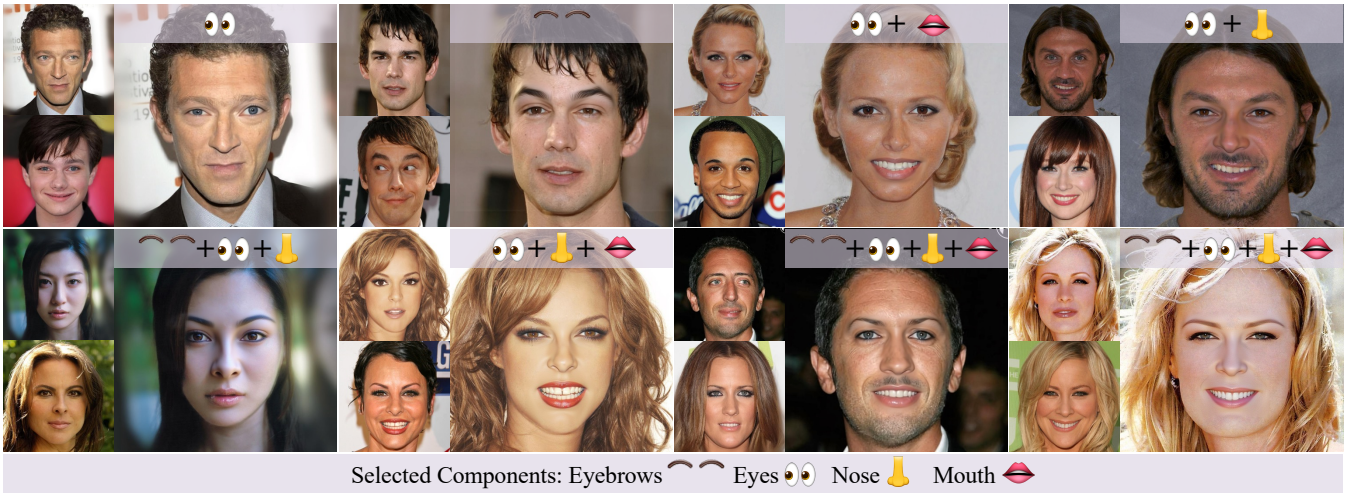
FaceComposer: Learning Coherent and Realistic Face Composition from References

Ling Li[1], Lanqing Guo[2*], Siyuan Yang[3], Yufei Wang[4], Qian Zheng[5], Alex C. Kot[1,6,7], Weisi Lin[1], Lihui Chen[1]
[1]Nanyang Technological University, Singapore [2]University of Texas at Austin, USA
[3]KTH Royal Institute of Technology, Sweden [4]SparcAI, Inc. [5]Zhejiang University, China
[6]Shenzhen MSU-BIT University, China [7]VinUniversity, Vietnam
{ling012, eackot, wslin, elhchen}@ntu.edu.sg, lanqing.guo@austin.utexas.edu,
siyuany@kth.se, yufei.wang@sparclab.ai, qianzheng@zju.edu.cn
*Corresponding author



Qualitative results of FaceComposer on CelebAMask-HQ [1]. In each example, the source (top left) and reference (bottom left) images are provided, and our method (right) transfers the specified components (eyebrows, eyes, nose, or mouth) while preserving fidelity, structural coherence, and photorealism. *Zoom in for details.*

***Abstract*—Reference-based face composition enables precise editing of individual facial components in a source portrait using a reference image, with applications in forensic sketching, facial surgery simulation, and digital media. Despite recent advances in diffusion models, achieving realistic and structurally consistent component-level editing remains challenging. To fill this gap, we propose Context-Aware Face Fusion, a generative framework that seamlessly integrates selected facial components into a source portrait while preserving both structural coherence and texture harmony. To handle mismatches in pose, scale, and cross-ethnic appearance variations, we introduce a Component-Aware Augmentation strategy that adaptively aligns reference features to the source. In addition, we design a Multi-Stage Training pipeline that mitigates the lack of paired training data by progressively transitioning from reconstruction to composition tasks. Extensive experiments demonstrate that our method produces high-fidelity composites that faithfully incorporate reference components while maintaining realism and global coherence.**

***Index Terms*—Reference-based Composition, Human Face Editing, Diffusion Models, Context-aware Generation**

I. Introduction

Recent advances in text-to-image diffusion models [2]–[4] have enabled high-fidelity image generation, forming the basis for tasks such as image editing [5]–[12], style transfer [13], [14], text-guided manipulation [15], [16], and reference-based image transformation [17], [18]. In the facial domain, identity-preserving generation [19], [20] and text-guided editing [21] mostly target global or attribute-level changes. Fine-grained component editing (e.g., nose or mouth) remains underexplored, despite its importance in applications such as forensic sketching and presurgical simulation.

Face composition aims to synthesize a composite portrait by integrating target facial components from a reference image into a source portrait. The objective is twofold: ensuring high-fidelity identity transfer and maintaining the source's global consistency in pose, illumination, and context. Unlike general image editing, this task is uniquely constrained by the rigid structural regularity of faces and the acute human sensitivity to facial distortions, which necessitate both geometric precision and seamless textural blending at the boundaries. Recent reference-based editing methods, including Gemini3 [22], FLUX.1-Kontext [3], ACE++ [18], and FreeEdit [23], support instruction- and reference-guided editing, providing flexible multimodal control within a unified diffusion backbone. MimicBrush [17] and AnyDoor [24] trans-

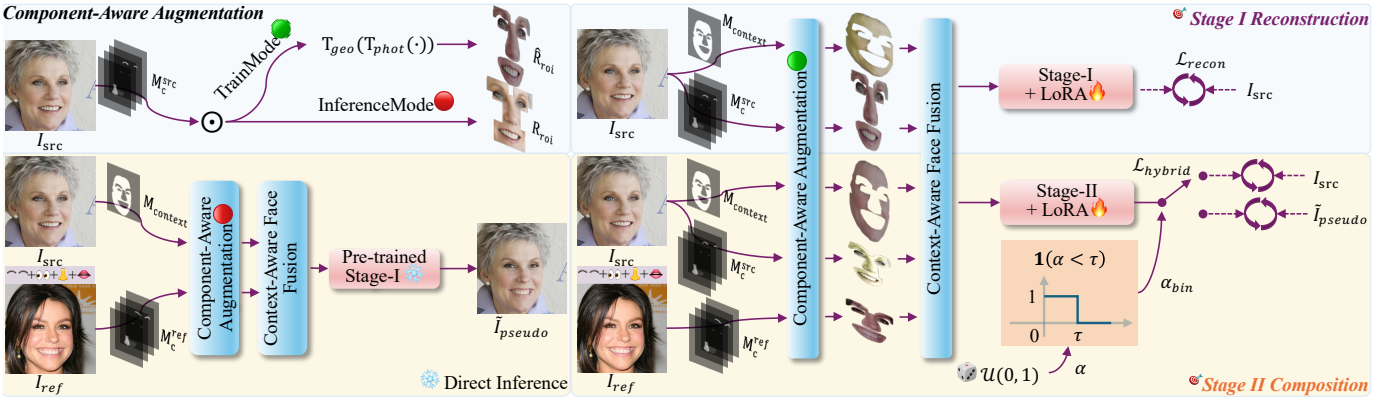


Fig. 1. **Pipeline overview**. Given a source I_{src} and reference I_{ref}, Component-Aware Augmentation (top left) generates context and component regions. Training has two stages: Stage I reconstructs the source, and its trained model is then frozen to produce pseudo composites $\tilde{I}_{\text{gen}}$ (bottom left) for pseudo supervision in Stage II, which integrates reference components with structural and textural consistency.

fer features from reference images into a source image using masks to specify editable regions. However, these approaches primarily target object-level inpainting and lack mechanisms to handle structure-sensitive tasks such as facial component editing. Multi-conditioned composition methods, such as UniCombine [25], DreamFuse [26], and OminiControl [6], enable compositional edits guided by references, while approaches like MADD [27], InsertAnything [28], and FLUX.1-Fill [3] leverage masks for object insertion. These methods perform well for natural images, where minor inconsistencies are tolerable, but often struggle with faces. Existing methods often suffer from boundary discontinuities, copy-paste effects and unfaithful transfers (Fig. 2), necessitating a specialized approach for reference-based face composition.

Addressing the challenges of faithful transfer and structural coherence, our Context-Aware Face Fusion framework explicitly models the interaction between reference components and the source context. This design effectively mitigates copy-paste effects and boundary seams, resulting in perceptually continuous and authentic face composition. To handle cross-identity variations in pose, scale, and texture, we introduce a Component-Aware Augmentation strategy that simulates realistic facial variations, allowing the model to learn robust feature integration across diverse appearances. Furthermore, in the absence of paired training data, we devise a Multi-Stage Training strategy that begins with self-reconstruction from masked inputs and gradually progresses to full face composition with hybrid pseudo-supervision. This progressive scheme stabilizes training, mitigates error accumulation, and ultimately enables the generation of high-fidelity, semantically faithful facial composites.

Our contributions are summarized as follows: (1) a fine-grained face composition framework for faithful attribute transfer with global structural consistency; (2) a Context-Aware Face Fusion framework that enables mutual adaptation and suppresses boundary artifacts and copy–paste effects; and (3) a Multi-Stage Training strategy with Component-Aware Augmentation enabling cross-identity synthesis without paired data. Extensive evaluations demonstrate superior identity preservation and photorealism.

II. Methodology

A. Overview

Given a source image $\mathbf{I}_{\text{src}} \in \mathbb{R}^{H\times W\times 3}$ providing the global portrait context, a reference image $\mathbf{I}_{\text{ref}} \in \mathbb{R}^{H\times W\times 3}$ containing target facial attributes, and a binary mask $\mathbf{M} \in \{0,1\}^{H\times W}$ specifying the facial components to be transferred, our goal is to learn a conditional generative model that synthesizes a composite image. Specifically, the synthesized regions defined by $\mathbf{M}$ preserve the identity-related characteristics of $\mathbf{I}_{\text{ref}}$, while all non-edited attributes, including pose, illumination, and facial expression, remain consistent with $\mathbf{I}_{\text{src}}$. Additionally, the model aims to ensure smooth and natural transitions between edited and non-edited regions, maintaining high-quality blending at the ROI–context boundaries.

To address these challenges, we introduce a framework composed of three complementary components operating at different levels of the pipeline (Fig. 1): (i) a *Component-Aware Augmentation Strategy* at the data level, which introduces controlled geometric and photometric variations to improve robustness under cross-ethnic compositions in the absence of paired supervision; (ii) a *Context-Aware Face Fusion Framework* at the model level, which employs cross-attention to integrate transferred components with surrounding facial context; and (iii) a *Multi-Stage Training Strategy* at the optimization level, which progressively transitions from self-reconstruction to reference-guided composition to balance component fidelity and global coherence.

B. Data Construction

1) Pairwise Data: We adopt CelebAMask-HQ [1] and FFHQ [29] due to their high-resolution facial images, rich attribute diversity, and broad ethnic coverage. To construct clean composition pairs, we remove samples with extreme poses or heavy occlusions, pair subjects across ethnicities, and generate 10,000 source–reference pairs for fine-tuning. For multi-reference settings, all references are filtered with the same criteria and are collectively referred to as "reference".

2) Component-Aware Augmentation Strategy: A fundamental challenge in reference-based face composition is seamlessly integrating transferred facial components under large

Dataset	Method	ROI LPIPS↓	ROI SSIM↑	ROI CLIP↑	Context LPIPS↓	Context SSIM↑	Context CLIP↑	Boundary SobelDiff↓	Boundary LapDiff↓	Global FID↓
CelebAMask-HQ	FLUX.1-Fill-dev [3]	0.0357	0.9648	0.9527	0.0798	0.9359	0.9923	0.0361	0.0193	48.35
	FLUX.1-Kontext-dev [3]	0.0392	0.9639	0.9375	0.3333	0.6691	0.9250	0.0662	0.0278	41.37
	InsertAnything [28]	0.0327	0.9679	0.9597	**0.0366**	**0.9773**	**0.9974**	0.0024	0.0021	46.52
	InsertAnythingFT [28]	0.0297	0.9701	**0.9660**	0.0402	0.9762	0.9972	0.0024	0.0022	45.71
	MADD [27]	0.0361	0.9652	0.9393	0.2635	0.7510	0.9812	0.0543	0.0212	47.2
	DreamFuse [26]	0.0396	0.9633	0.9467	0.3601	0.6705	0.8914	0.0684	0.0282	73.70
	Ours	**0.0207**	**0.9815**	0.9611	0.1142	0.8997	0.9799	**0.0022**	**0.0010**	**27.88**
FFHQ	FLUX.1-Fill-dev [3]	0.0273	0.9761	0.9475	0.0787	0.9417	0.9957	0.0338	0.0197	31.15
	FLUX.1-Kontext-dev [3]	0.0268	0.9774	0.9418	0.3544	0.6408	0.9649	0.0730	0.0249	32.20
	InsertAnything [28]	0.0248	0.9784	0.9555	**0.0446**	**0.9755**	**0.9988**	**0.0014**	**0.0011**	31.12
	InsertAnythingFT [28]	0.0228	0.9793	**0.9608**	0.0458	0.9748	**0.9988**	**0.0014**	**0.0011**	30.68
	MADD [27]	0.0277	0.9766	0.9350	0.2759	0.7983	0.9842	0.0577	0.0215	39.12
	DreamFuse [26]	0.0283	0.9754	0.9442	0.3159	0.7455	0.9451	0.0643	0.0273	49.18
	Ours	**0.0220**	**0.9808**	0.9607	0.1414	0.8971	0.9895	0.0025	0.0015	**25.68**

TABLE I

QUANTITATIVE EVALUATION ON CELEBAMASK-HQ AND FFHQ. METRICS ACROSS ROI, CONTEXT, AND BOUNDARY ASSESS REGIONAL FIDELITY AND TRANSITION SMOOTHNESS. OUR METHOD ACHIEVES THE BEST GLOBAL FID AND LEADING PERFORMANCE IN ROI AND BOUNDARY ZONES.[1]

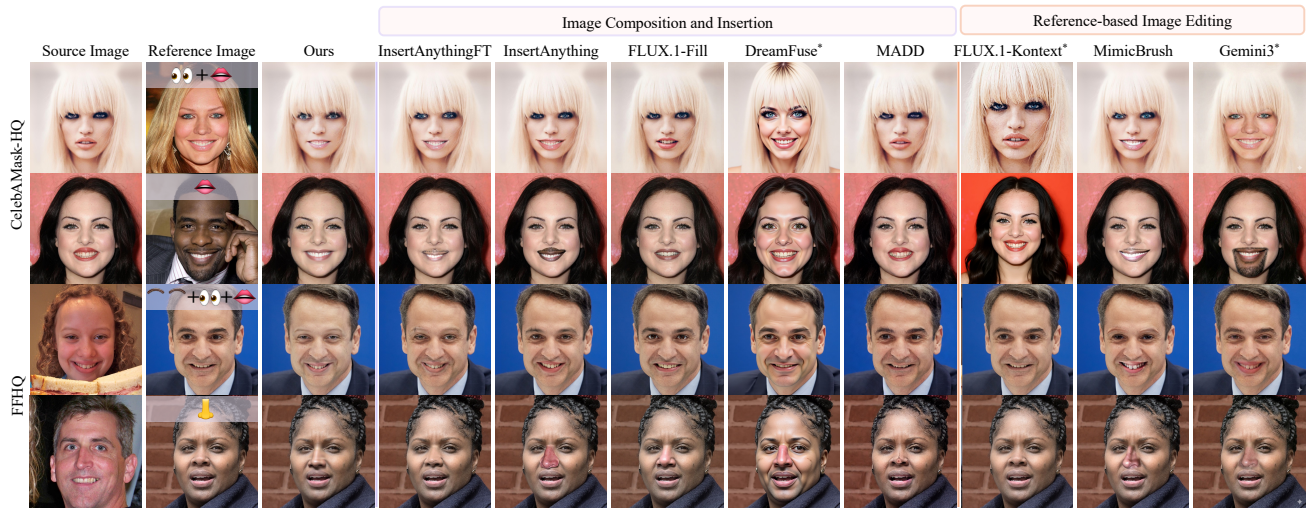


Fig. 2. **Visual results for reference-based face composition.** Our method preserves reference facial components while seamlessly integrating them into the source image, outperforming existing insertion and editing models that often distort component details or fail to maintain structural and textural consistency.

appearance discrepancies between subjects, particularly across ethnicities. Such discrepancies amplify misalignment and produce visible artifacts, including boundary seams and geometric distortions. To improve robustness and generalization under cross-ethnic and cross-appearance variations without paired training data, we propose a component-aware augmentation strategy using a two-stage transformation process.

Given a reference region $\mathcal{R} = \mathbf{M} \cdot \mathbf{I}_{\text{ref}}$, we apply sequential photometric and geometric augmentations to simulate cross-ethnic appearance and structural variations:

$$\hat{\mathcal{R}}_{\text{roi}} = \mathcal{T}_{\text{geo}}\big(\mathcal{T}_{\text{phot}}(\mathcal{R})\big).$$

Photometric augmentation $\mathcal{T}_{\text{phot}}$ perturbs the LAB color space where the asymmetric luminance shift $\Delta_L \sim \mathcal{U}(-8, 5)$ biases samples toward darker tones, while moderate chrominance shifts $\Delta_a \sim \mathcal{U}(-4, 4)$ and $\Delta_b \sim \mathcal{U}(-6, 6)$ introduce color diversity. The perturbed image is then converted to RGB space to produce the photometrically augmented component.

Following photometric augmentation, $\mathcal{T}_{\text{geo}}$ applies geometric transformations to simulate realistic spatial variations in facial component shape and placement. Specifically, a mild affine transformation is applied with rotations ($\pm 5°$), translations (up to $\pm 5\%$ of the image size), and isotropic scaling in $[0.95, 1.05]$ to account for pose variations. Together with photometric augmentation, this design disentangles appearance and geometric variations, improving generalization across ethnically diverse and structurally heterogeneous faces.

The non-edited facial context of the source image $\mathbf{I}_{\text{src}}$ is augmented using the same transformation pipeline, with $\mathbf{M}_{\text{context}}$ denoting its corresponding binary mask:

$$\hat{\mathcal{I}}^{\text{src}}_{\text{context}} = \mathcal{T}\big(\mathbf{M}_{\text{context}} \cdot \mathbf{I}_{\text{src}}\big), \quad \mathcal{T} = \mathcal{T}_{\text{geo}} \circ \mathcal{T}_{\text{phot}}.$$

C. Context-Aware Face Fusion Framework

Ensuring coherent structure and harmonious texture in composed faces requires careful integration of augmented references with the source context. After augmentation, the fusion input is defined as

$$\mathcal{F}_{\text{input}} = \hat{\mathcal{I}}^{\text{src}}_{\text{context}} \oplus \hat{\mathcal{R}}^{\text{ref}}_{\text{roi}},$$

[1]Unlike baselines that use rigid ROI masks to keep context pixels unchanged—yielding artificially high context scores—our *Context-Aware Face Fusion* treats the entire face as an adaptive synthesis area. This prioritizes global coherence over pixel-level preservation, ensuring seamless integration and superior realism as evidenced by our leading FID scores.

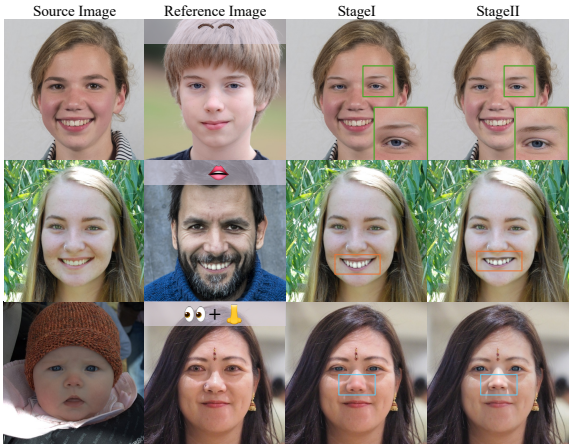


Fig. 3. **Stage I vs. Stage II comparison.** Each row shows the source, reference, results from both stages, and zoomed-in patches. Bounding boxes highlight artifacts in Stage I and corresponding improvements in Stage II.

where $\hat{\mathcal{R}}_{\text{roi}}$ is the aggregated region of interest:

$$\hat{\mathcal{R}}^{\text{ref}}_{\text{roi}} = \sum_{c\in\mathcal{C}} \mathcal{T}\Big(\mathbf{M}_{\text{c}} \odot \mathbf{I}^{(c)}_{\text{ref}}\Big),$$

with $\mathcal{C}$ the set of facial components, $\mathbf{M}_{\text{c}}$ the spatial mask of the selected component, and $\mathbf{I}^{(c)}_{\text{ref}}$ the corresponding reference. Here, $\oplus$ denotes channelwise concatenation of context and region features, and $\odot$ denotes componentwise masking of reference components, and $\sum$ denotes spatial aggregation of non-overlapping component regions.

Unlike hard-masked insertions, which often introduce seams and artifacts [28], our framework jointly processes $\hat{\mathcal{R}}^{\text{ref}}_{\text{roi}}$ and $\hat{\mathcal{I}}^{\text{src}}_{\text{context}}$ as complementary fusion components. Both are encoded by a frozen VAE into latent tokens via $\mathcal{Z} = \text{VAE}(\mathcal{F}_{\text{input}})$ and passed to DiT transformer blocks. Cross-attention modulation balances component fidelity with global consistency, enabling structure- and texture-coherent composition.

D. Multi-Stage Training Strategy

Reference-guided face composition is challenging without paired ground truth that preserves both reference identity and source structure. We tackle this with a two-stage curriculum that transitions from identity-preserving reconstruction to cross-identity composition under hybrid supervision.

1) Stage I: Self-Reconstruction: Given

$$\mathcal{F}^{\text{recon}}_{\text{input}} = \hat{\mathcal{I}}^{\text{src}}_{\text{context}} \oplus \hat{\mathcal{R}}^{\text{src}}_{\text{roi}}$$

constructed from pair $(\mathbf{I}_{\text{src}}, \mathbf{I}_{\text{src}})$ of the same source image, the diffusion model predicts noise ε_θ on latent $\mathbf{z}_t$ with the standard denoising loss:

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{t,\varepsilon}\Big[\, \|\varepsilon - \varepsilon_\theta(\mathbf{z}_t, t; \mathcal{F}^{\text{recon}}_{\text{input}})\|^2_2 \,\Big].$$

This stage guides the model to recover facial identity, structure, and appearance under perfect alignment, providing a stable initialization for subsequent reference-guided learning.

2) Stage II: Hybrid Pseudo-Supervision: Using the frozen Stage I model, pseudo composites $\tilde{\mathbf{I}}_{\text{pseudo}}$ are synthesized from source–reference pairs $(\mathbf{I}_{\text{src}}, \mathbf{I}_{\text{ref}})$ (bottom-left of Fig. 1).

To regulate the hybrid supervision, we sample $\alpha \sim \mathcal{U}(0,1)$ and define a binary indicator α_{bin} to determine the activation of pseudo-label supervision. Controlled by a threshold τ, the indicator is formulated as:

$$\alpha_{\text{bin}} = \mathbf{1}(\alpha < \tau), \tag{1}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. The hybrid training objective is thus defined as:

$$\mathcal{L}_{\text{hybrid}} = (1-\alpha_{\text{bin}})\,\mathcal{L}_{\text{recon}} + \alpha_{\text{bin}}\,\mathbb{E}_{t,\varepsilon}\Big[\, \|\varepsilon - \varepsilon_\theta(\mathbf{z}_t, t; \mathcal{F}_{\text{input}})\|^2_2 \,\Big].$$

By stochastically switching between self-reconstruction and pseudo-supervision, the model maintains the structural integrity of the source while learning realistic feature transfer. As illustrated in Fig. 3, this hybrid approach effectively rectifies the artifacts (e.g., boundary seams, color inconsistency and unfaithful transfer) exist in Stage I, leading to superior generation quality.

III. Experiments

Implementation Details. Our method builds on FLUX.1 Fill [3], an inpainting model based on DiT [30]. We fine-tune it for face composition using LoRA [31] with rank 256, enabling efficient updates while preserving the pre-trained backbone. Training proceeds on 10,000 source–reference image pairs using a two-stage protocol. Stage I operates at a resolution of 256×256 for 5,000 optimization steps, focusing on learning coarse alignment between facial components and the global facial context. From Stage I outputs, 9,705 source–reference pairs with corresponding pseudo composite images $\tilde{\mathbf{I}}_{\text{pseudo}}$ are filtered and selected as pseudo labels to provide supervision for Stage II. Stage II increases the resolution to 1024×1024 and continues training on the selected 9,705 pairs for 50,000 steps, enabling refinement of structural consistency and high-frequency details. All experiments are conducted on NVIDIA A100 and A40 GPUs, and quantitative evaluation is performed on 2,000 held-out source–reference pairs.

Baselines. We compare with InsertAnything [28] (also fine-tuned on our data), FLUX.1-Fill [3], DreamFuse [26], MADD [27] for image insertion, and FLUX.1-Kontext-dev [3], MimicBrush [17], Gemini2.5 [22] for reference-based editing.

Evaluation Metrics. We evaluate composition quality using LPIPS [32], SSIM [33], and CLIP similarity [34], computed over two semantic regions: the *ROI*, corresponding to the edited facial components, and the *Context*, covering the remaining facial area. Metrics on the ROI quantify alignment with the reference image, while Context metrics assess consistency with the source facial appearance. In addition, we report FID [35] to measure overall image realism at the dataset level.

To assess the smoothness of transitions at the compositing interface, we define a *Boundary* region as a narrow band surrounding the ROI. Within this boundary, we report two operator-based metrics: SobelDiff and LapDiff, which quantify structural discrepancies between the source and edited images:

$$\text{SobelDiff} = \frac{1}{|\mathcal{B}|}\sum_{p\in\mathcal{B}} \big|\, \|\nabla_{\text{Sobel}}\mathbf{I}_{src}(p)\| - \|\nabla_{\text{Sobel}}\mathbf{I}_{\text{edit}}(p)\| \,\big|,$$

$$\text{LapDiff} = \frac{1}{|\mathcal{B}|}\sum_{p\in\mathcal{B}} \big|\, \Delta\mathbf{I}_{src}(p) - \Delta\mathbf{I}_{\text{edit}}(p) \,\big|,$$

Module			LPIPS↓		SSIM↑		CLIP↑		FID↓
Photometric	Geometric	ContextFusion	ROI	CONTEXT	ROI	CONTEXT	ROI	CONTEXT	
			0.0215	**0.0244**	0.9812	**0.9902**	**0.9611**	**0.9976**	28.76
✓			0.0221	0.0266	0.9810	0.9896	0.9577	0.9974	28.74
✓	✓		0.0219	0.0253	0.9811	0.9899	0.9582	0.9974	29.46
✓	✓	✓	**0.0207**	0.1142	**0.9815**	0.8997	**0.9611**	0.9799	**27.88**

TABLE II

ABLATION STUDY ON DATA AUGMENTATION AND CONTEXT FUSION. BEST RESULTS ARE IN **BOLD**, SECOND-BEST UNDERLINED. COMBINING PHOTOMETRIC AND GEOMETRIC AUGMENTATIONS WITH CONTEXT FUSION YIELDS THE BEST TRADE-OFF, IMPROVING ROI METRICS AND FID.

τ	LPIPS↓		SSIM↑		FID↓
	ROI	CONTEXT	ROI	CONTEXT	
0.2	0.0196	0.1120	0.9835	0.9100	29.31
0.3	0.0199	0.1249	0.9823	0.8925	28.12
0.5	0.0207	0.1142	0.9815	0.8997	27.88
0.75	0.0211	0.1209	0.9811	0.8968	28.25
0.9	0.0226	0.1328	0.9804	0.8889	30.35

TABLE III

ABLATION STUDY ON THE MIXING THRESHOLD τ BETWEEN PSEUDO-LABELED AND REAL IMAGE PAIRS IN STAGE II TRAINING. THE **BEST** AND SECOND-BEST RESULTS FOR EACH METRIC ARE HIGHLIGHTED.

where $\mathcal{B}$ is the boundary region, $\nabla_{\text{Sobel}}(\cdot)$ is the Sobel gradient, and $\Delta(\cdot)$ is the Laplacian. SobelDiff measures edge discontinuities (first-order gradients), while LapDiff captures curvature variations (second-order structure). Lower values indicate smoother, more coherent transitions.

A. Quantitative results

Table I presents the quantitative results on CelebAMask-HQ and FFHQ. Our method consistently achieves the best **Global FID** (27.88 and 25.68), indicating superior overall realism and structural coherence. In the **ROI**, our approach delivers leading LPIPS and SSIM scores, demonstrating high fidelity in facial component preservation. Furthermore, the significantly lower **Boundary** discrepancy (e.g., 0.0010 LapDiff) confirms the effectiveness of our seamless integration. While some baselines show higher SSIM in the **Context**, this reflects our strategic *Context-Aware Fusion* design. By allowing the ROI and context to mutually adapt rather than enforcing rigid pixel preservation, our model prioritizes global composition harmony, ultimately yielding more natural results as evidenced by the state-of-the-art FID performance.

B. Qualitative results

Figure 2 shows qualitative comparisons against state-of-the-art insertion and editing baselines across diverse face composition scenarios. While existing methods largely preserve overall facial structure, they often exhibit unfaithful component transfer, boundary artifacts, and inadequate structural or textural blending. In contrast, our approach achieves seamless integration of reference components with surrounding regions, faithfully retaining fine details from the reference. These benefits are particularly pronounced in high-frequency regions such as eye corners, lip contours, and the nose bridge, where precise structure and identity cues are essential.

C. User Study

We conducted a user study with 26 participants to evaluate perceptual quality across four aspects: (i) component similarity, assessing how well the facial components match the reference image; (ii) identity preservation, judging whether the subject's context are maintained; (iii) blending quality, evaluating the seamlessness of transitions between edited regions and surrounding face; and (iv) overall realism, rating visual plausibility and aesthetic appeal. In the comparative evaluation, 30 image sets were shown, each including outputs from our method and seven state-of-the-art baselines. Participants selected preferred images, and our method was consistently favored: 74.7% for component fidelity, 54.7% for blending, and 51.8% for overall realism (Figure 4).

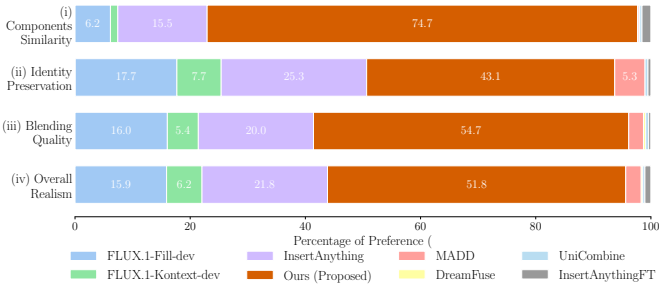


Fig. 4. User preferences in comparative evaluation across four aspects. Our method is consistently favored for components, identity, blending, and realism.

In the absolute evaluation, 50 samples generated by our method were rated on a 5-point Likert scale for the same four aspects. High satisfaction was observed: "excellent" ratings (5) reached 53.5% for component fidelity, 59.2% for identity preservation, 50.9% for blending, and 51.2% for overall realism. Considering positive ratings (Score ≥ 4), our method achieved over 80% satisfaction in fidelity and 87% in identity, demonstrating robust identity preservation and high fidelity.

D. Ablation study

We perform ablation studies to isolate the roles of the *Component-Aware Augmentation* strategy and the *Context-Aware Face Fusion* framework, and to assess the impact of training data composition.

Effect of Augmentation and Context-Aware Fusion. Table II shows that photometric and geometric augmentation improves robustness, while the Context-Aware Face Fusion mechanism further enhances boundary blending. Each step boosts structural coherence and visual fidelity, with the full pipeline yielding the best ROI LPIPS/SSIM and overall FID.

Mixing Ratio of Pseudo and Real Pairs. Table III shows that increasing the proportion of reconstructed ground truth pairs improves LPIPS and SSIM by preserving selected components and context, while the best FID is achieved with a balanced 50% pseudo-labeled / 50% real mix ($\tau = 0.5$).

IV. Conclusion

We tackled the challenging task of reference-based face composition, requiring precise component transfer with global structural and photometric consistency. Our *Context-Aware Face Fusion*, together with *Component-Aware Augmentation* and *Multi-Stage Training*, enables robust learning in the absence of paired training data by explicitly modeling spatial, appearance, and contextual relations. These designs strike a balance between local fidelity and global harmony, allowing FaceComposer to surpass existing methods in both realism and structural coherence. Beyond achieving state-of-the-art results, our work offers new insights into structure-sensitive face editing and opens avenues for broader applications in reference-based image composition.

Acknowledgment

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

[1] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[2] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Forty-first international conference on machine learning*, 2024.

[3] Black Forest Labs, "Flux," https://github.com/black-forest-labs/flux, 2024, accessed: 2025-07-15.

[4] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint:2307.01952*, 2023.

[5] G. Koo, S. Yoon, and C. D. Yoo, "Wavelet-guided acceleration of text inversion in diffusion-based image editing," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 4380–4384.

[6] Z. Tan, S. Liu, X. Yang, Q. Xue, and X. Wang, "Ominicontrol: Minimal and universal control for diffusion transformer," *arXiv preprint arXiv:2411.15098*, 2024.

[7] A. Cvejic, A. Eldesokey, and P. Wonka, "Partedit: Fine-grained image editing using pre-trained diffusion models," in *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, 2025, pp. 1–11.

[8] D. Garibi, S. Yadin, R. Paiss, O. Tov, S. Zada, A. Ephrat, T. Michaeli, I. Mosseri, and T. Dekel, "Tokenverse: Versatile multi-concept personalization in token modulation space," *ACM Transactions on Graphics (TOG)*, vol. 44, no. 4, pp. 1–11, 2025.

[9] S. Wu, M. Huang, W. Wu, Y. Cheng, F. Ding, and Q. He, "Less-to-more generalization: Unlocking more controllability by in-context generation," *arXiv preprint arXiv:2504.02160*, 2025.

[10] Y. Yu, Z. Zeng, H. Zheng, and J. Luo, "Omnipaint: Mastering object-oriented editing via disentangled insertion-removal inpainting," *arXiv preprint arXiv:2503.08677*, 2025.

[11] H. Wang, W. Feng, Y. Li, Z. Zhang, J. Lv, J. Shen, Z. Lin, and J. Shao, "Generate e-commerce product background by integrating category commonality and personalized style," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[12] C. Wang, J. Gu, P. Hu, Y. Guo, X. Dong, H. Xu, and X. Liang, "Dreamvideo: High-fidelity image-to-video generation with image retention and text guidance," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[13] J. Chung, S. Hyun, and J.-P. Heo, "Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 8795–8805.

[14] Z. Zhang, Q. Zhang, W. Xing, G. Li, L. Zhao, J. Sun, Z. Lan, J. Luan, Y. Huang, and H. Lin, "Artbank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, 2024, pp. 7396–7404.

[15] M. Brack, F. Friedrich, K. Kornmeier, L. Tsaban, P. Schramowski, K. Kersting, and A. Passos, "Ledits++: Limitless image editing using text-to-image models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 8861–8870.

[16] B. Liu, C. Wang, T. Cao, K. Jia, and J. Huang, "Towards understanding cross and self-attention in stable diffusion for text-guided image editing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 7817–7826.

[17] X. Chen, Y. Feng, M. Chen, Y. Wang, S. Zhang, Y. Liu, Y. Shen, and H. Zhao, "Zero-shot image editing with reference imitation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 84 010–84 032, 2024.

[18] C. Mao, J. Zhang, Y. Pan, Z. Jiang, Z. Han, Y. Liu, and J. Zhou, "Ace++: Instruction-based image creation and editing via context-aware content filling," *arXiv preprint arXiv:2501.02487*, 2025.

[19] L. Jiang, Q. Yan, Y. Jia, Z. Liu, H. Kang, and X. Lu, "Infiniteyou: Flexible photo recrafting while preserving your identity," *arXiv preprint arXiv:2503.16418*, 2025.

[20] H. Lin, "Dreamsalon: A staged diffusion framework for preserving identity-context in editable face generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8589–8598.

[21] K. Shiohara and T. Yamasaki, "Face2diffusion for fast and editable face personalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6850–6859.

[22] Google, "Gemini 2.5 Flash Image," AI Model and Service, 2025.

[23] R. He, K. Ma, L. Huang, S. Huang, J. Gao, X. Wei, J. Dai, J. Han, and S. Liu, "Freeedit: Mask-free reference-based image editing with multi-modal instruction," *arXiv preprint arXiv:2409.18071*, 2024.

[24] X. Chen, L. Huang, Y. Liu, Y. Shen, D. Zhao, and H. Zhao, "Anydoor: Zero-shot object-level image customization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 6593–6602.

[25] H. Wang, J. Peng, Q. He, H. Yang, Y. Jin, J. Wu, X. Hu, Y. Pan, Z. Gan, M. Chi *et al.*, "Unicombine: Unified multi-conditional combination with diffusion transformer," *arXiv preprint arXiv:2503.09277*, 2025.

[26] J. Huang, P. Yan, J. Liu, J. Wu, Z. Wang, Y. Wang, L. Lin, and G. Li, "Dreamfuse: Adaptive image fusion with diffusion transformer," *arXiv preprint arXiv:2504.08291*, 2025.

[27] J. He, W. Li, Y. Liu, J. Kim, D. Wei, and H. Pfister, "Affordance-aware object insertion via mask-aware dual diffusion," *arXiv preprint arXiv:2412.14462*, 2024.

[28] W. Song, H. Jiang, Z. Yang, R. Quan, and Y. Yang, "Insert anything: Image insertion via in-context editing in dit," *arXiv preprint arXiv:2504.15009*, 2025.

[29] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.

[30] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.

[31] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[32] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. [Online]. Available: https://ieeexplore.ieee.org/document/1261367

[34] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[35] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.