Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Qwen-Edit+: Scaling Image Editing with VLM-Guided Consistency and Aesthetic Preference Distillation

Fan Tang

tangfan2021@yeah.net

AI Laboratory, Chongqing Valiant Cat Technology https://orcid.org/0009-0008-7945-8693

Siyuan Li

AI Laboratory, Chongqing Valiant Cat Technology https://orcid.org/0000-0002-2354-4233

Research Article

Keywords: instruction-based image editing, vision-language models, structural consistency, aesthetic preference distillation, data curation

Posted Date: April 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9352857/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare potential competing interests as follows: Competing Interests Both authors are affiliated with Chongqing Valiant Cat Technology Co., Ltd., which organized this research. The authors declare that they have no additional competing financial or non-financial interests.

Qwen-Edit+: Scaling Image Editing with VLM-Guided Consistency and Aesthetic Preference Distillation

Fan Tang Siyuan Li
AI Laboratory, Chongqing Valiant Cat Technology Co., Ltd.
tangfan2021@yeah.net, qq497291093@gmail.com
* Correspondence: tangfan2021@yeah.net

Abstract

Instruction-based image editing has advanced substantially with the emergence of Diffusion Transformers (DiTs). However, a central challenge remains unresolved: how to accurately execute complex editing instructions while preserving the structural consistency and visual quality of the source image. Existing methods are primarily limited by three factors: noisy and imbalanced training data, insufficient structural supervision, and inadequate alignment with human aesthetic preferences. To address these issues, we propose Qwen-Edit+, a unified framework for image editing. Specifically, we first introduce Semantic-Consistency Aware Filtering (SCAF) and Distribution-Adaptive Sampling (DAS) to construct high-quality and category-balanced training data. We then propose a VLM-aware Consistency Loss (VCL), which exploits the hierarchical hidden states of Qwen2.5-VL to provide deep semantic and structural supervision. Finally, we incorporate Aesthetic Preference Distillation (APD) to further improve visual harmony and perceptual quality. In comparative experiments, our method achieved a CLIP Score of 0.347, an LPIPS of 0.219, a PSNR of 25.63, and an Aesthetic Score of 6.31 on Qwen-Consistent-Edit-1.2K, outperforming representative baselines in editability, structural fidelity, and visual quality.

Keywords: instruction-based image editing; vision-language models; structural consistency; aesthetic preference distillation; data curation

1 Introduction

Instruction-based image editing is emerging as a key interactive paradigm for creative production: users increasingly expect to perform complex visual modifications directly through natural language, rather than relying on precise masks or labor-intensive manual workflows. The central difficulty of this task is to satisfy two objectives simultaneously: accurately executing the requested edit (editability) while preserving the identity, structure, and texture of the source image (fidelity). We refer to this persistent conflict as the *Editing-Fidelity Paradox*.

Although diffusion models and Diffusion Transformers (DiTs) [1, 4, 5] have significantly improved image generation quality, stronger generative capacity does not automatically translate into more reliable editing behavior. Under intensive edits, existing instruction-based methods still exhibit systematic failure modes: when the model prioritizes instruction execution, it often suffers from identity drift and geometric distortion; when the model emphasizes fidelity preservation, it frequently degenerates into weakly edited outputs that fail to respond adequately to user intent. At a fundamental level, most existing supervision remains dominated by pixel reconstruction or global semantic matching, which is insufficient to jointly model the dual requirement that edited regions should change while non-target regions should remain stable.

The principal challenge is therefore structural drift: without transferable hierarchical geometric supervision, the model struggles to maintain consistency between local modifications and global layout. Two secondary challenges further aggravate this difficulty. First, noisy synthetic triplets amplify structural learning errors, because semantic mismatches in automatically generated data corrupt the supervision signal [6]. Second, preference consistency remains limited: for instructions involving aesthetic enhancement or global visual harmonization, objective losses do not fully reflect human perceptual judgments.

Motivated by this challenge decomposition, we propose **Qwen-Edit+**, a unified VLM-guided training framework with a hierarchical design. VCL serves as the core component for addressing structural drift. SCAF and DAS improve data reliability and stabilize

the supervision signal. APD enhances visual quality for aesthetically oriented instructions without compromising semantic fidelity or structural consistency. CCL stabilizes multi-objective optimization. This design follows a structure-first, aesthetics-second training strategy that reduces competition among objectives. Experimental results demonstrate consistent improvements in editability, structural fidelity, and visual quality.

Contributions.

- We propose a unified training framework for instruction-based image editing that integrates data curation, structural consistency supervision, and preference refinement into a single optimization pipeline to alleviate the editing-fidelity trade-off.
- We introduce a VLM-guided consistency supervision mechanism (VCL) that leverages hierarchical hidden features to provide dense structural constraints and establish a more robust balance between editing effectiveness and structural preservation.
- We present a coordinated data-side and preference-side design (SCAF + DAS + APD), and empirically demonstrate that this combination consistently improves instruction alignment, structural fidelity, and visual quality.

2 Related Work

2.1 Instruction-based Image Editing

Modern image editing was initially shaped by early diffusion-based methods such as SDEdit [3], which iteratively refined images via stochastic differential equations. As the field matured, the paradigm shifted toward instruction-driven models. For example, InstructPix2Pix [7] leveraged GPT-3 and Prompt-to-Prompt [15] to synthesize large-scale triplets, enabling direct learning from natural language editing instructions. Subsequent benchmarks such as MagicBrush [9] further improved the precision of localized edits through manual annotation.

Nevertheless, most early frameworks relied on U-Net backbones, which, despite their efficiency, struggle to model the long-range dependencies and multimodal interactions required for complex high-resolution editing. The recent transition to Diffusion Transformers (DiTs) [4, 14] has provided a more scalable and expressive architecture. However, the increased capacity of DiTs also makes them more prone to overfitting noisy, low-quality training data. Our work builds on this line of research by adopting a DiT-based architecture while introducing a dedicated data purification pipeline (SCAF + DAS) to ensure that the transformer backbone is trained on a high-fidelity, semantically balanced distribution [6].

2.2 Vision-Language Models for Visual Supervision

The integration of large vision-language models (VLMs) has fundamentally reshaped visual learning by bridging low-level perception and high-level reasoning. Models such as Qwen2.5-VL [10] and LLaVA [11] align visual tokens with large-scale linguistic knowledge. Traditionally, VLMs have been used in image editing only for auxiliary purposes, such as generating captions or serving as automated annotation engines.

In contrast, Qwen-Edit+ treats the VLM as a dense structural supervisor. The motivation is that hidden states from a pretrained VLM contain hierarchical semantic information: shallow layers capture local texture and edge cues, whereas deep layers encode abstract layouts and object relationships. Through the proposed VLM-aware Consistency Loss (VCL) [13], we exploit these intermediate representations to constrain structural continuity between the source and edited images. Compared with conventional perceptual metrics such as LPIPS or global text-image matching signals such as CLIP score, this supervisory signal is more effective at penalizing subtle geometric distortions during editing.

2.3 Human Preference Alignment and Aesthetics

Aligning generative models with human preferences is essential for narrowing the gap between mathematical loss functions (e.g., MSE) and subjective visual appeal [16]. Inspired by the success of RLHF and DPO [12] in large language models, recent studies have begun adapting preference alignment techniques to text-to-image diffusion models [17].

Even when instructions are followed accurately, standard diffusion outputs often exhibit an *aesthetic gap*, manifesting as flat lighting, dull colors, or insufficient textural clarity. Previous work has typically addressed this issue by filtering training data with static aesthetic predictors. Qwen-Edit+ instead introduces Aesthetic Preference Distillation (APD) [17]. By distilling knowledge from a differentiable aesthetic evaluator and adopting a margin-based ranking loss, our framework ensures that edited results remain semantically faithful to the instruction while better satisfying human preferences for harmony, realism, and overall visual appeal

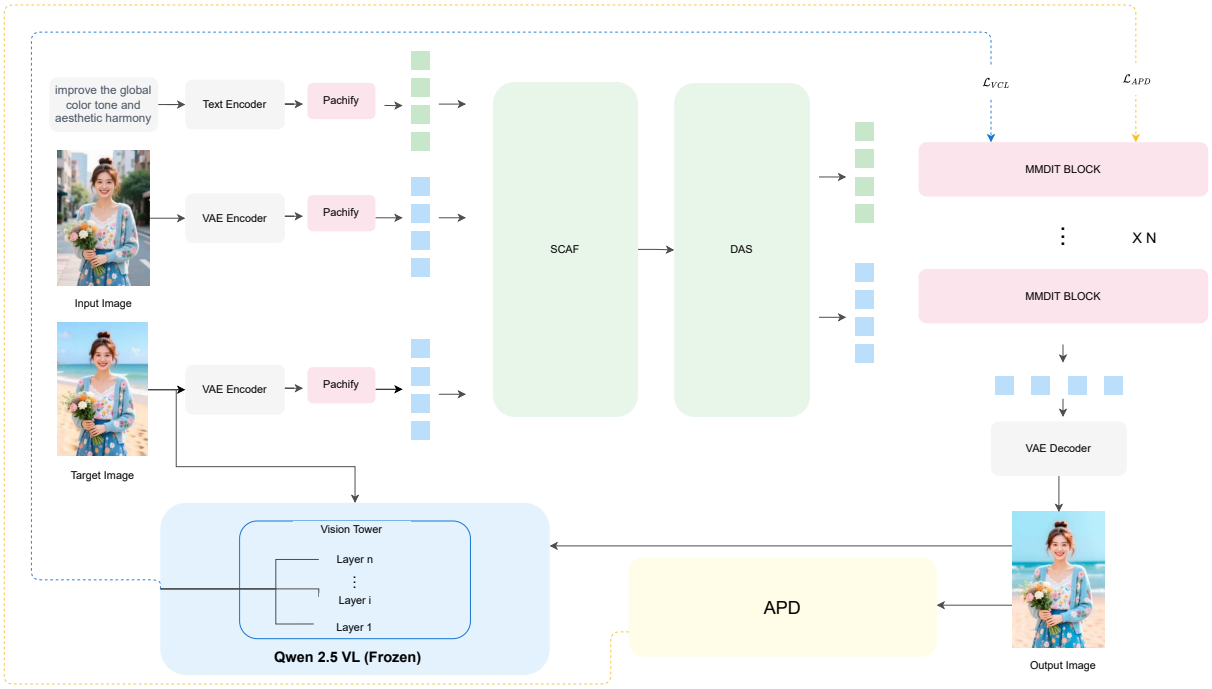


Figure 1: Overall architecture of **Qwen-Edit+**. Left: the data-centric optimization pipeline curates training triplets with SCAF and DAS. Center: the MMDiT backbone processes latent tokens under VLM-aware consistency supervision (L_{VCL}) derived from hierarchical hidden states of a frozen Qwen2.5-VL. Right: the Aesthetic Preference Distillation (APD) module refines the visual manifold through a margin-based ranking mechanism.

[8].

3 Method

3.1 System Overview and Training Pipeline

Training proceeds in five stages. Raw triplets are first purified by SCAF, which retains samples with high instruction-target semantic consistency S_{align} and strong non-edited-region identity preservation S_{id}, and are then rebalanced by DAS to mitigate long-tail edit intents. The source latent and noisy target latent are concatenated and processed by the MMDiT backbone for conditional noise prediction. During denoising, VCL imposes structural consistency using hierarchical hidden states from a frozen Qwen2.5-VL. After structural stability is established, APD improves perceptual quality, while CCL schedules the auxiliary objectives to stabilize optimization. This design follows a main-challenge-first strategy in which structural fidelity is enforced before aesthetic refinement.

3.2 Preliminaries: Instruction-Guided Latent Diffusion

Our model builds on the Latent Diffusion Model (LDM) [2] and the DiT framework [4]. Given a source image $x \in \mathbb{R}^{H\times W\times 3}$, a pretrained VAE encoder $\mathcal{E}$ maps it into a compressed latent $z_0 = \mathcal{E}(x)$. The forward diffusion process adds Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ according to the variance schedule $\bar{\alpha}_t$:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \tag{1}$$

The training objective is conditional noise prediction under dual conditioning on the instruction c_{txt} and the source-image representation c_{img}. In the MMDiT backbone, these conditions interact with z_t through cross-attention. The predicted noise is

$$\hat{\epsilon} = \epsilon_\theta(z_t, t, c_{txt}, c_{img}) \tag{2}$$

and the base diffusion objective is

$$\mathcal{L}_{diff} = \mathbb{E}_{z_0,t,\epsilon,c_{txt},c_{img}}\left[\|\epsilon - \hat{\epsilon}\|_2^2\right]. \tag{3}$$

Standard framework versus our modifications. The diffusion process itself is standard. Our modifications are fourfold: concatenating the source latent z_{src}

with the noisy latent z_t before the first DiT block, introducing VCL as the principal structural supervisor, applying SCAF and DAS on the data side, and adding APD together with CCL for quality refinement and stable convergence.

3.3 Data-Side Optimization: DAS and SCAF

To reduce noise and distributional bias in large-scale editing data, we apply two preprocessing steps before training.

3.3.1 Distribution-Adaptive Sampling (DAS)

Let $\mathcal{D} = \{(x_i, c_i)\}_{i=1}^{N}$ denote the raw training set. We encode each instruction with a frozen Qwen2.5-VL text encoder E_{txt} and perform K-means clustering with $K = 12$ on the resulting instruction embeddings, so that resampling is driven by edit intent rather than image appearance. Let N_k denote the number of samples in category k. The category sampling probability p_k is

$$w_k = \frac{1}{(N_k)^{\alpha}}, \quad p_k = \frac{w_k}{\sum_{j=1}^{K} w_j} \tag{4}$$

where $\alpha \in [0, 1]$ is the smoothing factor. We tune α over $\{0.25, 0.5, 0.75\}$ and adopt $\alpha = 0.5$, which provides a good trade-off between minority-category coverage and sample stability.

3.3.2 Semantic-Consistency Aware Filtering (SCAF)

Automatically generated editing triplets often contain semantic mismatches. Given $\mathcal{D} = \{(x_{src}^{(i)}, x_{tgt}^{(i)}, c_{txt}^{(i)})\}_{i=1}^{N}$, we use the Qwen2.5-VL encoder $f_{vlm}(\cdot)$ [10] to compute a joint semantic score S_{align}. Here, $f_{vlm}(x, c)$ uses joint image-text input, mean pooling over the final-layer hidden states, and L2 normalization. The instruction-conditioned source representation is

$$e_{query} = f_{vlm}(x_{src}, c_{txt}) \tag{5}$$

and the semantic score is

$$S_{align}(x_{src}, x_{tgt}, c_{txt}) = \cos\left(e_{query}, f_{vlm}(x_{tgt})\right). \tag{6}$$

To reject samples with severe background warping, we further introduce an identity preservation score S_{id}. Let $g(\cdot)$ denote a frozen vision tower, and let M_{edit} denote the edit-region mask derived from cross-attention maps. The non-edited regions are

$$\tilde{x}_{src} = (1 - M_{edit}) \odot x_{src}, \qquad \tilde{x}_{tgt} = (1 - M_{edit}) \odot x_{tgt} \tag{7}$$

and the identity score is

$$S_{id}(x_{src}, x_{tgt}) = \cos\left(g(\tilde{x}_{src}), g(\tilde{x}_{tgt})\right). \tag{8}$$

Filtering is performed in two stages:

$$\mathcal{D}_{align} = \{(x_{src}, x_{tgt}, c_{txt}) \in \mathcal{D} \mid S_{align} > \tau\} \tag{9}$$

$$\mathcal{D}_{clean} = \{(x_{src}, x_{tgt}, c_{txt}) \in \mathcal{D}_{align} \mid S_{id} > \tau_{id}\}. \tag{10}$$

Here, S_{align} measures instruction-consistent semantic change, while S_{id} preserves the stability of non-edited regions during filtering.

For thresholding, τ is set to the 70th percentile of S_{align} over all candidates, while τ_{id} is set to the 50th percentile of S_{id} within each DAS category to avoid over-filtering categories with inherently larger structural variation.

3.4 VLM-Guided Structural Consistency (VCL)

VCL is the core component of our method. It preserves local geometry and global layout continuity under intensive edits. Pixel reconstruction mainly constrains low-level appearance, whereas global semantic matching enforces text-image compatibility; neither captures the asymmetric requirement that edited regions may change while non-target regions should remain stable. By contrast, the hierarchical hidden states of Qwen2.5-VL provide dense supervision spanning both texture and layout cues [13].

Let $\Phi^{(l)}(x)$ denote the feature map from layer l of the VLM encoder. The per-layer consistency term is

$$\ell_{VCL}^{(l)} = \frac{1}{H_l W_l} \left\| \Phi^{(l)}(x_{edit}) - \Phi^{(l)}(x_{src}) \right\|_2^2 \tag{11}$$

and the multi-layer objective is

$$\mathcal{L}_{VCL} = \sum_{l \in \mathcal{L}_{layers}} \gamma_l \ell_{VCL}^{(l)}. \tag{12}$$

The mask M is the mechanism that balances editability and fidelity. Enforcing consistency over the entire image would suppress intended changes, so we derive an edit mask from instruction-conditioned cross-attention and define the masked per-layer term as

$$\ell_{VCL,mask}^{(l)} = \left\| (1 - M) \odot \left(\Phi^{(l)}(x_{edit}) - \Phi^{(l)}(x_{src}) \right) \right\|_2^2 \tag{13}$$

which gives

$$\mathcal{L}_{VCL}^{masked} = \sum_{l} \gamma_l \ell_{VCL,mask}^{(l)}. \tag{14}$$

Local and global edits require different consistency strengths. For localized instructions such as object replacement, M is typically sparse and the non-target constraint suppresses identity drift effectively. For global stylization or tone adjustment, however, M often has high spatial coverage, and a hard mask would become overly restrictive. We therefore define the edit coverage

$$r = \frac{1}{HW} \sum_{u,v} M_{u,v} \tag{15}$$

and modulate the layer weight by

$$\tilde{\gamma}_l = (1 - r)\gamma_l. \tag{16}$$

When r is large, we replace γ_l with $\tilde{\gamma}_l$ and use a soft mask $(1 - \rho M,\ 0 < \rho < 1)$, allowing global transformation while preserving structural anchors.

3.5 Aesthetic Preference Distillation (APD)

APD is an auxiliary objective used mainly for aesthetic enhancement, realism improvement, and global visual harmonization; for purely local semantic edits, it should not dominate the optimization direction. We therefore activate it only after structural consistency has largely stabilized. Let D_{aes} denote a differentiable aesthetic discriminator, and let $s_{edit} = D_{aes}(x_{edit})$ and $s_{src} = D_{aes}(x_{src})$ be the predicted aesthetic scores. The relative aesthetic margin is

$$\Delta_{aes} = s_{edit} - s_{src} - m \tag{17}$$

and the ranking loss is

$$\mathcal{L}_{aes} = \mathbb{E}\left[\log\left(1 + \exp\left(-\beta \cdot \Delta_{aes}\right)\right)\right]. \tag{18}$$

where β is the temperature coefficient and m is the desired improvement margin. To reduce the oversaturation artifacts commonly induced by aesthetic reward models, we additionally use a color-balance regularizer

$$\mathcal{L}_{reg} = \|\mu(x_{edit}) - \mu(x_{src})\|_2. \tag{19}$$

3.6 Consistency-Guided Curriculum Learning (CCL)

CCL stabilizes multi-objective optimization. If consistency and aesthetic constraints are activated too early, the model often exhibits suppressed edit magnitude. We group the auxiliary objectives as

$$\mathcal{L}_{aux}(t) = \lambda_{vcl}(t)\mathcal{L}_{VCL}^{masked} + \lambda_{aes}(t)\mathcal{L}_{aes} + \lambda_{vae}\mathcal{L}_{VAE} \tag{20}$$

and define the total objective by

$$\mathcal{L}_{total}(t) = \mathcal{L}_{diff} + \mathcal{L}_{aux}(t). \tag{21}$$

Here, $\mathcal{L}_{VCL}^{masked}$ adopts the adaptive weighting strategy described above for high-coverage instructions. The scheduling policy is phase-based:

1. **Phase 1 (Warm-up):** $t < T_1$. Only the diffusion objective is optimized.
2. **Phase 2 (Structure Fixation):** $T_1 \le t < T_2$. λ_{vcl} is gradually increased.
3. **Phase 3 (Aesthetic Refinement):** $t \ge T_2$. λ_{aes} is introduced and progressively increased.

The dynamic weights follow a sigmoid schedule:

$$\lambda_k(t) = \lambda_{max} \cdot \frac{1}{1 + \exp(-k \cdot (t - T_{start}))} \tag{22}$$

This schedule smooths transitions among objectives, reduces optimization oscillation, and maintains a stable balance between editability and structural fidelity.

4 Experiments

4.1 Dataset Construction and Curation

A high-quality dataset is the foundation of effective instruction-based image editing. To this end, we curate a dedicated benchmark, termed **Qwen-Consistent-Edit-1.2K**.

Data Collection and Source Integration. We first aggregate approximately 50,000 raw image-instruction triplets from diverse sources, including MagicBrush [9], Emu-Edit, and synthetic pairs generated by Stable Diffusion 3 with LLM-based prompting. To cover a broad range of editing tasks, we retain samples involving style transfer, object manipulation, and global attribute adjustment.

Data Purification via SCAF. To satisfy the high-fidelity requirements of Qwen-Edit+, we apply the proposed Semantic-Consistency Aware Filtering (SCAF). Specifically, we use the Qwen2.5-VL encoder to compute the joint semantic score S_{align} between the instruction-conditioned source representation and the target representation, and further combine it with the non-edited-region identity preservation score S_{id} to remove semantically mismatched samples or samples with severely damaged background structure. Samples with $S_{align} < \tau$ or $S_{id} < \tau_{id}$ are filtered out, yielding a refined set of 1,200 high-quality editing pairs with consistency-aware supervision.

Dataset Composition. The final 1,200 pairs are carefully balanced across four primary scenarios:

Table 1: Quantitative comparison of **Qwen-Edit+** against representative state-of-the-art baselines.

Method	Backbone	VLM Guide	CLIP ↑	LPIPS ↓	PSNR ↑	Aesthetic ↑
InstructPix2Pix [7]	U-Net	None	0.286	0.338	21.47	5.21
MagicBrush [9]	U-Net	None	0.298	0.312	22.36	5.34
OmniEdit [18]	DiT	GPT-4o	0.314	0.286	23.18	5.61
VIBE [19]	DiT	Qwen3-VL	0.326	0.255	24.07	5.88
VAREdit [20]	VAR	LLM-aware	0.333	0.244	24.41	5.95
Qwen-Edit+ (Ours)	**MMDiT**	**Qwen2.5-VL**	**0.347**	**0.219**	**25.63**	**6.31**

- **Identity Preservation (30%):** Portrait and object editing scenarios in which the core identity must remain invariant despite changes in texture or background.
- **Aesthetic Enhancement (25%):** Edits focusing on professional-grade lighting, color grading, and texture refinement.
- **Object-Centric Editing (25%):** Addition, removal, or replacement of specific semantic entities within complex scenes.
- **Global Stylization (20%):** Complex artistic style transfer tasks that require global structural understanding.

4.2 Implementation Details

We implement Qwen-Edit+ on top of the Diffusion Transformer (DiT) architecture [4]. The MMDiT backbone is trained with the AdamW optimizer using a learning rate of 5×10^{-5} and a weight decay of 0.01. Training is performed on 8 NVIDIA H800 GPUs for approximately 10,000 steps. In our implementation, a frozen Qwen2.5-VL plays two roles: it serves as the semantic consistency scorer in SCAF and as the hierarchical hidden-feature extractor in VCL. The data-side optimization stages of SCAF and DAS are performed before model training in latent space. The curriculum schedule uses $T_1 = 2000$ steps for structural warm-up and $T_2 = 5000$ steps for aesthetic refinement.

4.3 Main Quantitative Comparison

To evaluate Qwen-Edit+, we compare it with several state-of-the-art methods, including InstructPix2Pix [7], MagicBrush [9], and recent DiT-based editors. We report four key metrics: (1) CLIP Score for instruction adherence; (2) LPIPS for structural fidelity; (3) PSNR for pixel-level consistency; and (4) Aesthetic Score for visual appeal. The results are summarized in Table 1.

As shown in Table 1, Qwen-Edit+ achieves the best performance across all four metrics. Compared with the strongest baseline, VAREdit, our method improves the CLIP Score from 0.333 to 0.347 (+0.014), reduces LPIPS from 0.244 to 0.219 (-0.025), increases PSNR from 24.41 to 25.63 (+1.22 dB), and raises the Aesthetic Score from 5.95 to 6.31 (+0.36). These consistent gains on both editability-oriented and fidelity-oriented metrics indicate that our framework better resolves the editing-fidelity trade-off than prior methods.

4.4 Ablation Study

To verify the contribution of each component, we conduct ablation experiments on the 1.2K consistency-aware editing dataset by incrementally adding modules to the baseline MMDiT model.

Effect of SCAF and DAS. Starting from the baseline MMDiT (CLIP 0.319 / LPIPS 0.287 / Aesthetic 5.74), adding SCAF increases CLIP to 0.328 and reduces LPIPS to 0.268, confirming that cleaner supervision improves both instruction alignment and structural preservation. With the addition of DAS, the model further reaches CLIP 0.334 and LPIPS 0.251, showing that rebalancing long-tail editing categories improves robustness beyond pure filtering.

Impact of VCL. Adding VCL produces a clear structural gain: LPIPS drops from 0.251 to 0.229, while CLIP increases from 0.334 to 0.341. This indicates that hierarchical VLM features provide effective geometric constraints during editing, especially in identity-sensitive and layout-sensitive cases.

Role of APD. Adding APD on top of SCAF+DAS+VCL yields the largest aesthetic improvement (5.96 → 6.31, +0.35), while also improving CLIP (0.341 → 0.347) and LPIPS (0.229 → 0.219). This demonstrates that preference distillation enhances visual appeal without sacrificing editing faithfulness or fidelity.

4.5 Qualitative Analysis

Figure 2 presents a visual comparison between our method and current state-of-the-art baselines.

Table 2: Ablation study of the proposed components. Each module contributes to the final balance between fidelity and aesthetics.

SCAF	DAS	VCL	APD	CLIP ↑	LPIPS ↓	Aesthetic ↑
				0.319	0.287	5.74
✓				0.328	0.268	5.81
✓	✓			0.334	0.251	5.88
✓	✓	✓		0.341	0.229	5.96
✓	✓	✓	✓	**0.347**	**0.219**	**6.31**

Structural Consistency and Fidelity. In the portrait scenario (Figure 2b), our method exhibits superior consistency. As shown in the zoomed-in regions, the baseline produces subtle facial distortions, whereas our VCL module effectively preserves geometric layout through dense semantic guidance from Qwen2.5-VL.

Aesthetic Quality and Clarity. Cases (a) and (c) highlight the effect of APD. Compared with the relatively dim outputs of baseline models, our results are brighter and more visually harmonious.

Table 3: User preference win rate (%) comparison.

Method	Editability	Fidelity	Aesthetic
InstructPix2Pix	10.2	10.4	12.0
MagicBrush	15.6	16.4	16.0
Qwen-Edit+	**74.2**	**73.2**	**72.0**

4.6 User Preference Study

To further validate the effectiveness of our method in terms of subjective visual quality, we conducted a user study. Specifically, the study involved 50 pairs of edited images, and 5 independent evaluators were invited to perform pairwise preference judgments. As shown in Table 3, participants preferred the results generated by our method in more than 70% of the samples. This finding indicates that, beyond objective evaluation metrics, the proposed method also demonstrates strong competitiveness from the perspective of human subjective perception.



(a) Prompt: *"Improve the global color tone and aesthetic harmony"*



(b) Prompt: *"Refine skin tone and texture consistency"*



(c) Prompt: *"Increase photo realism and clarity without changing content"*

Figure 2: **Qualitative comparison of image editing results.** The first column is the original image, and the last column is the result after editing with our method.

5 Conclusion

We presented Qwen-Edit+, a unified instruction-guided image editing framework that combines data-centric curation (SCAF + DAS), VLM-guided structural supervi-

sion (VCL), and preference-aware refinement (APD) to better balance editability and fidelity. Experimental results show consistent improvements over representative baselines in instruction alignment, structural preservation, and visual quality, while the ablation study confirms that each module contributes to the final performance. Nevertheless, the current evaluation still depends on limited benchmarks and a relatively small user study. Future work will focus on broader cross-benchmark validation and more efficient high-resolution inference for practical interactive deployment.

Competing Interests

Both authors are affiliated with Chongqing Valiant Cat Technology Co., Ltd., which organized this research. The authors declare that they have no additional competing financial or non-financial interests.

References

[1] J. Ho et al. Denoising diffusion probabilistic models. In *Proc. NeurIPS*, 2020.

[2] R. Rombach et al. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, 2022.

[3] C. Meng et al. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *Proc. ICLR*, 2021.

[4] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *Proc. ICCV*, 2023.

[5] P. Esser et al. Scaling rectified flow transformers for high-resolution synthesis. *arXiv:2403.03206*, 2024.

[6] S. Y. Gadre et al. DataComp: In search of the next generation of multimodal datasets. In *Proc. ICCV*, 2023.

[7] T. Brooks et al. InstructPix2Pix: Follow image editing instructions. In *Proc. CVPR*, 2023.

[8] N. Tumanyan et al. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proc. CVPR*, 2023.

[9] K. Zhang et al. MagicBrush: A manually annotated dataset for editing. In *Proc. NeurIPS*, 2023.

[10] S. Bai et al. Qwen2.5-VL Technical Report. *arXiv:2502.13923*, 2025.

[11] H. Liu et al. Visual instruction tuning. *arXiv:2304.08485*, 2023.

[12] R. Rafailov et al. Direct preference optimization: Your language model is secretly a reward model. *arXiv:2305.18290*, 2024.

[13] Z. Li et al. Instructional image editing with VLM supervision. In *Proc. CVPR*, 2024.

[14] J. Chen et al. PixArt-α: Fast training of diffusion transformer. In *Proc. ICLR*, 2024.

[15] A. Hertz et al. Prompt-to-prompt image editing with cross attention control. In *Proc. ICLR*, 2023.

[16] Y. Kirstain et al. Pick-a-Pic: An open dataset of human preferences. In *Proc. NeurIPS*, 2023.

[17] X. Wu et al. Human-preference-aided diffusion for high-quality text-to-image generation. In *Proc. CVPR*, 2024.

[18] C. Wei et al. OmniEdit: Building image editing generalist models through specialist supervision. In *Proc. ICLR*, 2025.

[19] G. Alekseenko et al. VIBE: Visual instruction based editor. *arXiv:2601.02242*, 2026.

[20] Q. Mao et al. Visual autoregressive modeling for instruction-guided image editing. *arXiv:2508.15772*, 2025.