FACEKEEPER: MULTI-EXPERT LORA FOR HIGH-FIDELITY PORTRAIT RETOUCHING

Anonymous authors
Paper under double-blind review



Figure 1: We propose FaceKeeper, multi-expert LoRA framework that achieves a balance between blemish removal, detail preservation, and identity consistency.

ABSTRACT

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.

Portrait retouching must balance blemish removal, texture preservation, and identity consistency. Existing learning or diffusion methods often struggle to balance these three objectives: high-frequency details are lost, skin is overly smoothed, or multiple conflicting objectives are compressed into a single LoRA, leading to trade-offs and degradation. This paper proposes FaceKeeper, a multi-expert LoRA framework that internalizes these objectives into a unified training process and gradually resolves conflicts across three dimensions: In terms of supervision, we propose a semantically guided spatial-high-frequency coupled detail oversampling strategy that focuses on salient local regions; at the parameter level, we adopt a three-stage adaptation process:freezing, dense training, and sparse refinement,where the shared projection A is frozen, and the intent-specific factor B is merged into a lightweight adapter via mask-aware integration after dense training and sparse refinement; at the semantic level, we bridge the gap between pixel restoration and texture perception through multi-expert reward pre-alignment. Based on FLUX.2-klein-base-9B, our method outperforms dedicated retouching models and general-purpose editors on the self-built FaceKeeper-Bench and CelebA-HQ benchmarks for reference-free metrics, while remaining competitive for reference-based metrics. User studies further validate its excellent balance between blemish removal and the preservation of texture and identity.

1 INTRODUCTION

With the rapid advancement of technology, people's expectations for photo retouching have become increasingly sophisticated. From early incremental blemish-removal strategies (Xie et al., 2023) to the rise of diffusion models (Xu et al., 2025; Duan et al., 2025; Yang et al., 2026), the requirements for facial retouching have evolved from simple beautification towards a dual pursuit of natural texture and identity preservation. The goal is to remove blemishes without compromising details such as pores and texture, while maintaining identity consistency. The key to effective retouching lies in whether the output can balance the removal of imperfections with the complete preservation of characteristic details and textures. However, in existing methods for such tasks, models tend to erase textural details (Ronneberger et al., 2015; Isola et al., 2017; Wang et al., 2004; Zhang et al., 2018; Ding et al., 2020; Jiang et al., 2021), and the lack of spatially guided global optimization often leads to texture collapse (Wang et al., 2018a;b). The emergence of diffusion models has not alleviated this problem. Whether it is conditional diffusion models (Saharia et al., 2022; Lugmayr et al., 2022) or recent instruction-driven retouching methods (Xue et al., 2025; Wu et al., 2026b; Ye-Bin et al., 2025; Zeng et al., 2026; Labs et al., 2025), the MSE or flow-matching loss in their training objectives naturally tends to converge at low frequencies, causing the models to inevitably erase the fine textural details of the skin while removing imperfections (Xu et al., 2025; Yang et al., 2026) . Existing remedies (Yang et al., 2026; Zhou et al., 2022) have failed to fundamentally reconstruct the spatial geometry and frequency division of the supervision signal. The above analysis raises a question: Is it possible to reconstruct the supervision signal from the perspective of the coupling between spatial localization and frequency response, enabling the model to autonomously learn during training "where fine-grained reconstruction is needed and where moderate smoothing is acceptable?" Multi-expert LoRA aims to balance fine-tuning efficiency with multi-task generalization capabilities (Hu et al., 2021; Zhang et al., 2023b; 2025) , but existing approaches generally have limitations. On the one hand, some methods rely on complex architectures or external modules, raising the barrier to adoption (Xiao et al., 2025; Li et al., 2025; Peng et al., 2026; Simsar et al., 2025; Lee et al., 2025; Wu et al., 2026a; Liu et al., 2025); on the other hand, mainstream methods mostly adopt weight superposition or discrete routing strategies (Yadav et al., 2023) , failing to address the problem of multi-expert collaboration at the gradient level. Due to the lack of an adaptive gradient fusion mechanism, the models cannot effectively resolve gradient interference caused by conflicting styles, thereby limiting the full realization of multi-expert capabilities.

To prevent different skin intents from interfering with one another and to achieve a balanced harmony among blemish removal, texture detail preservation, and identity consistency, we propose the FaceKeeper framework, which introduces a multi-expert adaptation scheme where each intent is processed by an independent LoRA expert. Its core design is reflected in the following three aspects:

•We propose a semantically guided space-high-frequency coupling detail oversampling strategy that uses discrete set selection and a Gaussian semantic kernel to precisely focus the supervision signal on salient local regions, effectively alleviating the issues of gradient dominance and oversmoothing in global regression.

•We designed a multi-expert sparse adaptation architecture: we freeze the shared base model A to preserve general representations, perform intensive differential training on each expert to achieve functional differentiation, and then consolidate the unique parameter subspaces of each expert through amplitude-mask-guided sparse fine-tuning. This "freeze-dense-sparse" strategy naturally decouples the feature experts in the parameter space.

•We introduce a multi-expert reward pre-alignment mechanism to bridge the gap between pixel-level restoration and semantic-level texture, thereby achieving a unified approach to blemish removal, texture detail preservation, and identity consistency.

2 Related Work

2.1 Face Retouching

The core of facial beautification lies in balancing blemish removal, texture preservation, and identity consistency. When paired data is limited, incremental blemish removal (Xie et al., 2023) remains ineffective: blemishes, details, and identity are highly intertwined at the pixel level, and failure to separate them can result in missed corrections or unintended alterations to identity. Early learning-based retouching was constrained by deterministic mappings; although it preserved high-frequency contours (Ronneberger et al., 2015), L1/L2 losses tended to result in skin smoothing (Isola et al., 2017); perceptual metrics (Isola et al., 2017; Wang et al., 2004; Zhang et al., 2018; Ding et al., 2020; Jiang et al., 2021) could suppress texture collapse but, due to a lack of spatial guidance, still tended to smooth the entire face. Implicit alignment (Zhu et al., 2017) and residual compensation (Gharbi et al., 2017) reduce texture synthesis to a secondary task, neglecting the joint modeling of texture and structure; although subsequent high-frequency restoration (Wang et al., 2018a;b) has improved, the supervisory signal remains insufficient to precisely decouple blemishes from features that need to be preserved. To address this, CodeFormer (Zhou et al., 2022) introduces a discrete codebook and an adjustable fidelity parameter to achieve a trade-off between blemish removal and identity preservation.

Further developments in diffusion models have not completely resolved the aforementioned contradictions. Although improvements have been made in terms of conditional control and identity consistency (Saharia et al., 2022; Lugmayr et al., 2022), existing methods still struggle to balance global consistency with local details (Xu et al., 2025) and are prone to the problem of excessive smoothing (Duan et al., 2025). Specifically, the soft masking mechanism in RetouchFormer (Wen et al., 2024) operates only within the network, while the globally applied perceptual constraints actually reinforce the model's tendency toward smoothing. Furthermore, although BeautyGRPO (Yang et al., 2026) introduces ArcFace to emphasize identity consistency, it does not internalize this as pixel-level supervision.

In summary, the root cause of texture collapse lies in the inability of the supervisory signal to precisely decouple the artifacts to be removed from the features to be preserved. We apply the aforementioned objective directly to the three-stage multi-LoRA expert training process, rather than relying on post-processing.

2.2 The evolution of generative image modeling

Generative image editing backbone models have evolved from innovations in the generative paradigm (Xue et al., 2025; Rombach et al., 2021; Lipman et al., 2022; Liu et al., 2022; Peebles & Xie, 2023; Esser et al., 2024) to the intelligent implementation of command interfaces. Specifically, InstructPix2Pix (Brooks et al., 2023) established a natural language editing interface, and subsequent methods (Zhang et al., 2023a; Gu et al., 2025; Bai et al., 2026) enhanced data understanding and intent parsing. The introduction of large language models has further raised the bar, progressing from planning strategies (Xue et al., 2025) to agent-based decision-making (Wu et al., 2026b) and then to the exploration of code-based image editing (Ye-Bin et al., 2025), gradually establishing a closed-loop for aesthetic reasoning (Zeng et al., 2026). FLUX.1 Kontext (Labs et al., 2025), as a recent representative, has preliminarily achieved the unification of generation and editing. However, the external invocation paradigm suffers from information loss, serialization delays, and gradient disconnection. Furthermore, overreliance on MSE reconstruction loss with its low-frequency bias,whichtends to erase skin texture details, necessitating the introduction of an explicit frequency domain to balance editability and identity consistency.

2.3 Multi-Expert LoRA

LoRA (Hu et al., 2021) has been widely used for parameter-efficient fine-tuning, and LoRA-FA (Zhang et al., 2023b) further demonstrates that component A can be frozen while training component B. However, under multiple conflicting objectives, dense multi-LoRA architectures are prone to mutual interference; LoRI (Zhang et al., 2025), which reduces multi-task interference in LLMs by orthogonalizing the adapter subspaces, inspired us. However, the core challenge in merging is parameter-level interference (Yadav et al., 2023), meaning that simply averaging parameters across different tasks can easily cause the model to collapse into a smooth compromise solution. In diffusion and multimodal generation, the merging of multiple LoRAs still requires optimization. Existing approaches suffer from three main drawbacks: First, architectural coupling. Embedded routing or supernets (Xiao et al., 2025; Li et al., 2025) struggle to transfer across heterogeneous backbones; TARA's (Peng et al., 2026) token-level masking relies on specific cross-attention mechanisms, lacking generality. Second, external dependencies and retraining costs are high. LoRACLR (Simsar et al., 2025) relies on

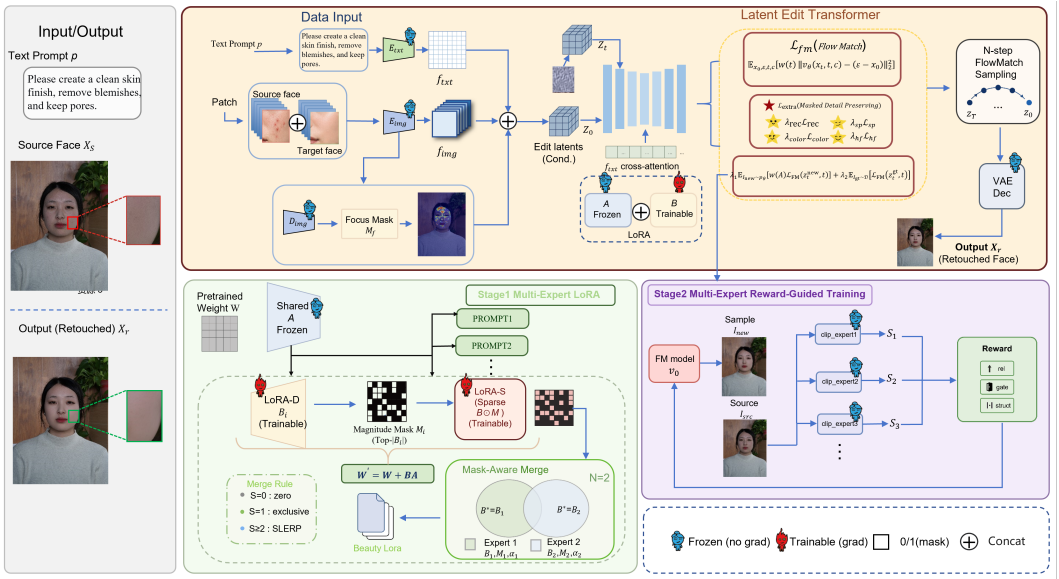

Figure 2: Overview of FaceKeeper.FaceKeeper couples space-frequency detail supervision with freeze-dense-sparse multi-expert LoRA and multi-expert reward-guided perceptual alignment for high-fidelity portrait retouching

contrastive alignment of intermediate representations, resulting in weakened signals when semantic similarity is high; K-StyleLoRA (Lee et al., 2025) depends on the external CLIP model; and CollectionLoRA (Wu et al., 2026a) requires re-running the distillation process whenever the set of adapters is modified. Finally, there is the issue of inference cost. Although FreeFuse (Liu et al., 2025) avoids the need to modify the architecture during training, it requires retaining all candidate LoRAs during inference, causing GPU memory usage to increase linearly with the number of adapters. In the supervised paradigm, diffusionOPD (Li et al., 2026) was the first to integrate diffusion-based multi-task learning, focusing on general-purpose image generation via online distillation; however, it cannot distinguish between modified and retained features at the pixel level. In summary, the current approach still struggles to reliably merge multiple conflicting intents into a single adapter without modifying the underlying architecture or relying on external modules.

3 METHOD

To address the trade-offs between blemish removal, texture preservation, and identity consistency in facial retouching, as well as the interdependencies among experts for different skin types, this paper proposes the FaceKeeper architecture. This architecture comprises three complementary core components: (1) a semantically guided, spatially high-frequency detail-coupled multi-sampling strategy that determines the details the model perceives. (2) A multi-expert adaptation scheme comprising three LoRA experts, combined with a three-stage "freeze-dense-sparse" adaptive training process, which determines which combination of merged parameters the model uses to express its intent. (3) Multi-expert reward pre-alignment, which determines how the model makes choices based on reward alignment. These three components work in concert to resolve the aforementioned conflicts. The main text focuses on a subset of texture-based skin experts and reports the results; the remainder is presented in Appendix B.

3.1 SPACE-FREQUENCY COUPLED SAMPLING

To address the challenges of balancing blemish removal, detail preservation, and identity consistency, we must reexamine the fundamental difficulties of high-resolution portrait editing. The blemishes to be removed and the target texture are sparsely distributed across the image and are primarily manifested as high-frequency components. During whole-image training, large areas of flat skin and the background dominate the loss gradients. More fundamentally, global regression faces a threefold contradiction: (1) low-frequency global context requires large receptive fields.(2) high-frequency local textural details require refined modeling.(3) random or fixed windows struggle to accommodate spatially separated defect regions such as the forehead, cheeks, and chin, often leading to coverage redundancy or gaps. These challenges urgently call for a joint optimization approach.

Inspired by the above analysis, we no longer process the entire graph or make point-by-point selections independently; instead, we model the selection of K patch anchors as a discrete set selection problem.

Specifically ,define the local saliency score as follows:

$$s(u;\mathcal{I}) = \log\left(1 + \int_{B(u,r)} \|\nabla I_s(v)\|_F^2 \, dv\right), \tag{1}$$

Where ∇I_s is the spatial gradient tensor, characterizing local texture abruptness and edge strength; $B(u,r)$ denotes the local spatial neighborhood centered at pixel u. And v is each pixel traversed within this local region To encode the semantic consistency of spatial locations, we introduce a geometry-aware spatial embedding e_u and a Gaussian semantic kernel(Appendix E.2):

$$\kappa(u,c) = \exp\Big(-\frac{\|e_u - e_c\|^2}{2\sigma^2}\Big), \tag{2}$$

whose construction and implementation are detailed in Appendix E. Let the candidate anchor set be $B = \{b_1, \ldots, b_M\} \subset \Omega$. The coverage utility of an anchor subset A with $|A| = K$ is defined as:

$$F(A) = \sum_{u \in B} \max_{c \in A} s(u;\mathcal{I}) \cdot \kappa(u,c). \tag{3}$$

The three formulas above represent, respectively: (1) the tendency of anchor points to cluster in regions of high salience; (2) the incorporation of semantic prior knowledge via a kernel function; and (3) the implicit implementation of spatial deduplication by the max operator. We employ a greedy algorithm to select K anchors from the candidate set B to maximize the submodular function $F(A)$, which guarantees a $(1 - 1/e)$ approximation ratio. The proof is provided in Appendix E.1.

3.2 Intent-Specific Sparse LoRA

When training multiple intents together on a single low-rank parameter, conflicting gradient directions significantly weaken the intent-specific responses, ultimately leading to an overly smooth, compromised output. To address this, we maintain a set of LoRA experts within the parameter space, allowing each intent to be optimized independently during training. The general-purpose encoder layers remain frozen, and the differences between intents are confined to the B-matrices, which can be merged. Set Definition.For the target linear layer in DiT, the standard low-rank update is

$$W' = W + \frac{\alpha}{r} BA, \quad A \in \mathbb{R}^{r \times d}, \quad B \in \mathbb{R}^{d \times r}, \tag{4}$$

where r is the rank and d is the number of hidden dimensions. We construct a set of experts:

$$S = \bigcup_{i=1}^{N} \{A, B_i, M_i\}, \tag{5}$$

Where N denotes the total number of intents. is shared and frozen across all experts, while all the discrepancies are captured in the trainable B_i. This design is based on three considerations:(1)Unified Basis: All experts reside in the same low-rank projection space, ensuring algebraic compatibility for subsequent parameter merging;(2)General-Purpose Code Reuse: A carries general-purpose directions independent of intent, avoiding the need to relearn each channel from scratch;(3)Training Efficiency: Only B_i receives gradients, consistent with the asymmetric adaptation paradigm of DiT-LORA, significantly reducing computational overhead.

Stage 1: Activation of Dense Intentions. Each expert is trained independently on its own exclusive intent data, without mixing batches across intents. In this stage, all parameters of B_i are tunable, with the goal of fully activating the exclusive feature responses for each intent. Formally, for the i-th expert, we optimize:

$$B_i^{(0)} = \arg\min_{B} \mathcal{L}_i \left(W + \frac{\alpha}{r} BA\right), \tag{6}$$

Here, $\mathcal{L}_i$ denotes the supervised loss for the i-th intent, This stage outputs the dense parameters $B_i^{(0)}$, whose components already carry the semantic information of the current intent. However, numerous small-magnitude components contribute negligibly to the intent and will introduce spurious noise during subsequent merging.

Stage 2: Amplitude-Driven Mask Construction. Using parameter magnitude as a measurable proxy for importance, we select the positions with the largest magnitudes for each $\{B_i^{(0)}\}$ according to a global retention ratio $p = 1 - s$ (where s is the sparsity ratio) to construct a Boolean mask. Let $N_{\text{total}} = \sum_j \text{numel}(B_j^{(0)})$ and $k = \text{round}(p \cdot N_{\text{total}})$. A single global threshold is obtained by top-k selection (equivalently, the $(1-p)$-quantile) over the concatenated magnitudes:

$$M_i = \mathbf{1}\Big\{\Big|B_i^{(0)}\Big| \geq \tau\Big\}, \quad \tau = \text{Quantile}_{1-p}\Big(\text{concat}_j \, \Big|B_j^{(0)}\Big|\Big). \tag{7}$$

Where τ denotes the $(1-p)$-quantile of the concatenated magnitudes of all B matrices, the positions with $M_i = 1$ correspond to the coordinates exhibiting the strongest responses in the dense phase and are viewed as the core support of the intent; the remaining positions, viewed as weak responses or noise support, are pruned by the mask. This mask is not randomly sparse but anchors important positions on high-magnitude support.

Stage 3:Sparse Refinement. Continue training under the mask constraint while constraining parameter updates and gradient backpropagation:

$$B_i \leftarrow B_i \odot M_i, \quad \nabla B_i \leftarrow \nabla B_i \odot M_i, \tag{8}$$

Specifically, updates occur only at the core support points where $M_i = 1$. Components outside the mask are forced to zero and no longer receive gradients, causing the support set to contract. At this stage, each expert gradually converges to the manifold corresponding to its specific intent. Following sparsification, the intersection of the support sets of different experts in parameter space is minimized, thereby mitigating gradient conflicts during the merging process,which is a prerequisite for low-interference multi-intent merging.

Mask-Aware Merging. To resolve directional conflicts among intents that weaken instruction discriminability, we perform mask-aware merging at the *channel row-vector* level. Let $\mathbf{b}_j(u) = B_j(u,:) \in \mathbb{R}^r$ denote the u-th output-channel row of expert j, and let $M_j^{\mathrm{row}}(u)$ be the corresponding row-level indicator aggregated from the element-wise mask. For each channel we count the number of active experts, $s(u) = \sum_i M_i^{\mathrm{row}}(u)$, and apply a case-wise strategy: (1) if $s(u) = 0$, the row is treated as noise and set to zero; (2) if $s(u) = 1$, we directly keep the unique active expert's row vector, preserving exclusive capability without degradation; (3) if $s(u) \geq 2$, we merge the active row vectors via SLERP-Tree, avoiding the directional cancellation and sign flipping of element-wise linear averaging. The merged matrix B^* is defined as:

$$B^*(u,:) = \begin{cases} 0, & s(u) = 0, \\ \mathbf{b}_1(u), & s(u) = 1, \\ \mathrm{SLERP\text{-}Tree}\big(\{\mathbf{b}_j(u) \mid M_j^{\mathrm{row}}(u) = 1\}; \{w_j(u)\}\big), & s(u) \geq 2. \end{cases} \tag{9}$$

Details of mask aggregation and weight normalization are provided in Appendix C.

3.3 MULTI EXPERT REWARD LEARNING

We optimize the Flow Matching model on its own online generations using multi-expert CLIP feedback, and use this feedback to modulate the update strength of each sample. For each attribute expert k, we define the similarity $S_k(I) = \langle \Phi_k^{\mathrm{img}}(I), \Phi_k^{\mathrm{text}}(y_k) \rangle_{\cos}$ and the improvement $\delta_k = S_k(I_{\mathrm{new}}) - S_k(I_{\mathrm{src}})$. Let c denote the target attribute. A competitive margin $m = \delta_c - \rho\,\mathrm{Agg}(\{\delta_k\}_{k \neq c})$ rewards the target attribute while penalizing competing ones. An absolute gate $g = \mathbf{1}[S_c(I_{\mathrm{new}}) > \tau]$ and a multi-scale L_1 distance $\mathcal{D}_{dis}(I_{\mathrm{new}}, I_{\mathrm{src}})$further suppress false positives and preserve source-image structure, yielding the reward $\xi = m + \eta\, g - \gamma\, \mathcal{D}_{dis}(I_{\mathrm{new}}, I_{\mathrm{src}})$. We normalize ξ to $\tilde{\xi}$, center it with an EMA baseline b, and advantage $A = \tilde{\xi} - b$, weight $w = \exp\Big(\mathrm{clip}\Big(\frac{A}{\beta}, -4, 4\Big)\Big)$, where β is a temperature hyperparameter. The final objective is $\lambda_1\, \mathbb{E}_{I_{\mathrm{new}} \sim p_\theta}\big[w(A)\, \mathcal{L}_{\mathrm{FM}}(z_t^{\mathrm{new}}, t)\big] + \lambda_2\, \mathbb{E}_{I_{\mathrm{gt}} \sim \mathcal{D}}\big[\mathcal{L}_{\mathrm{FM}}(z_t^{\mathrm{gt}}, t)\big]$, the first term is the advantage-weighted loss, and the second term is the supervised anchor loss.

3.4 LOSS FUNCTIONS

To achieve a balance between defect removal and fidelity preservation, we incorporate a differentiable supervised signal into the adapter's training process. Specifically, in addition to the main loss, we introduce a pixel-level perceptual loss during the later stages of training: first, the clean latent variables are decoded into pixel space, and then spatial weighting is applied using the focus mask M_f, enabling the reconstruction and color to more closely approximate the ground truth in defect-related regions. The overall objective, main loss, and additional loss are defined as follows:

$$\mathcal{L} = \mathcal{L}_{\text{fm}} + \lambda \mathcal{L}_{\text{extra}}, \tag{10}$$

$$\mathcal{L}_{\mathrm{fm}}(\theta) = \mathbb{E}_{x_0, \varepsilon, t, c}\Big[w(t)\, \|v_\theta(x_t, t, c) - (\varepsilon - x_0)\|_2^2\Big], \quad x_t = (1 - \sigma_t)x_0 + \sigma_t \varepsilon \tag{11}$$

$$\mathcal{L}_{\text{extra}} = \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{color}} \mathcal{L}_{\text{color}} + \lambda_{\text{hf}}(\mathcal{L}_{\text{edge}} + \mathcal{L}_{\text{freq}}) + \lambda_{\text{sp}} \mathcal{L}_{\text{sp}} \tag{12}$$

Where $\mathcal{L}_{\text{rec}}$ is the mean squared error within the focal region, driving the reconstruction of defective areas; $\mathcal{L}_{\text{color}}$ is the channel-level cosine color bias loss, which corrects color deviations; $\mathcal{L}_{\text{edge}}$ and $\mathcal{L}_{\text{freq}}$ work together to preserve high-frequency details and edge sharpness; and $\mathcal{L}_{\text{sp}}$ is the perceptual loss, used to suppress texture collapse and maintain natural perceptual quality.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Setup. We used FLUX.2-Klein-base-9B as our pre-trained backbone and trained the model using instruction-paired image retouching triplets (source image, target image, and text instruction); see the appendix D for details on the training process. The overall training is divided into two stages, as shown in Figure 2. All experiments were conducted on 6 NVIDIA H20 GPUs.

Benchmark. To ensure a fair and comprehensive evaluation, we constructed a composite dataset comprising approximately 2,000 8K portraits (FaceKeeper-Bench), which is gender-balanced and includes multiple ethnicities and scenarios. In addition, we included over 2,300 images from the CelebA-HQ 1024 dataset to further enhance facial diversity and identity coverage.

Parameters. The pixel-level loss weights are set to $\lambda_{\text{rec}} = 0.1$, $\lambda_{\text{color}} = 0.1$, $\lambda_{\text{hf}} = 0.1$, and $\lambda_{\text{sp}} = 0.01$, so that reconstruction, color, and high-frequency terms are of comparable magnitude, while the spatial-perceptual term remains comparatively weak, helping preserve skin tone and texture details during blemish removal.

Evaluation Metrics. To comprehensively evaluate generation performance, we employ a set of multidimensional metrics. For reference-based image quality, we use PSNR and SSIM to measure pixel reconstruction and structural similarity, and LPIPS (Zhang et al., 2018) to assess perceptual fidelity; for reference-free image quality, we use NIMA (Talebi & Milanfar, 2018), MANIQA (Yang et al., 2022), NQRM (Ma et al., 2017), and TOPIQ (Chen et al., 2024) to evaluate image quality in terms of artifacts, naturalness, and technical quality.

Baselines. Baselines.We compared the method proposed in this paper with seven representative state-of-the-art baselines, covering both dedicated retouching models and general-purpose image editing frameworks. These include RetouchFormer (Wen et al., 2024), RetouchGPT (Xue et al., 2025), ABPN (Lei et al., 2022), Qwen-Image-Edit-2511 (Wu et al., 2025), FLUX.2-klein-base-9B (Black Forest Labs, 2025),Nano-banana-Pro (Comanici et al., 2025), and GPT-Image-Edit2 (OpenAI, 2026) .

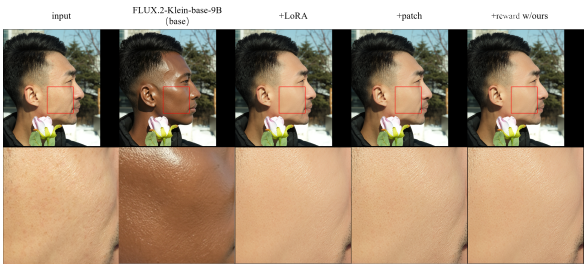

Figure 3: Face retouching results of FLUX.2-klein-base-9B with various training stages.

4.2 Experimental Results

4.2.1 Quantitative Evaluation

Table 1 presents the results of the quantitative comparison. Our method outperforms other methods across the board in terms of perceptual, aesthetic, and reference-free metrics, while maintaining highly competitive pixel fidelity. In contrast, although RetouchFormer and RetouchGPT achieved slightly higher PSNR values, their LPIPS scores dropped significantly; ABPN, while achieving the lowest FID, also left noticeable white artifacts due to over-optimization of pixels (see the first row of Figure 4), **demonstrating that high pixel-level metrics do not necessarily equate to high perceptual quality.** In contrast, our method achieves significant perceptual improvements at a minimal cost to PSNR. Furthermore, we outperform all other methods on reference-free metrics such as NIMA and MANIQA, and achieve a highly competitive second-best performance on FID, significantly outperforming all baselines except ABPN, indicating that the generated image distribution closely resembles that of real data. In summary, our method achieves the optimal balance between pixel accuracy, perceptual quality, and distributional fidelity.

Table 1: Quantitative comparison on FaceKeeper-Bench. Best / second-best are in red / blue.

Method	PSNR↑	SSIM↑	LPIPS↓	NIMA↑	MANIQA↑	NRQM↑	TOPIQ↑	FID↓
ABPN Lei et al. (2022)	29.641	0.869	0.024	5.089	0.437	6.797	0.606	2.169
RetouchFormer Wen et al. (2024)	30.063	0.858	0.237	4.433	0.284	3.535	0.335	28.985
RetouchGPT Xue et al. (2025)	31.079	0.853	0.254	4.380	0.279	3.130	0.312	18.020
Qwen-Image-Edit-2511 Wu et al. (2025)	23.060	0.708	0.215	5.168	0.371	5.605	0.490	22.064
Nano-Banana-Pro Comanici et al. (2025)	21.434	0.608	0.227	4.767	0.410	6.297	0.553	14.211
GPT-Image-Edit2 OpenAI (2026)	20.948	0.623	0.212	5.038	0.389	6.233	0.537	11.075
FLUX.2-klein-base-9B Black Forest Labs (2025)	21.099	0.632	0.237	5.104	0.375	6.143	0.499	24.564
Ours	29.908	0.864	0.055	5.249	0.455	6.957	0.609	4.805

4.2.2 Qualitative Evaluation

Figure 4 shows a comparison of results from various state-of-the-art methods on the task of facial blemish removal and detail preservation. Among dedicated retouching models, ABPN not only fails to remove blemishes thoroughly but also leaves noticeable white retouching marks in the processed areas; RetouchFormer and RetouchGPT similarly suffer from insufficient removal. General-purpose image editing models (such as Qwen-image-edit-2511 and Flux.2-klein-base-9B) tend to over-beautify faces, resulting in a strong artificial look and a lack of realistic texture. Furthermore, Nano-banana-Pro and Gpt-Image-Edit2 exhibit a phenomenon where generated pseudo-textures coexist with residual blemishes. In contrast, our method effectively removes facial blemishes while fully preserving pore details, natural luster, and even skin tone, and maintains personalized features such as moles, demonstrating excellent identity preservation performance.

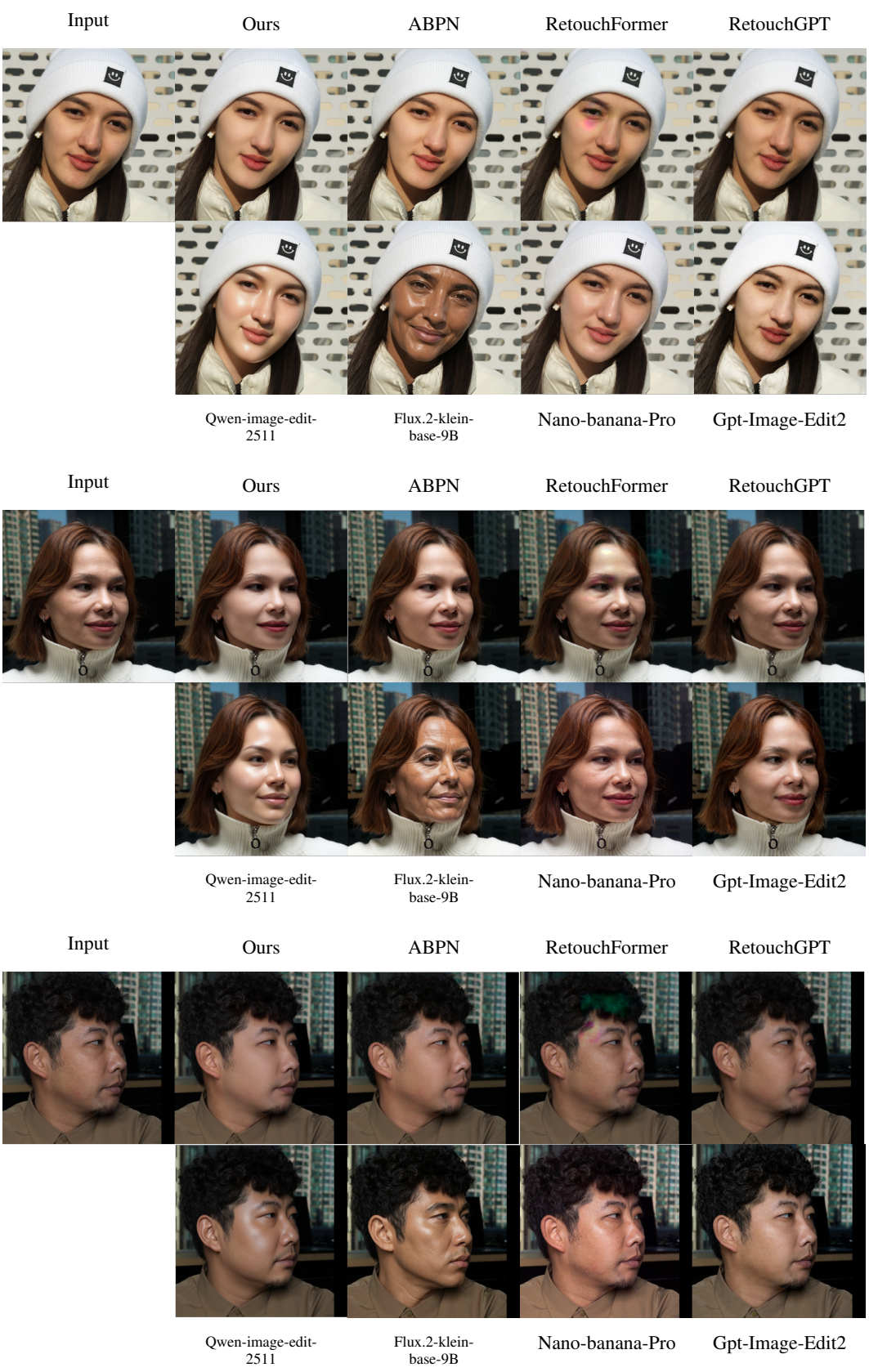

Figure 4: A comparison of visual results from eight different methods on our FaceKeeper-Bench. Our model effectively removes imperfections while preserving texture details and identity consistency, ranking as the best among the various models (more results can be found in Appendix B).

4.3 User Study

We randomly selected 100 representative raw portrait images from the test set and invited 5 evaluators to assess them. Scoring was conducted using a five-point scale ranging from 1 to 5, covering skin smoothness, blemish removal, texture quality, clarity, and identity preservation. To alleviate visual fatigue, each raw image and its retouched version were displayed side-by-side on the same screen in a split-screen layout to facilitate direct comparison; The final score was calculated as the average of the evaluators' ratings. As shown in Figure 5, our method demonstrated the best overall performance, significantly outperforming all comparison methods, particularly in blemish removal. At the same time, it also received high scores in skin smoothness, texture clarity, and identity preservation, clearly indicating that it achieves a good balance between blemish removal, detail preservation, and identity consistency.

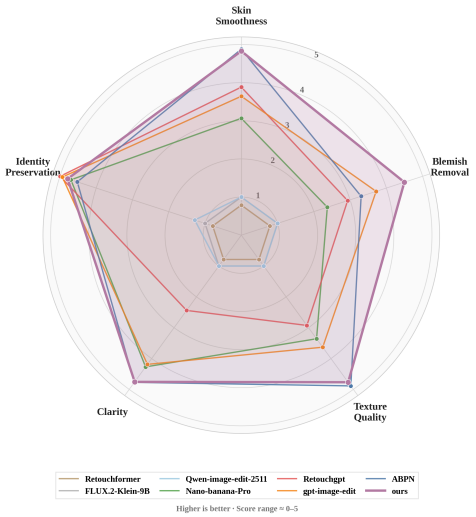

Figure 5: User study dimensional comparison across five metrics.

4.4 Ablation Study

Table 2 presents the ablation experiment results for each module, and Figure 3 shows the corresponding qualitative results. Compared to the Base model, the model with +LoRA demonstrates significant improvements across all metrics, validating the effectiveness of parameter-efficient fine-tuning.Upon further introducing patch-based anchor sampling, the FID decreased from 5.07 to 4.04, indicating that concentrating supervision on prominent high-frequency regions can enhance distribution realism. Although some metrics, such as NIMA, showed a slight decline at this stage, we infer that this is due to patch sampling reallocating capacity toward local texture fidelity, thereby incurring a certain marginal cost in global perceptual scores. Given the significant improvement in FID and the qualitative enhancement of pore-level details shown in Figure 3, we consider this an acceptable trade-off. Finally, reward-guided alignment (the method proposed in this paper) further improves the perceptual and aesthetic metrics that are most sensitive to human perceived quality. Although objective fidelity metrics such as PSNR and SSIM showed slight declines, reference-free perceptual and aesthetic metrics,including NIMA, NRQM, and TOPIQ,which are all improved, with MANIQA reaching its optimal value. This indicates that the reward-guided model focuses on high-dimensional perceptual semantics and achieves an effective trade-off between pixel fidelity and perceptual quality, a finding corroborated by the user study in Section 4.3.

Table 2: Ablation study on the FaceKeeper-Bench. Best / second-best are in red / blue.

Method	PSNR↑	SSIM↑	LPIPS↓	NIMA↑	MANIQA↑	NRQM↑	TOPIQ↑	FID↓
FLUX.2-klein-base-9B (base)	21.0989	0.6321	0.2374	5.1037	0.3753	6.1427	0.4987	24.5638
+LoRA	29.5525	0.8623	0.0512	5.2581	0.4528	6.9622	0.6110	5.0745
+patch	30.0504	0.8650	0.0540	5.2439	0.4533	6.9553	0.6056	4.0408
+reward(ours)	29.9078	0.8640	0.0545	5.2489	0.4546	6.9569	0.6090	4.8054

5 Conclusion

This paper proposes FaceKeeper, a multi-expert LoRA framework for high-fidelity facial retouching that balances blemish removal, texture preservation, and identity consistency. To address excessive smoothing caused by low-frequency bias in supervision and conflicts arising from multi-intent adaptation, FaceKeeper introduces three improvements: (i) spatio-frequency-coupled detail block sampling to reconstruct supervision in salient facial regions.(ii) a freeze-dense-sparse multi-expert LoRA architecture, merged into a single adapter via mask-aware fusion.(iii) Multi-expert reward learning, bridging pixel-level restoration and semantic texture preferences. On our FaceKeeper-Bench and CelebA-HQ, its no-reference metrics outperform dedicated retouching models and general-purpose editors, while its full-reference metrics remain competitive; qualitative analysis and five-dimensional user studies confirm its ability to effectively remove blemishes without compromising pore texture or facial identity.

REFERENCES

Xuehai Bai, Xiaoling Gu, Akide Liu, Hangjie Yuan, YiFan Zhang, and Jack Ma. Mcie: Multimodal llm-driven complex instruction image editing with spatial guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2399–2407, 2026.

Black Forest Labs. FLUX.2: Analyzing and enhancing the latent space of FLUX – representation comparison, 2025. URL https://bfl.ai/research/representation-comparison.

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, 2023.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, 2025. URL https://arxiv.org/abs/2507.06261.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

Zheng-Peng Duan, Jiawei Zhang, Zheng Lin, Xin Jin, XunDong Wang, Dongqing Zou, Chun-Le Guo, and Chongyi Li. Diffretouch: Using diffusion to retouch on the shoulder of experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2825–2833, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017.

Xiaojie Gu, Ziying Huang, Jia-Chen Gu, and Kai Zhang. Ultraedit: Training-, subject-, and memory-free lifelong editing in language models. *arXiv preprint arXiv:2505.14679*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5967–5976. Ieee, 2017.

Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13899–13909. IEEE, 2021.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Soyoung Lee, Hyoungseo Cho, Myungjoo Kang, and YoungJoon Yoo. K-stylelora: Information-guided image generation via selective feature learning. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4524–4531. IEEE, 2025.

Biwen Lei, Xiefan Guo, Hongyu Yang, Miaomiao Cui, Xuansong Xie, and Di Huang. Abpn: Adaptive blend pyramid network for real-time local retouching of ultra high-resolution photo. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2098–2107. IEEE, 2022.

Mengtian Li, Jinshu Chen, Wanquan Feng, Bingchuan Li, Fei Dai, Songtao Zhao, and Qian He. Hyperlora: Parameter-efficient adaptive generation for portrait synthesis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13114–13123. IEEE, 2025.

Quanhao Li, Junqiu Yu, Kaixun Jiang, Yujie Wei, Zhen Xing, Pandeng Li, Ruihang Chu, Shiwei Zhang, Yu Liu, and Zuxuan Wu. Diffusionopd: A unified perspective of on-policy distillation in diffusion models. *arXiv preprint arXiv:2605.15055*, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Yaoli Liu, Yao-Xiang Ding, and Kun Zhou. Freefuse: Multi-subject lora fusion via adaptive token-level routing at test time. *arXiv preprint arXiv:2510.23515*, 2025.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11451–11461. IEEE, 2022.

Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

OpenAI. Introducing gpt-image-2 — available today in the api and codex. OpenAI Developer Community, April 2026. URL https://community.openai.com/t/1379479. Announcement, released April 21, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Yuqi Peng, Lingtao Zheng, Yufeng Yang, Yi Huang, Mingfu Yan, Jianzhuang Liu, and Shifeng Chen. Tara: Token-aware lora for composable personalization in diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8385–8393, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjrn Ommer. High-resolution image synthesis with latent diffusion models. 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022.

Enis Simsar, Thomas Hofmann, Federico Tombari, and Pinar Yanardag. Loraclr: Contrastive adaptation for customization of diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13189–13198. IEEE, 2025.

Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 8798–8807. Ieee, 2018a.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *European conference on computer vision*, pp. 63–79. Springer, 2018b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Xue Wen, Lianxin Xie, Le Jiang, Tianyi Chen, Si Wu, Cheng Liu, and Hau San Wong. Retouchformer: Semi-supervised high-quality face retouching transformer with prior-based selective self-attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5903–5911, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Shengming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report. Technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Fangtai Wu, Hailong Guo, Shijie Huang, Jiayi Song, Yubo Huang, Mushui Liu, Zhao Wang, Yunlong Yu, Jiaming Liu, and Ruihua Huang. Collectionlora: Collecting 50 effects in 1 lora via multi-teacher on-policy distillation. *arXiv preprint arXiv:2605.25378*, 2026a.

Qiucheng Wu, Jing Shi, Simon Jenni, Kushal Kafle, Tianyu Wang, Shiyu Chang, and Handong Zhao. Retouchiq: Mllm agents for instruction-based image retouching with generalist reward. *arXiv preprint arXiv:2602.17558*, 2026b.

Jinqi Xiao, Qing Yan, Liming Jiang, Zichuan Liu, Hao Kang, Shen Sang, Tiancheng Zhi, Jing Liu, Cheng Yang, Xin Lu, et al. Instructmole: Instruction-guided mixture of low-rank experts for multi-conditional image generation. *arXiv preprint arXiv:2512.21788*, 2025.

Lianxin Xie, Wen Xue, Zhen Xu, Si Wu, Zhiwen Yu, and Hau San Wong. Blemish-aware and progressive face retouching with limited paired data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5599–5608, 2023.

Zhidan Xu, Xiaoqin Zhang, and Shijian Lu. Face retouching with diffusion data generation and spectral restorement. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14722–14731. IEEE, 2025.

Wen Xue, Chun Ding, Ruotao Xu, Si Wu, Yong Xu, and Hau-San Wong. Retouchgpt: Llm-based interactive high-fidelity face retouching via imperfection prompting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9059–9067, 2025.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in neural information processing systems*, 36:7093–7115, 2023.

Jiachen Yang, Xianhui Lin, Yi Dong, Zebiao Zheng, Xing Liu, Hong Gu, and Yanmei Fang. Beautygrpo: Aesthetic alignment for face retouching via dynamic path guidance and fine-grained preference modeling. *arXiv preprint arXiv:2603.01163*, 2026.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1190–1199. IEEE, 2022.

Moon Ye-Bin, Roy Miles, Tae-Hyun Oh, Ismail Elezi, and Jiankang Deng. Retouchllm: Training-free code-based image retouching with vision language models. *arXiv preprint arXiv:2510.08054*, 2025.

Ying Zeng, Miaosen Luo, Guangyuan Li, Yang Yang, Ruiyang Fan, Linxiao Shi, Qirui Yang, Jian Zhang, Chengcheng Liu, Siming Zheng, et al. Smartphotocrafter: Unified reasoning, generation and optimization for automatic photographic image editing. *arXiv preprint arXiv:2604.19587*, 2026.

Juzheng Zhang, Jiacheng You, Ashwinee Panda, and Tom Goldstein. Lori: Reducing cross-task interference in multi-task low-rank adaptation. *arXiv preprint arXiv:2504.07448*, 2025.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in neural information processing systems*, 36:31428–31449, 2023a.

Longteng Zhang, Lin Zhang, Shaohuai Shi, Xiaowen Chu, and Bo Li. Lora-fa: Memory-efficient low-rank adaptation for large language models fine-tuning. *arXiv e-prints*, pp. arXiv–2308, 2023b.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022.

Jun Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *IEEE*, 2017.

A Dataset Building

A.1 Details on Test Set Building

To ensure a fair and comprehensive evaluation, we have constructed a composite test benchmark consisting of our in-house high-definition portrait dataset, FaceKeeper-Bench, and a curated subset of CelebA-HQ. Both datasets have undergone automated screening, rule-based filtering, and manual review to ensure that the samples are suitable for evaluation in terms of clarity, face scale, skin texture discernibility, and identity diversity.

FaceKeeper-Bench. After an initial collection of multi-scenario portrait images, we combined face detection with texture and color quality screening to remove samples that were too blurry, too flat, had excessive highlights, or exhibited abnormal colors. We also limited the number of people per image and ensured that the face region was large enough for editing. Subsequently, a manual review was conducted to prioritize samples that were clear and displayed visible skin imperfections, while removing those with severe occlusions or where identity was difficult to discern; clusters were then formed based on similarity to reduce duplication of the same individuals. The final collection contains approximately 2,000 portraits, primarily in 8K HD resolution, with a balanced gender ratio and coverage of Caucasian, Asian, and African American individuals. Scenes include family gatherings, weddings, graduation ceremonies, and everyday life, featuring both single-person and group compositions.

CelebA-HQ subset. Under the same core quality standards, over 2,300 images were selected from CelebA-HQ to supplement facial diversity and identity coverage, while ensuring consistency in evaluation criteria between the private and public datasets.

A.2 Training Data

To support controllable retouching for multiple skin types, we constructed an imperative training dataset aligned with three experts, corresponding to water skin, matte skin, and textured skin, respectively. Each sample is a triplet $\langle \text{source_img}, \text{gt_img}, \text{text} \rangle$, consisting of the original image to be edited, the target retouched result, and the editing instructions. The three categories contain 18,800, 17,000, and 17,000 sample pairs, respectively. Each expert was trained exclusively on data within their respective category. The instructions are categorized as follows: (1) Create a high-end water complexion that enhances hydration while preserving visible pores and natural skin texture, with the subject's identity, pose, hairstyle, and background unchanged. (2) Create a refined matte makeup effect where high-definition skin texture and fine pores remain visible under the matte finish, while maintaining identity and background. (3) Natural beautification: Remove acne, blemishes, and fine lines while enhancing the realistic texture of pores; make the skin smoother without losing biological texture, and slightly slim the face while maintaining identity. Matching data comes from FFHQ-derived faces, self-built and commercial before-and-after retouched portraits, as well as curated online portrait editing data. To support portrait retouching and controllable editing across multiple skin texture categories, we constructed an instruction-based training dataset organized by skin texture category. Each sample is formalized as a triplet $\langle \text{source_img}, \text{gt_img}, \text{text} \rangle$, where source_img is the original image to be edited, gt_img is the corresponding target image, and text specifies the editing instruction.

B Evaluation

B.1 Evaluation Metrics

We adopted a set of multidimensional evaluation metrics. For perceived naturalness, we used NIQE (↓). For aesthetic quality, we used NIMA (↑). For perceived and text quality, we used NRQM (↑), MANIQA (↑), and TOPIQ (↑). For diagnostic metrics relative to the reference image, we also report PSNR (↑), SSIM (↑), LPIPS (↓), and FID (↓), but do not consider them primary evaluation criteria.

Full-reference metrics are not suitable as primary evaluation criteria for face retouching. Given the perception-distortion trade-off, optimizing PSNR/SSIM often leads to excessive skin smoothing and replicates imperfections present in the reference images. Our method may deviate from these reference images to enhance naturalness and aesthetic appeal; therefore, these metrics are used only as supplementary evaluation tools.

B.2 Quantitative Evaluation on celebA-HQ

We further evaluate the generalization ability of our method on the CelebA-HQ dataset, as shown in Table 3. Compared to the results on our self-built dataset, we observe distinct discrepancies, which can be attributed to the inherent characteristics of CelebA-HQ and the generation mechanisms of different methods.

Hallucinated Details and NR-IQA Bias. Images in CelebA-HQ are inherently clean, aligned, and contain minimal degradations. Under such conditions, generative methods (Nano-Banana-Pro, GPT-Image-Edit2,

FLUX.2-klein-base-9B) tend to hallucinate high-frequency pseudo-details to enhance the visual appearance. Notably, no-reference image quality assessment (NR-IQA) metrics such as MANIQA, NRQM, and TOPIQ are highly sensitive to high-frequency information. Consequently, they are easily deceived by these hallucinated textures, yielding artificially inflated scores for generative models.

Superior Fidelity and Perception of Our Method. Unlike generative models that prioritize texture synthesis at the cost of structural distortion, our method focuses on faithful restoration. Despite facing the aforementioned metric bias, our method achieves the best NIMA and the second-best SSIM and LPIPS This demonstrates that our method successfully maintains structural integrity and identity consistency while delivering highly competitive perceptual quality, avoiding the notorious identity shift and structural hallucinations common in generative approaches.

Table 3: Quantitative comparison on the CelebA-HQ dataset. Best / second-best are in red / blue.

Method	PSNR↑	SSIM↑	LPIPS↓	NIMA↑	MANIQA↑	NRQM↑	TOPIQ↑	FID↓
ABPN Lei et al. (2022)	26.837	0.793	0.035	5.281	0.402	6.513	0.499	2.052
RetouchFormer Wen et al. (2024)	29.527	0.831	0.103	5.001	0.346	4.394	0.413	3.488
RetouchGPT Xue et al. (2025)	29.506	0.831	0.108	5.029	0.348	4.290	0.412	3.029
Qwen-image-edit-2511 Wu et al. (2025)	24.521	0.727	0.164	5.426	0.417	6.554	0.585	15.893
Nano-banana-Pro Comanici et al. (2025),	25.633	0.712	0.131	5.103	0.387	7.239	0.527	14.666
Gpt-Image-Edit2 OpenAI (2026)	25.511	0.701	0.158	5.268	0.452	7.540	0.619	95.428
FLUX.2-klein-base-9B Black Forest Labs (2025)	20.889	0.592	0.247	5.255	0.467	7.338	0.675	47.775
Ours	27.323	0.806	0.054	5.500	0.391	6.503	0.502	5.004

B.3 Qualitative Evaluation on CelebA-HQ

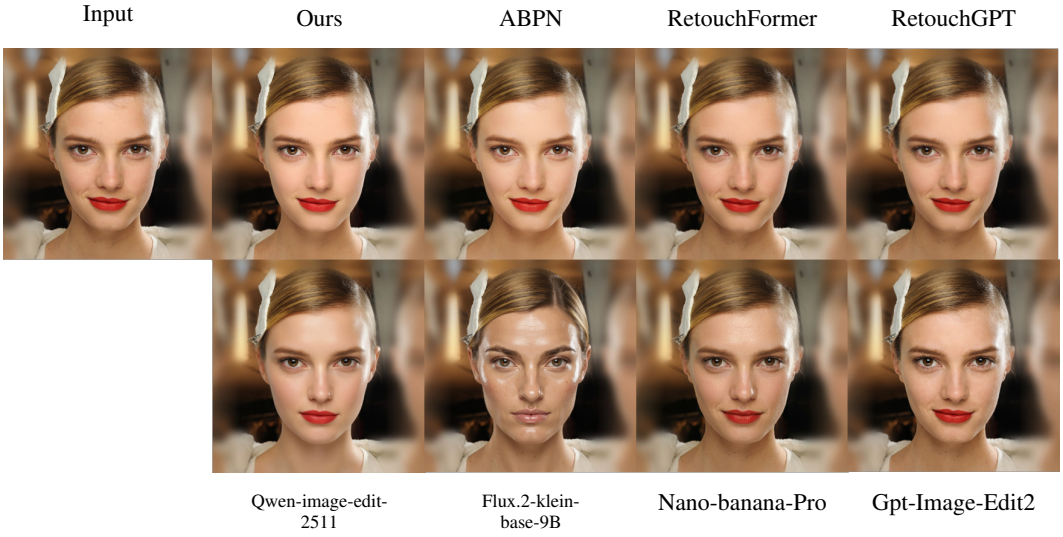

Figure 6: Qualitative Evaluation on CelebA-HQ

Figure 6 shows a qualitative comparison on CelebA-HQ. Unlike FaceKeeper-Bench, the CelebA-HQ input is inherently cleaner and better aligned; the main challenge lies in avoiding excessive editing while preserving identity and fine textures. Dedicated retouching models (ABPN, RetouchFormer, RetouchGPT) often leave residual artifacts, whereas general-purpose editing (Qwen-Image-Edit-2511, FLUX.2-klein-base-9B, Nano-Banana-Pro, GPT-Image-Edi2) tend to over-enhance images, resulting in plastic-looking skin, overly intense highlights, or slight identity shifts. In contrast, our method comes closer to GT.It gently cleanses the skin while preserving pore-level texture and facial identity, demonstrating better generalization on public benchmarks.

B.4 More Qualitative Results

In this section, we present additional qualitative results demonstrating the image enhancement effects of our method on a variety of images, as shown in Figure 7.

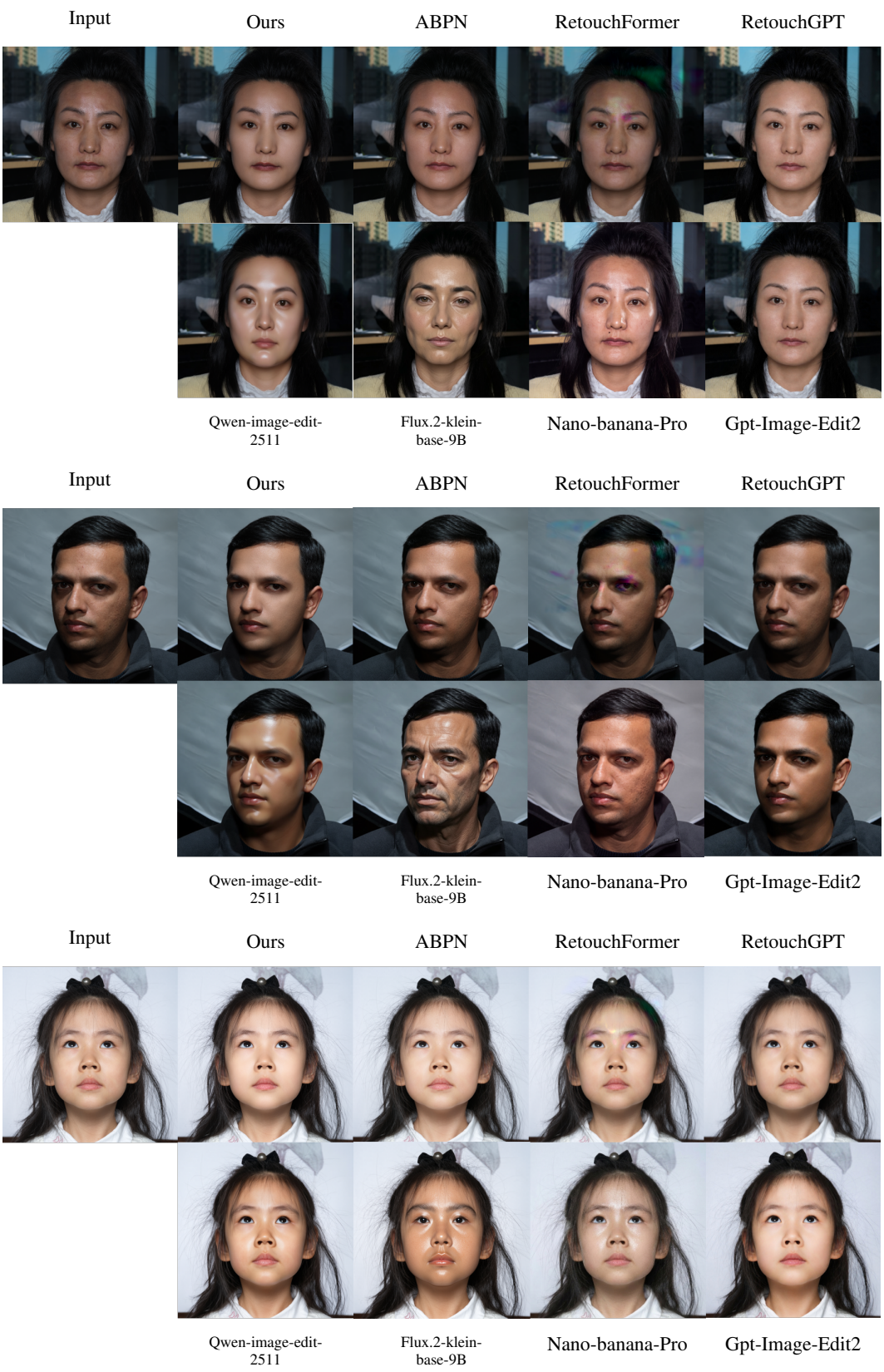

Figure 7: Extended qualitative results: FaceKeeper beauty enhancement on diverse images.

B.5 Quantitative and Qualitative Results for Water and Matte Skin

Table 4: Qualitative comparison of Matte and Water on the CelebA-HQ dataset.

Method / Dataset	PSNR↑	SSIM↑	LPIPS↓	NIMA↑	MANIQA↑	NRQM↑	TOPIQ↑	FID↓
Ours (Matte)	34.832	0.951	0.041	5.722	0.335	4.683	0.428	9.468
Ours (Wate)	28.600	0.880	0.068	5.857	0.349	4.785	0.452	12.476



Figure 8: Qualitative results of matte skin



Figure 9: Qualitative results of water skin

As shown in Table 4, we observed a clear trade-off between fidelity and perceived quality. The matte skin effect performed exceptionally well in maintaining image fidelity, making it well-suited for applications requiring high structural consistency. On the other hand, the water skin effect prioritized aesthetic enhancement and achieved the best performance on reference-free perceived quality metrics, demonstrating that it can generate more visually appealing skin sheen even with greater pixel deviation.

Figures 8 and 9 show the qualitative results for the matte and water skin finishes, respectively. They illustrate the matte and water luster, respectively.

C Detailed Description of Mask-Aware Merging

This appendix provides the complete definition of spherical linear interpolation (SLERP) for multi-expert fusion, including the choice of vector granularity, degeneration handling, and the multi-expert tree combination strategy.

C.1 Fusion Granularity and Weight Normalization

Element-wise SLERP admits only angles in $\{0, \pi\}$ and thus lacks geometric degrees of freedom. We therefore lift the fusion granularity to output-channel row vectors. For $u \in \{1, \dots, d\}$, the contribution of expert i is

$$\mathbf{b}_i(u) = B_i(u, :) \in \mathbb{R}^r. \tag{13}$$

The row-level activation indicator is obtained from the element-wise mask $M_i(u, v)$ with threshold ρ :

$$M_i^{\text{row}}(u) = \mathbf{1}\left[\frac{1}{r}\sum_{v=1}^{r} M_i(u, v) \geq \rho\right]. \tag{14}$$

Let the active-expert set be $\mathcal{J}(u) = \{j \mid M_j^{\mathrm{row}}(u) = 1\}$ with $n(u) = |\mathcal{J}(u)| \geq 2$. Fusion weights are normalized by the relative row-vector magnitudes:

$$w_j(u) = \frac{\|\mathbf{b}_j(u)\|_2}{\sum_{k\in\mathcal{J}(u)} \|\mathbf{b}_k(u)\|_2}, \qquad \sum_{j\in\mathcal{J}(u)} w_j(u) = 1. \tag{15}$$

C.2 Two-Vector SLERP and Degeneration Handling

For nonzero vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^r$ and weight $t \in [0, 1]$, the standard SLERP is

$$\mathrm{SLERP}(\mathbf{x}, \mathbf{y}; t) = \frac{\sin((1-t)\theta)}{\sin\theta}\,\mathbf{x} + \frac{\sin(t\theta)}{\sin\theta}\,\mathbf{y}, \qquad \theta = \arccos\left(\frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|}\right). \tag{16}$$

Degeneration rules.

- **Zero vector:** if one argument is $\mathbf{0}$, return the other; if both are $\mathbf{0}$, return $\mathbf{0}$.
- **Nearly parallel** ($\theta < \epsilon_0$): degenerate to linear interpolation $(1-t)\mathbf{x} + t\mathbf{y}$.
- **Nearly antipodal** ($\theta > \pi - \epsilon_\pi$): the geodesic is non-unique; keep the direction of the larger-weight side ($t < 1/2$ keeps $\mathbf{x}$, otherwise $\mathbf{y}$) and apply linear interpolation only to the norms.

Combining the above yields the complete piecewise operator

$$\mathrm{SLERP}^*(\mathbf{x}, \mathbf{y}; t) = \begin{cases} \mathbf{y} \text{ or } \mathbf{x} \text{ or } \mathbf{0}, & \mathbf{x} = \mathbf{0} \text{ or } \mathbf{y} = \mathbf{0}, \\ (1-t)\mathbf{x} + t\mathbf{y}, & \theta < \epsilon_0, \\ \dfrac{\sin((1-t)\theta)}{\sin\theta}\,\mathbf{x} + \dfrac{\sin(t\theta)}{\sin\theta}\,\mathbf{y}, & \epsilon_0 \leq \theta \leq \pi - \epsilon_\pi, \\ \big[(1-t)\|\mathbf{x}\| + t\|\mathbf{y}\|\big]\,\hat{\mathbf{d}}, & \theta > \pi - \epsilon_\pi. \end{cases} \tag{17}$$

C.3 Multi-Expert SLERP-Tree Combination and Mask Integration

When $n(u) \geq 3$, we use a weighted pairwise tree combination:

1. Sort $\mathcal{J}(u)$ by $w_j(u)$ in descending order as $e_1, \ldots, e_n$.
2. Initialize $\mathbf{r} \leftarrow \mathbf{b}_{e_1}(u)$ and $w_{\mathrm{acc}} \leftarrow w_{e_1}(u)$.
3. For $i = 2, \ldots, n$, set $t = \dfrac{w_{e_i}(u)}{w_{\mathrm{acc}} + w_{e_i}(u)}$, update $\mathbf{r} \leftarrow \mathrm{SLERP}^*(\mathbf{r}, \mathbf{b}_{e_i}(u); t)$, and accumulate $w_{\mathrm{acc}} \leftarrow w_{\mathrm{acc}} + w_{e_i}(u)$.
4. Output $\mathbf{b}^*(u) = \mathbf{r}$.

D Parameter Settings

Table 5 shows the hyperparameter settings shared by the three expert models.We build upon FLUX.2-klein-base-9B and train three skin-appearance experts: water-glow, matte, and texture. All experts share a randomly initialized LoRA down-projection A, while each expert-specific B_i is optimized. We use LoRA with rank 64 and scaling factor 64, inserting LoRA layers into the query, key, value, output, and MLP projections of the DiT backbone. Each expert is trained in two stages. In the dense stage, we initialize B_i to zero and optimize all of its entries. We then derive a binary support mask from the resulting dense weights. In the sparse stage, B_i is initialized from a selected checkpoint; entries outside the support set are set to zero, and their gradients remain masked throughout optimization. Both stages use AdamW with a constant learning rate of 5×10^{-5}, weight decay 0.01, bfloat16 precision, a per-GPU batch size of 1, and one gradient-accumulation step. Gradient checkpointing is enabled. Checkpoints and validation results are generated every 500 optimization steps.

Table 5: Shared hyperparameters used for all three experts.

Hyperparameter	Value
Backbone	FLUX.2-klein-base-9B
Trainable component	DiT LoRA up-projection B_i
Shared component	Random frozen LoRA down-projection A
LoRA rank / scaling factor	64 / 64
Initialization	A: seed 42; B_i: zero in LoRA-D
Optimizer	AdamW
Learning-rate schedule	Constant
Learning rate	5×10^{-5}
Weight decay	0.01
Precision	bfloat16
Batch size	1 per GPU
Gradient accumulation	1
Checkpoint / validation interval	500 steps

E Detail of Space-Frequency Coupled Sampling

This appendix provides theoretical support and implementation details for the anchor sampling strategy in Section 3.1. Section E.1 proves the submodularity of the coverage utility $F(A)$ and establishes the theoretical guarantee of the greedy algorithm; Section E.2 details the construction of the geometry-aware spatial embedding (multi-band positional encoding) and the semantic kernel.

E.1 Submodularity and Optimization Guarantee

Proposition. Let $w(u, c) = s(u; \mathcal{I}) \cdot \kappa(u, c) \geq 0$ and

$$F(A) = \sum_{u \in B} \max_{c \in A} w(u, c).$$

Then F is a monotone nondecreasing submodular set function on B.

Proof. For each $u \in B$, define $f_u(A) = \max_{c \in A} w(u, c)$ and $f_u(\emptyset) = 0$. Monotonicity of f_u is immediate. To show submodularity, take $A \subseteq D \subseteq B \setminus \{e\}$ and write $m_A = \max_{c \in A} w(u, c)$, $m_D = \max_{c \in D} w(u, c)$. Then $m_A \leq m_D$, and the marginal gains are

$$f_u(A \cup \{e\}) - f_u(A) = \max\big(0,\, w(u, e) - m_A\big),$$

$$f_u(D \cup \{e\}) - f_u(D) = \max\big(0,\, w(u, e) - m_D\big).$$

Since $t \mapsto \max(0, w(u, e) - t)$ is nonincreasing in t, the marginal gain on A is at least that on D. Thus each f_u is submodular. As $F = \sum_u f_u$ is a sum of nonnegative submodular functions, F is submodular.

By Nemhauser, Wolsey, and Fisher (Nemhauser et al., 1978), the greedy solution A_g satisfies $F(A_g) \geq (1 - 1/e)F(A^\star)$. A naive greedy update costs $O(KM^2)$ and can be accelerated by lazy evaluation. This proof guarantees that selecting anchors via the greedy algorithm yields a $(1 - 1/e)$-approximate optimal solution, providing a solid theoretical foundation for the discrete set selection in the main text.

E.2 Geometric Perception, Spatial Embedding, and Semantic Kernels

Geometric spatial embeddings are used for patch anchor selection, with the aim of incorporating facial features into the optimization of overlapping submodels.

First, the normalized coordinates must be encoded using multi-band position encoding, as follows:

$$\gamma(u) = [\sin(2^0 \pi u), \cos(2^0 \pi u), \ldots, \sin(2^{L-1} \pi u), \cos(2^{L-1} \pi u)]. \tag{18}$$

Here, L represents the number of frequency bands. Multiband encoding is used to enhance the MLP's ability to represent fine-grained spatial variations. Feeding $\gamma(u)$ into a lightweight MLP f_θ yields the spatial embedding

$$e_u = f_\theta(\gamma(u)). \tag{19}$$

The semantic affinity between the anchor c and a candidate u is defined as:

$$\kappa(u, c) = \exp\left(-\frac{\|e_u - e_c\|^2}{2\sigma^2}\right). \tag{20}$$

σ controls the kernel width. When σ is too large, the kernel becomes nearly constant and anchor selection degenerates into sorting by significance alone; when σ is too small, the kernel becomes nearly one-hot and soft deduplication fails.

F LIMITATION

Although our dense-to-sparse adaptation retains salient parameters, consolidates expert-exclusive supports, and applies mask-aware merging to reduce interference in overlapping regions, increasing the number of LoRA experts may still enlarge support overlap and merge conflicts, raising the risk of quality degradation. Our study is validated under a limited expert set; scaling to more intents with stronger conflict control remains future work.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we did not use generative AI tools for any tasks. All text, claims, and artifacts in this paper were produced solely by the authors. We take full responsibility for the final content of this work.