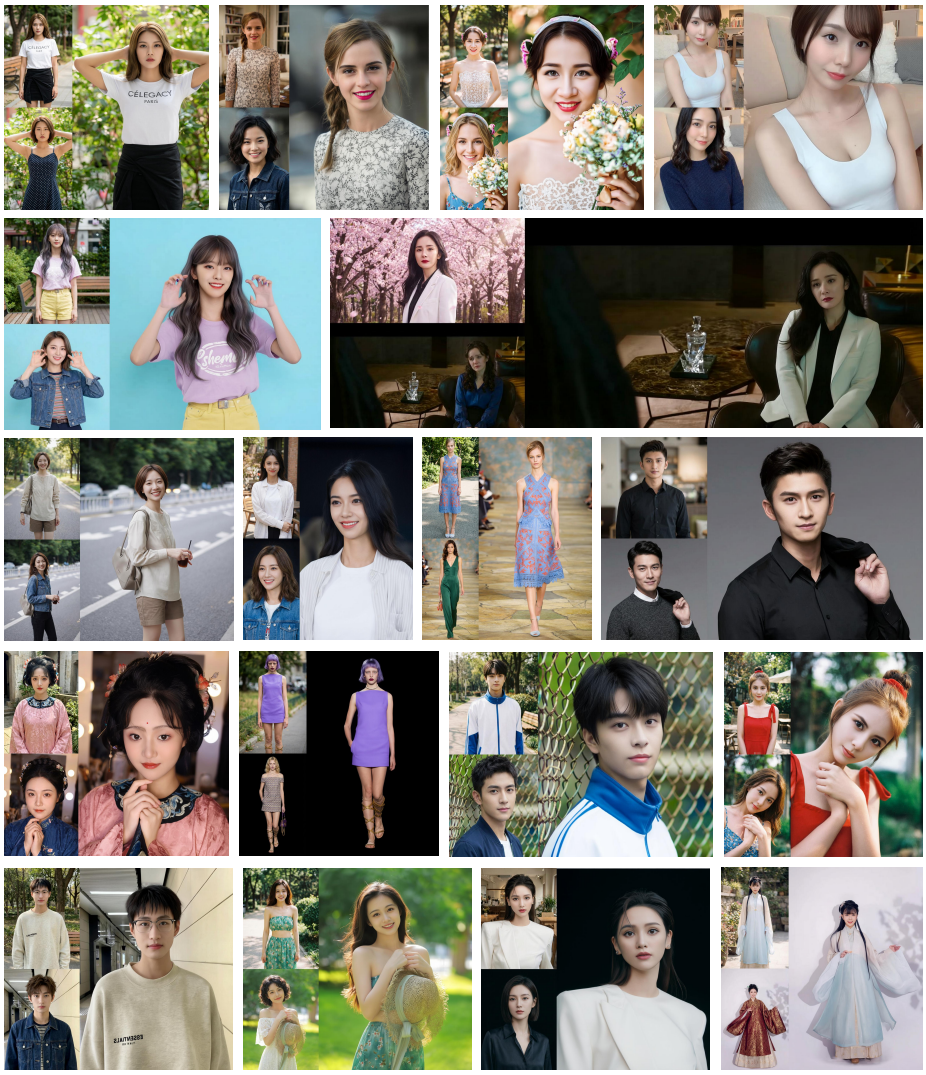

Fig. 1: Qualitative results of SIPA on diverse identity-swapping cases. For each example, the top-left image is the reference identity I_{ref}, the bottom-left image is the source scene I_{src} (providing background, illumination, composition, and pose), and the right image is the SIPA output $\hat{I}$. SIPA consistently transfers identity cues from I_{ref} while preserving the source background and geometric structure across varied scenes, viewpoints, lighting conditions, and appearances.

traditional face swapping [22, 32]. While face swapping predominantly focuses on localized facial region replacement, character swapping requires a holistic transformation that replaces the identity and clothing simultaneously, all while strictly preserving the original body pose and global scene context. Unlike un-

conditional generation, this task faces highly challenging constraints: achieving seamless identity fusion while strictly avoiding unintended modifications to background textures, lighting, and non-target structures (e.g., clothing contours, complex occlusions) [14, 31].

Despite significant progress in image editing driven by diffusion models [17, 20], existing methods still struggle with an inherent trade-off among identity consistency, background fidelity, and pose stability. State-of-the-art general image editing models, such as GPT-Image 1.5 [16], NanoBanana Pro [23], Qwen-Image-Edit [27], and Flux v2.0 [9], exhibit notable deficiencies in this highly constrained setting. These models often fail to maintain structural integrity, frequently yielding "copy-and-paste" artifacts where the reference identity is rigidly superimposed onto the target, severely disrupting the original pose. Furthermore, models pursuing extreme identity similarity often fall into over-editing, leading to uncontrolled background alterations and structural degradation. This contradiction becomes particularly acute in complex scenarios involving profile faces, severe occlusions, or drastic pose variations.

Our core observation is that the effective supervision signals for character swapping are inherently localized in space (concentrated on the face and foreground) and stage-sensitive in time. However, current diffusion fine-tuning paradigms based on Direct Preference Optimization (DPO) [19] typically rely on a global mean squared error to fit preference gradients. Such global objectives are easily dominated by large-area background differences, severely diluting crucial identity preference gradients and inducing spurious correlations. Consequently, existing preference learning heavily relies on ideal pairs (with only local differences) that are difficult to construct at scale, leading to noisy signals, slow convergence, and limited generalization.

To address these limitations, we propose the Spatio-Temporal Identity Preference Alignment (SIPA) framework. Spatially, we introduce Region-Weighted Preference (RWP), which effectively suppresses background interference by explicitly weighting critical regions during the DPO denoising error calculation. This mechanism enables the model to obtain high signal-to-noise ratio identity preference gradients even without ideal pairs. Temporally, we introduce Time-Gated Identity Preference (TIP). Given that diffusion models dictate global layout in early stages and high-frequency details in later stages, TIP injects identity supervision based on ArcFace [5] features exclusively during the low-noise stages. This continuously reinforces identity consistency without disrupting the early generation trajectory. Furthermore, we curate a two-stage dataset and an accompanying training paradigm covering controllable synthesis to real-world degradation, ensuring stable convergence for both supervised and preference learning.

The main contributions of this paper are summarized as follows:

- We propose the **SIPA** framework, which reconstructs diffusion preference optimization from spatial focusing and temporal gating dimensions, breaking the trade-off deadlock among identity alignment, background fidelity, and structural stability.

- We introduce the **RWP** mechanism, which utilizes explicit spatial mask weighting to guide preference gradients toward identity-critical regions, significantly reducing reliance on ideal preference pairs and improving training robustness.
- We present the **TIP** strategy, which injects perception-level identity difference penalties at specific low-noise stages of diffusion sampling, achieving extreme identity fidelity without sacrificing pose and layout.
- We construct a two-stage dataset and a corresponding training paradigm customized for character swapping, providing systematic data support for future research.

2 Related Work

2.1 Identity-conditioned Generation and Editing

Subject-driven personalization adapts generative models to a reference identity, enabling controllable synthesis and editing under new prompts or conditions. For diffusion models, prior work spans per-identity fine-tuning (e.g., DreamBooth [21]) and plug-and-play identity injection via image prompts/adapters (e.g., IP-Adapter [30], PhotoMaker [11], InstantID [25]). While effective for identity-consistent generation, these methods are not tailored to the stricter character swapping setting, where the desired changes are spatially localized (mostly on the face/foreground) but preservation constraints dominate the remaining pixels; consequently, globally aggregated training signals can be overwhelmed by non-identity regions, leading to over-editing or insufficient identity transfer.

2.2 Face Swapping and Reenactment

A closely related line is face swapping, which replaces only the facial identity while largely ignoring full-body appearance, pose/contour constraints, and background preservation. Most pipelines operate on detected/aligned facial regions and rely on synthesis-plus-blending to insert the swapped face back to the target image. Representative methods include FSGAN [15], which performs subject-agnostic face swapping and reenactment; FaceShifter [10], which improves high-fidelity swapping under occlusions; and SimSwap [3] along with its stronger in-the-wild variant SimSwap++ [4], which emphasize image-agnostic, high-fidelity identity transfer. While these methods achieve impressive face fidelity, their objective is fundamentally face-centric: they typically do not explicitly optimize for whole-body identity consistency, global pose/geometry stability, or strict background preservation, which are crucial in character swapping and storyboard-like generation.

2.3 Preference Optimization for Generative Models

Preference-based post-training is an effective way to align generative models with human judgments; Direct Preference Optimization (DPO) offers a simple

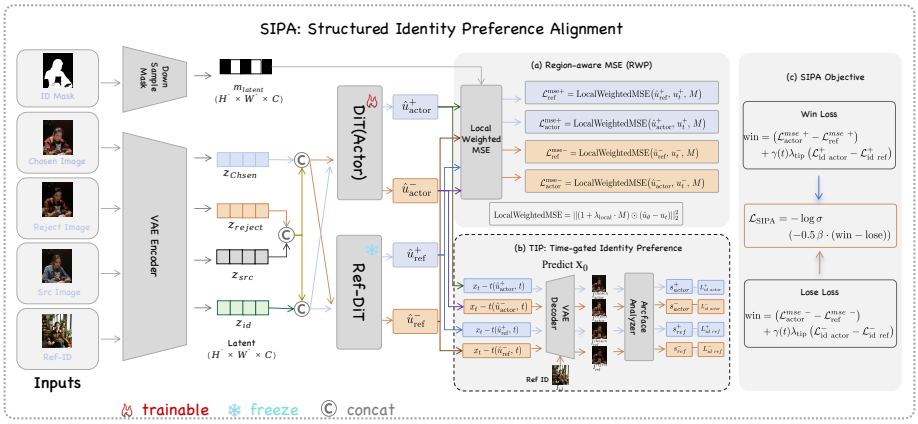

Fig. 2: The pipeline of SIAP. **(a) Region-Weighted Preference (RWP)** computes ROI-aware denoising errors using a latent-space mask and forms actor–reference differences for preference learning. **(b) Time-Gated Identity Preference (TIP)** reconstructs $\hat{\mathbf{x}}_0$ from x_t and applies ArcFace-based identity supervision only at low-noise steps. **(c) Loss integration** combines RWP and gated TIP into the SIAP win/lose terms and optimizes the final log-sigmoid DPO objective.

objective without explicit reward modeling [19]. Recent studies extend preference learning to diffusion models via RL-style tuning [1] and DPO-like objectives specialized for diffusion alignment [24,28]. However, for character swapping, globally aggregated denoising/flow errors can be dominated by background/geometry, weakening face-centric identity gradients. This motivates structured preference design that specifies where to focus (identity-critical regions) and when supervision is reliable (noise-stage aware).

3 Method

As illustrated in Fig. 2, we propose the Spatio-Temporal Identity Preference Alignment (SIPA) framework. To achieve natural and high-fidelity character swapping, SIPA decouples the learning process into two stages. The model establishes foundational feature mappings through Supervised Fine-Tuning (SFT) and subsequently incorporates spatial focusing and temporal gating mechanisms into the Flow Matching [12] framework during the Direct Preference Optimization (DPO) [19] stage. The remainder of this section briefly outlines the two-stage data curation strategy supporting this framework. We then derive the Region-Weighted Preference (RWP) and the Time-Gated Identity Preference (TIP) within the mathematical formulation of velocity field matching, culminating in the complete spatio-temporal unified optimization objective of SIPA.

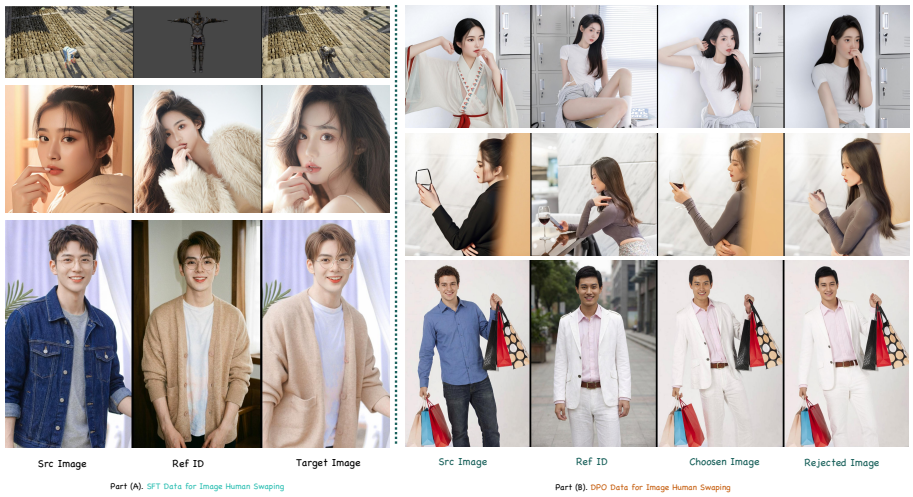

Fig. 3: Training data illustration. Part (A) shows the SFT training triplets, and Part (B) shows the DPO training quadruplets.

3.1 Data Curation

Reliable data underpins high-quality swapping mappings. During the SFT stage, the model learns the foundational alignment between the source scene and the target identity using the triplet $\langle I_{\text{Src\_Image}}, I_{\text{Ref\_ID}}, I_{\text{Target\_Image}} \rangle$. To achieve structural precision while maintaining distribution diversity, we curate training data from Unreal Engine controllable rendering, pseudo-supervision from closed-source models, and sketch-guided synthesis. We ensure the purity of the supervision signals by subjecting all samples to automated geometric consistency verification and strict manual screening. This systematic filtering removes noisy data that exhibit pose distortions or failed identity transfers, as illustrated in Fig. 3A.

During the DPO stage, we construct preference quadruplets $\langle I_{\text{Src\_Image}}, I_{\text{Ref\_ID}}, I_{\text{Chosen\_Image}}, I_{\text{Reject\_Image}} \rangle$ for contrastive learning. The chosen sample $I_{\text{Chosen\_Image}}$ corresponds to the strictly filtered high-fidelity target image from the SFT stage, whereas the rejected sample $I_{\text{Reject\_Image}}$ is generated by the SFT-trained baseline model under identical conditions. To improve the attributability of the preference signals, we apply a secondary filtering step to the quadruplets to discard suboptimal samples exhibiting corrupted background layouts or significant pose shifts. This strategy prevents preference noise caused by structural inconsistencies and ensures the optimization gradients converge precisely toward identity feature alignment.

3.2 Region-Weighted Preference, RWP

Based on the constructed quadruplets, we instantiate $I_{\text{Chosen\_Image}}$ and $I_{\text{Reject\_Image}}$ as the positive sample x_w and the negative sample x_l for preference learning.

Within the Flow Matching framework [12,13], given a data point $\mathbf{x}_0$ and a noise endpoint $\mathbf{x}_1$, the linear interpolation path constructed by optimal transport and its target velocity field are defined as:

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \qquad \mathbf{u}_t = \mathbf{x}_1 - \mathbf{x}_0, \quad t \in [0,1]. \tag{1}$$

The objective of the policy model $\mathbf{u}_\theta(\mathbf{x}_t, t)$ is to approximate this velocity field.

In character swapping tasks, core preference discrepancies naturally concentrate on foreground subjects such as the face and body. Relying directly on a global mean squared error for preference alignment [19] allows minor perturbations in background textures to overwhelm the crucial identity gradients. We propose Region-Weighted Preference to address this issue. This approach explicitly introduces a spatial binary mask $\mathbf{M} \in \{0,1\}^{H\times W}$ to delineate the region of interest, reformulating the standard velocity matching error into a spatially aware weighted error:

$$\text{LocalWeightedMSE}\left(\mathbf{u}_\theta, \mathbf{u}_t, \mathbf{M}\right) = \left\|(\mathbf{1} + \lambda_{\text{local}}\mathbf{M}) \odot (\mathbf{u}_\theta(\mathbf{x}_t, t) - \mathbf{u}_t)\right\|_2^2, \tag{2}$$

where $\odot$ denotes the Hadamard product, and $\lambda_{\text{local}} > 0$ modulates the penalty strength within the region of interest.

Following the implicit reward formulation, we define the local preference margin for a sample x as the difference in weighted errors between the policy model $\mathbf{u}_{actor}$ and the frozen reference model $\mathbf{u}_{\text{ref}}$:

$$\Delta\mathcal{L}_{\text{RWP}}(x) = \mathbb{E}_{t,\mathbf{x}_1}\left[\text{LocalWeightedMSE}_{actor}(x) - \text{LocalWeightedMSE}_{\text{ref}}(x)\right]. \tag{3}$$

This spatial weighting mechanism forces the optimization trajectory to prioritize the velocity field highly correlated with the subject. The mechanism significantly amplifies the signal-to-noise ratio of the identity gradients, alleviating the severe reliance on image pairs with perfect local differences.

3.3 Time-Gated Identity Preference, TIP

Relying solely on velocity field matching is insufficient to guarantee perceptual identity fidelity. However, in Flow Matching, identity discriminative signals such as ArcFace [5] features only become reliable when the sample is sufficiently denoised and approaches the data endpoint ($t \to 0$). Enforcing perceptual losses during high-noise stages induces erroneous attribution because the facial structure is not yet formed. To address this issue, we propose Time-Gated Identity Preference TIP.

To implement this mechanism, we utilize a one-step denoising reconstruction [29]. Based on the velocity field prediction $\mathbf{u}_\theta(\mathbf{x}_t, t) \approx \mathbf{x}_1 - \mathbf{x}_0$ in Flow Matching [12], we extrapolate along the generation path to conduct a lightweight one-step data endpoint estimation. We then obtain the predicted image through the VAE decoder D_{VAE}:

$$\hat{\mathbf{x}}_0 = \mathbf{x}_t - t\,\mathbf{u}_\theta(\mathbf{x}_t, t), \qquad \hat{\mathbf{I}} = D_{\text{VAE}}(\hat{\mathbf{x}}_0). \tag{4}$$

We then construct a gated relative identity advantage. Let $\mathbf{I}_{id}$ denote the reference identity image. We compute the ArcFace similarity to define the identity loss $\mathcal{L}_{id}(\hat{\mathbf{I}}, \mathbf{I}_{id}) = -s(\hat{\mathbf{I}}, \mathbf{I}_{id})$. To maintain the relative alignment nature of DPO and prevent model over-editing, we define the identity advantage difference between the policy model and the reference model as:

$$\Delta^{id}(x) = \mathcal{L}_{\text{id,actor}}(x) - \mathcal{L}_{\text{id,ref}}(x). \tag{5}$$

We simultaneously introduce a time gating function $\gamma(t) = \mathbb{I}(t \leq \tau)$. This function ensures that feature-level identity supervision is activated exclusively within a low-noise safe window, providing stable and high-confidence supervision without disrupting the early layout generation.

3.4 Unified Spatio-Temporal Optimization Objective

Finally, we integrate the spatial RWP margin and the temporal TIP advantage into the preference model. Using the constructed quadruplet dataset, we formulate the joint relative reward terms for the positive sample x_w and the negative sample x_l:

$$\text{win} = \Delta\mathcal{L}_{\text{RWP}}(x_w) + \gamma(t)\lambda_{\text{tip}}\Delta^{id}(x_w), \tag{6}$$

$$\text{lose} = \Delta\mathcal{L}_{\text{RWP}}(x_l) + \gamma(t)\lambda_{\text{tip}}\Delta^{id}(x_l), \tag{7}$$

where λ_{tip} balances the weight of the identity preference. The complete optimization objective of SIPA is formulated as follows:

$$\mathcal{L}_{\text{SIPA}} = -\mathbb{E}_{(x_w, x_l)}\left[\log\sigma\left(-\frac{\beta}{2}\left(\text{win} - \text{lose}\right)\right)\right]. \tag{8}$$

4 Experiments

4.1 Implement Details

Basic Information. We divide the training process into two stages. The first stage employs Unreal Engine data for Supervised Fine-Tuning to equip the model with fundamental character swapping capabilities. The second stage conducts Direct Preference Optimization [19], enhanced by the proposed RWP and TIP algorithms, on a custom-collected quadruplet dataset. Both stages utilize Low-Rank Adaptation [7] for parameter-efficient fine-tuning based on the Qwen-Image-Edit 2509 [27] model and utilizing 16 NVIDIA H20 GPUs. During Stage I, we perform Supervised Fine-Tuning using 12,000 Unreal Engine rendered triplets for 3 epochs. We employ the AdamW optimizer with a learning rate of 1×10^{-4}, a batch size of 32, a maximum image resolution of 1024×1024, and a LoRA rank of 32. Building upon the model fine-tuned in Stage I, Stage II conducts Direct Preference Optimization using 4,451 real-world preference quadruplets over 1,950 steps. For this stage, we apply a learning rate of 1×10^{-5} and a batch size of

16 under bfloat16 mixed precision, while maintaining the identical resolution and LoRA rank. Finally, we configure the core preference hyperparameters as follows: $\beta = 10$, $\alpha_{\text{sft}} = 10$, $\lambda_{\text{local}} = 2$, $\lambda_{\text{tip}} = 3$, and the time gating threshold $\tau = 0.1$.

Evaluation Metrics. We evaluate the character swapping performance across multiple dimensions. To assess identity transfer, we compute the cosine similarity in the ArcFace [5] feature space between the generated and reference images to measure identity preservation (ID-Sim). We complement this measurement with foreground-specific metrics, namely FG-CLIP [18] and FG-SSIM [26], to verify the semantic and structural accuracy of foreground feature migration. For background and pose fidelity, we utilize BG-CLIP [18] and BG-SSIM [26] to quantify the semantic and structural retention of background regions, while calculating the human keypoint alignment error (PoseErr) based on a standard 2D pose estimator [2] to evaluate pose consistency. Finally, we report SSIM [26], PSNR [8] and FID [6] to assess the overall visual quality and generation fidelity.

Benchmark. To evaluate identity swapping performance, we construct a diverse validation set of 321 high-quality samples spanning various age groups, ethnicities, and visual styles. These images are curated from movies, television series, and animations. We employ a rigorous filtering pipeline, combining automated facial detection and manual verification, to ensure data quality. The final dataset strictly consists of single-subject images with a minimum resolution of 1080p.

Baseline. Given the current absence of dedicated image identity swapping models, we benchmark the proposed framework against state-of-the-art general image editing models. Specifically, we evaluate two open-source models, Qwen-Image-Edit [27] and Flux v2.0 [9], alongside two closed-source commercial models, GPT-Image 1.5 [16] and NanoBanana Pro [23]. We select these baselines because their robust generation stability and high accessibility make them the most suitable proxies for the identity swapping objective. We deliberately exclude purely text-instruction-based editing methods to prevent unfair comparisons under distinct task configurations. Detailed implementation configurations for all evaluated baselines are provided in the appendix.

4.2 Quantitative Comparison

Table 1: Quantitative results of SIPA compared with baselines, including GPT-Image1.5, NanoBanana Pro, Flux v2.0, and Qwen-Image-Edit.

Method	SSIM↑	PSNR↑	FID↓	ID-Sim↑	FG-CLIP↑	FG-SSIM↑	BG-CLIP↑	BG-SSIM↑	PoseErr↓
Flux v2	0.7771	13.3266	89.4054	0.5192	0.7507	0.8091	0.6908	0.7450	416.2634
Nanobanana Pro	<u>0.8731</u>	<u>17.9291</u>	78.6922	**0.7044**	0.8618	<u>0.8571</u>	0.7109	<u>0.8971</u>	181.9931
GPT-Image	0.8652	17.0812	<u>65.3418</u>	<u>0.6133</u>	<u>0.8649</u>	0.8447	<u>0.7306</u>	0.8857	<u>163.0843</u>
Qwen-Image Edit 2509	0.7906	13.8196	86.7862	–	0.7975	0.8160	0.6817	0.7651	220.1150
Ours (SIPA)	**0.8957**	**19.5424**	**51.8000**	0.5975	**0.8788**	**0.8636**	**0.7328**	**0.9277**	**100.0920**

As shown in Table 1, the proposed SIPA framework demonstrates strong performance in overall image quality, structural consistency, and fidelity. Specifically, SIPA achieves the highest scores in foreground and background semantic and structural preservation. It outperforms both open-source baselines, including Flux v2.0 [9] and Qwen-Image-Edit [27] , and closed-source models GPT-Image 1.5 and Nanobanana Pro, across metrics such as FG-CLIP, BG-CLIP, FG-SSIM, and BG-SSIM. Furthermore, SIPA yields the highest global SSIM (0.8957)and PSNR(19.5424) alongside the lowest FID (51.8000). These results confirm the capability of the method to maintain global structural stability during identity transfer, mitigating the structural degradation and detail distortion commonly observed in models like Qwen-Image-Edit.

We particularly note the trade-off between identity similarity (ID-Sim) and pose alignment (PoseErr). Although Qwen-Image-Edit 2509 achieves a higher ID-Sim (0.7143), after excluding the copy-paste samples, its ID-Sim is 0.4573. And its PoseErr reaches 220.11. This discrepancy indicates severe copy-and-paste artifacts from the reference identity, which significantly disrupt the target pose. In contrast, our TIP strategy introduces identity supervision exclusively during the stable denoising stages. This design allows SIPA to maintain a competitive ID-Sim (0.5975) while significantly reducing the PoseErr to 100.09, representing an error reduction of over 50% compared to the Qwen model. Additionally, the improved FID score compared to GPT-Image 1.5 demonstrates the effectiveness of the synergistic RWP and TIP modules in suppressing generative artifacts. Overall, while SIPA exhibits a marginal gap in absolute identity metrics compared to the evaluated closed-source models, it achieves a more robust balance regarding pose preservation, layout consistency, and background stability.

4.3 Qualitative Results

Figure 4 illustrates the qualitative comparison between the proposed method and several baseline models on the character swapping task. The results indicate that our approach achieves a favorable balance between identity consistency and pose preservation. Under complex scenes and variable lighting conditions, the generated images effectively retain the spatial layout and anatomical structure of the source images. In contrast, although GPT-Image 1.5 generates natural human appearances, it exerts limited constraints on the original pose, which frequently leads to positional shifts. NanoBanana Pro tends to perform global repainting, causing background alterations and distorted human proportions that hinder controlled identity swapping. Furthermore, the sixth and seventh columns demonstrate that Flux v2 and Qwen-Image-Edit exhibit instability in complex environments, often yielding splicing artifacts and identity confusion. Qwen-Image-Edit specifically suffers from conspicuous copy-and-paste artifacts. Overall, the proposed framework attains a more robust trade-off among visual naturalness, structural consistency, and identity matching.To evaluate robustness, we add challenging cases covering viewpoint, illumination, occlusion, expression, and body-structure variations in Fig. 5. And more qualitative results are shown in Fig. 1.

Fig. 4: Qualitative comparison with GPT-Image1.5, NanoBananaPro, fluxv2, Qwen-Image-Edit.

Fig. 5: Additional Qualitative Results on Hard Cases

4.4 Ablation Study

Table 2: Ablation of SIPA framework on our evaluation benchmark.

Configuration	FG-CLIP↑	BG-CLIP↑	FG-SSIM↑	BG-SSIM↑	FG-PSNR↑	BG-PSNR↑	PSNR↑	PoseErr↓	ID-Sim↑	FID↓
Stage I. Necessity of DPO Training										
SFT Base Model	0.8439	0.7156	0.8478	0.8274	14.59	19.86	16.39	144.47	0.4244	66.88
Continual SFT	0.8544	0.7184	0.8547	0.8621	15.27	22.06	17.85	120.59	0.4356	59.02
Diffusion DPO	**0.8689**	**0.7470**	**0.8638**	**0.9259**	**16.01**	**26.17**	**19.42**	**107.59**	**0.4959**	**51.49**
Stage II. Effect of Region-Weighted Preference (RWP)										
w/o RWP	0.8689	**0.7470**	**0.8638**	0.9259	16.01	26.17	19.42	107.59	0.4959	51.49
w/ RWP	**0.8792**	0.7416	0.8628	**0.9199**	**16.14**	**25.78**	**19.52**	**99.10**	**0.5168**	**50.65**
Stage III. Effect of Time-Gated Identity Preference (TIP)										
w/o TIP	**0.8792**	**0.7416**	0.8628	0.9199	**16.14**	25.78	19.52	**99.10**	0.5168	**50.65**
Full SIPA	0.8788	0.7328	**0.8636**	**0.9277**	16.10	**26.17**	**19.54**	100.09	**0.5975**	51.80

Effectiveness of Preference Optimization. To isolate the performance gains of the proposed SIPA framework from the benefits of mere data augmentation, we investigate the specific impact of the optimization paradigm. We establish a Continual SFT baseline by further fine-tuning the initial SFT base model using exclusively the chosen samples from our preference dataset as ground-truth targets. To evaluate the fundamental advantage of preference learning, we also implement a standard Diffusion DPO [19] baseline utilizing the complete preference pairs. As indicated by Stage I in Table 2 and the qualitative comparisons in Figure 7, the standard DPO formulation significantly outperforms the Continual SFT approach. This performance gap confirms that the improvements stem primarily from the preference optimization mechanism rather than the mere exposure to additional high-quality data, thereby validating the rationale of our preference-based training strategy.

Effectiveness of Region-Weighted Preference RWP. The evaluation of RWP demonstrates its critical role in enhancing identity consistency and pose stability by explicitly localizing preference signals. As indicated in Stage II of Table 2, integrating the RWP mechanism yields substantial improvements across all consistency metrics. The visual comparisons in Figure 7(A) further corroborate this trend, confirming the capacity of the mechanism to decouple background variations and preserve original poses.

Impact of Time-Gated Identity Preference TIP. Regarding identity fidelity, TIP effectively overcomes the preservation bottleneck during the generation process. Quantitative analysis reveals an increase in the ID-Sim metric from 0.5168 to 0.5975. Figure 7(B) confirms the robustness of the TIP mechanism in strengthening facial feature consistency, thereby validating the value of the gated injection strategy in avoiding generative artifacts.

4.5 Limitations and Failure Modes

While the proposed SIPA framework demonstrates robust performance in identity swapping, it encounters certain limitations in extreme scenarios. First, when

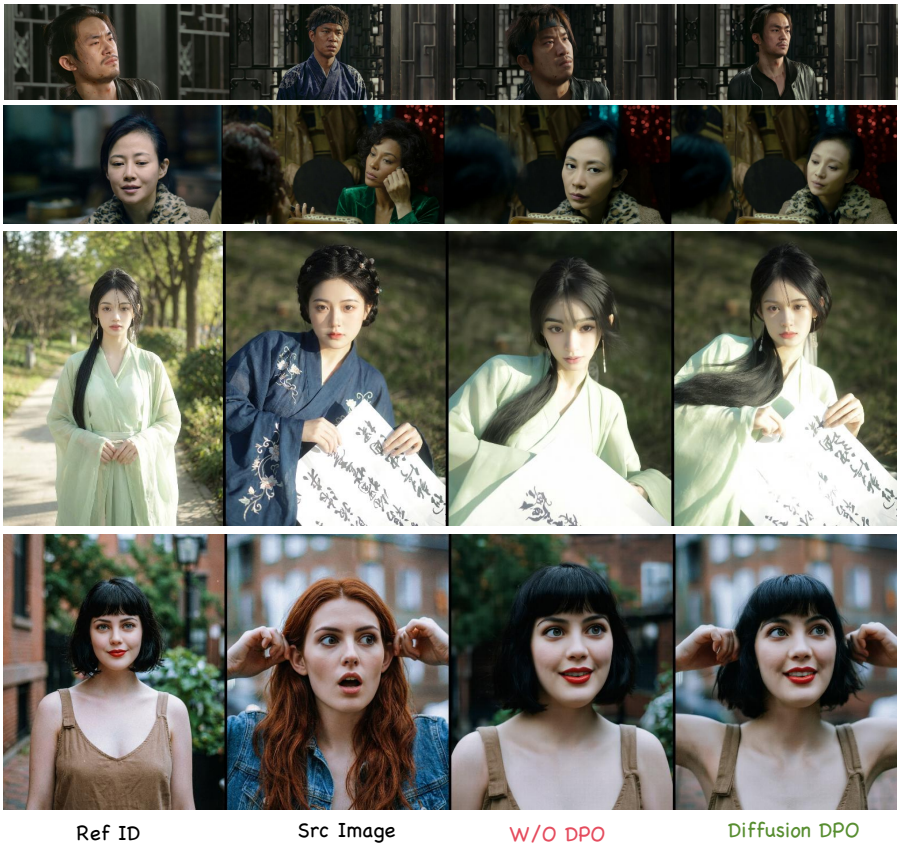

Fig. 6: Qualitative ablation results of DPO

the source image contains severe facial occlusions or presents an extreme pose discrepancy relative to the reference identity, the model occasionally struggles to accurately infer the unseen facial geometries, leading to structural artifacts or degraded identity resemblance. Second, under highly complex or dramatic illumination conditions, balancing the target lighting atmosphere with the intrinhasic skin tone of the reference identity remains challenging, sometimes resulting in unnatural color blending. Finally, the reliance on high-quality preference quadruplets imposes a stringent requirement on the data curation pipeline. Future work will focus on integrating explicit 3D facial priors to enhance pose extrapolation and exploring illumination-aware conditioning mechanisms to further improve blending naturalness.

5 Conclusion and Acknowledgments

Conclusion. In this paper, we propose the Spatio-Temporal Identity Preference Alignment framework to address the inherent trade-offs among identity consis-

Fig. 7: Qualitative ablation results of SIPA. (A) Visual comparisons of Region-Weighted Preference (RWP). (B) Visual comparisons of Time-Gated Identity Preference (TiP).

tency, background preservation, and structural stability in character swapping. By restructuring diffusion preference optimization [19], we introduce the Region-Weighted Preference mechanism. This mechanism explicitly localizes optimization gradients onto identity-critical regions, reducing the reliance on perfectly aligned image pairs. Concurrently, the Time-Gated Identity Preference strategy injects perceptual identity supervision exclusively during the low-noise stages, ensuring reliable facial fidelity without disrupting the overall layout generation. To support this framework, we construct a customized two-stage dataset and an accompanying training paradigm. Extensive experiments confirm that the proposed method effectively enhances background and pose preservation, achieving state-of-the-art performance across comprehensive metrics compared to existing image editing baselines.

Acknowledgments. This work was supported by HUJING Digital Media Entertainment Group through HUJING Digital Media Entertainment Research Intern Program.

References

1. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301 (2023)
2. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
3. Chen, R., Chen, X., Ni, B., Ge, Y.: Simswap: An efficient framework for high fidelity face swapping. In: Proceedings of the 28th ACM international conference on multimedia. pp. 2003–2011 (2020)
4. Chen, X., Ni, B., Liu, Y., Liu, N., Zeng, Z., Wang, H.: Simswap++: Towards faster and high-quality identity swapping. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(1), 576–592 (2023)
5. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4690–4699 (2019)
6. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems (NeurIPS) (2017), arXiv:1706.08500
7. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), `https://openreview.net/forum?id=nZeVKeeFYf9`
8. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of psnr in image/video quality assessment. In: Electronics Letters (2008), dOI: 10.1049/el:20080522
9. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2` (2025)
10. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Faceshifter: Towards high fidelity and occlusion aware face swapping. arXiv preprint arXiv:1912.13457 (2019)
11. Li, Z., Cao, M., Wang, X., Qi, Z., Cheng, M.M., Shan, Y.: Photomaker: Customizing realistic human photos via stacked id embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8640–8650 (2024)
12. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: arXiv preprint (2022), arXiv:2210.02747
13. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (2023)
14. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: 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. vol. 38, pp. 4296–4304 (2024)
15. Nirkin, Y., Keller, Y., Hassner, T.: FSGAN: Subject agnostic face swapping and reenactment. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7184–7193 (2019)

16. OpenAI: GPT Image 1.5 Model. https://platform.openai.com/docs/models/gpt-image-1.5 (2025), accessed: 2026-01-28. Snapshot listed: gpt-image-1.5-2025-12-16
17. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4199–4209 (2023)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML) (2021), arXiv:2103.00020
19. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36 (2023)
20. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022)
21. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22500–22510 (2023)
22. Shao, H., Wang, S., Zhou, Y., Song, G., He, D., Qin, S., Zong, Z., Ma, B., Liu, Y., Li, H.: Vividface: A diffusion-based hybrid framework for high-fidelity video face swapping. arXiv preprint arXiv:2412.11279 (2024)
23. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
24. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8228–8238 (2024)
25. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A.: Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (2024)
26. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. In: IEEE Transactions on Image Processing (2004), dOI: 10.1109/TIP.2003.819861
27. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), https://arxiv.org/abs/2508.02324
28. Yang, K., Tao, J., Lyu, J., Ge, C., Chen, J., Shen, W., Zhu, X., Li, X.: Using human feedback to fine-tune diffusion models without any reward model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8941–8951 (2024)
29. Yang, X., Cheng, C., Yang, X., Liu, F., Lin, G.: Text-to-image rectified flow as plug-and-play priors. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=SzPZK856iI

30. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
31. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3836–3847 (2023)
32. Zhao, W., Rao, Y., Shi, W., Liu, Z., Zhou, J., Lu, J.: Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. CVPR (2023)