HARMONIA: OPTIMIZING AESTHETIC GAINS WITHIN THE SEMANTIC ELASTICITY BOUNDARY FOR AESTHETIC PHOTO RECONSTRUCTION

Anonymous authors
Paper under double-blind review

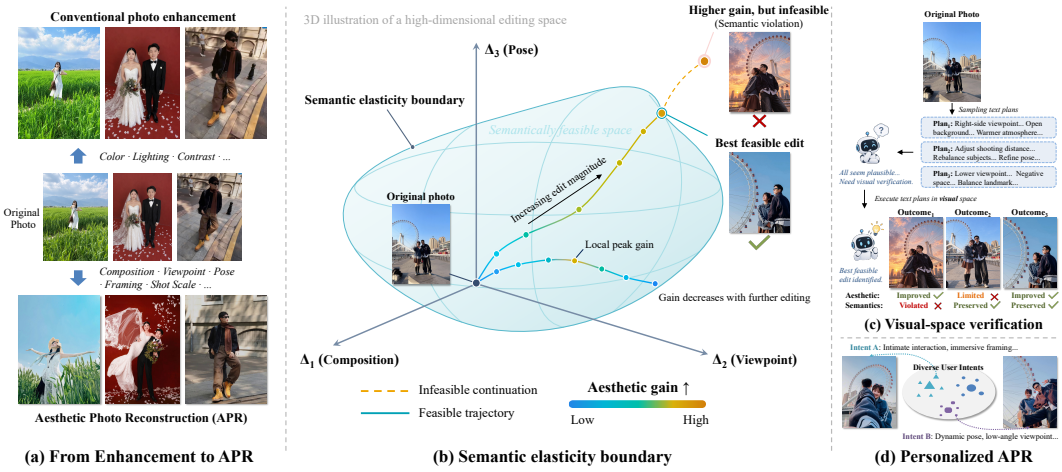

Figure 1: (a) APR goes beyond conventional photo enhancement by reconstructing photographic structure. (b) APR exhibits an inherent aesthetic–semantic tension between aesthetic gains and semantic preservation. We therefore introduce a photo-dependent **semantic elasticity boundary** and formulate APR as maximizing aesthetic gains within this feasible space. (c) Text-only plans describe intended edits but cannot reveal their realized visual outcomes, making it difficult to assess aesthetic gains and semantic feasibility without visual-space verification. (d) Different user intents may correspond to different valid reconstructions, motivating personalized APR.

ABSTRACT

Aesthetic Photo Reconstruction (APR) improves photo aesthetics through structural edits while preserving the original subjects and scene. However, APR inherently faces an **aesthetic–semantic tension**: larger structural changes may yield greater aesthetic gains but also increase the risk of altering subject identity and scene semantics. We therefore formulate APR as maximizing aesthetic gains within a photo-dependent **semantic elasticity boundary**. Existing methods optimize reconstruction plans only in text space, without feedback from their realized visual outcomes, making it difficult to assess semantic feasibility and optimize aesthetics within this boundary. They also lack support for intent-conditioned personalized APR. To address these limitations, we propose **Harmonia**, a two-stage APR framework: (1) Structured Photographic Planning Learning, which decomposes reconstruction plans into atomic photographic operations and jointly learns reference-free, reference-guided, and intent-conditioned planning; and (2) Feasible-Space Visual Optimization, which introduces **Visual Aesthetic GRPO (VA-GRPO)** to execute candidate plans and optimize the planner from their realized visual outcomes within the semantic elasticity boundary. We further construct **Harmonia-Bench** for evaluating both default and personalized APR. Experiments show that Harmonia outperforms AesFormer using only 30% of its training data even without fine-tuning the image editor, while additionally supporting personalized APR for the first time.

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.

1 INTRODUCTION

Aesthetics plays a fundamental role in how humans perceive and interpret the world and serves as an important guide for artistic and visual creation (Huang et al., 2024b; Cao et al., 2026; Du et al., 2026b). Enabling machines to understand human aesthetics, uncover the principles underlying visual beauty, and further support aesthetic creation has long been a central topic in image aesthetics research (Murray et al., 2012; Kong et al., 2016; Yang et al., 2022; Zhou et al., 2024; Huang et al., 2024a; Yang et al., 2026b). With recent advances in multimodal understanding and generative models (Li et al., 2023; Liu et al., 2023; Wu et al., 2025), this field is gradually evolving from *assessing and understanding* beauty toward *enhancing and creating* beauty. Accordingly, research has progressed from image aesthetic scoring to **Aesthetic Photo Reconstruction (APR)**, which improves photo aesthetics by reconstructing photographic structure, such as composition, camera viewpoint, and human pose, while preserving subject identity and scene semantics (Du et al., 2026a), as shown in Figure 1(a). Unlike conventional photo enhancement (Dutt et al., 2025; Lin et al., 2026a) that adjusts color or appearance, APR moves beyond appearance-level enhancement toward reconstructing photographic decisions, improving photo aesthetics at a more fundamental structural level.

AesFormer (Du et al., 2026a) first defined and studied APR, establishing a basic paradigm for the task. It mines poor–good photo pairs depicting the same subjects and scenes from online photography tutorial videos, trains a vision-language model (VLM) to predict textual reconstruction plans from an input photo, and then executes these plans with a fine-tuned image editing model to produce the reconstructed photo. To account for the inherently one-to-many nature of aesthetics (Yang et al., 2022; Sheng et al., 2025), AesFormer further adopts GRPO (Shao et al., 2024) to sample multiple plans and explore diverse photographic decisions. However, rewarding and optimizing reconstruction plans solely in text space leaves a more fundamental challenge of APR unresolved.

APR inherently exhibits an **aesthetic–semantic tension**: (1) Substantial aesthetic improvement often requires larger structural changes in composition, viewpoint, or human pose. (2) As these changes become more extensive, they increasingly risk altering subject identity and scene semantics, or even producing photos that could not plausibly be captured in the original physical scene (Kim et al., 2025). We define the boundary of structural variation that a photo can accommodate while maintaining acceptable core semantics as its **semantic elasticity boundary**, as shown in Figure 1(b). This raises a central question for APR: **How far should we edit?** For a given photo, how much should the model adjust each photographic dimension to maximize aesthetic gains while staying within this boundary? Crucially, this boundary cannot be reliably inferred from textual plans alone, because a plan specifies only the intended edit rather than its realized visual consequence (Figure 1(c)) (Zhou et al., 2025; Lin et al., 2026b; Wu et al., 2026). The same plan may improve one photo but degrade another, and its actual aesthetic gain and semantic preservation can only be assessed after the plan is executed to produce an image (Wu et al., 2026; Li et al., 2026). However, the existing APR method samples and compares different plans only in text space, without feedback from their visual outcomes. Consequently, it struggles to determine whether an edit remains within the semantic elasticity boundary and to maximize aesthetic quality within it. Moreover, aesthetic preferences vary across individuals (Ren et al., 2017; Kim et al., 2026), whereas existing APR learns a single strategy and therefore cannot support intent-conditioned personalized APR (Figure 1(d)).

To address these challenges, we propose **Harmonia**, a two-stage training framework for APR: (1) Structured Photographic Planning Learning, which decomposes reconstruction plans into atomic photographic operations and jointly learns reference-free, reference-guided, and intent-conditioned planning, enabling fine-grained and personalized reconstruction; and (2) Feasible-Space Visual Optimization, which introduces **Visual Aesthetic GRPO (VA-GRPO)** to execute candidate plans into reconstructed images and applies a multidimensional semantic gate to retain only outcomes within the semantic elasticity boundary. Aesthetic optimization is then performed within this feasible space, where visual rewards jointly capture structural improvement, final aesthetic quality, and user intent satisfaction. We further construct **Harmonia-Bench** for comprehensive APR evaluation, covering both default APR under general aesthetic preferences and personalized APR conditioned on specific user intents. Experiments show that Harmonia outperforms AesFormer (Du et al., 2026a) using only approximately 30% of its training pairs and even without fine-tuning the image editor, while enabling personalized APR for the first time.

The main contributions are summarized as follows:

- We identify the inherent **aesthetic–semantic tension** in APR and introduce the concept of a **semantic elasticity boundary**, framing APR as maximizing aesthetic gains within a photo-dependent space of semantically acceptable structural changes.
- **Harmonia** is proposed as a two-stage framework that first decomposes reconstruction plans into atomic photographic operations for fine-grained planning, and then introduces **VA-GRPO** to execute candidate plans and optimize the planner within the semantic elasticity boundary.
- **Harmonia-Bench** is constructed for evaluating both default and personalized APR. Extensive experiments show that Harmonia outperforms AesFormer using only 30% of its training pairs and even without fine-tuning the image editor, while enabling personalized APR for the first time.

2 Related Work

Photo Enhancement and APR. Conventional photo enhancement mainly falls into two categories: (1) photo retouching (Bychkovsky et al., 2011; Liang et al., 2021; Ouyang et al., 2023; Dutt et al., 2025; Duan et al., 2025; Lin et al., 2026a;b) improves tonal appearance through adjustments such as exposure, contrast, and shadows, while (2) portrait enhancement (Xie et al., 2023; Wen et al., 2024; Jin et al., 2024; Xue et al., 2025; Xu et al., 2025; Yang et al., 2026a) focuses on facial and skin-level refinement. These methods remain constrained by the original photographic structure: when composition, viewpoint, or human pose is flawed, appearance-level enhancement alone cannot substantially improve aesthetics, as shown in Figure 1 (a). To address this, APR improves aesthetics through structural photographic edits while preserving the original subjects and scene semantics (Du et al., 2026a). However, existing APR optimizes reconstruction plans only in text space, where text-based rewards cannot reliably reflect realized aesthetic gains or determine whether an edit remains within the semantic elasticity boundary.

Image Editing. Recent generative models have substantially advanced image editing, enabling diverse content modifications through natural-language instructions and providing an execution basis for APR (Brooks et al., 2023; Zhang et al., 2023; Yu et al., 2025; Xiao et al., 2025). However, these models generally rely on explicit editing instructions (Labs et al., 2025; Liu et al., 2025; Zhao et al., 2026), whereas APR must first decide *how* a photo should be edited. Given a low-aesthetic photo, the model needs to identify deficiencies across photographic dimensions and determine appropriate editing directions and magnitudes. Thus, image editing models alone are insufficient for APR without additional photographic planning capabilities (Du et al., 2026a).

Aesthetic VLMs. With the rapid development of VLMs, image aesthetics research has evolved from aesthetic scoring (Murray et al., 2012) toward fine-grained aesthetic perception and description (Huang et al., 2024b). Recent works (Zhou et al., 2024; Huang et al., 2024a; Qi et al., 2025; Cao et al., 2026) further improve the understanding of aesthetic factors such as composition, color, and lighting. However, APR requires not only understanding, but also executable and physically plausible reconstruction strategies that improve aesthetics under diverse user preferences. Such capabilities remain largely underexplored in existing aesthetic VLMs.

3 Method

3.1 Overview

In this section, we present Harmonia. As illustrated in Figure 2, Harmonia consists of two stages: **(1) Structured Photographic Planning Learning**, which decomposes reconstruction plans into atomic photographic operations and jointly learns reference-free, reference-guided, and intent-conditioned planning for fine-grained and personalized reconstruction; and **(2) Feasible-Space Visual Optimization**, which introduces VA-GRPO to sample and execute diverse candidate plans, retain semantically feasible outcomes using a multidimensional semantic gate, and optimize the photographic planner with visual rewards within the semantic elasticity boundary.

3.2 Structured Photographic Planning Learning

APR requires making appropriate photographic decisions for an input photo and translating them into plans executable by an image editor. Given a low-aesthetic-quality photo I_p, a photographic

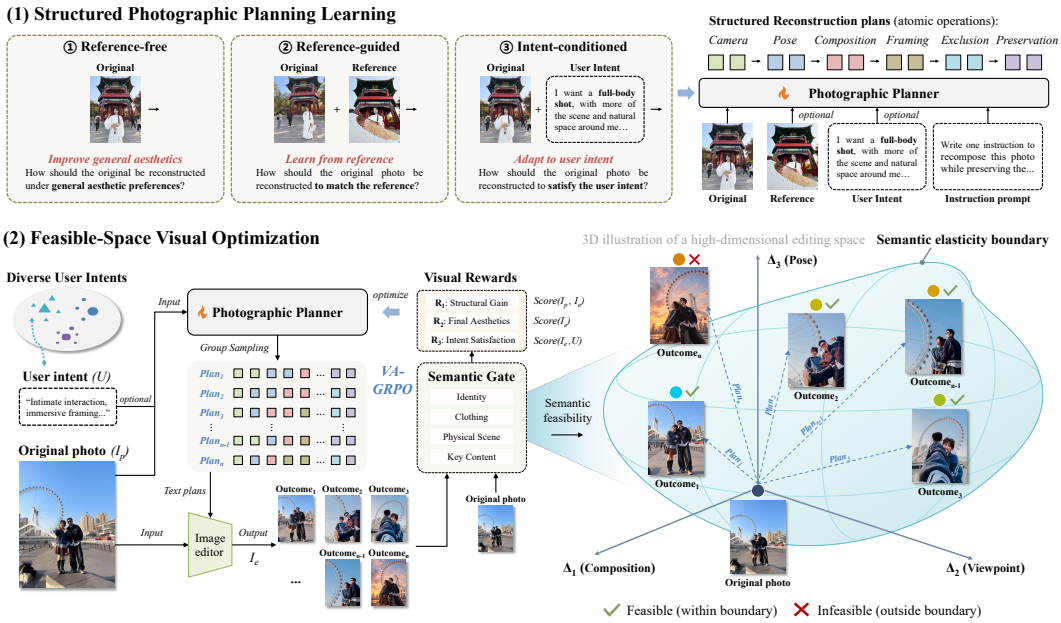

Figure 2: Overview of the Harmonia framework. **(1) Structured Photographic Planning Learning** decomposes reconstruction plans into atomic operations and learns multi-task planning. **(2) Feasible-Space Visual Optimization** introduces **VA-GRPO** to execute candidate plans and optimize the planner from feasible visual outcomes within the semantic elasticity boundary.

planner π_θ generates a reconstruction plan R conditioned on C, which is then executed by an image editor $\mathcal{E}$ to produce the reconstructed image I_e:

$$R = \pi_\theta(I_p, C), \qquad I_e = \mathcal{E}(I_p, R). \tag{1}$$

Here, $C \in \{\varnothing, I_g, U\}$ corresponds to reference-free, reference-guided, and intent-conditioned planning, respectively; I_g denotes a reference image with more desirable photographic structure from the training set, and U a user intent.

Existing APR learns a single strategy aligned with general aesthetic preferences, making it difficult to accommodate diverse user needs (Du et al., 2026a). Harmonia further incorporates user intent U, enabling diverse reconstructions of the same photo under different preferences. Meanwhile, we adopt a deliberately constrained training setting: (1) we use only 2,779 poor–good pairs from AesRecon, approximately 30% of AesFormer's training pairs; and (2) we keep the image editor $\mathcal{E}$ fully frozen and optimize only the photographic planner. This setting isolates improvements in photographic planning from editor adaptation to the training data, while decoupling the planner from a specific editor so that stronger editors can be incorporated directly as they become available.

Structured Reconstruction Plans. APR requires decisions across multiple photographic factors, such as camera viewpoint, human pose, composition, and framing. Free-form instructions may entangle these decisions and make policy exploration dominated by linguistic variation rather than meaningful photographic actions (Du et al., 2026a). We therefore decompose each plan into atomic operations across six dimensions: *Camera*, *Pose*, *Composition*, *Framing*, *Exclusion*, and *Preservation*, forming an explicit and compositional action space for subsequent VA-GRPO. We construct target plans R^* from poor–good pairs (I_p, I_g). Specifically, Gemini 3 (Comanici et al., 2025) first infers their structural differences and generates an initial plan following the atomic-operation format. Qwen3.7-Plus (Bai et al., 2025) then verifies its structural fidelity, semantic preservation and executability. Failed plans are automatically regenerated until all criteria are satisfied.

Multi-Task Planning Data. We construct three complementary planning tasks. For **(1) reference-free**, the model predicts $I_p \rightarrow R^*$ under general aesthetic preferences using all 2,779 pairs. For **(2) reference-guided**, we randomly select 1,000 pairs and provide I_g as an additional visual condition, i.e., $(I_p, I_g) \rightarrow R^*$, to learn transformations toward an explicit target photographic structure. For **(3) intent-conditioned**, formulated as $(I_p, U) \rightarrow R^*$, we randomly select 200 poor–good pairs. Gemini

3 generates a grounded intent consistent with each paired good image, yielding 200 samples using the original R^*. We further generate an alternative plausible intent for each I_p using Qwen2.5-VL-72B (Bai et al., 2025), and Nano Banana Pro (Comanici et al., 2025) produces the corresponding reconstruction I_u, yielding another 200 samples. Generated I_u are strictly filtered for image quality, aesthetics, semantic preservation, and intent consistency using dedicated scoring models (Wu et al., 2023; Schuhmann et al., 2022) and Qwen2.5-VL-72B (Bai et al., 2025), and only samples passing all criteria are retained. For each accepted (I_p, I_u) pair, we derive and verify its structured target plan using the same pipeline above.

Multi-Task Fine-Tuning. We jointly fine-tune the photographic planner on the three planning tasks. Let $\mathcal{D}_{\text{free}}$, $\mathcal{D}_{\text{ref}}$, and $\mathcal{D}_{\text{intent}}$ denote the corresponding training subsets, containing 2,779, 1,000, and 400 samples, respectively, yielding 4,179 planning samples in total. Using $C \in \{\varnothing, I_g, U\}$ to denote the task-specific condition, we optimize the planner with the standard autoregressive objective:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(I_p, C, R^*)\sim\mathcal{D}_{\text{SFT}}}\Big[\sum_{t=1}^{|R^*|} \log \pi_\theta(x_t^* \mid I_p, C, x_{<t}^*)\Big]. \tag{2}$$

where x_t^* denotes the t-th token of the target reconstruction plan R^*.

$Stage_1$ equips the photographic planner with initial reconstruction capabilities, but text-space supervision cannot capture actual visual consequences. The planner therefore cannot reliably assess aesthetic gains or whether the resulting reconstructions remain feasible (Lin et al., 2026b). In the next section, we introduce VA-GRPO to optimize reconstruction strategies based on their realized visual outcomes, maximizing aesthetic gains within the semantic elasticity boundary.

3.3 Feasible-Space Visual Optimization

After Stage 1, the planner can generate structured reconstruction plans and support initial personalized APR. However, textual plans specify only intended operations and cannot reveal their realized visual consequences (Lin et al., 2026b). A seemingly reasonable plan may yield different aesthetic gains across photos, or even alter subject identity and scene semantics when changes become too extensive. We therefore introduce **Visual Aesthetic GRPO (VA-GRPO)**, which executes candidate plans into reconstructed images and optimizes the photographic planner based on their realized visual outcomes. VA-GRPO is applied to both reference-free and intent-conditioned training.

Crucially, we do not treat semantic consistency as another reward to be maximized: unchanged or overly conservative reconstructions may preserve semantics well while providing little aesthetic improvement. Instead, we treat semantic preservation as a constraint, first identifying semantically feasible outcomes and then maximizing aesthetic gains within this space. Accordingly, VA-GRPO consists of three steps: (1) visual rollout, (2) semantic feasibility gating, and (3) visual optimization.

(1) Visual Rollout. For each training sample, we sample $K = 8$ candidate reconstruction plans from the current photographic planner and execute them using the frozen image editor $\mathcal{E}$:

$$R_i \sim \pi_\theta(\cdot \mid I_p, C), \qquad I_e^i = \mathcal{E}(I_p, R_i), \qquad i = 1, \ldots, K, \tag{3}$$

where $C = \varnothing$ denotes reference-free APR and $C = U$ denotes intent-conditioned APR. Rewards are grounded in the realized visual outcomes I_e^i rather than directly in the textual plans, allowing the planner to learn from actual execution results.

(2) Semantic Feasibility Gating. Before visual optimization, we determine whether each reconstruction remains within the semantic elasticity boundary of the input photo. Specifically, a semantic gate evaluates consistency between the original and reconstructed images along four dimensions: *Identity*, *Clothing*, *Physical Scene*, and *Key Content*. Each dimension is scored on a 1–5 scale, with higher scores indicating better consistency. Let $s_{i,d}^{\text{sem}}$ denote the score of the i-th reconstruction on semantic dimension d. A reconstruction is considered feasible only if all four scores are at least 4:

$$G_i = \mathbb{I}\left[\min_{d\in\mathcal{S}} s_{i,d}^{\text{sem}} \geq 4\right], \tag{4}$$

where $\mathcal{S}$ denotes the four semantic dimensions and $G_i \in \{0, 1\}$ indicates whether the candidate passes the gate. Candidates with $G_i = 0$ receive zero final reward, and their subsequent visual rewards need not be computed. The semantic gate thus operationalizes the photo-dependent semantic

elasticity boundary, treating semantic preservation as a prerequisite for aesthetic optimization rather than a competing objective.

(3) Visual Optimization. For candidates that pass the semantic gate, we evaluate reconstruction quality from three complementary aspects: *Structural Gain*, *Final Aesthetics*, and *Intent Satisfaction*, with all rewards normalized to $[0, 1]$. (i) *Structural Gain* measures photographic-structure improvement relative to the input. Qwen3.7-Plus (Bai et al., 2025) compares I_p and I_e^i along composition, pose, viewpoint, and spatial relationship, with each dimension scored as $s_{i,d}^{\mathrm{str}} \in \{-2, -1, 0, 1, 2\}$ from substantial degradation to substantial improvement, yielding a total score in $[-8, 8]$. (ii) *Final Aesthetics* measures the overall aesthetic quality of I_e^i using an aesthetic scoring model (Cao et al., 2026), producing $s_i^{\mathrm{aes}} \in [0, 100]$. (iii) *Intent Satisfaction* evaluates how faithfully I_e^i satisfies the given user intent U using Qwen3.7-Plus (Bai et al., 2025), producing $s_i^{\mathrm{int}} \in \{1, \ldots, 5\}$. The three rewards are normalized as:

$$r_i^{\mathrm{str}} = \frac{\sum_{d\in\mathcal{P}} s_{i,d}^{\mathrm{str}} + 8}{16}, \qquad r_i^{\mathrm{aes}} = \frac{s_i^{\mathrm{aes}}}{100}, \qquad r_i^{\mathrm{int}} = \frac{s_i^{\mathrm{int}} - 1}{4}, \tag{5}$$

where $\mathcal{P}$ denotes the four photographic structure dimensions above. For reference-free APR, we equally combine Structural Gain and Final Aesthetics. For intent-conditioned APR, Intent Satisfaction is additionally included:

$$\tilde{r}_i = \begin{cases} \dfrac{1}{2}\left(r_i^{\mathrm{str}} + r_i^{\mathrm{aes}}\right), & C = \varnothing, \\ \dfrac{1}{3}\left(r_i^{\mathrm{str}} + r_i^{\mathrm{aes}} + r_i^{\mathrm{int}}\right), & C = U. \end{cases} \tag{6}$$

Combined with the semantic gate, the final visual reward for the i-th candidate is defined as

$$r_i = G_i\, \tilde{r}_i. \tag{7}$$

Thus, semantically infeasible reconstructions receive zero reward regardless of their aesthetic quality, while feasible outcomes are rewarded according to their realized visual quality. For the $K = 8$ candidates, we compute group-relative advantages from the final rewards:

$$A_i = \frac{r_i - \mu_r}{\sigma_r + \epsilon}. \tag{8}$$

where μ_r and σ_r denote the mean and standard deviation of rewards within the group. We then update the planner using the clipped GRPO objective with these visually grounded group-relative advantages. For the t-th token $x_{i,t}$ in candidate plan R_i, the policy ratio is defined as

$$\rho_{i,t}(\theta) = \frac{\pi_\theta\left(x_{i,t} \mid I_p, C, x_{i,<t}\right)}{\pi_{\theta_{\mathrm{old}}}\left(x_{i,t} \mid I_p, C, x_{i,<t}\right)}. \tag{9}$$

The VA-GRPO objective is formulated as

$$\mathcal{J}_{\text{VA-GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\frac{1}{|R_i|}\sum_{t=1}^{|R_i|}\min\left(\rho_{i,t}(\theta)A_i, \mathrm{clip}\left(\rho_{i,t}(\theta), 1-\epsilon, 1+\epsilon\right)A_i\right)\right] - \beta D_{\mathrm{KL}}\left(\pi_\theta \| \pi_{\mathrm{ref}}\right) \tag{10}$$

where $\pi_{\theta_{\mathrm{old}}}$ denotes the rollout policy and π_{ref} the reference policy, with KL regularization constraining deviation from π_{ref}. Through group-relative optimization, the planner increasingly favors reconstruction plans that yield semantically feasible outcomes with higher visual gains.

Overall, VA-GRPO shifts APR optimization from textual plans to their realized visual consequences. By learning directly from visual feedback, the planner can explore diverse photographic decisions and maximize aesthetic gains within the semantic elasticity boundary.

4 Experiments

4.1 Harmonia-Bench

Benchmark Construction. Existing APR evaluation relies on AesRecon (Du et al., 2026a), whose test images all undergo image editing and are essentially synthetic, covering only default APR under

Table 1: Quantitative comparison on Harmonia-Bench across synthetic/real-world and Default/Personalized APR settings. All methods use the same image editor except for AesFormer, and all scores range from 0–100. Best and second-best results are in **bold** and underlined, respectively.

Method	Synthetic (AesRecon)					Real-World (PARA)				
	Default		Personalized			Default		Personalized		
	$S_1 \uparrow$	$S_2 \uparrow$	$S_1 \uparrow$	$S_2 \uparrow$	$S_3 \uparrow$	$S_1 \uparrow$	$S_2 \uparrow$	$S_1 \uparrow$	$S_2 \uparrow$	$S_3 \uparrow$
Proprietary General VLMs										
GPT-5.6 *(OpenAI)*	57.44	39.67	60.73	39.19	55.89	56.78	53.19	58.54	52.97	57.90
Gemini 3 *(Google)*	60.76	40.36	63.39	40.22	60.10	57.96	53.49	59.69	53.24	64.99
Doubao 2.0 *(ByteDance)*	60.13	40.59	63.09	39.71	58.73	58.88	54.24	60.43	53.56	59.69
DeepSeek V4.1 *(DeepSeek)*	60.62	39.17	60.93	38.55	61.07	59.09	53.33	58.24	52.83	60.45
Open-Source General VLMs										
Gemma 3 12B *(Google)*	53.00	40.66	55.63	38.25	45.67	51.91	52.63	53.58	52.03	44.76
Qwen3-VL-8B *(Alibaba)*	55.23	38.30	57.07	37.64	58.34	55.21	52.33	55.55	52.05	53.41
MiniCPM-o 2.6 *(ModelBest)*	54.60	41.60	59.12	39.34	50.53	53.54	53.30	57.13	52.18	48.29
InternVL2.5-8B *(OpenGV)*	52.02	38.35	57.34	38.09	57.23	50.51	52.19	55.41	52.56	54.18
Aesthetic VLMs										
UNIAA *(arXiv'24)*	54.64	38.19	54.61	39.29	39.67	54.27	52.12	53.44	52.49	36.60
AesExpert *(ACM MM'24)*	51.67	44.33	53.33	41.77	35.47	51.61	51.51	52.81	51.71	34.60
Venus *(CVPR'26)*	56.40	43.60	55.95	42.92	33.43	55.85	53.67	54.39	52.99	33.51
AesThinker *(ICML'26)*	70.85	48.09	70.90	46.70	59.90	68.77	55.80	67.46	54.72	64.75
APR Models										
AesFormer *(ICML'26)*	67.82	47.76	68.33	47.02	54.22	61.70	54.07	65.59	54.10	54.10
Harmonia (ours)	**76.58**	**48.64**	**75.29**	**47.07**	**63.21**	**75.18**	**57.89**	**74.03**	**57.61**	**74.51**

general aesthetic preferences. We therefore construct **Harmonia-Bench** along two axes: synthetic vs. real-world photos and default vs. personalized reconstruction. The synthetic subset inherits all 903 AesRecon test images (Du et al., 2026a), while the real-world subset contains 658 portrait images with clear and unambiguous subjects curated from the PARA dataset (Yang et al., 2022), yielding 1,561 images in total. For each image, we evaluate *Default APR*, conditioned only on I_p, and *Personalized APR*, additionally conditioned on a user intent U. For personalized APR, three intents per image are jointly written by human and Claude Sonnet 4.5, followed by manual filtering.

Evaluation Protocol. We evaluate APR along three metrics: $\mathbf{S}_1$: Structural Improvement, measuring photographic-structure improvement over the input; $\mathbf{S}_2$: Aesthetic Quality, measuring aesthetic quality of the final reconstruction; and $\mathbf{S}_3$: Intent Fulfillment, measuring how faithfully the reconstruction satisfies the user intent for Personalized APR. Scores are evaluated by GPT-5.6, which is never used during training, and normalized to 0–100, ensuring an independent evaluation. A user study in the Appendix A.1 further confirms that these evaluation scores are well aligned with human judgments. For ablations, we additionally report the **Semantic Feasibility Rate (SFR)**: a reconstruction is feasible only if its identity, clothing, physical scene, and key content scores are all at least 4 on a 1–5 scale, as scored by GPT-5.6.

4.2 Experimental Setup

Baselines. We compare Harmonia with three groups of baselines: (1) **General VLMs**, including proprietary GPT-5.6, Gemini 3, Doubao Seed 2.0, and DeepSeek V4.1 (Xu et al., 2026), as well as open-source Gemma 3 12B (Team et al., 2025), Qwen3-VL-8B (Bai et al., 2025), MiniCPM-o 2.6 (Yao et al., 2024), and InternVL2.5-8B (Chen et al., 2024). (2) **Aesthetic VLMs**, including UNIAA (Zhou et al., 2024), AesExpert (Huang et al., 2024a), Venus (Du et al., 2026b), and AesThinker (Du et al., 2026a). (3) **APR model**, AesFormer (Du et al., 2026a). For fair comparison, all reconstruction plans are executed using the same FLUX.2 [klein] 9B image editor (Labs, 2025) except for AesFormer, which retains its original configuration.

Implementation Details. Harmonia uses Qwen3-VL-8B (Bai et al., 2025) as the photographic planner. $Stage_1$ performs SFT on 4,179 multi-task samples with LoRA rank 16 for 3 epochs, using

Table 2: Ablation studies of key components in Harmonia.

Configuration	Synthetic (AesRecon)					Real-World (PARA)				
	Default		Personalized			Default		Personalized		
	$S_1 \uparrow$	$S_2 \uparrow$	$S_1 \uparrow$	$S_2 \uparrow$	$S_3 \uparrow$	$S_1 \uparrow$	$S_2 \uparrow$	$S_1 \uparrow$	$S_2 \uparrow$	$S_3 \uparrow$
Baseline (Qwen3-VL-8B)	55.23	38.30	57.07	37.64	58.34	55.21	52.33	55.55	52.05	53.41
$Stage_1$ (w/o atomic operations)	61.94	40.96	62.58	40.25	37.84	58.19	53.85	59.84	53.17	38.97
$Stage_1$	72.26	41.66	71.20	41.62	44.76	70.64	54.50	68.89	54.58	55.09
$Stage_{1+2}$ (Consistency Reward)	75.82	47.36	74.58	46.25	61.90	74.63	56.58	73.12	57.04	71.30
Harmonia (ours)	**76.58**	**48.64**	**75.29**	**47.07**	**63.21**	**75.18**	**57.89**	**74.03**	**57.61**	**74.51**

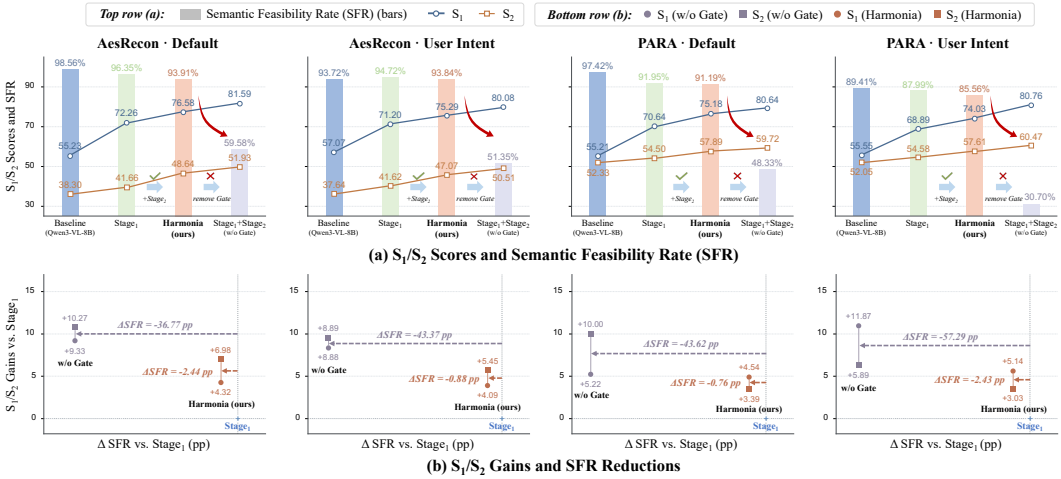

Figure 3: Effect of the semantic gate in VA-GRPO on aesthetic gains and semantic feasibility.

a learning rate of 5×10^{-5}, 10% warm-up, and BF16. $Stage_2$ performs VA-GRPO with $K = 8$ candidates per input, LoRA rank 64, 9 epochs, and a learning rate of 5×10^{-6}. We use Qwen3.7-Plus for semantic gating. All experiments are conducted on 10 NVIDIA A40 GPUs (48GB each).

4.3 Main Results

Default APR. As shown in Table 1, Harmonia achieves the best results across both synthetic and real-world subsets under Default and Personalized APR. Under Default APR, Harmonia achieves S_1/S_2 scores of 76.58/48.64 on AesRecon and 75.18/57.89 on PARA. Compared with AesFormer, Harmonia improves S_1 by 8.76 and 13.48 on the two subsets, respectively. It also surpasses AesThinker by 5.73 and 6.41 in S_1, while outperforming strong proprietary VLMs such as GPT-5.6 and Gemini 3. These improvements are accompanied by the best S_2 scores, showing that stronger reconstruction translates into higher final aesthetic quality.

Personalized APR. Harmonia achieves $S_1/S_2/S_3$ scores of 75.29/47.07/63.21 on AesRecon and 74.03/57.61/74.51 on PARA under Personalized APR. In particular, its S_3 on real-world photos exceeds the strongest Gemini 3 by 9.52 (74.51 vs. 64.99). Compared with AesFormer, Harmonia improves S_3 by 8.99 and 20.41 on the synthetic and real-world subsets, respectively, while also achieving higher S_1 and S_2. These results show that Harmonia can adapt photographic decisions to diverse intents without sacrificing reconstruction quality. Notably, these gains are achieved using only approximately 30% of AesFormer's training pairs and without fine-tuning the image editor.

Qualitative Results. Figure 4 further illustrates these findings. Under Default APR, several baselines either make only limited structural improvements or produce aggressive edits that violate subject identity or scene semantics. In contrast, Harmonia achieves substantial yet semantically feasible reconstruction, improving aesthetics while preserving subject and scene. Under Personalized APR, Harmonia reconstructs the same input into a full-body, left-positioned composition for Intent A and

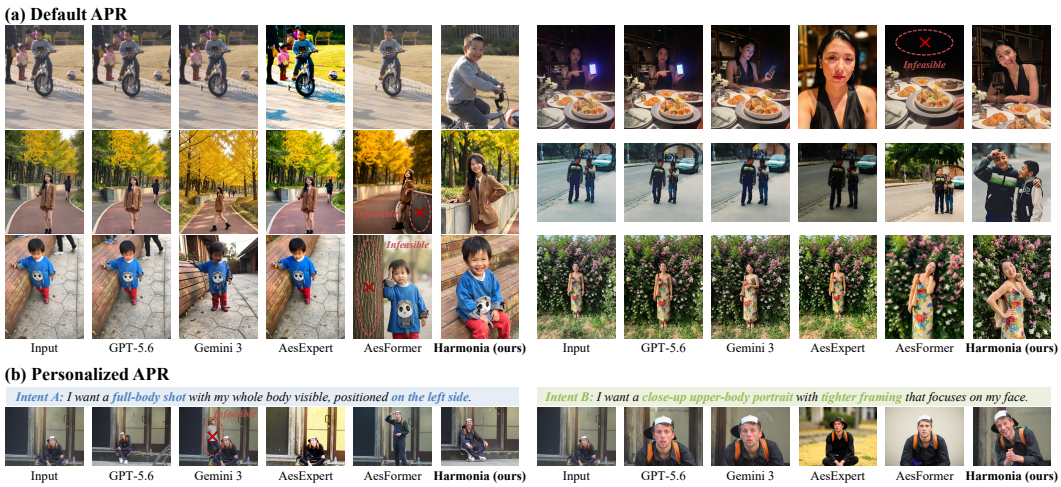

Figure 4: Qualitative comparison on Default and Personalized APR. Harmonia yields (a) greater aesthetic gains with feasible edits and (b) distinct reconstructions aligned with different user intents.

a tighter upper-body portrait for Intent B, demonstrating its ability to adapt photographic structure to different user intents rather than applying a fixed reconstruction strategy.

4.4 Ablation Studies

Do atomic operations improve planning? As shown in Table 2, $Stage_1$ consistently outperforms the Qwen3-VL-8B baseline across both subsets and settings, demonstrating the effectiveness of structured photographic planning. Removing atomic operations leads to clear performance degradation, indicating that decomposing free-form plans into explicit and composable photographic operations facilitates more effective fine-grained reconstruction decisions.

Does VA-GRPO further improve APR? Building on $Stage_1$, $Stage_2$ further improves all metrics. Under Default APR, it improves S_1/S_2 over $Stage_1$ by 4.32/6.98 on AesRecon and 4.54/3.39 on PARA, demonstrating the benefit of optimizing the planner from realized visual outcomes. Directly incorporating semantic consistency into the reward performs consistently worse. Since unchanged or conservative reconstructions can achieve high semantic consistency with little aesthetic gain, semantic preservation is better treated as a feasibility constraint rather than an objective to be maximized.

Why is the semantic gate necessary? As shown in Figure 3, removing the gate yields larger S_1/S_2 gains but causes a substantial drop in SFR. On AesRecon under Default APR, SFR decreases from 96.35% after $Stage_1$ to 59.58% without the gate, while Harmonia retains 93.91% SFR together with S_1/S_2 gains of 4.32/6.98. The same pattern holds across all other settings, where removing the gate reduces SFR by 43.37–57.29 pp, compared with only 0.76–2.43 pp for Harmonia. Although Harmonia incurs a slight SFR decrease, it yields substantial aesthetic gains while maintaining high semantic feasibility, and user studies in Appendix A.2 and A.3 further show that human participants consistently prefer its reconstructions in real-world settings. These results demonstrate the effectiveness of VA-GRPO in maximizing aesthetic gains within the semantic elasticity boundary.

5 Conclusion

In this work, we propose **Harmonia** for Aesthetic Photo Reconstruction, addressing the inherent aesthetic–semantic tension by maximizing aesthetic gains within a photo-dependent **semantic elasticity boundary**. Harmonia decomposes reconstruction plans into atomic photographic operations and introduces **VA-GRPO** to optimize the planner from realized visual outcomes while preserving semantic feasibility. We further construct **Harmonia-Bench** for evaluating both default and personalized APR. Experiments demonstrate that Harmonia outperforms existing methods using only approximately 30% of AesFormer's training pairs and even without fine-tuning the image editor, while additionally supporting personalized APR for the first time.

AI USE STATEMENT

In this work, we used generative AI tools to assist in constructing a subset of the training data and to polish parts of the manuscript at an early stage. All AI-assisted content was reviewed and verified by the authors. We did not use generative AI tools for other tasks such as developing theoretical or conceptual frameworks or designing research methodology or experiments. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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.

Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011*, pp. 97–104. IEEE, 2011.

Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, et al. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15313–15322, 2026.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 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. *arXiv preprint arXiv:2507.06261*, 2025.

Tianxiang Du, Hulingxiao He, and Yuxin Peng. Aesformer: Transform everyday photos into beautiful memories. *arXiv preprint arXiv:2605.22126*, 2026a.

Tianxiang Du, Hulingxiao He, and Yuxin Peng. Venus: Benchmarking and empowering multimodal large language models for aesthetic guidance and cropping. *arXiv preprint arXiv:2602.23980*, 2026b.

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.

Niladri Shekhar Dutt, Duygu Ceylan, and Niloy J Mitra. Monetgpt: Solving puzzles enhances mllms' image retouching skills. *ACM Transactions on Graphics (TOG)*, 44(4):1–12, 2025.

Yipo Huang, Xiangfei Sheng, Zhichao Yang, Quan Yuan, Zhichao Duan, Pengfei Chen, Leida Li, Weisi Lin, and Guangming Shi. Aesexpert: Towards multi-modality foundation model for image aesthetics perception. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5911–5920, 2024a.

Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Haoning Wu, Pengfei Chen, Yuzhe Yang, Leida Li, and Weisi Lin. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. *arXiv preprint arXiv:2401.08276*, 2024b.

Qiaoqiao Jin, Xuanhong Chen, Meiguang Jin, Ying Chen, Rui Shi, Yucheng Zheng, Yupeng Zhu, and Bingbing Ni. Toward tiny and high-quality facial makeup with data amplify learning. In *European Conference on Computer Vision*, pp. 340–356. Springer, 2024.

Jinwoo Kim, Jihye Yoo, and Seon Joo Kim. Learning personalized photographic style from pairwise user preferences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1134–1144, 2026.

Yoonjeon Kim, Soohyun Ryu, Yeonsung Jung, Hyunkoo Lee, Joowon Kim, June Yong Yang, Jaeryong Hwang, and Eunho Yang. Preserve or modify? context-aware evaluation for balancing preservation and modification in text-guided image editing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23474–23483. IEEE, 2025.

Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European conference on computer vision*, pp. 662–679. Springer, 2016.

Black Forest Labs. Flux. 2: Frontier visual intelligence. *https://bfl. ai/blog/flux-2*, 2025.

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.

Fan Li, Chonghuinan Wang, Lina Lei, Yuping Qiu, Jiaqi Xu, Jiaxiu Jiang, Xinran Qin, Zhikai Chen, Fenglong Song, Zhixin Wang, et al. Hp-edit: A human-preference post-training framework for image editing. *arXiv preprint arXiv:2604.19406*, 2026.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

Jie Liang, Hui Zeng, Miaomiao Cui, Xuansong Xie, and Lei Zhang. Ppr10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 653–661. IEEE, 2021.

Yunlong Lin, Zixu Lin, Kunjie Lin, Jinbin Bai, Panwang Pan, Chenxin Li, Haoyu Chen, Zhongdao Wang, Xinghao Ding, Wenbo Li, et al. Jarvisart: Liberating human artistic creativity via an intelligent photo retouching agent. *Advances in Neural Information Processing Systems*, 38: 52088–52130, 2026a.

Yunlong Lin, Linqing Wang, Kunjie Lin, Zixu Lin, Kaixiong Gong, Wenbo Li, Bin Lin, Zhenxi Li, Shiyi Zhang, Yuyang Peng, et al. Jarvisevo: Towards a self-evolving photo editing agent with synergistic editor-evaluator optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27291–27302, 2026b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 2408–2415. IEEE, 2012.

Wenqi Ouyang, Yi Dong, Xiaoyang Kang, Peiran Ren, Xin Xu, and Xuansong Xie. Rsfnet: A white-box image retouching approach using region-specific color filters. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12126–12135. IEEE, 2023.

Daiqing Qi, Handong Zhao, Jing Shi, Simon Jenni, Yifei Fan, Franck Dernoncourt, Scott Cohen, and Sheng Li. The photographer's eye: Teaching multimodal large language models to see and critique like photographers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24807–24816. IEEE, 2025.

Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J Foran. Personalized image aesthetics. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 638–647. IEEE, 2017.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Xiangfei Sheng, Pangu Xie, Weidong Zou, Pengfei Chen, Tong Zhu, and Leida Li. Instructcrop: Teaching multimodal large language models to crop aesthetic images. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6830–6839, 2025.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5903–5911, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

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*, 2026.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13294–13304. IEEE, 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.

Anyi Xu, B Li, Bangcai Lin, Bing Xue, BingCheng Xian, Bingzheng Xu, Bochao Wu, Bowei Zhang, Boyi Deng, CC Yu, et al. Deepseek-v4.1-flash: Pushing the limits of kv cache compression. *arXiv preprint arXiv:2609.19969*, 2026.

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.

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*, 2026a.

Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. Personalized image aesthetics assessment with rich attributes. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19829–19837. IEEE, 2022.

Zhichao Yang, Jianjie Wang, Zhixianhe Zhang, Pangu Xie, Xiangfei Sheng, Pengfei Chen, and Leida Li. Fine-grained image aesthetic assessment: Learning discriminative scores from relative ranks. *arXiv preprint arXiv:2603.03907*, 2026b.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26125–26135. IEEE, 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, 2023.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, et al. Qwen-image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026.

Jun Zhou, Jiahao Li, Zunnan Xu, Hanhui Li, Yiji Cheng, Fa-Ting Hong, Qin Lin, Qinglin Lu, and Xiaodan Liang. Fireedit: Fine-grained instruction-based image editing via region-aware vision language model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13093–13103. IEEE, 2025.

Zhaokun Zhou, Qiulin Wang, Bin Lin, Yiwei Su, Rui Chen, Xin Tao, Amin Zheng, Li Yuan, Pengfei Wan, and Di Zhang. Uniaa: A unified multi-modal image aesthetic assessment baseline and benchmark. *arXiv preprint arXiv:2404.09619*, 2024.

A Appendix

A.1 Human Validation of Automatic Evaluation Metrics

APR is intended for real-world photographic applications, where reconstruction quality ultimately depends on human judgments of photographic structure, aesthetic quality, and satisfaction of personalized intents. To validate the evaluation protocol of Harmonia-Bench, we conduct a pairwise user study on $\mathbf{S}_1$: Structural Improvement, $\mathbf{S}_2$: Aesthetic Quality, and $\mathbf{S}_3$: Intent Fulfillment.

Experimental Setup. We construct 400 reconstruction pairs from Harmonia-Bench, evenly covering four evaluation settings: AesRecon Default, AesRecon Personalized, PARA Default, and PARA Personalized, with 100 pairs per setting. For each pair, given the same input image I_p, we select two reconstructions I_e^A and I_e^B generated by different methods:

$$\mathcal{P} = (I_p, I_e^A, I_e^B). \tag{11}$$

For Personalized APR, the two candidates additionally share the same user intent U:

$$\mathcal{P} = (I_p, U, I_e^A, I_e^B). \tag{12}$$

We sample results from methods with different performance levels and include pairs spanning different automatic-score gaps, avoiding evaluation dominated by either trivially separable or nearly indistinguishable cases. Participants then compare the two reconstructions separately under each

Table 3: Human–evaluator agreement across different Harmonia-Bench settings. S_1 and S_2 are evaluated on all 400 pairs, while S_3 is evaluated on 200 Personalized APR pairs.

Dataset	Setting	# Pairs	S1	S2	S3
AesRecon	Default	100	88.0%	85.0%	–
AesRecon	Personalized	100	90.0%	87.0%	91.0%
PARA	Default	100	87.0%	86.0%	–
PARA	Personalized	100	89.0%	88.0%	92.0%
Overall	–	–	**88.5%**	**86.5%**	**91.5%**

Table 4: Human–evaluator agreement under different automatic-score gaps.

Metric	Small	Medium	Large	Overall
S1: Structural Improvement	81.2%	89.5%	**94.8%**	88.5%
S2: Aesthetic Quality	78.2%	87.2%	**94.0%**	86.5%
S3: Intent Fulfillment	83.6%	93.9%	**97.0%**	91.5%

metric. (1) For S_1, they determine which reconstruction achieves greater photographic-structure improvement over the input, considering factors such as composition, camera viewpoint, human pose, framing, and spatial relationships. (2) For S_2, they select the reconstruction with higher overall aesthetic quality. (3) For Personalized APR, S_3 is additionally evaluated by asking which reconstruction better satisfies the given user intent U.

Participants. A total of 10 participants take part in the study, including general users, photography enthusiasts, computer vision researchers, PhD students in AI, and imaging professionals from major technology companies. Each pair is independently evaluated by three participants using a forced binary choice, and the final human preference is determined by majority voting. Method identities and automatic scores are hidden, and the presentation order of the two candidates is randomized to reduce positional bias. For a metric S_m, if $S_m^A > S_m^B$ and the majority of participants prefer I_e^A, or vice versa, the pair is counted as a human–evaluator agreement.

Overall Agreement. As shown in Table 3, all three metrics achieve high agreement with human judgments, supporting the reliability of the Harmonia-Bench evaluation protocol. S_1 and S_2 achieve overall agreement rates of 88.5% and 86.5% over all 400 pairs, respectively, while S_3 reaches 91.5% over the 200 Personalized APR pairs. The results are also consistent across both datasets and settings, indicating that the agreement is not specific to a particular data source or task setting. Specifically, (1) the 88.5% agreement of S_1 shows that the automatic evaluator can reliably distinguish which reconstruction provides greater photographic-structure improvement. (2) S_2 achieves 86.5% agreement; despite the subjectivity of overall aesthetic quality, its induced ranking remains highly consistent with majority human judgments. (3) S_3 achieves the highest agreement of 91.5%, showing that it can effectively assess how well different reconstructions satisfy a given user intent. Overall, S_1, S_2, and S_3 consistently reflect human judgments of the relative quality of APR reconstructions, validating their use as automatic evaluation metrics in Harmonia-Bench.

Agreement under Different Score Gaps. We further analyze how human–evaluator agreement varies with the difference between automatic scores. For each metric S_m, we define the score gap between the two reconstructions as

$$\Delta S_m = \left| S_m^A - S_m^B \right|. \tag{13}$$

Based on ΔS_m, we divide the pairs into *Small*, *Medium*, and *Large* groups, corresponding to cases with small, moderate, and large score differences, respectively. As shown in Table 4, agreement with human judgments consistently increases with the score gap. For Small-gap pairs, the agreement rates of S_1, S_2, and S_3 are 81.2%, 78.2%, and 83.6%, respectively. They increase to 89.5%, 87.2%, and 93.9% for Medium-gap pairs, and further reach 94.8%, 94.0%, and 97.0% for Large-gap pairs. This trend shows that when the evaluator identifies a clearer quality difference between two reconstructions, human observers are more likely to perceive the same preference ordering. Notably, even for the most challenging Small-gap pairs, all three metrics maintain agreement rates close to or above 80%, while their overall agreement rates all exceed 86%. These findings demonstrate that

Table 5: Human preference and semantic acceptability of Harmonia compared with $Stage_1$ across different Harmonia-Bench settings.

Dataset	Setting	# Pairs	Harmonia Preference	Semantic Acceptability
AesRecon	Default	50	84.0%	94.0%
AesRecon	Personalized	50	86.0%	92.0%
PARA	Default	50	82.0%	94.0%
PARA	Personalized	50	90.0%	90.0%
Overall	–	200	**85.5%**	**92.5%**

S_1, S_2, and S_3 provide accurate and stable estimates of human preferences over APR reconstruction quality, further supporting the reliability of the Harmonia-Bench automatic evaluation protocol.

A.2 Human Preference

Figure 3 shows that, compared with $Stage_1$, VA-GRPO substantially improves reconstruction quality while causing only a slight decrease in SFR. Since APR ultimately targets real users, we further conduct a user study to examine whether the additional aesthetic gains brought by VA-GRPO make the final Harmonia reconstructions more preferred than $Stage_1$ outputs.

Experimental Setup. We sample 200 cases from the four Harmonia-Bench settings, with 50 cases from each setting. For each input image I_p, we present the reconstructions produced by $Stage_1$ and the full Harmonia model, denoted as I_e^{S1} and I_e^{H}, respectively. For Personalized APR, the corresponding user intent U is additionally provided. We evaluate from two perspectives: *(1) Overall Preference:* participants consider both aesthetic improvement and preservation of the original subject, clothing, and scene, and select which reconstruction they prefer as the final result. *(2) Semantic Acceptability:* participants judge whether the Harmonia result can still reasonably be regarded as a new photograph of the same subject and real-world scene, directly assessing whether its semantic changes remain acceptable to human observers. Each case is independently evaluated by three participants, and the final judgment is determined by majority voting.

Results. As shown in Table 5, Harmonia is consistently preferred by humans across all four settings when considering both aesthetic improvement and semantic preservation, achieving an overall Harmonia Preference of **85.5%**. This shows that users prefer the reconstructions further optimized by VA-GRPO over those produced by $Stage_1$ alone, indicating that its additional aesthetic gains are meaningful in human perception. Meanwhile, **92.5%** of Harmonia reconstructions are judged acceptable by human participants, showing that the vast majority can still be reasonably regarded as new photographs of the same subject and scene. Overall, Harmonia substantially improves human preference while maintaining high acceptability, further supporting our design of treating semantic preservation as a feasibility constraint and maximizing aesthetic gains within the feasible space.

A.3 Human Preference Comparison with AesFormer

To further verify whether the improvements of Harmonia align with human aesthetic preferences, we conduct an additional human preference study against AesFormer (Du et al., 2026a).

Evaluation Data. We construct an evaluation set of 200 samples. Specifically, 100 samples are randomly selected from Harmonia-Bench, and another 100 are real-world photographs collected from users and taken in everyday-life settings. Each collected photograph is manually annotated with at least one explicit aesthetic issue, such as suboptimal composition, improper subject placement, undesirable framing, or unnatural pose. This additional set complements Harmonia-Bench with photographs captured under practical, real-world shooting conditions.

Evaluation Protocol. Three participants independently compare the reconstruction results of Harmonia and AesFormer (Du et al., 2026a). For each input photo, participants are shown the two reconstructed images in a blind pairwise comparison and asked to select the one with better overall photographic quality, considering whether the original aesthetic issues are effectively addressed, the final aesthetic quality, and the preservation of key semantic content.

Results. Harmonia achieves a human preference rate of **78%**, compared with **22%** for AesFormer (Du et al., 2026a). The clear preference for Harmonia across both Harmonia-Bench samples and collected real-world photographs provides further evidence that the improvements observed under automatic evaluation are consistent with human judgments, and demonstrates the effectiveness of Harmonia in addressing structural aesthetic issues in practical photography scenarios.