PerTouch: VLM-Driven Agent for Personalized and Semantic Image Retouching

Zewei Chang[1], **Zheng-Peng Duan**[1], **Jianxing Zhang**[2], **Chun-Le Guo**[1,4], **Siyu Liu**[1],
Hyungju Chun[3], **Hyunhee Park**[3], **Zikun Liu**[2], **Chongyi Li**[1,4*]
[1]VCIP, CS, Nankai University
[2]Samsung R&D Institute China - Beijing (SRC-B)
[3]The Department of Camera Innovation Group, Samsung Electronics
[4]NKIARI, Shenzhen Futian
changzewei@mail.nankai.edu.cn, adamduan0211@gmail.com,
{guochunle, liusiyu29, lichongyi}@nankai.edu.cn, {jx2018.zhang, hyungju.chun, inextg.park, zikun.liu}@samsung.com

Abstract

Image retouching aims to enhance visual quality while aligning with users' personalized aesthetic preferences. To address the challenge of balancing controllability and subjectivity, we propose a unified diffusion-based image retouching framework called **PerTouch**. Our method supports semantic-level image retouching while maintaining global aesthetics. Using parameter maps containing attribute values in specific semantic regions as input, PerTouch constructs an explicit parameter-to-image mapping for fine-grained image retouching. To improve semantic boundary perception, we introduce semantic replacement and parameter perturbation mechanisms during training. To connect natural language instructions with visual control, we develop a VLM-driven agent to handle both strong and weak user instructions. Equipped with mechanisms of feedback-driven rethinking and scene-aware memory, PerTouch better aligns with user intent and captures long-term preferences. Extensive experiments demonstrate each component's effectiveness and the superior performance of PerTouch in personalized image retouching.

Code — https://github.com/Auroral703/PerTouch

1 Introduction

With the increasing accessibility of photography devices and the lowering threshold for photo-taking, capturing images has become an essential medium for personal expression. However, due to the lack of professional photography knowledge and the uncontrollable shooting environment, raw photos often fail to achieve satisfactory visual quality. To bridge this gap, image post-processing has become a crucial technique for enhancing photo quality and improving visual expressiveness. While professional software such as Adobe Lightroom (Adobe Inc. 2024a) and Photoshop (Adobe Inc. 2024b) provides powerful tools for image retouching, these systems typically require expert knowledge and involve complex workflows, making them less accessible to ordinary users, especially for batch processing or personalized style editing at scale.

Therefore, a number of deep learning-based image retouching methods have been proposed. Yet, existing approaches still face several limitations. The limitations of current approaches mainly fall into three categories. (1) **Lack of subjectivity modeling**: Most methods adopt deterministic architectures that generate a single fixed result for a given input, failing to account for the diversity and subjectivity of user preferences. (2) **Lack of region-level control**: While some works introduce controllable parameters and reference images to control the image retouching style (Duan et al. 2025; Ouyang et al. 2023), they often fail to support flexible local editing. Attempts that incorporate external segmentation maps are sensitive to segmentation quality and tend to produce visually unnatural results. (3) **Lack of user interaction modeling and personalization**: Existing methods typically cannot interpret vague user instructions and ignore the need to memorize long-term editing preferences, resulting in poor adaptability and high user burden, particularly in batch or repeated editing scenarios.

To address these challenges, we propose **PerTouch**, a unified framework for fine-grained and personalized image retouching. Our method leverages the powerful diffusion prior to learn a diverse and high-quality retouching distribution, enabling the generation of globally aesthetic yet regionally consistent images conditioned on user intent. To enable semantic-aware regional editing, we propose a novel data preprocessing strategy that incorporates semantic replacement and parameter perturbation during training, helping the model better perceive semantic boundaries and mitigate overfitting to segmentation information existing in inputs. Furthermore, we design an agent driven by a vision language model (VLM) to lower the barrier for user interaction. The agent supports both strong and weak natural language prompts and can interpret vague instructions by inferring parameters in context. In addition, we introduce a scene memory mechanism to record the user's editing preferences under different semantic scenarios, enabling personalized and context-aware retouching over long-term usage.

In summary, our contributions are as follows:

- We propose a semantic-aware region adjustment strategy based on diffusion priors, enabling globally aesthetic and locally consistent image retouching.
- We design a data preprocessing scheme combining

*Corresponding author.

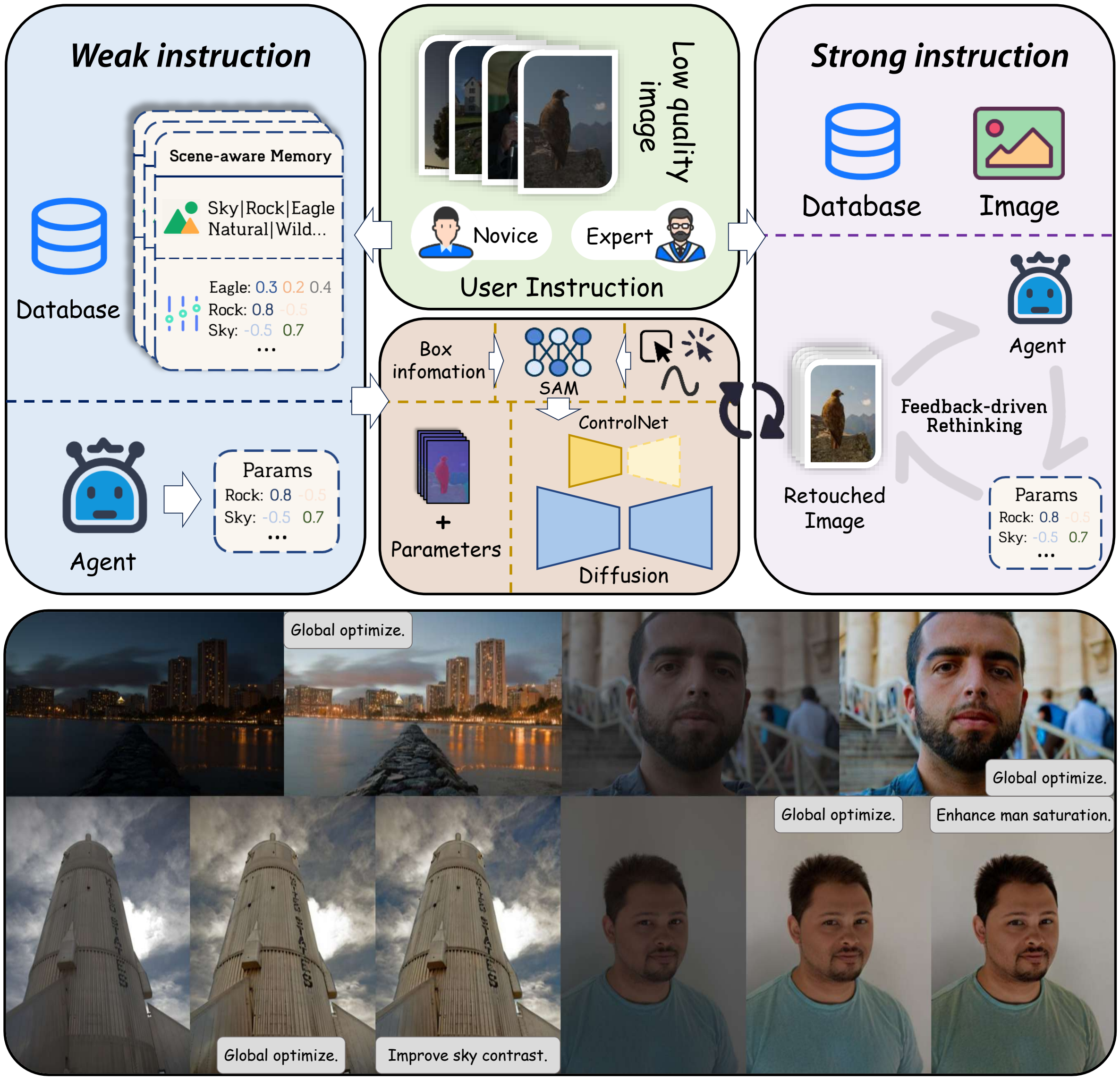

Figure 1: Overview of our **PerTouch** pipeline. Our method supports region-level personalized retouching with long-term user memory. Given images and natural language instruction, PerTouch determines the strength of the instruction, and then leverages the scene-aware memory to adaptively perform corresponding retouching operations based on the user's historical preferences. The final result is retouched by a fine-tuned diffusion model, ensuring globally pleasing and finely controlled region-level edits. The examples at the bottom demonstrate the system's ability to perform both global retouching and fine-grained regional adjustment across various instruction types.

semantic replacement and parameter perturbation to improve semantic boundary perception and parameter learning.

- We develop a VLM-driven agent with a scene memory mechanism to model long-term user preferences as well as enable personalized and context-aware retouching.

2 Related Work

2.1 Image Retouching in Deep Learning

Recent advances in deep learning, alongside the availability of high-quality datasets (Bychkovsky et al. 2011; Liang et al. 2021), have driven significant progress in automated image retouching. Early approaches predominantly adopt Fully Convolutional Networks (FCNs) for end-to-end image-to-image translation (Chen et al. 2018a,b; He et al. 2020; Kim et al. 2021; Sun et al. 2021), while others incorporate photographic priors such as Retinex theory (Liu et al. 2021; Zhu et al. 2020; Wang et al. 2019), 3D-LUTs (Zeng et al. 2020; Yang et al. 2022; Wang et al. 2021), or curve and grid-based operations (Moran, McDonagh, and Slabaugh 2021; Gharbi et al. 2017; Moran et al. 2020; Song, Qian, and Du 2021) to enhance interpretability and controllability.

To address aesthetic diversity, style transfer methods (Kim, Koh, and Kim 2020; Kim and Lee 2024; Song, Qian, and Du 2021) enable multi-style outputs but often rely on reference exemplars, which increases user burden. More recently, diffusion-based methods such as DiffRetouch (Duan et al. 2025) have shown promise in modeling the complex distribution of expert-retouched styles via interpretable attribute control. However, most existing methods lack regional controllability or overly depend on external masks (Ouyang et al. 2023), which may lead to unnatural artifacts. To this end, we propose PerTouch which enables semantic-aware regional retouching while preserving global aesthetic quality. Our method introduces explicit region-to-parameter mapping and supports fine-grained control and user interaction, addressing the limitations of both deterministic and reference-driven approaches.

2.2 Agent in Low-level Vision

With the rise of vision language models, researchers have begun leveraging their strong visual priors to perceive and invoke external tools, driving substantial progress in agent-based systems for low-level vision. Recent studies (Li et al. 2025; Chen et al. 2024; Zhu et al. 2025; Jiang et al. 2025) employ agents to tackle various degradation tasks toward achieving all-in-one restoration capabilities. Parallel to this, a line of work (Chen et al. 2025; Lin et al. 2025; Dutt, Ceylan, and Mitra 2025) explores using vision language models as agents to interact with image retouching toolchains like Lightroom, enabling automated photo retouching through exposure, contrast, and tone curve adjustments guided by language instructions. However, current retouching systems typically rely on fixed tool invocation pipelines and lack adaptability to individual user preferences. To address this, we propose a scene memory mechanism that stores users' editing history, infers personalized preferences, and generates retouching results aligned with user intent, enabling truly personalized and preference-aware photo retouching.

3 Methodology

3.1 Overview

Given a low-quality input image X, the objective of image retouching is to generate a high-quality image that aligns with human aesthetic preferences while preserving original details. We propose PerTouch, a diffusion-based approach to address the underexplored challenges of semantic-aware, region-level adaptive retouching and personalized enhancement based on user preferences. An overview of our framework is illustrated in Figure 1. Similar to DiffRetouch (Duan et al. 2025), to assist users in retouching image styles that match their aesthetics, we provide four predefined image attributes (colorfulness, contrast, color temperature, and brightness) to facilitate intuitive user control. Our method is extensible: once a region-level score can be computed for a new attribute, our framework can incorporate it to enable controllability over that attribute. To fully leverage the diffusion prior, we adopt Stable Diffusion (Rombach et al. 2021) as the backbone and introduce ControlNet (Zhang, Rao, and Agrawala 2023) to inject region-level attribute information, enabling control over the generation process. Detailed model architecture is presented in Section 3.2.

To facilitate user interaction with the system and preserve editing preferences across users, we introduce a VLM-based agent. We categorize user instructions into strong instructions and weak instructions, and invoke the retouching algorithm process accordingly. To incorporate users' historical editing preferences, we store scene-aware memory for each editing session, which helps guide the agent's decision-making. Furthermore, we design a feedback-driven rethinking mechanism to help the agent understand the relationship between parameter changes and image variations, enabling multi-stage decision-making that leads to results more aligned with user intent. The agent design and workflow is detailed in Section 3.3.

3.2 Architecture

Data Preparation Our primary dataset is the MIT-Adobe FiveK dataset, which consists of 5,000 RAW images, each accompanied by five expert-retouched reference versions (A/B/C/D/E). Additional dataset details are presented in the Supplementary Material. To enable the model to learn the relationship between regional attribute values and visual changes, we provide the model with paired data of parameter maps and corresponding images for supervised training.

To obtain coarse semantic segmentation maps as auxiliary guidance, we leverage the panoptic segmentation capability of SAM. SAM automatically samples a set of evenly distributed points across the image and generates multiple segmented regions by treating each point as an independent prompt. A series of post-processing steps, including non-maximum suppression (NMS), is applied to obtain a high-confidence panoptic segmentation map. Once the segmentation map is obtained, we evaluate each semantic region using predefined regional scoring methods to assign attribute-specific scores. The segmentation and scoring information is then fused into a single parameter map by embedding the scores into the segmentation map and extending its channel dimensions to match the number of controllable attributes. This parameter map serves as guidance for attribute-aware image retouching. The detailed data preparation pipeline is marked in blue in Figure 2.

We observe that directly injecting the parameter maps into the network leads to over-reliance on the information it encodes, which is contrary to our objective. Rather than enforcing strict adherence to the parameter maps, we aim to use them as a soft hint, allowing the diffusion prior to play a central role in generating aesthetically pleasing results. To mitigate the model's over-reliance on the parameter map, we introduce two mechanisms.

First, we find that the model struggles to perceive region boundaries based on the injected parameter map, often defaulting to global retouching. This is because the input map contains only coarse semantic scores, which are not spatially continuous across regions, while real images often exhibit spatially coherent color transitions. This discrepancy makes it difficult for the model to correlate the parameter maps with the image structure. To address this, we propose a semantic replacement module. During training, a small subset of sam-

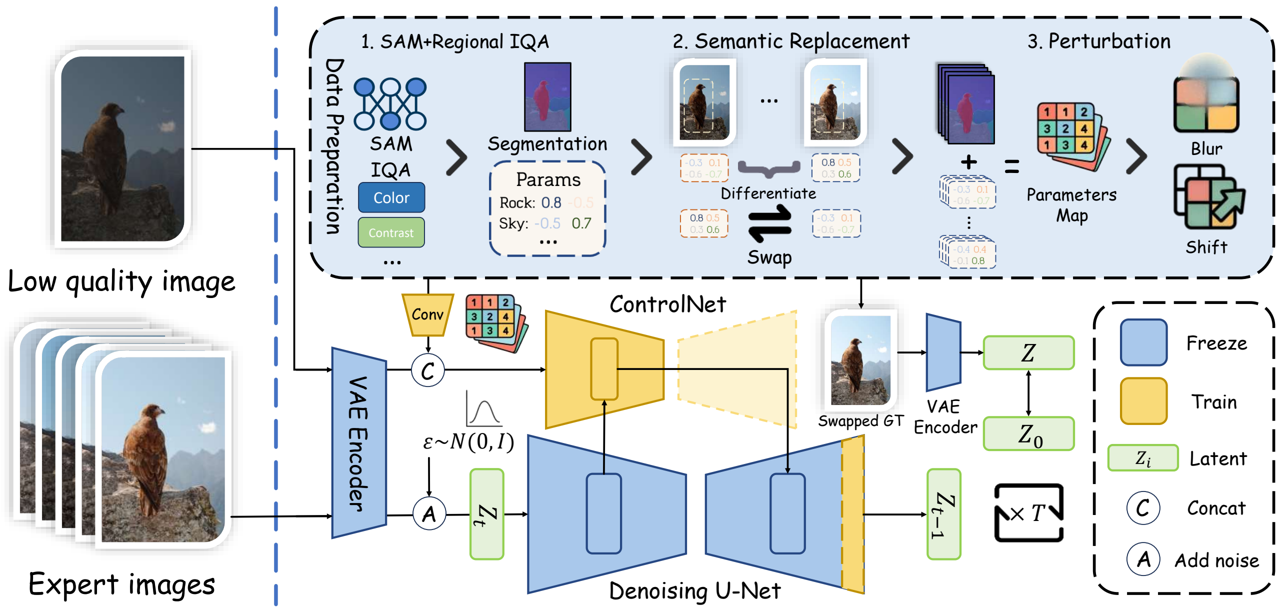

Figure 2: Dataset construction and training pipeline of PerTouch. To enable region-level controllable retouching, we construct training samples by generating parameter maps that transform low-quality input images into expert-retouched ground truth results. Specifically, we 1. extract semantic masks using SAM and estimate corresponding attribute parameters for each region; 2. introduce the Semantic Replacement Module to help the model perceive semantic regions by constructing diverse yet semantically consistent samples; and 3. apply the Perturbation Mechanism to prevent overfitting to segmentation boundaries and improve overall visual quality. The final parameter maps are injected into ControlNet alongside the original images, enabling the model to balance the global aesthetic consistency provided by diffusion priors and the regional guidance from parameter maps, thereby producing high-quality region-aware retouching outputs.

ples is selected, and a region is randomly chosen based on semantic area size as probability. The selected region is replaced with a region from another sample with the most divergent attributes in the parameter space. This artificial manipulation encourages the model to detect regional discrepancies and thus develop fine-grained retouching capabilities.

Second, although the semantic replacement module facilitates local retouching, the model tends to ignore global coherence, resulting in visually inconsistent and aesthetically unpleasing outputs. This suggests that the model is overly sensitive to the segmentation information contained in the parameter maps and not fully utilizing the strong generative capacity of the diffusion prior. To alleviate this, we introduce perturbations to the parameter maps along multiple dimensions, such as channel shifts and blurring, thereby enforcing the treatment of the segmentation as soft guidance rather than a rigid structure. This encourages the model to interpret and respond to semantic boundaries implicitly during the generation process. The effectiveness of both modules is further examined in Section 4.3.

Baseline Our baseline model builds upon Stable Diffusion, which extends denoising diffusion probabilistic models (DDPM) by operating in a learned latent space rather than directly in pixel space. The denoising model $\varepsilon_\theta(Z_t, t, m)$ is trained to reverse the noise process in the latent space, where Z_t denotes the noised latent at timestep t, and m represents conditioning signals.

To enable fine-grained control over regional image attributes, we integrate ControlNet into the Stable Diffusion framework. To accommodate multi-attribute conditioning, we expand the region-level attribute scores into a multi-channel guidance map $C = \{C^1, C^2, \ldots, C^K\}$ of the same spatial resolution as the segmentation map, where each channel C^k encodes the spatial distribution of a specific image attribute (e.g., colorfulness, contrast, color temperature, brightness). Each pixel value in C^k reflects the score of its corresponding region in that attribute. By adjusting the values of specific semantic regions within C, the model outputs corresponding retouching styles for the associated attributes, while simultaneously maintaining global image aesthetics. The coefficient values are adjusted within the range $[-1, 1]$, where each value corresponds to a visual style learned within the distribution of high-quality images. The specific parameter maps injection method and training details are given in the Supplementary Material.

3.3 Agent

Instruction Types and Agent Strategies To accommodate diverse user demands and editing intentions in personalized photo retouching, we design an interactive and preference-aware agent. Our agent supports two types of instruction parsing: weak instructions and strong instructions, aimed at simulating user needs ranging from casual, quick edits to professional, fine-grained retouching.

Weak instructions are designed for non-expert users who

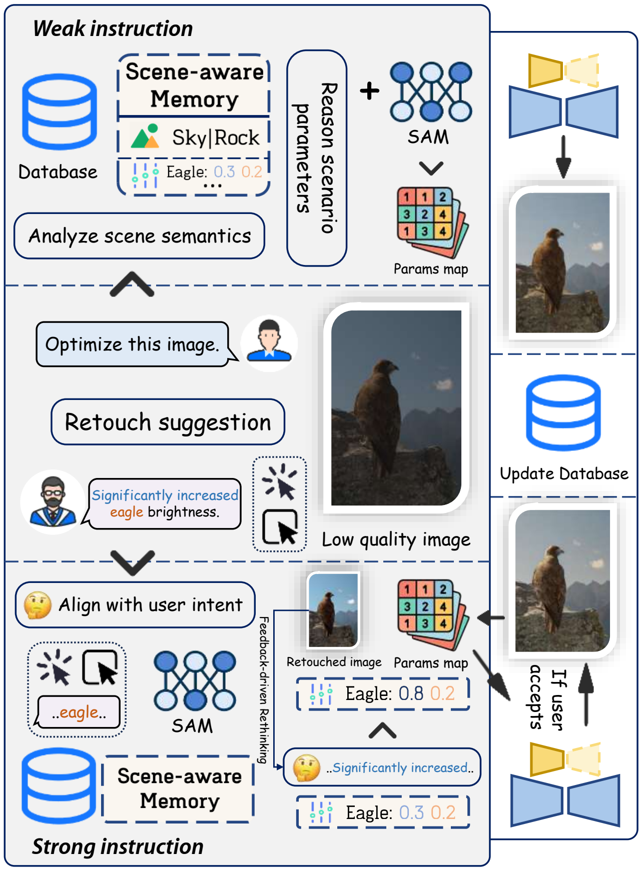

Figure 3: Agent workflow in PerTouch. Our unified agent framework adaptively parses user instructions of varying strength. For weak instructions (e.g., "Optimize this image."), the agent leverages scene-aware memory to retrieve long-term user preferences and generates editable parameter maps based on historical behavior. For strong instructions (e.g., "Significantly increased eagle brightness."), the agent further adopts a feedback-driven rethinking mechanism to iteratively refine vague or unsatisfactory outputs. This adaptive instruction-following pipeline allows PerTouch to support both global and region-level personalized retouching under natural language commands.

prefer minimal interaction. In this mode, the agent automatically constructs the multi-channel parameter maps $C = C^1, C^2, \ldots, C^K$ using the midpoint of each image attribute as the default value. While some methods (Song, Qian, and Du 2021) allow users to provide reference images to indicate their preferred style, we adopt the midpoint-based initialization to reduce user learning costs and improve usability. The parameter maps are then further tailored based on the user's editing history and preferences. This guidance map is then injected into the ControlNet to enable rapid and accurate region-aware retouching, producing visually appealing results that align with the user's historical aesthetic tendencies, without requiring explicit manual input.

In contrast, strong instructions are intended for users with clearer editing goals or professional demands. In this mode, users can specify the target region, the attribute dimension(s) to be modified, and the desired retouching strength. Upon receiving instructions, the agent leverages the VLM's object detection capability to identify the region of interest and invokes the SAM model to obtain a coarse segmentation mask. Based on this, a revised guidance map is constructed via a feedback-driven Rethinking mechanism, applying precise modifications to the designated region while retaining the adjustments derived from the weak instruction elsewhere. This design enables accurate local retouching while preserving global coherence and aesthetic consistency.

Feedback-driven Rethinking In real-world usage scenarios, users typically lack a precise understanding of the parameter space. As a result, they often do not provide exact adjustment values, but instead use vague or subjective expressions such as "slightly increase", "increase significantly", or "reduce a bit". This introduces a challenge for the model to interpret the intended degree of retouching and to determine whether the generated result aligns with the user's instruction.

To address this, we propose a Feedback-driven Rethinking mechanism. At the initial stage, the model estimates a control value c_0 by conditioning on the user's instruction I, the model's prior knowledge $\mathcal{P}$, and the user's historical preferences $\mathcal{H}_u$:

$$c_0 \sim p(c \mid I, \mathcal{P}, \mathcal{H}_u), \quad \hat{X}_0 = \mathcal{G}(X, c_0) \tag{1}$$

Here, $\hat{X}_0$ denotes the first-round retouched image produced by the generative model $\mathcal{G}$, and X is the original image. Importantly, the instruction I implicitly corresponds to an ideal control value c^*, which would generate a preferred image X^*. However, since X^* is not directly accessible, the system initiates an iterative rethinking process. At each step t, the current output $\hat{X}_t$ is sent to the agent alongside the original image X and the instruction I, allowing the multi-modal model to assess whether the result satisfies the intended semantic adjustment. If not, the agent revises the control variable by incorporating feedback from the previous result:

$$c_t \sim p(c \mid I, \hat{X}_{t-1}, \mathcal{P}, \mathcal{H}_u), \quad \hat{X}_t = \mathcal{G}(X, c_t) \tag{2}$$

This forms a closed loop of parameter refinement and visual retouching, where the system progressively adjusts the parameters to bring the output $\hat{X}_t$ closer to the latent target X^* aligned with the user's instructions. This mechanism not only enables the model to handle ambiguous user expressions, but also facilitates the construction of a learned mapping between language-level adjustment cues, control values, and perceptual visual outcomes. Ultimately, it improves the alignment between user intent and retouching results.

Scene-aware Memory To further enhance the model's capacity for capturing long-term user preferences, we introduce a mechanism called Scene-aware Memory. After the image retouching operation, the agent extracts the scene semantics of the image $\mathcal{F}(\cdot)$, yielding $f_t = \mathcal{F}(I_t)$, and stores them alongside the final confirmed editing parameter $\hat{c}_t$ in the personalized memory bank $\mathcal{M}$. As the number of user interactions increases, the memory gradually encodes a distribution of user preferences conditioned on different scenes.

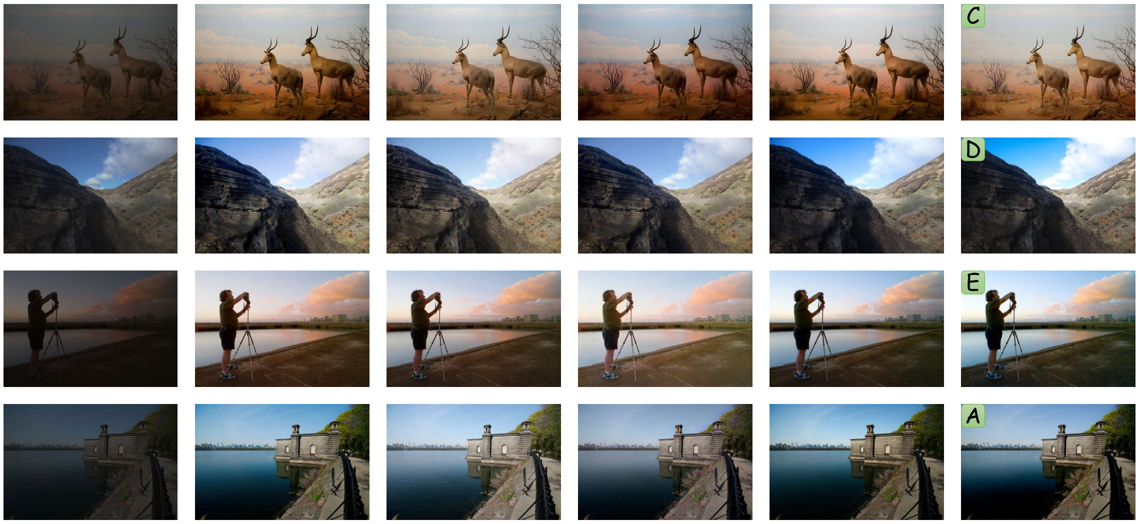

Figure 4: Qualitative comparison with other methods.

Method	A		B		C		D		E	
	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
PIENet	21.5184	0.1265	25.9065	0.0912	25.1927	0.0975	22.8989	0.1119	24.1171	0.1131
TSFlow	20.6123	0.1037	25.2474	0.0716	25.6243	**0.0630**	22.3720	0.0894	23.5393	0.0822
StarEnhancer	20.7100	0.1057	25.7296	0.0738	25.5198	0.0645	23.3875	0.0803	24.4558	0.0834
Diffretouch	24.5082	0.0812	26.1473	**0.0672**	25.9148	0.0684	24.5087	**0.0768**	24.7373	**0.0776**
PerTouch	**25.1430**	**0.0798**	**27.4733**	0.0687	**26.7510**	0.0844	**25.9726**	0.0823	**25.6602**	0.0792

Table 1: Quantitative comparisons on the MIT-Adobe FiveK dataset. Evaluations are conducted on five expert retouching versions (A/B/C/D/E) in the test set, with each model provided the appropriate condition for generating expert-style outputs. Best results are shown in **bold**, and second-best are underlined.

When the user later edits a new image I_q, the agent first extracts its scene semantics $f_q = \mathcal{F}(I_q)$. Based on these semantics and the memory bank $\mathcal{M}$, the agent estimates a conditional preference distribution $p(c \mid f_q; \mathcal{M})$, and samples a parameter vector to guide the editing process:

$$\tilde{c}_q \sim p(c \mid f_q; \mathcal{M}) \tag{3}$$

The sampled parameter $\tilde{c}_q$ serves as the control signal for downstream image editing modules, enabling the model to generate outputs that better align with the user's long-term aesthetic tendencies. This mechanism allows the system to maintain consistent personalization across diverse users and scene types, while significantly reducing the burden of manual parameter tuning and improving both interaction efficiency and visual coherence.

4 Experiments

4.1 Settings

Detailed dataset information and experimental settings are presented in the Supplementary Material.

4.2 Comparisons

We compare our PerTouch with several existing image retouching methods, focusing on approaches that support diverse retouching styles. These include DiffRetouch, StarEnhancer, TSFlow and PIE-Net, which adopt a single model trained on retouched results from multiple experts. During inference, these models can generate style-specific outputs either by providing control parameters or extracting the style from reference images to emulate different expert preferences. To evaluate the multi-style retouching capability of PerTouch, we follow the same data preparation pipeline as in the training phase to generate expert-specific supervision on the test set. For each low-quality input image in the MIT-Adobe FiveK test set, we construct a set of expert guidance maps $C = \{C^1, C^2, \ldots, C^K\}$ according to the procedure in Sec. 3.2, enabling the model to produce multiple expert-style outputs under different conditions. Since all compared methods support multi-style generation, we evaluate retouching results across all five experts. Both qualitative and quantitative comparisons are shown in Figure 4 and Table 1. Our method, while introducing region-level retouching, main-

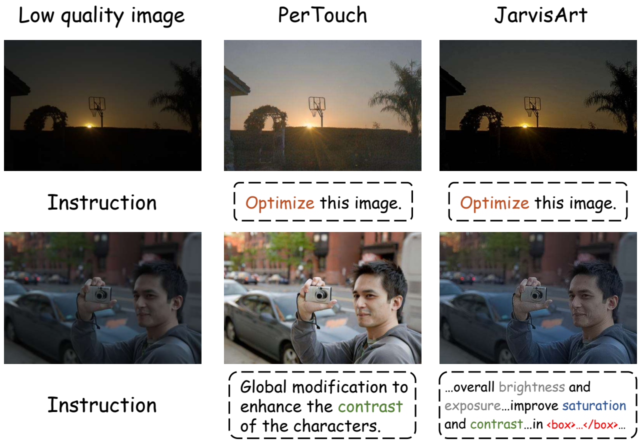

Figure 5: Comparison with Jarvis Art.

tains or even surpasses the global retouching performance of existing state-of-the-art methods in terms of objective evaluation, demonstrating the effectiveness of PerTouch.

In addition, recent works explore using vision language models (VLMs) as agents to control photo-editing toolchains like Adobe Lightroom. However, the MIT-Adobe FiveK dataset does not contain detailed retouching path descriptions between low-quality inputs and ground-truth images, making direct quantitative evaluation of such systems infeasible. Therefore, we include a qualitative comparison for reference, as illustrated in Figure 5.

To further validate the effectiveness of our approach, we conduct a user study to assess human preferences over PerTouch and other SOTA baselines including DiffRetouch, StarEnhancer, TSFlow, and JarvisArt. Detailed results are presented and analyzed in the Supplementary Material.

4.3 Ablation Studies

Semantic Replacement Module This module is designed to enhance the model's understanding of semantic regions within the image and improve the accuracy and consistency of semantic-aware local retouching. We observe that the provided parameter maps are often spatially discrete, while real images exhibit natural spatial continuity. Directly injecting such discrete control signals into the model often leads to ambiguity around semantic boundaries. In the ablation experiment, we removed this module while keeping all other settings unchanged and retrained the model. As shown in the left column of Figure 6, the absence of the semantic replacement module significantly degrades the model's ability to localize retouching, leading to spillover effects, where local edits undesirably affect global regions. This confirms the necessity of the semantic replacement module in improving region-level control precision and generalization.

Perturbation Mechanism We found that directly injecting parameter maps obtained after semantic replacement can cause the model to overly rely on externally encoded segmentation boundaries, resulting in overfitting to these semantic borders. This behavior deviates from our goal, which is to allow the model to internally balance the global aesthetic guided by the diffusion prior with the localized control suggested by segmentation cues. To this end, we introduce a perturbation mechanism that encourages the model to perceive how different parameter values influence semantic boundaries. As shown in the right column of Figure 6, removing this mechanism causes the model to overfit external segmentation structures, resulting in reduced global visual coherence during user-guided manipulation. This demonstrates the importance of the perturbation mechanism in enhancing semantic awareness and improving user experience.

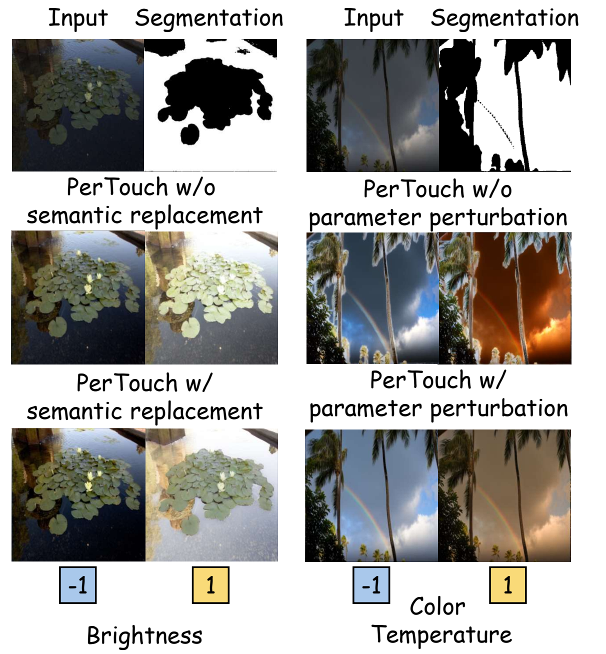

Figure 6: Ablation studies on key components of PerTouch. The left column compares results with and without the Semantic Replacement Module, showing its effectiveness in improving semantic region control and reducing undesired global spillover. The right column compares results with and without the Perturbation Mechanism, demonstrating its role in mitigating overfitting to segmentation boundaries and enhancing global visual quality. We only set the parameter values of the masked region dimensions to 1/-1, leaving all others to their default values.

5 Conclusion

In this paper, we propose PerTouch, a unified diffusion-based framework for personalized image retouching. By introducing an explicit image-to-parameter mapping mechanism, along with semantic replacement and parameter perturbation modules, our method enables fine-grained, region-aware image retouching. To further align with user intent, we incorporate an agent that supports prompt-based control, iterative feedback refinement, and long-term preference modeling through scene-aware memory. Extensive experiments validate the effectiveness of each component and demonstrate that PerTouch generates high-quality results consistent with user preferences.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62306153, 62225604), the Natural Science Foundation of Tianjin, China (24JCJQJC00020), the Young Elite Scientists Sponsorship Program by CAST (YESS20240686), the Fundamental Research Funds for the Central Universities (Nankai University, 070-63243143), and Shenzhen Science and Technology Program (JCYJ20240813114237048). This work was also funded by Samsung R&D Institute China-Beijing (SRC-B). The computational devices is supported by the Supercomputing Center of Nankai University (NKSC).

References

Adobe Inc. 2024a. Adobe Lightroom. Available at https://www.adobe.com/products/photoshop-lightroom.html.

Adobe Inc. 2024b. Adobe Photoshop. Available at https://www.adobe.com/products/photoshop.html.

Bychkovsky, V.; Paris, S.; Chan, E.; and Durand, F. 2011. Learning photographic global tonal adjustment with a database of input / output image pairs. In *CVPR*, 97–104.

Chen, C.; Chen, Q.; Xu, J.; and Koltun, V. 2018a. Learning to See in the Dark. In *CVPR*, 3291–3300.

Chen, H.; Li, W.; Gu, J.; Ren, J.; Chen, S.; Ye, T.; Pei, R.; Zhou, K.; Song, F.; and Zhu, L. 2024. RestoreAgent: Autonomous Image Restoration Agent via Multimodal Large Language Models. In *NeurIPS*, 110643–110666.

Chen, H.; Tao, K.; Wang, Y.; Wang, X.; Zhu, L.; and Gu, J. 2025. PhotoArtAgent: Intelligent Photo Retouching with Language Model-Based Artist Agents. *arXiv preprint arXiv:2505.23130*.

Chen, Y.-S.; Wang, Y.-C.; Kao, M.-H.; and Chuang, Y.-Y. 2018b. Deep Photo Enhancer: Unpaired Learning for Image Enhancement from Photographs with GANs. In *CVPR*, 6306–6314.

Duan, Z.-P.; Zhang, J.; Lin, Z.; Jin, X.; Wang, X.; Zou, D.; Guo, C.-L.; and Li, C. 2025. DiffRetouch: Using Diffusion to Retouch on the Shoulder of Experts. In *AAAI*, 2825–2833.

Dutt, N. S.; Ceylan, D.; and Mitra, N. J. 2025. MonetGPT: Solving Puzzles Enhances MLLMs' Image Retouching Skills. *ACM TOG*, 1–12.

Gharbi, M.; Chen, J.; Barron, J. T.; Hasinoff, S. W.; and Durand, F. 2017. Deep bilateral learning for real-time image enhancement. *ACM TOG*, 1–12.

He, J.; Liu, Y.; Qiao, Y.; and Dong, C. 2020. Conditional sequential modulation for efficient global image retouching. In *ECCV*, 679–695.

Jiang, X.; Li, G.; Chen, B.; and Zhang, J. 2025. Multi-Agent Image Restoration. *arXiv preprint arXiv:2503.09403*.

Kim, H.; Choi, S.-M.; Kim, C.-S.; and Koh, Y. J. 2021. Representative Color Transform for Image Enhancement. In *ICCV*, 4439–4448.

Kim, H.; and Lee, K. M. 2024. Learning Controllable ISP for Image Enhancement. *IEEE TIP*, 867–880.

Kim, H.-U.; Koh, Y. J.; and Kim, C.-S. 2020. PieNet: Personalized Image Enhancement. In *ECCV*, 374–390.

Li, B.; Li, X.; Lu, Y.; and Chen, Z. 2025. Hybrid Agents for Image Restoration. *arXiv preprint arXiv:2503.10120*.

Liang, J.; Zeng, H.; Cui, M.; Xie, X.; and Zhang, L. 2021. PPR10K: A Large-Scale Portrait Photo Retouching Dataset with Human-Region Mask and Group-Level Consistency. In *CVPR*, 653–661.

Lin, Y.; Lin, Z.; Lin, K.; Bai, J.; Pan, P.; Li, C.; Chen, H.; Wang, Z.; Ding, X.; Li, W.; and Yan, S. 2025. JarvisArt: Liberating Human Artistic Creativity via an Intelligent Photo Retouching Agent. In *NeurIPS*.

Liu, R.; Ma, L.; Zhang, J.; Fan, X.; and Luo, Z. 2021. Retinex-inspired Unrolling with Cooperative Prior Architecture Search for Low-light Image Enhancement. In *CVPR*, 10561–10570.

Moran, S.; Marza, P.; McDonagh, S.; Parisot, S.; and Slabaugh, G. 2020. DeepLPF: Deep Local Parametric Filters for Image Enhancement. In *CVPR*, 12823–12832.

Moran, S.; McDonagh, S.; and Slabaugh, G. 2021. CURL: Neural Curve Layers for Global Image Enhancement. In *ICPR*, 9796–9803.

Ouyang, W.; Dong, Y.; Kang, X.; Ren, P.; Xu, X.; and Xie, X. 2023. RSFNet: A white-Box image retouching approach using region-specific color filters. In *ICCV*, 12160–12169.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 10684–10695.

Song, Y.; Qian, H.; and Du, X. 2021. StarEnhancer: Learning Real-Time and Style-Aware Image Enhancement. In *ICCV*, 4126–4135.

Sun, X.; Li, M.; He, T.; and Fan, L. 2021. Enhance images as you like with unpaired learning. In *IJCAI*, 1011–1017.

Wang, R.; Zhang, Q.; Fu, C.-W.; Shen, X.; Zheng, W.-S.; and Jia, J. 2019. Underexposed Photo Enhancement Using Deep Illumination Estimation. In *CVPR*, 6842–6850.

Wang, T.; Li, Y.; Peng, J.; Ma, Y.; Wang, X.; Song, F.; and Yan, Y. 2021. Real-time Image Enhancer via Learnable Spatial-aware 3D Lookup Tables. In *ICCV*, 2471–2480.

Yang, C.; Jin, M.; Jia, X.; Xu, Y.; and Chen, Y. 2022. AdaInt: Learning Adaptive Intervals for 3D Lookup Tables on Real-time Image Enhancement. In *CVPR*, 17522–17531.

Zeng, H.; Cai, J.; Li, L.; Cao, Z.; and Zhang, L. 2020. Learning Image-Adaptive 3D Lookup Tables for High Performance Photo Enhancement in Real-Time. *IEEE TPAMI*, 2058–2073.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*, 3836–3847.

Zhu, A.; Zhang, L.; Shen, Y.; Ma, Y.; Zhao, S.; and Zhou, Y. 2020. Zero-Shot Restoration of Underexposed Images via Robust Retinex Decomposition. In *ICME*, 1–6.

Zhu, K.; Gu, J.; You, Z.; Qiao, Y.; and Dong, C. 2025. An Intelligent Agentic System for Complex Image Restoration Problems. In *ICLR*, 6419–6447.