EMPOWERING A MULTIMODAL AESTHETIC AGENT FOR PHOTOGRAPHIC REFERENCE GRADING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reference-guided photographic color grading helps users realize their aesthetic intent by selectively drawing on the color and tonal characteristics of a reference photograph. Doing so requires understanding which characteristics users wish to adopt and translating these preferences into controllable edits. However, image-only reference methods align overall appearance without instruction-grounded selection, while pixel-space generative editors typically produce flattened images without editable parameters, limiting subsequent refinement. To address these challenges, we introduce ***JarvisAura***, a multimodal aesthetic agent that formulates photographic color grading as structured decision-making in an editable parameter space. JarvisAura offers three key advantages: (1) an aesthetic agent paradigm that integrates visual attribute perception, strategy-level reasoning, and parameterized execution for interpretable and controllable photographic color grading; (2) a unified reasoning-to-execution pipeline that connects aesthetic planning with backend rendering and editable parameter refinement; and (3) a three-stage training strategy comprising Parameter-Grounded Continual Pre-Training for parameter–appearance grounding, Reference-Grounded Supervised Fine-Tuning for reference-guided reasoning and parameter decisions, and Reference-Aligned Group Relative Policy Optimization for rendered-outcome refinement with multidimensional rewards and dynamic data derivation. On Ref-Bench, JarvisAura achieves the best scores among methods with available results across all reported automatic metrics and a 74% human-rated task-success rate, exceeding GPT-Image-2.5 by 9 percentage points.

1 INTRODUCTION

Photographic color grading helps users realize their aesthetic intent by shaping the color and atmosphere of their images. Reference photographs provide concrete examples of color relationships, saturation, and tonal appearance Pitié & Kokaram (2007); Ke et al. (2023); Larchenko et al. (2025). Yet users may want to borrow only part of a reference's appearance, such as its warm palette but not its strong contrast. Natural-language instructions can express these preferences. We study reference-guided photographic color grading in which instructions determine which reference characteristics to adopt and how to adapt them to a source photograph. The resulting adjustments should remain editable so that users can continue refining the image.

Existing methods support different parts of this process. Image-only reference transfer methods use color distribution matching, statistical transformation, or color-space mapping Larchenko et al. (2025); Huang et al. (2023); Pouli & Reinhard (2010); Pitié & Kokaram (2007), as well as deep feature transformation Wen et al. (2023); Chiu & Gurari (2022); Ho & Zhou (2021), to change an image according to a reference. They generally lack natural-language input, leaving users unable to specify which reference characteristics they want to adopt. Generative image editors can use both reference images and textual instructions Team et al. (2023); Hurst et al. (2024); Wu et al. (2025b), but pixel synthesis may introduce unintended content changes or local inconsistencies. Their outputs also typically lack editing parameters, making further parameter-level adjustments difficult within a photographic post-processing workflow. Text-driven parameterized editors produce editable operations from natural-language requests Lin et al. (2025a;b); Wu et al. (2026), but generally do not use a visual reference. Users must therefore describe the desired appearance in words alone, which can be difficult when their preferences involve subtle color relationships or tonal effects.

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.

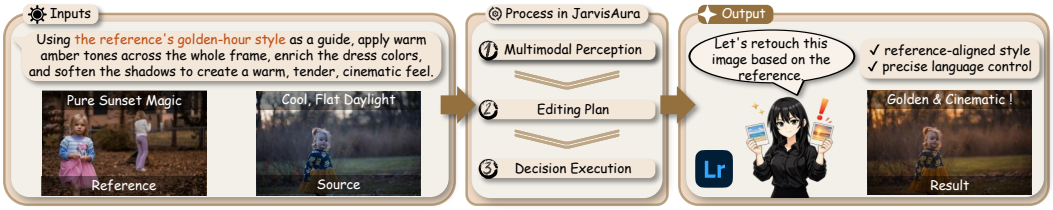


Figure 1: JarvisAura takes a source image, a reference image, and a user instruction, and produces a reference-driven color-grading result with an editable parameter representation. It identifies instruction-relevant grading attributes in the reference, determines which attributes to emphasize, and generates executable editing operations.

We introduce ***JarvisAura***, a multimodal aesthetic agent for reference-guided photographic color grading. Given a source image, a reference photograph, and a user instruction, JarvisAura reasons about which reference characteristics to adopt and which adjustments are needed to achieve them on the source image. It then converts the grading plan into executable, editable parameters, as shown in Fig. 1. Adobe Lightroom serves as a representative rendering backend. The predicted parameters remain available for inspection and adjustment, allowing users to refine the rendered result.

We train JarvisAura in three stages. Parameter-Grounded Continual Pre-Training (PG-CPT) establishes how grading parameters affect image appearance. For this stage, we construct Ref-CPT with parameter-anchored descriptions of visual state changes. Reference-Grounded Supervised Fine-Tuning (RG-SFT) then teaches the agent to choose parameters for a source–reference pair under a user instruction. Its training dataset, Ref-SFT, provides expert reasoning trajectories and editing configurations that connect the intended appearance to specific grading decisions.

Reference-Aligned Group Relative Policy Optimization (RA-GRPO) further refines these decisions using their rendered outcomes. We construct Ref-RL, an expert-guided reinforcement-learning dataset, and train with multidimensional outcome rewards. This stage also addresses a difference between expert-derived training data and practical user requests. In expert data, instructions and results are linked through known editing parameters. In practice, users describe the appearance they want without specifying the underlying parameters. RA-GRPO therefore incorporates dynamic data derivation to create additional training examples from rendered images. Acceptable rollouts that remain insufficiently aligned with the reference are reused as new source states and paired with expert-refined results. New instructions are generated from the resulting image triplets without access to editing parameters, grounding the instructions in visible appearance changes.

On Ref-Bench, JarvisAura achieves the best performance on all reported automatic metrics among methods with available results. Its human-rated task-success rate reaches 74%, exceeding GPT-Image-2.5 by 9 percentage points. Our contributions are summarized as follows:

- We introduce JarvisAura, a multimodal aesthetic agent that uses reference photographs and user instructions to plan adjustments for source images and produce executable, editable grading parameters.
- We design two data-construction pipelines for Ref-CPT and Ref-SFT, providing parameter-anchored state-change descriptions and expert reference-guided reasoning trajectories with editing configurations.
- We develop a three-stage training strategy combining PG-CPT, RG-SFT, and RA-GRPO. The reinforcement-learning stage uses multidimensional outcome rewards and dynamic data derivation to refine rendered results and address the mismatch between parameter-linked expert supervision and appearance-based user requests.

2 Related Work

Reference-guided Photographic Color Grading. Existing reference-guided methods for photographic color grading mainly follow two directions: image-only transfer and multimodal generative editing. Image-only methods formulate transfer through color distribution matching, statisti-

cal transformation, color-space mapping Park et al. (2016); Huang et al. (2023); Larchenko et al. (2025); Pouli & Reinhard (2010); Pitié & Kokaram (2007), or goal-conditioned optimization Wu et al. (2024), while deep methods transform features or learn color-style correspondences Ho & Zhou (2021); Wen et al. (2023); Hong et al. (2021); Ke et al. (2023); Yoo et al. (2019); Chiu & Gurari (2022); Li et al. (2017; 2018); An et al. (2020). Although they exploit references for global appearance adjustment, they provide limited language interaction and attribute-level control. Multimodal generative editors condition on multiple images and textual instructions Wu et al. (2025b); Team et al. (2023); Hurst et al. (2024), with recent work further incorporating reasoning Yang et al. (2026). However, pixel-space synthesis may alter content, introduce local inconsistencies, and produce unstable results. These methods also typically output flattened images rather than executable parameters, hindering professional workflow integration and subsequent refinement. Thus, existing approaches do not jointly support visual references, instruction-guided emphasis, and editable parameterized outputs.

Task-specific Agents. The development of multimodal foundation models has advanced task-specific agents in visual interaction, knowledge acquisition, and image processing. GUI agents execute software operations based on screen content and natural-language instructions Zhou et al. (2025); Yan et al. (2025); Liu et al. (2025); Wang et al. (2025a); Qin et al. (2025), while knowledge acquisition and data analysis agents produce structured results through retrieval, reasoning, and tool use Zhang et al. (2025); Narayan et al. (2025); Wu et al. (2025c). In image processing, recent studies have explored operation awareness Dutt et al. (2025), parameterized editing workflows Chen et al. (2025), natural-language-driven editable parameter generation Lin et al. (2025a), reward-based editing optimization Wu et al. (2026), and editor–evaluator co-evolution Lin et al. (2025b). However, existing methods mainly rely on textual instructions or predefined objectives, limiting their ability to exploit visual grading evidence from reference images. We address this gap by generating executable, parameterized grading results from multimodal references for photographic color grading.

3 METHOD

3.1 AGENT FORMULATION

JarvisAura is a multimodal aesthetic agent for Reference-guided photographic reference Grading, as shown in Fig 1. Formally, the agent is defined as:

$$(C, P) = F_\theta(I_{\mathrm{src}}, I_{\mathrm{ref}}, Q), \tag{1}$$

where $(I_{\mathrm{src}}, I_{\mathrm{ref}})$ denotes the source–reference image pair, Q the user instruction, C the grading analysis and plan, P the executable parameter configuration, and F_θ JarvisAura. The configuration is rendered by an editing backend as:

$$I_{\mathrm{out}} = \mathcal{R}_{\mathcal{B}}(I_{\mathrm{src}}, P), \tag{2}$$

where I_{out} denotes the rendered image and $\mathcal{R}_{\mathcal{B}}$ the renderer of backend $\mathcal{B}$. We instantiate $\mathcal{B}$ with Adobe Lightroom, while accommodating other backends by mapping P to their parameter spaces or adapting the agent via data construction.

This formulation shows the limitations of existing methods. Image-only transfer methods Larchenko et al. (2025) omit Q, limiting control over user-specified attributes. Generative editors Hurst et al. (2024) directly synthesize I_{out}, potentially altering source content, without providing editable parameters P for subsequent refinement. Text-driven parameterized methods Lin et al. (2025a) accept Q and predict P but lack I_{ref}, limiting visual grounding of target grading attributes.

3.2 DATA CONSTRUCTION

We build Ref-CPT and Ref-SFT through dedicated data-generation pipelines for continual pre-training and supervised fine-tuning, respectively. Ref-RL consists of samples refined by human experts, with data collection and quality-control details provided in Appendix A.

3.2.1 DATA GENERATION FOR CONTINUAL PRE-TRAINING.

To ground parameter–appearance correspondences, Ref-CPT combines broad parameter-space coverage with parameter-grounded state-change captions, as illustrated in Fig. 2.

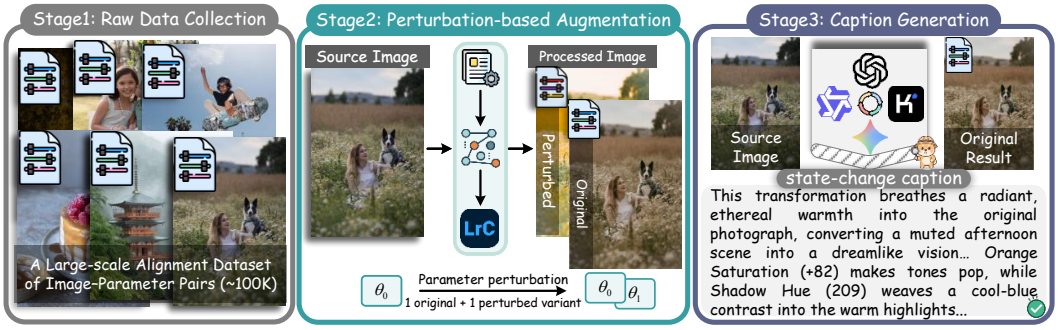


Figure 2: Ref-CPT construction. The pipeline collects community edit pairs with configurations, renders results under original and perturbed settings, and generates parameter-anchored state-change captions for continual pre-training.

Data Collection and Parameter Space Enhancement. To obtain large-scale data, we collect approximately 100K image–configuration pairs from online creator communities, in compliance with their terms of use and for academic research only. To scale up samples and broaden parameter-space coverage, we perturb editing parameters and clamp them to valid ranges, then render results under both original and perturbed settings via an open-source protocol Lin et al. (2025a). Through this augmentation, we construct about 200K high-quality triplets of original images, parameterized configurations, and resulting images. This scale sufficiently covers the parameter space, enabling the dataset to capture how parameter values affect appearance in trend and relative strength, and supporting the model's learning of professional parametric retouching knowledge.

Automated State-Transition Captioning. With large-scale sample pairs, we use advanced MLLMs for automated annotation. To generate professionally grounded, information-dense captions, we design concise prompts (see Appendix A) that jointly analyze the original image, parameterized configuration, and the rendered result. Unlike generic semantic image descriptions, Ref-CPT annotations focus on a fine-grained characterization of the retouching trajectory. To improve annotation quality and diversity, we use four multimodal models, Kimi-K2.5 Team et al. (2026), Qwen3-VL-235 Bai et al. (2025), Gemini-3-Pro Team et al. (2023), and GPT-5.2 Hurst et al. (2024), with Qwen3-VL-235 from the same family as our backbone to align annotation style and terminology and reduce supervision mismatch. We rotate among these models when generating captions and enforce a unified prompting scheme for output format and terminology, yielding high-quality, diverse state-change captions. To ensure annotation consistency, we use a shared VLM-as-Judge verifier to assess caption validity and faithfulness to the underlying edits, with details provided in Appendix A.

3.2.2 Data Generation for Supervised Fine-tuning.

Following Fig. 3, we construct Ref-SFT to model expert-style reasoning and parameter decisions for SFT. Using the notation above, each sample is represented as $\langle I_{\text{src}}, I_{\text{ref}}, I_{\text{out}}, Q, C, P \rangle$.

Data Generation and Reference Construction. This stage builds visually informative reference relations from large-scale data. We integrate MIT-Adobe FiveK Bychkovsky et al. (2011), PPR10K Liang et al. (2021), other licensed sources, obtaining image pairs together with their parameterized configurations. To ensure content match and style transferability, we use a two-stage retrieval strategy: 1) **In semantic retrieval**, we generate content summaries with Gemini-3-Pro and use Qwen3-VL-Embedding Li et al. (2026) to encode fused image-summary representations, retrieving the top 100 candidates. 2) **In style reranking**, we summarize candidates' style attributes and compute similarity using image-style embeddings, selecting the top 5 as reference originals. Finally, we apply the source parameterized configuration to the selected reference originals to obtain the reference image I_{ref}. To incorporate human aesthetic judgment and ensure data quality, we apply human-in-the-loop screening to filter mismatched or low-quality samples, completing $\langle I_{\text{src}}, I_{\text{ref}}, I_{\text{tgt}}, P \rangle$.

User-Instruction Generation. In this stage, we use Gemini-3.1-Pro Team et al. (2023) with role-playing prompting to elicit user-like instructions. To promote expression diversity, we add a simple sliding-window memory that discourages repetitive phrasing when generating reference-grounded

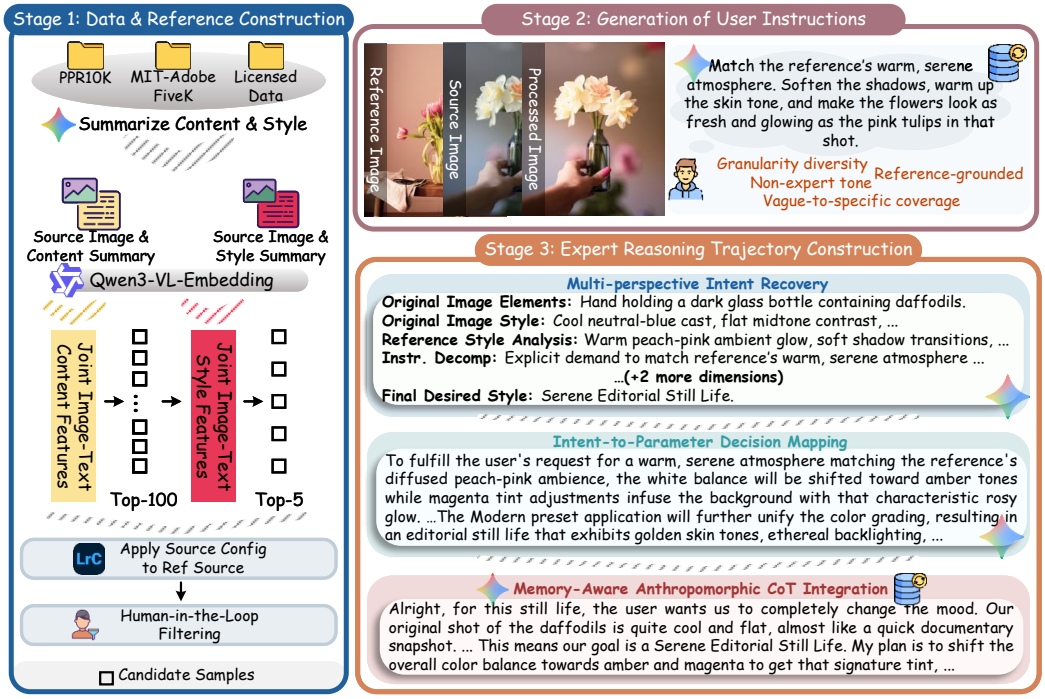


Figure 3: Ref-SFT construction. The pipeline builds source–reference pairs, generates reference-grounded instructions, and synthesizes expert trajectories for planning and parameter decisions.

instructions. These instructions vary in granularity, wording, and emphasis, covering interactions from vague aesthetic requests to specific functional edits. Prompt templates are provided in Appendix A. In this step, we map $\langle I_{\text{src}}, I_{\text{ref}}, I_{\text{tgt}}, P\rangle$ to $\langle I_{\text{src}}, I_{\text{ref}}, I_{\text{tgt}}, Q, P\rangle$.

Expert Reasoning Trajectory Construction. This key stage aims to capture expert reasoning via progressive prompting. First, the model decomposes the task using cues such as source content, reference style, and user instruction to infer the underlying intent. Then, using the parameterized configuration, we map the inferred intent into structured decision logic that specifies parameter edits and expected visual effects, linking numeric adjustments to aesthetic goals. Finally, role-playing prompts integrate the intent analysis and decision logic into human-like reasoning trajectories. We introduce a memory mechanism, similar to instruction generation, to avoid templated outputs and keep CoTs structured while diversifying colloquial phrasing. These steps yield approximately 50K CoT-supervised Ref-SFT samples of the form $\langle I_{\text{src}}, I_{\text{ref}}, I_{\text{out}}, Q, C, P\rangle$.

3.3 Training Strategy

3.3.1 Parameter-Grounded Continual Pre-training (PG-CPT).

Pre-training corpora for general-purpose multimodal models underrepresent professional photographic color grading, leaving them without reliable parameter–appearance correspondences. Following the *Superficial Alignment Hypothesis* Zhou et al. (2023); Zhuang et al. (2025), we inject this knowledge through Parameter-Grounded Continual Pre-training on Ref-CPT. State-change captions provide explicit supervision by describing appearance changes induced by parameterized adjustments. This enables the model to learn: 1) visual effects of each parameter across values; 2) parameter interactions under joint adjustment; and 3) content-dependent variation in parameter effects.

3.3.2 Reference-Grounded Supervised Fine-tuning (RG-SFT).

Parameter knowledge alone cannot turn the source image and multimodal reference into grading decisions. We therefore perform Reference-Grounded Supervised Fine-Tuning (RG-SFT) on Ref-SFT, supervising reasoning trajectories and executable parameters to learn reference-guided decisions.

3.3.3 REFERENCE-ALIGNED GROUP RELATIVE POLICY OPTIMIZATION (RA-GRPO).

Improving parameter accuracy and reducing instruction mismatch with limited expert annotation remain challenging. Building on CPT and SFT, we propose RA-GRPO, which combines multidimensional outcome rewards with dynamic data derivation, as shown in the lower half of Fig. 4.

Reinforcement Learning for Accurate Decisions. User evaluation is anchored to the reference image, placing stringent demands on parameter accuracy. SFT primarily fits demonstrated decisions and does not directly optimize them through rendering feedback. We therefore apply Group Relative Policy Optimization (GRPO) to optimize candidate outputs sampled for each multimodal input. Given an input x, we sample G candidate outputs $\{y_i\}_{i=1}^{G}$ from the old policy $\pi_{\theta_{\text{old}}}$ and compute rewards from their rendered results. The objective is defined as:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min\left(\rho_{i,t}(\theta)A_i,\ \text{clip}\left(\rho_{i,t}(\theta), 1-\epsilon, 1+\epsilon\right)A_i\right)\right], \tag{3}$$

where $\rho_{i,t}(\theta)$ is the probability ratio between the new and old policies for the t-th token, and A_i is the relative advantage computed from the group rewards:

$$A_i = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^{G})}{\text{std}(\{R_j\}_{j=1}^{G}) + \epsilon}. \tag{4}$$

The reward comprises three components, measuring task alignment, visual quality, and output validity, respectively, and is defined as $R_i = R_i^{\text{task}} R_i^{\text{qual}} R_i^{\text{fmt}}$.

1) Task-alignment reward. It combines reference matching with instruction following:

$$R_i^{\text{task}} = \underbrace{\text{RankScore}\Big(d\big(I_i^{\text{out}}, I_{\text{tgt}}\big); \big\{d\big(I_j^{\text{out}}, I_{\text{tgt}}\big)\big\}_{j=1}^{G}\Big)}_{\text{reference matching}} + \underbrace{\mathcal{V}_{\text{inst}}\Big(I_{\text{src}}, I_{\text{ref}}, I_i^{\text{out}}, Q, P^{(i)}\Big)}_{\text{instruction following}}, \tag{5}$$

where I_i^{out} and P_i denote candidate i's rendered result and predicted configuration, respectively, and I_{tgt} the target. The distance $d(\cdot,\cdot)$ is the Earth Mover's Distance between normalized Lab a–b color histograms. $\text{RankScore}(\cdot)$ maps distances to within-group scores, assigning higher scores to smaller distances. $\mathcal{V}_{\text{inst}}(\cdot)$ provides parameter-aware instruction feedback, grounding evaluation in executable decisions and mitigating hallucinations. Detailed protocols appear in Appendix B.

2) Visual-quality reward. It evaluates the intrinsic visual quality:

$$R_i^{\text{qual}} = Q_i = \frac{1}{100K}\sum_{k=1}^{K} q_{i,k}, \tag{6}$$

where $q_{i,k}$ is the score for the k-th quality criterion and $K = 14$ is the total number of criteria. The rubric uses four score levels, with their definitions and aggregation detailed in Appendix B.

3) Format-validity reward. To encourage structured and executable outputs, R_i^{fmt} evaluates output-format compliance and parameter validity. At the response level, it checks the structure of `<think>...</think>` and `<tool_call>`P`</tool_call>` blocks and the parseability of configuration P. At the parameter level, it verifies that parameter names are valid and values fall within their permitted ranges. Detailed checks and constraints are provided in Appendix B.

Dynamic Data Derivation. We use human-annotated expert samples for reinforcement learning, with the detailed construction procedure provided in the Appendix A. These samples encode an ideal parameter-mediated link between instructions and results. Real users, however, provide instructions without a predefined parameter link to the desired result. To bridge this gap, we derive training samples from rollout candidates. We retain candidates whose visual-quality reward satisfies $R_i^{\text{qual}} > 0.95$ and that show no apparent defect under any quality rubric. Among them, we select the

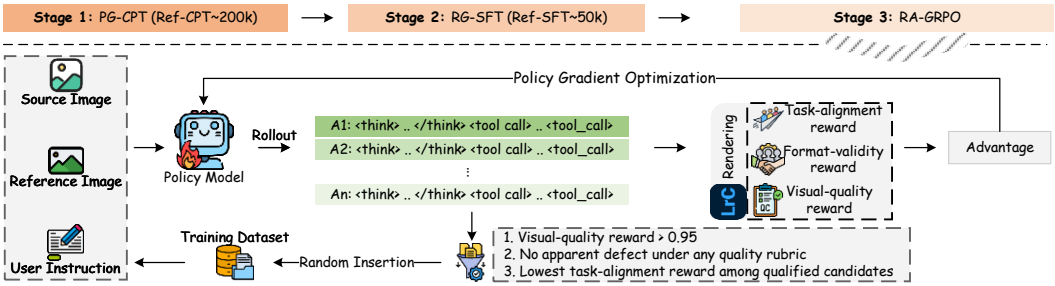


Figure 4: JarvisAura post-training framework: we first perform PG-CPT to ground color-grading parameter knowledge, then conduct RG-SFT to learn reference-guided reasoning and parameter decisions, and finally apply RA-GRPO with multidimensional rewards and dynamic data derivation to improve parameter accuracy and alignment with reference images and user instructions.

Table 1: Quantitative comparison on Ref-Bench (300 samples). We highlight the **best** and **second-best** results. SC, PQ, and O are evaluated by Gemini-3.1-Pro. Methods in the Ref.-Transfer group do not accept textual instructions.

Method	Instr.	$L1_{\times 10^2} \downarrow$	$L2_{\times 10^3} \downarrow$	CIEDE2000 ↓	SC ↑	PQ ↑	O ↑	HTSR ↑
ModFlows Larchenko et al. (2025)	✗	11.86	**16.89**	**11.16**	–	–	–	–
RLPixTuner Wu et al. (2024)	✗	17.16	51.46	18.84	–	–	–	–
UniPic-3 Wang et al. (2025b)	✓	22.12	89.15	23.61	4.87	3.86	4.21	0.09
FireRed-Image-Edit Team (2026)	✓	12.73	33.02	14.17	5.95	5.97	5.83	0.43
Qwen-Image-2.1 Wu et al. (2025a)	✓	13.60	36.76	15.03	5.89	6.20	5.91	0.38
JoyAI-Image Song et al. (2026)	✓	18.87	70.40	20.23	5.23	4.00	4.41	0.15
Nano-Banana-2 Team et al. (2023)	✓	12.59	32.07	13.90	6.20	6.96	6.49	0.55
GPT-Image-2.5 Hurst et al. (2024)	✓	**10.23**	20.47	11.67	**6.72**	7.18	**6.90**	**0.65**
Seedream-5-pro Seedream et al. (2025)	✓	10.97	24.04	12.37	6.43	**7.21**	6.75	0.58
JarvisAura	✓	**8.92**	**13.02**	**10.09**	**6.97**	**7.51**	**7.19**	**0.74**

candidate with the lowest task-alignment reward. Its rendered result is treated as a new source state and paired with the reference image and expert-refined result to form an image triplet for generating instructions without relying on editing parameters. Although refinement eases visual correction, removing parameter anchoring preserves intent-inference uncertainty and training value. The derived sample is randomly added to the training pool.

4 Experiment

4.1 Experimental Setup

Implementation details. We develop JarvisAura through three-stage post-training of Qwen3-VL-8B-Instruct Bai et al. (2025). (1) For PG-CPT, we train on Ref-CPT (∼200K state-change captions) for one epoch using ms-swift Zhao et al. (2024), with batch size 1 and learning rate 5×10^{-5}. (2) For RG-SFT, we fine-tune the PG-CPT checkpoint on Ref-SFT (∼50K CoT-supervised samples) for two epochs using LLaMAFactory Zheng et al. (2024), with batch size 1 and learning rate 5×10^{-6}. (3) For RA-GRPO, we train on Ref-RL (∼2K samples) using verl Sheng et al. (2025) with batch size 8, learning rate 5×10^{-6}, and 8 rollouts per input. Further details are provided in Appendix C.

Evaluation Setup. Ref-Bench contains 300 test instances with consensus-validated, expert-edited ground-truth targets, covering common photographic scenes and disjoint from the training data. We measure fidelity to expert targets using L1/L2 and CIEDE2000 (ΔE_{00}). Following prior evaluation protocols Lin et al. (2025a); Zhang et al. (2023), Gemini-3.1-Pro Team et al. (2023) scores Semantic Consistency (SC) and Perceptual Quality (PQ) on a 0–10 scale, with $O = \sqrt{\text{SC} \times \text{PQ}}$. With expert targets withheld, a blinded user study measures Human-rated Task Success Rate (HTSR) for

Table 2: A/B preference ratio for JarvisAura on cases where both JarvisAura and the compared baseline produce successful outputs. The ratio is defined as $(N_{\text{win}} + N_{\text{tie}})/(N_{\text{loss}} + N_{\text{tie}})$.

	GPT-Image-2.5	Seedream-5-pro	Nano-Banana-2	FireRed-Image-Edit
Preference ratio for ours ↑	1.12	1.30	1.47	1.53

JarvisAura and input-compatible generative baselines based on fine-grained instruction fulfillment, alignment with requested reference attributes, visual quality, and content preservation. More details are provided in Appendix C.

Baselines. We compare JarvisAura with two categories of baselines. Image-only transfer methods, including ModFlows Larchenko et al. (2025) and RLPixTuner Wu et al. (2024), receive only source–reference image pairs. Multimodal generative editors additionally receive user instructions and include Skyword Wang et al. (2025b), FireRed-Image-Edit Team (2026), Qwen-Image-2.1 Wu et al. (2025a), JoyAI-Image Song et al. (2026), Nano-Banana-2 Team et al. (2023), GPT-Image-2.5 Hurst et al. (2024), and Seedream-5-pro Seedream et al. (2025).

4.2 Experimental Results

Quantitative comparison. As shown in Table 1, JarvisAura achieves the best results across all reported automatic metrics. Lower L1, L2, and CIEDE2000 values indicate closer agreement with consensus-validated, expert-refined targets in terms of pixel-level error and perceptual color difference. Relative to ModFlows, JarvisAura reduces L2 and CIEDE2000 by approximately 22.9% and 9.6%, respectively. Compared with GPT-Image-2.5 under the same multimodal inputs, it reduces CIEDE2000 by approximately 13.5% while achieving higher SC (6.97 vs. 6.72) and O (7.19 vs. 6.90). Its PQ of 7.51 also exceeds the second-best score of 7.21 from Seedream-5-pro. These results demonstrate stronger target fidelity alongside improved semantic consistency and perceptual quality.

Qualitative comparison. As shown in Fig. 5, image-only transfer methods lack textual conditioning to capture users' emphasis on reference attributes. Some generative outputs blend source and reference content and deviate from the requested color attributes. In contrast, JarvisAura more accurately transfers instruction-specified reference attributes while preserving source content.

Human evaluation. To account for subjective aesthetic preferences, we assess task success and pairwise preference through a blinded user study, with expert targets withheld. JarvisAura achieves an HTSR of 74%, exceeding the strongest baseline, GPT-Image-2.5, by 9 percentage points (Table 1). For each baseline, we additionally conduct A/B evaluation on cases where both methods produce successful outputs. Preference ratios range from 1.12 to 1.53, favoring JarvisAura across all four comparisons (Table 2). These findings complement automatic metrics with human judgments of task success and relative preference. Detailed protocols are provided in Appendix C.

5 Ablation Study

Post-training stages. We compare RG-SFT alone, PG-CPT followed by RG-SFT, and the complete three-stage pipeline to examine the role of each training stage. RG-SFT establishes reference-grounded reasoning and parameter supervision. PG-CPT addresses the parameter–appearance correspondences required for accurate color and tone adjustment. Its inclusion reduces CIEDE2000 from 12.46 to 11.77 and raises O from 6.16 to 6.53. Relative to the PG-CPT and RG-SFT configuration, the complete pipeline further reduces L2 from 21.59 to 13.02 and CIEDE2000 from 11.77 to 10.09, while increasing O from 6.53 to 7.19. The concurrent gains in target fidelity and overall semantic–perceptual quality indicate that outcome-level policy optimization complements supervised parameter prediction by refining decisions according to their rendered results.

Dynamic data derivation. Compared with training without dynamic data derivation, training on both original and derived data reduces L2 from 17.45 to 13.02 and raises O from 6.96 to 7.19, with consistent improvements in CIEDE2000, SC, and PQ. This result indicates that derived samples provide complementary training signal alongside the original parameter-grounded data. To isolate the standalone contribution of derivation, we construct a derived-only control from the collected



Figure 5: Visual comparison of different methods on Ref-Bench. JarvisAura produces faithful retouching results that closely match the reference style while preserving source content.

Table 3: Ablation studies on Ref-Bench.

Configurations	$L1_{\times 10^2}\downarrow$	$L2_{\times 10^3}\downarrow$	CIEDE2000 ↓	SC ↑	PQ ↑	O ↑
Training stages						
RG-SFT only	11.04	25.00	12.46	5.60	6.83	6.16
PG-CPT + RG-SFT	10.25	21.59	11.77	6.13	7.02	6.53
PG-CPT + RG-SFT + RA-GRPO (Ours)	8.92	13.02	10.09	6.97	7.51	7.19
Dynamic data derivation						
Derived data only	9.60	16.64	10.46	6.59	7.24	6.88
w/o data derivation	9.55	17.45	10.34	6.68	7.31	6.96
Original + derived data (Ours)	8.92	13.02	10.09	6.97	7.51	7.19

derived samples, supplementing them with SFT-rollout samples under the same derivation procedure to match the size of the no-derivation training set. As shown in Table 3, comparable performance to the no-derivation control suggests that the derived samples retain meaningful training value despite excluding the original data.

6 Conclusion

We present JarvisAura, a multimodal reference agent that maps a source–reference pair and a user instruction to executable, editable parameters for photographic color grading. We construct Ref-CPT with approximately 200K parameter-grounded state-change captions, Ref-SFT with approximately 50K reference-grounded CoT samples, and Ref-RL with expert-guided reinforcement-learning samples. Its training pipeline comprises three stages: (1) PG-CPT grounds parameter–appearance correspondences, (2) RG-SFT learns reference-grounded reasoning and parameter decisions, and (3) RA-GRPO optimizes rendered outcomes through multidimensional rewards and dynamic data derivation. On the benchmark, JarvisAura achieves the best performance on all reported automatic metrics. In the blinded human evaluation, it attains the highest task-success rate among the evaluated input-compatible methods. Pairwise comparisons with four selected leading generative baselines favor JarvisAura on jointly successful cases. Ablation results further support the effectiveness of the overall approach. Collectively, these findings support JarvisAura as an effective approach to multimodal, reference-guided photographic color grading.

AI USE STATEMENT

We used generative AI tools in the research pipeline and manuscript preparation. In the research pipeline, these tools generated state-change captions, user instructions, and reasoning trajectories, assisted with data-quality verification, and provided training rewards and automated evaluation, including Semantic Consistency (SC) and Perceptual Quality (PQ) scores. Details are provided in the methodology and relevant appendices. In manuscript preparation, AI assisted with language polishing, translation, and title and paragraph organization. The authors take responsibility for the final content, research claims, and experimental results, including AI-assisted text, data, and other research artifacts.

REFERENCES

Jie An, Haoyi Xiong, Jun Huan, and Jiebo Luo. Ultrafast photorealistic style transfer via neural architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10443–10450, 2020.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 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.

Haoyu Chen, Keda Tao, Yizao Wang, Xinlei Wang, Lei Zhu, and Jinjin Gu. Photoartagent: Intelligent photo retouching with language model-based artist agents. *arXiv preprint arXiv:2505.23130*, 2025.

Tai-Yin Chiu and Danna Gurari. Photowct2: Compact autoencoder for photorealistic style transfer resulting from blockwise training and skip connections of high-frequency residuals. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2868–2877, 2022.

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.

Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023.

Man M Ho and Jinjia Zhou. Deep preset: Blending and retouching photos with color style transfer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2113–2121, 2021.

Kibeom Hong, Seogkyu Jeon, Huan Yang, Jianlong Fu, and Hyeran Byun. Domain-aware universal style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14609–14617, 2021.

Junqing Huang, Haihui Wang, Andreas Weiermann, and Michael Ruzhansky. Optimal image transport on sparse dictionaries. *arXiv preprint arXiv:2311.01984*, 2023.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Zhanghan Ke, Yuhao Liu, Lei Zhu, Nanxuan Zhao, and Rynson WH Lau. Neural preset for color style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14173–14182, 2023.

Maria Larchenko, Alexander Lobashev, Dmitry Guskov, and Vladimir Vladimirovich Palyulin. Color transfer with modulated flows. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4464–4472, 2025.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017.

Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 453–468, 2018.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 653–661, 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. *arXiv preprint arXiv:2506.17612*, 2025a.

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. *arXiv preprint arXiv:2511.23002*, 2025b.

Yuhang Liu, Pengxiang Li, Zishu Wei, Congkai Xie, Xueyu Hu, Xinchen Xu, Shengyu Zhang, Xiaotian Han, Hongxia Yang, and Fei Wu. Infiguiagent: A multimodal generalist gui agent with native reasoning and reflection. *arXiv preprint arXiv:2501.04575*, 2025.

Kartik Narayan, Yang Xu, Tian Cao, Kavya Nerella, Vishal M Patel, Navid Shiee, Peter Grasch, Chao Jia, Yinfei Yang, and Zhe Gan. Deepmmsearch-r1: Empowering multimodal llms in multimodal web search. *arXiv preprint arXiv:2510.12801*, 2025.

Jaesik Park, Yu-Wing Tai, Sudipta N Sinha, and In So Kweon. Efficient and robust color consistency for community photo collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 430–438, 2016.

François Pitié and Anil Kokaram. The linear monge-kantorovitch linear colour mapping for example-based colour transfer. In *4th European conference on visual media production*, pp. 1–9. IET, 2007.

Tania Pouli and Erik Reinhard. Progressive histogram reshaping for creative color transfer and tone reproduction. In *Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering*, pp. 81–90, 2010.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, et al. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*, 2025.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Lin Song, Wenbo Li, Guoqing Ma, Wei Tang, Bo Wang, Yuan Zhang, Yijun Yang, Yicheng Xiao, Jianhui Liu, Yanbing Zhang, et al. Joyai-image: Awaking spatial intelligence in unified multimodal understanding and generation. *arXiv preprint arXiv:2605.04128*, 2026.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Super Intelligence Team. Firered-image-edit-1.0 technical report. 2026. URL https://arxiv.org/abs/2602.13344.

Haoming Wang, Haoyang Zou, Huatong Song, Jiazhan Feng, Junjie Fang, Junting Lu, Longxiang Liu, Qinyu Luo, Shihao Liang, Shijue Huang, et al. Ui-tars-2 technical report: Advancing gui agent with multi-turn reinforcement learning. *arXiv preprint arXiv:2509.02544*, 2025a.

Peiyu Wang, Yi Peng, Yimeng Gan, Liang Hu, Tianyidan Xie, Xiaokun Wang, Yichen Wei, Chuanxin Tang, Bo Zhu, Changshi Li, et al. Skywork unipic: Unified autoregressive modeling for visual understanding and generation. *arXiv preprint arXiv:2508.03320*, 2025b.

Linfeng Wen, Chengying Gao, and Changqing Zou. Cap-vstnet: Content affinity preserved versatile style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18300–18309, 2023.

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

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*, 2025b.

Jiarui Wu, Yujin Wang, Lingen Li, Fan Zhang, and Tianfan Xue. Goal conditioned reinforcement learning for photo finishing tuning. *Advances in Neural Information Processing Systems*, 37: 46294–46318, 2024.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. Mmsearch-r1: Incentivizing lmms to search. *arXiv preprint arXiv:2506.20670*, 2025c.

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.

Haolong Yan, Jia Wang, Xin Huang, Yeqing Shen, Ziyang Meng, Zhimin Fan, Kaijun Tan, Jin Gao, Lieyu Shi, Mi Yang, et al. Step-gui technical report. *arXiv preprint arXiv:2512.15431*, 2025.

Melany Yang, Yuhang Yu, Diwang Weng, Jinwei Chen, and Wei Dong. Seminft: Learning to transfer presets from imitation to appreciation via hybrid-sample reinforcement learning. *arXiv preprint arXiv:2602.08582*, 2026.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025.

Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9036–9045, 2019.

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.

Shaolei Zhang, Ju Fan, Meihao Fan, Guoliang Li, and Xiaoyong Du. Deepanalyze: Agentic large language models for autonomous data science. *arXiv preprint arXiv:2510.16872*, 2025.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning, 2024. URL https://arxiv.org/abs/2408.05517.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL http://arxiv.org/abs/2403.13372.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

Hanzhang Zhou, Xu Zhang, Panrong Tong, Jianan Zhang, Liangyu Chen, Quyu Kong, Chenglin Cai, Chen Liu, Yue Wang, Jingren Zhou, et al. Mai-ui technical report: Real-world centric foundation gui agents. *arXiv preprint arXiv:2512.22047*, 2025.

Yuchen Zhuang, Jingfeng Yang, Haoming Jiang, Xin Liu, Kewei Cheng, Sanket Lokegaonkar, Yifan Gao, Qing Ping, Tianyi Liu, Binxuan Huang, et al. Hephaestus: Improving fundamental agent capabilities of large language models through continual pre-training. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6041–6068, 2025.

APPENDICES

The appendices are organized into four sections covering dataset details in Section A, methodological details in Section B, additional experimental details in Section C, and additional qualitative results in Section D.

- **Section A: Dataset Details.** Statistics and composition, comparison with existing datasets, data construction, representative samples, and data-generation prompt figures.
- **Section B: Methodological Details.** Data quality control, task-alignment rewards, visual-quality rewards, format-validity rewards, and dynamic data derivation.
- **Section C: Additional Experimental Details.** Training and baseline configurations, Ref-Bench construction, automatic evaluation, human evaluation, and computational cost.
- **Section D: Additional Qualitative Results.** Additional visual comparisons and intermediate reasoning and parameter outputs.

A DATASET DETAILS

A.1 STATISTICS AND COMPOSITION

RefGrade is a multimodal dataset for Instruction-Selective Photographic Reference Grading, organized into three complementary components. Ref-CPT provides approximately 200K parameter-grounded state-change captions for Parameter-Grounded Continual Pre-Training. Ref-SFT contains approximately 50K reference-grounded samples pairing user instructions with reasoning trajectories and parameter configurations for Reference-Grounded Supervised Fine-Tuning. Ref-RL comprises approximately 2K expert-refined samples for Reference-Aligned Group Relative Policy Optimization. Table S1 summarizes their sizes and training roles.

The upper-left panel of Figure S1 illustrates the scene composition of the source-image collection. The four principal categories are portrait (32.7%, e.g., headshots, full-body portraits, and action shots), scene (26.4%, e.g., landscapes, waterscapes, and forests), still life (24.2%, e.g., food, dishes, and pets), and city (13.6%, e.g., streets, facades, and transit). These categories encompass fine-grained subcategories spanning diverse photographic subjects and settings.

The upper-right panel of Figure S1 summarizes the vocabulary used to express grading intent after removing common stopwords and generic instruction terms. The remaining vocabulary covers color attributes, lighting conditions, atmosphere, and visual style. The lower panels present the instruction and chain-of-thought (CoT) length distributions. Instruction lengths are primarily concentrated between approximately 10 and 45 words, covering common forms of expression ranging from brief requests to more detailed descriptions. CoT lengths are mainly concentrated between approximately 130 and 210 words, with two distinct peaks near 145 and 180 words.

A.2 COMPARISON WITH EXISTING DATASETS

Table S2 compares datasets in terms of photographic data, combined reference-image and textual conditioning, reasoning annotations, Lightroom configurations, non-fixed resolution, and continual pre-training support. Attributes of the Ref datasets are reported collectively and need not be present in every individual sample.

A.3 ADDITIONAL DATA CONSTRUCTION DETAILS

Annotation Quality Control for Ref-CPT. To promote annotation consistency, the four captioning models follow a shared prompting scheme with common output-format and terminology requirements (Figure S11). All generated captions undergo validation by a shared VLM-as-Judge verifier conditioned on the source image, editing parameters, and rendered result. The verifier assesses three aspects: **1) Format compliance.** Whether captions conform to the prescribed output structure. **2) Descriptive specificity.** Whether captions characterize parameter-induced appearance changes rather than generic image semantics or aesthetic judgments. **3) Editing faithfulness.** Whether the described changes are supported by the source and rendered images and consistent with the editing



Figure S1: Source-image scene composition, instruction word cloud, and instruction and reasoning length distributions.

parameters. The verification prompt is provided in Figure S5. Captions that fail validation are regenerated and re-evaluated for up to three rounds. Remaining failures undergo manual review, while clearly invalid outputs are discarded. This process applies consistent quality-control criteria across annotation sources while preserving diversity in wording and expression.

Human-in-the-Loop Aesthetic Screening for Ref-SFT. Screening is performed after reference-image construction and before user-instruction and reasoning-trajectory generation. Each candidate sample is independently reviewed by five volunteers against two criteria: **1) Reference alignment.** Whether the constructed reference and target retouched image exhibit consistent key visual attributes, and whether the reference provides transferable cues for the source image, following Table S4. **2) Aesthetic quality.** Whether the rendered results are aesthetically acceptable, without noticeable deviations from human aesthetic preferences. A sample proceeds to subsequent annotation if at least three reviewers judge it to satisfy both criteria. Otherwise, the underlying image assets are returned to the data generation and reference construction stage for reuse, and newly constructed candidates undergo the same screening.

Expert Refinement and Consensus Validation for Ref-RL. We curate a reference library from high-quality photographs purchased with licenses covering the intended uses. A free color-grading service collects real user requests, allowing users to select a reference and submit a source image with natural-language instructions. Submissions are used only after disclosure of the research purpose and explicit user authorization. Experts perform parameterized editing according to each request, retaining configurations that reproduce the corresponding target images.

Table S1: Overview of the training resources. Sizes are approximate and exclude dynamically derived samples.

Dataset	Samples	Training stage	Principal content
Ref-CPT	200K	PG-CPT	Parameter-grounded state-change captions
Ref-SFT	50K	RG-SFT	Reference-grounded instructions, reasoning, and parameters
Ref-RL	2K	RA-GRPO	Expert-refined cases for rendered-outcome optimization

Table S2: Comparison with existing datasets. Res. denotes non-fixed resolution and CPT denotes continual pre-training support.

Dataset	Real photos	Reference and text	CoT labels	Lightroom config.	Res.	CPT
InstructPix2Pix Brooks et al. (2023)	No	No	No	No	No	No
MagicBrush Zhang et al. (2023)	Yes	No	No	No	No	No
MGIE Fu et al. (2023)	Yes	No	No	No	No	No
ImgEdit Ye et al. (2025)	Yes	No	No	No	Yes	No
MIT-Adobe FiveK Bychkovsky et al. (2011)	Yes	No	No	Yes	Yes	No
PPR10K Liang et al. (2021)	Yes	No	No	Yes	Yes	No
MMArt Lin et al. (2025a)	Yes	No	Yes	Yes	Yes	No
ArtEdit Lin et al. (2025b)	Yes	No	Yes	Yes	Yes	No
Ref datasets (ours)	Yes	Yes	Yes	Yes	Yes	Yes

Each candidate is independently assessed by 20 reviewers spanning different age groups and levels of imaging expertise, including general users, photography enthusiasts, and imaging professionals. Editors do not review their own outputs. Reviewers examine the source image, reference image, user instruction, and edited result against two criteria: **1) Reference and instruction alignment.** Whether the result draws on the reference's overall style and tone while fulfilling the user's attribute-specific preferences and requested adjustments. **2) Aesthetic quality.** Whether the result is visually coherent and free of conspicuous defects. Both criteria receive binary pass/fail judgments, and acceptance requires unanimous approval on both. Consensus denotes panel agreement on task fulfillment and visual acceptability, rather than a uniquely optimal edit.

Failed candidates are returned to the editor with specific feedback and revised without altering the user request. The same panel independently reassesses revisions, and only unanimously approved samples are retained. Ref-RL comprises approximately 2K consensus-validated, expert-refined samples, disjoint from Ref-Bench. Samples dynamically derived during training are counted separately.

A.4 Representative Data Samples

Figures S2 and S3 illustrate scene coverage and Ref-CPT sample structure. Figures S4 present Ref-SFT examples.

A.5 Prompt Templates for Data Construction

Figure S11 presents the state-change captioning prompt. Figures S9 and S10 provide the summary prompts used in reference retrieval. Figures S6, S7, and S8 present the prompt templates for initial user-instruction generation, reference-grounded refinement, and reference-grounded local refinement, respectively. Figures S12, S13, and S14 correspond to intent analysis, parameter-decision mapping, and reasoning-trajectory integration. Quality-verification, reward, derivation, and evaluation prompts are referenced separately in their respective methodological sections.

B METHODOLOGICAL DETAILS

B.1 TASK-ALIGNMENT REWARD

The task-alignment reward combines instruction following and reference matching.

1) Instruction Following. A vision-language evaluator assesses instruction fulfillment using the source image, reference image, rendered output, user instruction, and executed parameter configuration. The reference provides stylistic and tonal context, while the instruction determines the attributes and adjustments to evaluate. Parameter settings provide auxiliary evidence of adjustment direction and scope, but fulfillment depends on visible outcomes rather than parameter values alone.

Using the prompt in Figure S16, the evaluator directly scores four dimensions: color, light and shadow, technical adjustments, and high-level style. Each applicable dimension receives a score of 100, 70, 35, or 0, corresponding to full fulfillment, fulfillment with minor discrepancies, partial fulfillment, or failure, respectively. Dimensions without corresponding requirements are marked inapplicable, whereas those lacking sufficient evidence are marked unassessable. Both are excluded from aggregation, with the reason recorded. The instruction-following component $\mathcal{V}_{\text{inst}}$ is the equal-weight mean of the remaining scores divided by 100, yielding a value in $[0, 1]$. If no dimension is assessable, the evaluation is flagged as invalid and assigned zero. Quality attributes not specified in the instruction are assessed separately by the visual-quality reward.

2) Reference Matching. Reference matching compares each rendered output with the expert-refined target. Before color conversion, images whose longest side exceeds 512 pixels are proportionally downsampled using area interpolation; smaller images retain their original resolution.

Color-space conversion. Each sRGB channel $C \in [0, 1]$ is first linearized:

$$C_{\text{lin}} = \begin{cases} C/12.92, & C \le 0.04045, \\ \left((C + 0.055)/1.055\right)^{2.4}, & C > 0.04045. \end{cases} \tag{S1}$$

The linear RGB values are converted to CIE XYZ under the D65 illuminant:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.412453 & 0.357580 & 0.180423 \\ 0.212671 & 0.715160 & 0.072169 \\ 0.019334 & 0.119193 & 0.950227 \end{bmatrix} \begin{bmatrix} R_{\text{lin}} \\ G_{\text{lin}} \\ B_{\text{lin}} \end{bmatrix}. \tag{S2}$$

CIELAB coordinates are then obtained as

$$\begin{aligned} L^* &= 116 f(Y/Y_n) - 16, \\ a^* &= 500\left[f(X/X_n) - f(Y/Y_n)\right], \\ b^* &= 200\left[f(Y/Y_n) - f(Z/Z_n)\right], \end{aligned} \tag{S3}$$

where $(X_n, Y_n, Z_n) = (0.950456, 1, 1.088754)$ and

$$f(t) = \begin{cases} t^{1/3}, & t > \delta^3, \\ t/(3\delta^2) + 4/29, & t \le \delta^3, \end{cases} \qquad \delta = 6/29. \tag{S4}$$

The implementation uses OpenCV's Lab conversion, with the chromatic channels offset by 128 in the 8-bit representation.

Histogram distance. Only the a and b channels are used for matching. Each channel is discretized into 24 bins, producing a 24×24 joint histogram normalized to unit mass. Bins with probability mass greater than 10^{-8} are retained as weighted two-dimensional bin centers. For candidate i, the matching distance is

$$d_i = \text{EMD}\left(H(I_i^{\text{out}}), H(I_{\text{tgt}})\right), \tag{S5}$$

where $H(\cdot)$ denotes the normalized chromatic histogram. The Earth Mover's Distance is computed using Euclidean distances between bin centers as the transport cost. Neither the L channel nor pixel coordinates enter this calculation.

Within-group scoring. For each input, the eight rollout candidates are ranked in ascending order of d_i. Equal distances are resolved by preserving the original rollout order. A measurable candidate with rank $r_i \in \{1, \ldots, 8\}$ receives

$$R_i^{\text{match}} = \frac{9 - r_i}{8}. \tag{S6}$$

Table S3: Four-level scoring scheme for visual quality.

Level	Numerical score	Definition
Level 1	100	The property is visibly preserved from the source, with a concrete supporting observation.
Level 2	70	A trace or mild defect is visible only on close inspection, or the property cannot be conclusively judged from the supplied images.
Level 3	35	A defect is clearly visible under normal viewing and materially affects the property.
Level 4	0	A severe or critical defect makes the property unacceptable.

Thus, scores decrease from 1 to 0.125 in increments of 0.125. Candidates without a measurable distance because of invalid model outputs receive the minimum score of 0.125. Infrastructure failures are excluded from optimization rather than treated as low-ranked candidates.

Finally, the reference-matching score and the instruction-following score are added:

$$R_i^{\text{task}} = R_i^{\text{match}} + \mathcal{V}_{\text{inst}}\left(I_{\text{src}}, I_{\text{ref}}, I_i^{\text{out}}, Q, P^{(i)}\right). \tag{S7}$$

B.2 Visual-Quality Reward

The visual-quality reward measures edit-induced damage independently of task alignment. Most of the $K = 14$ criteria assess preservation properties, and a criterion is considered violated only when the edit visibly degrades the corresponding property relative to the source image. The source image therefore serves as the baseline for identifying editing damage, while global tonal or stylistic changes are not penalized unless they cause a concrete criterion-specific defect.

The evaluator receives only the source image and the rendered output, without access to the reference image, user instruction, or executed parameter configuration. Following the prompt in Figure S15, the evaluator is invoked independently once for each criterion, covering tonal integrity, color fidelity, edge and material fidelity, geometric integrity, and residual artifacts (Table S10). Each criterion receives a score of 100, 70, 35, or 0, corresponding to different levels of degradation, or `null` when no relevant subject is present (Table S3). Applicable scores must be supported by concrete comparisons between the source and rendered images. An inapplicable criterion incurs no penalty and is assigned a score of 100 during aggregation.

The evaluator does not compute an aggregate score. Instead, the visual-quality reward is calculated externally as

$$R_i^{\text{qual}} = \frac{1}{100K}\sum_{k=1}^{K} q_{i,k}, \qquad K = 14, \tag{S8}$$

where $q_{i,k} \in \{0, 35, 70, 100\}$ and `null` outputs are mapped to 100, yielding $R_i^{\text{qual}} \in [0, 1]$.

B.3 Format-Validity Reward

The format-validity reward has two independently evaluated parts. Throughout this paper, the response is written as `<think>...</think>` followed by `<tool_call>`P`</tool_call>`.

1. **Response and XML parsing.** Let F denote the parsing score. It is the sum of four independent binary checks:

$$F = 0.2I_{\text{think}} + 0.2I_{\text{tool_call}} + 0.4I_{\text{xml}} + 0.2I_{\text{root}}. \tag{S9}$$

The reasoning and tool-call tag pairs must each occur exactly once, with the opening tag before the closing tag. The XML declaration is removed before parsing. The content inside `tool_call` is parsed with the project's sign-aware XML converter, and its root element must be literally `root`. A failed sub-check lowers F without necessarily making the entire parsing score zero.

Table S4: Reference-inspection checklist for Ref-SFT quality control.

Dimension	Inspection focus
Semantic correspondence	Correspondence of relevant content such as sky, skin, vegetation, water, and architectural regions.
Overall color tone	Dominant color tendencies, white balance, and overall color coherence.
Warm–cool balance	Relationships among highlights, midtones, and shadows.
Lighting atmosphere	Compatibility of the reference's lighting appearance with the intended grading changes.
Tonal structure and contrast	Highlight transitions, shadow detail, midtone gradation, and contrast.
Local texture and clarity	Relevant textural characteristics without conflating grading with content replacement.
Subject–background separation	Tonal and color relationships contributing to foreground and background separation.
Style coherence	Consistency of the reference attributes with the intended grading direction.

2. **Parameter validation.** Let P denote the parameter-validation score. Missing or empty parameters, a fatal structural or type error, or a failed parameter-structure check yields $P = 0$ and `valid`=0. Otherwise, the validator checks registered enumerations, the nested structures handled explicitly by the implementation, and registered top-level numeric fields. For a registered numeric parameter with legal interval $[l, u]$ and value v, define the normalized excess

$$e(v; l, u) = \frac{\max\{0,\, l - v,\, v - u\}}{u - l}. \tag{S10}$$

The corresponding numeric score is 1.0 when $e = 0$, 0.9 when $0 < e \leq 0.1$, 0.7 when $0.1 < e \leq 0.5$, and 0.2 when $e > 0.5$. Numeric scores are averaged within each populated category and combined with the implementation's category weights {basic:0.40, color:0.20, detail:0.10, advanced:0.05, mask:0.25}; an empty category defaults to 1.0.

Let $\mathcal{S}$ be the supported top-level parameter schema and let $\mathcal{K}$ be the set of top-level keys in the output. Define $\mathcal{U} = \{k \mid k \in \mathcal{K},\, k \notin \mathcal{S}\}$. Unknown names are treated as anomalous output: define the name gate $I_{\text{name}} = \mathbf{1}[\mathcal{U} =]$. Likewise, let I_{struct} indicate that the mapping is non-empty and passes all registered enum, nested-structure, and type checks. The parameter-validation score is then defined as

$$P = I_{\text{name}}\, I_{\text{struct}}\, R_{\text{range}}, \tag{S11}$$

where R_{range} is the category-weighted score from the registered numeric range checks above. Thus, any unknown parameter name or fatal structural error gives $P = 0$, while a registered numeric value outside its range can retain graded partial credit.

Table S5 summarizes both parts of the format-validity reward.

C Additional Experimental Details

C.1 Detailed Experimental Setup

JarvisAura is developed from Qwen3-VL-8B-Instruct through PG-CPT, RG-SFT, and RA-GRPO. Table S6 summarizes the three-stage training configuration.

C.2 Construction of Ref-Bench

Ref-Bench comprises 300 test instances and follows the same data collection, parameterized expert-editing, and consensus-validation protocol as the approximately 2K-sample Ref-RL dataset. The two

Table S5: Format-validity reward checks.

Check	Implemented property	Numerical rule
Reasoning block	Exactly one ordered `<think>` pair	0.20 in F
Tool-call block	Exactly one ordered `<tool_call>` pair	0.20 in F
XML parsing	Inner XML parses with sign-aware `xml_to_dict`	0.40 in F
Root tag	Parsed XML root is exactly `root`	0.20 in F
Parameter structure	Non-empty mapping; registered enums and explicitly handled nested structures/types are valid	$P = 0$ on fatal failure
Parameter names	Every top-level name must belong to the supported parameter schema	Any unknown name gives $P = 0$
Parameter values	Registered top-level numeric fields in the range schema are compared with their declared ranges	$1.0/0.9/0.7/0.2$ by excess

Table S6: Hyperparameters for the three training stages of JarvisAura.

Hyperparameter	**PG-CPT**	**RG-SFT**	**RA-GRPO**
Model & Data Configuration			
Max Sequence Length	24576	24576	24576
Train Batch Size (per GPU)	1	1	1
Training Samples	200K	50K	2K + derived (28%)
Optimizer & Scheduler			
Optimizer	AdamW	AdamW	AdamW
Learning Rate	5×10^{-5}	1×10^{-5}	5×10^{-6}
LR Scheduler	Cosine	Cosine	Constant
Warmup Ratio	0.1	0.1	0
Training Details			
Training Epochs	1	2	1
Precision	bf16	bf16	bf16
RA-GRPO Configuration			
Max Prompt Length	–	–	17900
Max Response Length	–	–	6384
Advantage Estimator	–	–	GRPO
Rollout Backend	–	–	vLLM (async)
Rollouts per Prompt	–	–	8
Sampling Temperature	–	–	0.75
Sampling Top-p	–	–	0.92
GPU Count	8×A100	8×A100	8×A100

datasets are independently collected and have no overlap in users, source images, user instructions, expert-edited targets, or reference images. In particular, Ref-Bench uses a separate reference-image library, and its expert editors are distinct from those involved in Ref-RL construction. Its samples further cover broader ranges of user ages, imaging expertise, and photographic scenes. Thus, Ref-Bench evaluates generalization to unseen cases within the same task domain.

C.3 Automatic Evaluation Protocols

Target-based fidelity. To measure how closely each edited image agrees with the corresponding expert consensus result, we compare it with the expert-refined target. The edited image and target are independently center-cropped to a square using the shortest image side and then resized to 512×512 pixels with Lanczos interpolation. RGB values are represented in normalized sRGB, with values in $[0, 1]$. For an edited image I and target T, L1 is the mean absolute RGB error over all pixels and

Table S7: Preprocessing and aggregation settings for automatic metrics.

Setting	Configuration
Image resizing and alignment	Independently center-crop the edited image and target to the shortest-side square; resize both to 512×512 with Lanczos interpolation
Pixel range and normalization	RGB values represented as normalized sRGB in $[0, 1]$
Color profile and transfer function	IEC 61966-2-1 sRGB transfer function
Lab conversion and reference white	sRGB-to-CIELAB conversion with the D65 reference white
CIEDE2000 implementation and factors	Standard CIEDE2000 formulation with $k_L = k_C = k_H = 1$
Spatial and channel reduction	L1/MSE averaged over all $H \times W \times 3$ RGB samples; CIEDE2000 averaged over all $H \times W$ pixels
Dataset-level aggregation	Arithmetic mean over valid test cases
Invalid or missing outputs	Excluded from aggregation

channels,

$$\mathrm{L1}(I,T) = \frac{1}{3HW} \sum_{x=1}^{W} \sum_{y=1}^{H} \sum_{c \in \{R,G,B\}} |I_{x,y,c} - T_{x,y,c}| \,, \tag{S12}$$

where $H = W = 512$. L2 denotes the mean squared error (MSE), rather than the root mean squared error,

$$\mathrm{L2}(I,T) = \mathrm{MSE}(I,T) = \frac{1}{3HW} \sum_{x=1}^{W} \sum_{y=1}^{H} \sum_{c \in \{R,G,B\}} (I_{x,y,c} - T_{x,y,c})^2 \,. \tag{S13}$$

CIEDE2000 measures perceptual color difference. We convert normalized sRGB values to CIELAB using the D65 reference white and compute the standard per-pixel ΔE_{00} with $k_L = k_C = k_H = 1$:

$$\mathrm{CIEDE2000}(I,T) = \frac{1}{HW} \sum_{x=1}^{W} \sum_{y=1}^{H} \Delta E_{00}(\mathrm{Lab}(I_{x,y}), \mathrm{Lab}(T_{x,y})) \,. \tag{S14}$$

For each method, we first compute each metric per valid test case and then report the arithmetic mean over valid cases; missing or invalid result images are excluded from aggregation. Lower values indicate better target fidelity. Main-text tables report L1 multiplied by 10^2 and L2 (MSE) multiplied by 10^3. Table S7 summarizes preprocessing and aggregation settings for these metrics.

Reference-guided style consistency and preservation quality. To evaluate the perceptual quality of reference-guided editing without relying on a ground-truth image, we use Gemini-3.1-Pro to assess each result from three inputs in a fixed order: the original source image, the style reference image, and the edited image. The model assigns two independent scores on a 0–10 scale using half-point increments. Style Consistency (SC) measures how faithfully the edited image transfers the reference's applicable color and tonal characteristics, while Preservation Quality (PQ) measures preservation of the source content together with the technical quality of the edited image. The overall score is computed per sample as the geometric mean $O = \sqrt{\mathrm{SC} \times \mathrm{PQ}}$, after which we aggregate the per-sample scores across the evaluation set. The evaluation prompts are shown in Figure S19. Training and evaluation use separate prompts: the former is task-specific and requires configuration parameters, whereas the latter is fixed across methods to ensure fair comparison.

C.4 Human Evaluation Protocols

Human evaluation comprises task-success assessment and pairwise preference assessment. Both evaluations are blinded, and expert target images are withheld. All evaluations are conducted on the same latest-generation MacBook Air under a consistent display setup.

Human-rated task success rate. HTSR is measured in a blinded user study with expert target images withheld. The study involves 30 participants spanning different age groups and levels of

Table S8: Training time for each stage of JarvisAura, measured in GPU hours on a single A100.

Stage	**Configuration**	**Samples**	**GPU Hours (A100)**
PG-CPT	1 epoch	200K	$\sim$500
RG-SFT	2 epochs	50K	$\sim$160
RA-GRPO	1 epoch	2K	$\sim$ 280
Total			$\sim$940

Table S9: Inference configurations and computational costs.

Inference framework	vLLM	Transformers
End-to-end time (s/sample)	38.18	47.75
Model generation time (s/sample)	29.36	36.24
Backend rendering time (s/sample)	8.75	11.22
Time to first token (s)	4.19	1.38
Throughput (tokens/s)	71.32	29.73
Generated tokens per sample	2,094	1078

imaging expertise, including both imaging professionals and non-professional users. None of the participants were involved in training-data screening or consensus validation. Using the same latest-generation MacBook Air, participants independently assign one of three labels: successful, uncertain, or unsuccessful. No fine-grained scoring rubric is imposed; participants make an overall judgment under a realistic user setting while considering fine-grained instruction fulfillment, alignment with requested reference attributes, visual quality, and content preservation. The final label is determined by majority vote, and a case is counted as successful only when successful receives the majority.

Pairwise A/B preference. Using the same participants and device, JarvisAura is compared separately with GPT-Image-2.5, Seedream-5-pro, Nano-Banana-2, and FireRed-Image-Edit. Each comparison uses the intersection of cases judged successful for JarvisAura and that baseline, not a common intersection across all methods. For each case, the outcome is determined by majority vote: JarvisAura is counted as winning only when it receives more votes than the compared baseline, while an equal vote count is recorded as a tie. The reported preference ratio is

$$\mathrm{PR} = \frac{N_{\mathrm{win}} + N_{\mathrm{tie}}}{N_{\mathrm{loss}} + N_{\mathrm{tie}}}, \tag{S15}$$

where wins and losses are defined from JarvisAura's perspective. A value above one favors JarvisAura. This quantity is a ratio, not a preference percentage or percentile.

C.5 Training and Inference Time

Tables S8 and S9 summarize training and inference costs, respectively.

D Additional Qualitative Results

D.1 Intermediate Reasoning and Parameter Outputs

Figure S18 presents source and reference images, the user instruction, generated reasoning, executable parameter configurations, and rendered outcomes. These examples expose the agent's intermediate outputs and their relation to the resulting edits.

D.2 More Visual Comparisons with Generative Methods

Figure S20 presents additional visual comparisons with input-compatible generative methods. These methods receive the same source image, reference image, and user instruction as JarvisAura, en-

abling a controlled and fair comparison. The examples focus on instruction-specified reference attributes, grading accuracy, visual quality, and source-content preservation.

Table S10: Criteria for the visual-quality reward.

Criterion	Name	Definition and defect condition
1	Highlight and overexposure integrity	Highlights retain roll-off, headroom, and texture relative to the source; broad near-white crowding or a blown key subject is a defect.
2	Shadow and dark-detail integrity	Shadows retain boundaries, local contrast, and material separation; newly flattened, featureless dark regions are a defect.
3	Contrast, dynamic-range, and veil integrity	Global and local contrast and tonal separation remain credible; a broad white or colored veil or harsh tonal compression is a defect.
4	Subject-environment tonal integration	A key subject's brightness and local contrast remain consistent with its surrounding scene; an implausible spotlighted or cutout appearance is a defect.
5	Gradient, banding, and posterization integrity	Smooth or gently graded regions (sky, water, walls) remain continuous; visible banding, stepped contours, or posterization is a defect.
6	Local chroma continuity and blocking	Color within a single material or semantic region stays spatially continuous; abrupt within-region color steps or blocky chroma cells are a defect.
7	Edge, mask, and transition integrity	Boundaries between subjects and regions remain natural; visible halos, color fringing, mask seams, or a cutout appearance are a defect.
8	Detail and texture integrity	Fine surface detail and micro-contrast (skin, fabric, foliage) are preserved; waxy, over-smoothed, or flattened texture is a defect.
9	Material rendering fidelity	Expected material cues (specularity, transparency, grain, moisture) remain credible; loss of a material's characteristic surface response is a defect.
10	Human face and skin chroma naturalness	Visible face, neck, or hand skin tones remain plausible and internally consistent; patchy, contaminated, or implausible skin chroma is a defect.
11	Neutral and monochrome chromatic integrity	Neutral, gray, or black-and-white regions keep a plausible, spatially consistent color identity; an unexplained localized color cast is a defect.
12	Non-human subject color plausibility	Colors of food, vegetation, sky, water, buildings, or objects remain semantically and physically plausible; fluorescent or identity-changing color shifts are a defect.

continued on next page

Table S10 continued from previous page.

Criterion	Name	Definition and defect condition
13	Geometry and semantic structure integrity	Shapes, boundaries, and the recognizability of faces, bodies, and objects are preserved; new warping, merging, or structural corruption is a defect.
14	Residual processing cleanliness	No newly introduced noise, oversharpening, aliasing, or compression artifact appears beyond what the other thirteen criteria already describe.

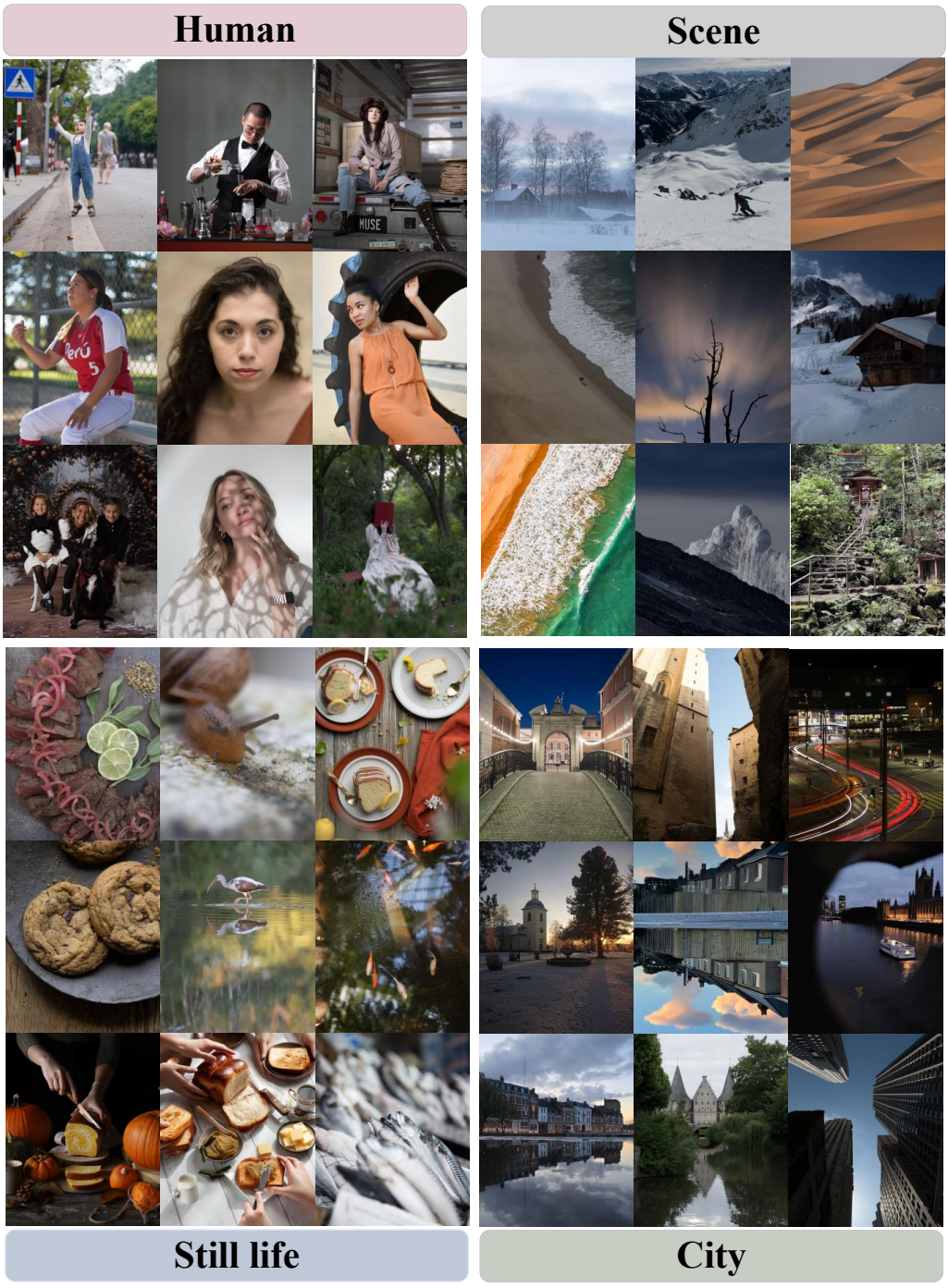
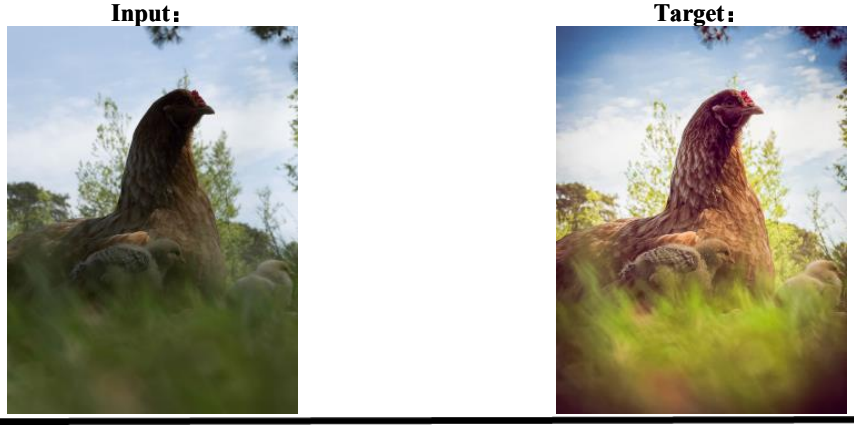


Figure S2: Representative photographic scenes in the Ref datasets.

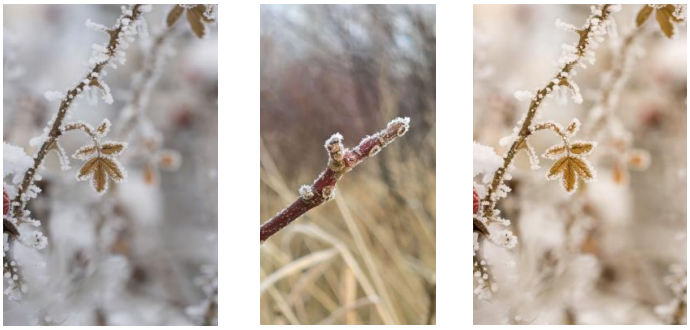
Input: **Target:**



This image, once a softly lit, naturalistic scene of a hen and her chicks nestled in the grass, has been transformed into a vivid, emotionally resonant portrait through a series of deliberate and synergistic adjustments. The overall effect is one of heightened drama and warmth, with a nostalgic, almost painterly quality achieved through a combination of exposure enhancement, selective color manipulation, and atmospheric toning. The sky, once pale and washed-out, now boasts a rich, deep blue that contrasts beautifully with the golden-brown plumage of the hen and the vibrant greens of the foliage. The foreground grass, previously a muted green, now glows with a luminous, almost ethereal yellow-green, drawing the viewer's eye into the heart of the scene. The foundation of this transformation lies in the Exposure (Exposure2012) value of 1.22, which dramatically brightens the entire image, lifting shadow detail and infusing the scene with a sense of radiant daylight. This is not a simple overexposure, however; it is carefully balanced by the Contrast (Contrast2012) set to 51, which deepens the tonal separation, giving the image a richer, more three-dimensional quality. The highlights have been aggressively pulled down with Highlights (Highlights2012) set to -100, preventing any blown-out areas in the sky and ensuring the whites remain clean and detailed rather than harsh. This is complemented by Whites (Whites2012) at -40, which further refines the brightest tones, adding a subtle coolness to the sky that contrasts with the warmth of the subject. In the shadows, Shadows (Shadows2012) at 8 and Blacks (Blacks2012) at 31 work in tandem to preserve texture in the hen's feathers and the chicks' downy fur without losing depth, while also contributing to the overall luminance. The Color Temperature (Temperature) has been shifted to 5736 Kelvin, introducing a warm, golden cast that bathes the entire scene in a sun-drenched glow. This warmth is subtly enhanced by the Tint (Tint) value of 10, which adds a gentle magenta undertone, particularly noticeable in the sky and the hen's comb. This warmth is not merely a global shift but is selectively amplified by the Color Grading (ColorGradeMidtoneHue) set to 221, a deep, rich purple that subtly influences the midtones, lending a sophisticated, almost vintage feel to the image. This is further refined by the Split Toning (SplitToningShadowHue) at 283 (a deep purple) and (SplitToningHighlightHue) at 327 (a cool blue), creating a beautiful tonal contrast where the shadows take on a moody, purplish depth and the highlights retain a crisp, cool clarity. The (SplitToningBalance) at 59 ensures this split toning is more pronounced in the highlights, which is particularly effective in the sky and the bright areas of the hen's feathers. Texture (Texture) at 48 and Clarity (Clarity2012) at 35 work together to enhance the intricate details of the hen's feathers and the softness of the chicks' down, adding a tactile quality to the image. This is not an aggressive sharpening, but rather a mid-tone contrast enhancement that gives the subject a more defined, almost sculpted presence. The Dehaze (Dehaze) parameter, set to -23, introduces a deliberate atmospheric haze, softening the background and creating a sense of depth and distance, making the hen and chicks stand out as the clear focal point. This effect is further supported by the Vignette (PostCropVignetteAmount) at -58, which darkens the corners and edges, drawing the viewer's gaze inward toward the central subject. The color palette has been carefully tuned through the HSL/Color adjustments. The Yellow (HueAdjustmentYellow) is shifted to -31, deepening the golden tones in the hen's feathers and the grass, while the Yellow (LuminanceAdjustmentYellow) at -31 and (SaturationAdjustmentYellow) at 25 enhance its richness without oversaturating. The Green (HueAdjustmentGreen) is shifted to 24, making the foliage appear more vibrant and natural, and its (SaturationAdjustmentGreen) at 36 and (LuminanceAdjustmentGreen) at 57 ensure the grass glows with a luminous, healthy green. The Blue (SaturationAdjustmentBlue) at 54 and (LuminanceAdjustmentBlue) at -43 deepen the sky and add a cool, crisp quality to the highlights, creating a beautiful counterpoint to the warm midtones. Finally, the image is given a subtle filmic texture through the Grain (GrainAmount) at 4, adding a delicate, organic grain that enhances the nostalgic, analog feel. The sharpening (Sharpness) at 62, with a (SharpenRadius) of 1.1 and (SharpenDetail) at 25, is applied with precision, enhancing edge definition without introducing harshness, while (SharpenEdgeMasking) at 46 ensures the sharpening is concentrated on the subject's edges. The overall effect is a harmonious blend of technical precision and artistic vision, transforming a simple farmyard moment into a visually compelling and emotionally evocative piece. The interplay between the warm, golden midtones, the cool, deep shadows, and the vibrant, glowing foreground creates a dynamic, layered image that feels both natural and profoundly stylized. <parametric retouching operation configuration>

Figure S3: Ref-CPT examples with source images, rendered outcomes, editing configurations, and state-change captions.

Input: **Reference:** **Output:**



User InStruction:
Can you apply the same style as the reference photo to mine? I love that cozy, warm filter effect that makes the whole picture feel gentle.

Answer with CoT annotations:

<think>Alright, check out the mood shift we're dealing with here—we've got this stark, frozen leaf that feels properly chilly with those crisp, cold blue tones. The user is asking for that "cozy, warm filter" from the reference to make it "gentle," and looking at that target style, I can see they aren't just asking for a temperature bump. They're pointing at that specific creamy, monochromatic beige wash that completely transforms the snow from icy to inviting. Reading between the lines, "gentle" implies they want to strip away the harsh reality of the cold to create a Soft & Warm Winter Fine Art aesthetic. My approach to nail this dreamlike vibe is to aggressively shift the white balance, swapping those blues for golden-amber hues. Then, I'll lift the shadows substantially and compress the highlights to flatten the contrast, which will give us that airy, ethereal softness that makes the reference look so comforting.</think>

<answer>HasSettings = true, ProcessVersion = "11.0", Look = { Name = "Adobe Color", Amount = 1, UUID = "B952C231111CD8E0ECCF14B86BAA7077", SupportsAmount = false, SupportsMonochrome = false, SupportsOutputReferred = false, Copyright = "© 2018 Adobe Systems, Inc.", Group = { ["x-default"] = "Profiles" }, Parameters = { Version = "15.1", ProcessVersion = "11.0", ConvertToGrayscale = false, CameraProfile = "Adobe Standard", LookTable = "E1095149FDB39D7A057BAB208837E2E1", ToneCurvePV2012 = { [1] = 0, [2] = 0, [3] = 22, [4] = 16, [5] = 40, [6] = 35, [7] = 127, [8] = 127, [9] = 224, [10] = 230, [11] = 240, [12] = 246, [13] = 255, [14] = 255 }, ToneCurvePV2012Red = { [1] = 0, [2] = 0, [3] = 255, [4] = 255 }, ToneCurvePV2012Green = { [1] = 0, [2] = 0, [3] = 255, [4] = 255 }, ToneCurvePV2012Blue = { [1] = 0, [2] = 0, [3] = 255, [4] = 255 } } }, WhiteBalance = "As Shot", Temperature = 6000, Tint = 12, Exposure2012 = 0.17, Contrast2012 = 7, Highlights2012 = -58, Shadows2012 = 37, Whites2012 = 25, Blacks2012 = -43, Texture = 0, Clarity2012 = 0, Dehaze = 0, Vibrance = 15, Saturation = 2, GrainAmount = 0, PostCropVignetteAmount = 0, OverrideLookVignette = false, Sharpness = 40, SharpenRadius = 1, SharpenDetail = 25, SharpenEdgeMasking = 0, ParametricShadows = 0, ParametricDarks = 0, ParametricLights = 0, ParametricHighlights = 0, ParametricShadowSplit = 25, ParametricMidtoneSplit = 50, ParametricHighlightSplit = 75, HueAdjustmentRed = 0, HueAdjustmentOrange = 0, HueAdjustmentYellow = 0, HueAdjustmentGreen = 0, HueAdjustmentAqua = 0, HueAdjustmentBlue = 0, HueAdjustmentPurple = 0, HueAdjustmentMagenta = 0, SaturationAdjustmentRed = 0, SaturationAdjustmentOrange = 0, SaturationAdjustmentYellow = 0, SaturationAdjustmentGreen = 0, SaturationAdjustmentAqua = 0, SaturationAdjustmentBlue = 0, SaturationAdjustmentPurple = 0, SaturationAdjustmentMagenta = 0, LuminanceAdjustmentRed = 0, LuminanceAdjustmentOrange = 0, LuminanceAdjustmentYellow = 0, LuminanceAdjustmentGreen = 0, LuminanceAdjustmentAqua = 0, LuminanceAdjustmentBlue = 0, LuminanceAdjustmentPurple = 0, LuminanceAdjustmentMagenta = 0, SplitToningShadowHue = 0, SplitToningShadowSaturation = 0, SplitToningHighlightHue = 0, SplitToningHighlightSaturation = 0, SplitToningBalance = 0, ColorGradeMidtoneSat = 0, ColorGradeShadowLum = 0, ColorGradeMidtoneLum = 0, ColorGradeHighlightLum = 0, ColorGradeBlending = 50, ColorGradeGlobalHue = 0, ColorGradeGlobalSat = 0, ColorGradeGlobalLum = 0, ShadowTint = 0, RedHue = 0, RedSaturation = 0, GreenHue = 0, GreenSaturation = 0, BlueHue = 0, BlueSaturation = 0, ToneCurveName2012 = "Linear", ToneCurvePV2012Red = { [1] = 0, [2] = 0, [3] = 255, [4] = 255 }, ToneCurvePV2012Green = { [1] = 0, [2] = 0, [3] = 255, [4] = 255 }, ToneCurvePV2012Blue = { [1] = 0, [2] = 0, [3] = 255, [4] = 255 }, ColorNoiseReduction = 25, ColorNoiseReductionDetail = 50, ColorNoiseReductionSmoothness = 50, ColorGradeMidtoneHue = 0, LuminanceSmoothing = 0, AutoLateralCA = 0, LensProfileEnable = 0, LensManualDistortionAmount = 0, VignetteAmount = 0, DefringePurpleAmount = 0, DefringePurpleHueLo = 30, DefringePurpleHueHi = 70, DefringeGreenAmount = 0, DefringeGreenHueLo = 40, DefringeGreenHueHi = 60, PerspectiveUpright = 0, PerspectiveVertical = 0, PerspectiveHorizontal = 0, PerspectiveRotate = 0, PerspectiveAspect = 0, PerspectiveScale = 100, PerspectiveX = 0, PerspectiveY = 0, CameraProfile = "Adobe Standard", CameraProfileDigest = "A4DC20C59DE4E2DF67946B80BFDA0801", AlreadyApplied = false, ToneCurvePV2012 = { [1] = 0, [2] = 0, [3] = 255, [4] = 255 }, AutoToneDigest = "2F3D0CABC073FFAE10A280EBBD95B35B", AutoToneDigestNoSat = "7D4C283D1890060BF0821E17E8B7F6BE"</answer>

Figure S4: Ref-SFT examples with source and reference images, instructions, reasoning trajectories, parameter configurations, and rendered outcomes.

Prompt for Parameter-Grounded State-Change Caption Verification

You are a professional annotation reviewer specializing in photographic post-processing. Review the candidate state-change caption using the original image, edited image, editing parameters, and parameter reference table.

Your task is to determine whether the caption follows the shared annotation guidelines, specifically explains parameter-related appearance changes, and provides explanations supported by the images and parameter information.

Inputs

1. Original image: {original image}
2. Edited image: {edited image}
3. Editing parameters: {Lightroom parameter configuration}
4. Parameter reference table: {parameter names, professional terms, and functional descriptions}
5. Candidate caption: {state-change caption}

Review Criteria

1. Format Compliance

- The caption must be written in English as coherent paragraphs, without bullet points or tables.
- When a parameter is first mentioned, use the format "Professional Term (Parameter Name)." The correspondence must match the parameter reference table.
- For subsequent mentions of the same parameter, use the corresponding professional term.
- The content must focus on overall appearance changes and parameter effects, without unrelated explanations.

2. Descriptive Specificity

- The caption must describe specific appearance changes between the original and edited images, rather than merely restating image content or offering generic aesthetic judgments.
- For the main non-default parameters that affect appearance, the caption must explain the associated visual effects and their scope in relation to the actual parameter values or settings.
- The description may address tonality, color, texture, spatial impression, and atmosphere, but these aspects must be linked to specific image changes.
- Combined effects may be explained through multiple parameters; a separate, pronounced visual change need not be assigned to every parameter.
- Do not invent pronounced editing effects for parameters whose effects are not apparent or that are not enabled.

3. Editing Faithfulness

- The reported direction, degree, and affected regions of changes must agree with the comparison between the original and edited images.
- Cited parameter names, values, units, and functional explanations must match the configuration and parameter reference table.
- The associations between parameters and visual effects must be reasonable and account for the actual outcome of multiple parameters acting together.
- Do not infer changes absent from the images solely from parameter definitions, or mechanically apply parameter explanations while ignoring visual evidence.
- Do not invent object additions or removals, scene changes, local edits, or unsupported claims about the precise scope of an effect.
- Effects that cannot be confirmed from the inputs must be expressed cautiously rather than stated as established facts.

Decision Rules

- Assign an overall pass only when all three criteria pass.
- If any criterion has a clear violation, assign an overall fail and provide specific revision suggestions.
- Flag unsupported definitive claims that affect the central explanation as having insufficient evidence.
- Allow reasonable differences in wording, sentence structure, and presentation order; do not reject a caption solely because of stylistic differences.
- Do not base the decision on whether the edited result matches personal aesthetic preferences.
- Only review the supplied caption. Do not rewrite it or generate new editing parameters.

Output Requirements

Return only valid JSON containing the following fields:

```
{
  "format_pass": true,
  "specificity_pass": true,
  "faithfulness_pass": true,
  "overall_pass": true,
  "issues": []
}
```

Set the Boolean values according to the actual review findings. overall_pass is true only when all three preceding fields are true.

If the caption fails the review, each entry in issues must follow this structure:

```
{
  "criterion": "Format Compliance / Descriptive Specificity / Editing Faithfulness",
  "text": "The problematic excerpt from the caption; for an omission, specify what is missing",
  "problem": "The specific problem and the corresponding visual evidence, parameter information, or format
    requirement",
  "suggestion": "A revision suggestion to guide caption regeneration"
}
```

If the caption passes the review, return an empty array for issues.

Figure S5: Prompt template for verifying state-change caption validity and faithfulness.

User Instruction Generation: Initial Draft

Role and Objective

You are a professional image editing intent analysis expert. Your core task is to analyze provided image information and infer and simulate user photo editing intentions.

Input Information

You will receive the following input:

- Original Image (Image 1)
- Edited Version of Image (Image 2)
- Editing Configuration Parameters (Lua) (e.g., brightness, saturation, sharpness values)

Analysis and Inference Requirements

1. Visual Change Analysis: Analyze the visual changes and adjustments made from Image 1 to Image 2.
2. Intent Inference: Based on these visual changes, infer the user's photo editing instructions or intent for the transformation from Image 1 to Image 2.
3. Non-Technical User Perspective:
   - Simulate the voice of a real customer, describing their needs directly in everyday English.
   - Imagine you are the actual user, explaining your editing needs to a trusted partner.
   - Do not mention any technical parameters, numerical values, or professional jargon (e.g., "brightness +20", "color temperature 5500K").
   - Translate technical parameters into approachable visual effect descriptions.

Characteristics of Simulated User Intent

The simulated user intent sentences you generate must meet the following criteria:

- High Frequency and Practicality: Use vocabulary commonly found in everyday photo editing scenarios.
- Non-Technical: Strictly avoid any professional editing terminology.
- Broad Applicability: Suitable for common daily image processing situations, including:
  - Social Media Sharing: Photos of people (selfies, group shots), landscapes (nature, urban), street scenes, still life (food, products), pets, and more.
  - Mobile Photo Post-Processing: Quick enhancements for everyday photos taken on mobile phones.
  - Personal Mementos: Travel photos, family photos, party photos, individual portrait close-ups, etc.

Vocabulary Selection Criteria

When generating user intent, please refer to and comprehensively cover the following categories of vocabulary:

- Mainstream App Features/Filter Names:
  - Include and emulate names of built-in filters or features from popular photo editing apps (e.g., "vibe" from Xingtu, "vibrance" from Lightroom, "film look" from VSCO).
  - Cover common preset style names (e.g., "black and white", "vintage", "cinematic", "Japanese aesthetic", "fresh and clean").
- Basic Visual Effect Descriptions:
  - Color: Regarding color temperature ("a bit warmer", "cooler tones"), tint ("more green", "more pink"), saturation ("richer colors", "less intense colors"), vibrancy ("more vivid", "dull colors").
  - Light & Contrast: Regarding brightness ("brighter", "darker"), contrast ("stronger contrast", "softer look"), highlights ("dull the highlights", "brighten the highlights"), shadows ("deepen shadows", "lift shadows").
  - Clarity & Detail: Regarding sharpness ("sharper", "blurry details"), texture ("add texture", "reduce grain").
  - Exposure: Regarding overall brightness ("underexposed", "overexposed").
- Mood and Atmosphere Expressions:
  - Emotional Association Words: Vocabulary conveying specific emotions or atmospheres (e.g., "cool and lonely", "warm", "melancholy", "vibrant", "serene", "dreamy", "premium feel").
  - Scene Atmosphere Words: Vocabulary simulating specific scene atmospheres (e.g., "sunset evening", "rainy day blues", "summer sunshine").
- Local Adjustment Instructions:
  - Local editing intents that might arise from advanced operations like masks or brushes (e.g., "brighten the face", "blur the background", "make the sky bluer", "whiten the skin", "remove distractions").
- Specific Subject Descriptions:
  - Vocabulary directly referring to specific objects in the frame for modification (e.g., "make the eyes pop", "make the clothes more vibrant", "change the background color").
  - General Adjectives/Adverbs:
  - Common modifiers such as "more", "a bit", "slightly", "very", "completely", "natural", "clean", "transparent".

Intent Types and Proportions

Based on the provided image changes, please prioritize generating the following three categories of user intent and allocate them according to the specified proportions:

1. Basic Color Adjustments (34%):
   - Focus: Instructions for adjusting fundamental parameters like color temperature, tint, saturation, light/darkness, and contrast.
   - Examples:
     - "Make it warmer."
     - "Darken the highlights."
     - "Boost the colors."
     - "Cool down the tones."
     - "Reduce the color intensity."
     - "Brighten it up a bit."
     - "Make it a bit darker."
     - "Increase the contrast."
     - "Soften the shadows."
     - "Make the blacks deeper."
     - "Lighten the overall image."
     - "Make the colors pop more."
     - "Desaturate the image slightly."
     - "Give it a less vibrant look."
     - "Adjust the tint towards green."
     - "Make the whites brighter."
     - "Reduce the harshness of the light."
     - "Give it a more balanced exposure."
     - "Make the blues more vivid."
     - "Soften the overall color palette."
2. Stylized Filters (33%):
   - Focus: Covering preset filter styles in popular photo editing apps or expressions that achieve similar effects.
   - Examples:
     - "Give it an ins minimalist vibe."
     - "Make it look like Fuji film."
     - "Add a cyberpunk feel."
     - "Give it a cinematic look."
     - "Make it feel like old Hong Kong movies."
     - "Convert it to black and white."
     - "Add a Japanese fresh style."
     - "Give it a vintage feel."
     - "Make it look like Kodak film."
     - "Add a dreamy bokeh effect."
     - "Give it a bright and airy feel."
     - "Make it look like a magazine cover."
     - "Add a sepia tone."
     - "Give it a grungy look."
     - "Make it moody and dramatic."
     - "Add a pastel effect."
     - "Give it an oil painting texture."
     - "Make it look like a hand-drawn sketch."
     - "Add a vibrant pop of color."
     - "Give it a soft focus look."
     - "Make it feel like a warm summer day."
     - "Add a cool winter ambiance."
     - "Give it a glowing skin effect."
     - "Make the colors muted and gentle."
     - "Add a classic portrait style."
     - "Give it a street photography vibe."
     - "Make it look like a polaroid picture."
     - "Add a matte finish."
     - "Give it a cinematic blue-green tint."
     - "Make it feel clean and crisp."
3. Mood and Atmosphere (33%):
   - Focus: Color expressions based on emotional associations that convey specific moods or atmospheres.
   - Examples:
     - "Make it feel cold and lonely."
     - "Give it a vibrant and joyful look."
     - "Add a vintage nostalgic touch."
     - "Make it more peaceful and calm."
     - "Give it an intense and powerful feel."
     - "Make it mysterious and intriguing."
     - "Add a dreamy and ethereal touch."
     - "Give it a cozy and comfortable vibe."
     - "Make it exciting and energetic."
     - "Add a melancholic touch."
     - "Make it feel fresh and natural."
     - "Give it a serious and solemn mood."
     - "Make it playful and whimsical."
     - "Add a festive and celebratory feel."
     - "Give it a clean and minimalist look."
     - "Make it feel dark and edgy."
     - "Add a soft and delicate touch."
     - "Give it a classic and timeless appeal."
     - "Make it feel authentic and real."
     - "Add a cheerful and optimistic mood."
     - "Give it a tranquil and serene atmosphere."
     - "Make it feel dramatic and impactful."

Output Requirements and Format

- Overall Style Intent: You must include the user's editing intent for the image's overall style.
- Quantity: You must provide four possible user intents for the edited image.
- Diversity: These four possible user intents may convey similar meanings, but their sentence structures and vocabulary must be different.
- Word Limit: Each user intent description MUST be within <word_number> words.
- Local Adjustments Integration: Inferred local adjustment intents should be adaptively and naturally integrated throughout the description text.
- Strict Output Format: Please strictly adhere to the following JSON format for your output:

JSON

```
{
"User Intent 1": "<first user intent>",
"User Intent 2": "<second user intent>",
"User Intent 3": "<third user intent>",
"User Intent 4": "<fourth user intent>"
}
```

The provided information is as follows:

Figure S6: Prompt template for initial user-instruction generation.

Reference-Grounded Instruction Refinement

Role and Objective

You are a professional image editing intent refinement expert.
Your task is to refine a draft user editing instruction into a clearer instruction for a reference-style image editing task.

The final instruction should mainly describe the dominant visual changes, while grounding those changes in transferable style characteristics shown in the Reference Image.
The connection to the Reference Image must be concrete and visually specific, but the wording should remain flexible and natural.
The draft instruction must remain the main content source.

Input Information

You will receive the following input:

- Reference Image: the image whose transferable style characteristics should guide the edit.
- Edited Result Image: an edited image that already reflects the intended visual change.
- Editing Configuration Parameters: the lua configuration corresponding to the Edited Result Image.
- Draft User Instruction: an initial instruction inferred from the Edited Result Image and the Editing Configuration Parameters. It may describe the main edit direction, but it may not clearly express the connection to the Reference Image.

Refinement Requirements

1. Preserve the Main Intent:
   - The refined instruction must be mainly based on the Draft User Instruction.
   - Do not replace it with a completely new editing goal.
   - Do not introduce new subjects, scene content, layout changes, or object changes that are not supported by the inputs.
2. Describe the Visual Changes Clearly:
   - The main content of the refined instruction should describe the dominant or task-relevant visual changes.
   - These changes may be global or local, as long as they are important to the edit and supported by the images.
   - Avoid incidental details that do not help describe the intended edit.
   - Keep the instruction change-oriented and directly useful for image editing.
3. Use the Reference Image as Style Evidence:
   - Compare the Edited Result Image with the Reference Image to identify their shared transferable style traits.
   - Use those shared traits to make the Draft User Instruction more accurate and reference-aware.
   - The refined instruction must connect the edit to transferable style characteristics from the Reference Image.
   - At least one concrete transferable style characteristic must be explicitly presented as coming from, being guided by, or being aligned with the Reference Image.
   - Do not rely on a broad reference relationship alone. The instruction must state what transferable style characteristics from the Reference Image guide the change.
   - Do not focus on reference-image details that are specific to that sample and not useful as an editing style.
   - The wording should not follow a fixed template; choose natural phrasing that fits the images and the Draft User Instruction.
   - Most of the sentence should still describe the actual visual changes.
   - Do not ask to copy objects, people, background layout, or scene content from the Reference Image.
4. Use Configuration as Supporting Evidence:
   - Use the Editing Configuration Parameters to better understand the actual visual adjustment direction.
   - Treat the configuration as supporting information, not as text to be copied.
   - If the configuration conflicts with the visible similarity between the Edited Result Image and the Reference Image, prioritize the visible similarity.
5. Natural User Expression:
   - Write from the perspective of a real user giving an image editing request.
   - Use natural everyday English.
   - Avoid technical parameters, numerical values, lua field names, and professional editing jargon.
   - Translate any technical configuration into plain visual descriptions.
6. Reference Wording Diversity:
   - If sample-specific Reference Wording Guidance is provided, follow that guidance for this output.
   - Integrate the reference relationship naturally into the visual change description.
   - Keep the concrete transferable style characteristic clear wherever the reference relationship appears.

Output Requirements

- Output one refined English user instruction.
- The instruction must be concise, natural, and directly usable as an image editing request.
- Keep the content mainly aligned with the Draft User Instruction.
- Describe the dominant or task-relevant visual changes as the main narrative.
- Ground the visual changes in concrete transferable style characteristics from the Reference Image.
- Make the source relationship explicit for at least one transferable style characteristic.
- Avoid broad reference mentions that do not specify the actual transferable style characteristic being followed.
- Do not turn the instruction into a generic "copy the reference" request.
- Do not mention file names such as reference.jpg, finish.jpg, or finish.lua.
- Do not output analysis, explanations, markdown, or JSON.
- Word limit: within <word_number> words.

The provided information is as follows:

Figure S7: Prompt template for reference-grounded user-instruction refinement.

Reference-Grounded Local Instruction Refinement

Role and Objective

You are a professional local image editing intent refinement expert.
Your task is to refine a draft user editing instruction into a clear, natural, and local instruction for a reference-style image editing task.

The final instruction must describe an adjustment to a specific visible subject, object, or region in the image, not a global transformation of the whole image.
It should preserve the main visual direction of the Draft User Instruction, localize that direction to the most appropriate target region, and ground the local change in transferable style characteristics shown in the Reference Image.

The connection to the Reference Image must be concrete and visually specific.
The wording should remain flexible, natural, and individually fitted to the current image content.

Input Information

You will receive the following input:

- Reference Image: the image whose transferable style characteristics should guide the local edit.
- Edited Result Image: an edited image that already reflects the intended visual change.
- Editing Configuration Parameters: the lua configuration corresponding to the Edited Result Image.
- Draft User Instruction: an initial instruction inferred from the Edited Result Image and the Editing Configuration Parameters. It may describe the main edit direction globally, but may not clearly identify a local target or express the connection to the Reference Image.

Refinement Requirements

1. Preserve the Main Intent:
   - The refined instruction must be mainly based on the Draft User Instruction.
   - Preserve its supported visual direction, such as making an area brighter, warmer, softer, clearer, more colorful, or more atmospheric.
   - Do not replace it with a completely new editing goal.
2. Identify a Clear Local Editing Target:
   - The refined instruction must clearly identify a specific local target in the image.
   - The target may be a person, face, skin, clothing, sky, water, plant, building, foreground, background, product, food item, pet, vehicle, or another clearly visible subject or region.
   - Prefer one primary target region.
   - Do not use vague targets such as "the image", "the photo", or "everything" when a more specific subject or region can be identified.
3. Describe the Local Visual Change Clearly:
   - Focus on one main adjustment or two closely related adjustments within the selected local target.
   - Describe visible results rather than technical operations.
   - Keep the instruction change-oriented and directly useful for image editing.
4. Use the Reference Image as Style Evidence:
   - Compare the Edited Result Image with the Reference Image to identify transferable visual characteristics relevant to the selected local target.
   - At least one concrete transferable style characteristic must be explicitly linked to the Reference Image.
   - Do not refer to the Reference Image only in a broad or generic way; state what specific visual quality is being followed.
   - Do not ask to copy people, objects, backgrounds, or composition from the Reference Image.
5. Use Configuration as Supporting Evidence:
   - Use the Editing Configuration Parameters to better understand the likely adjustment direction.
   - If the configuration conflicts with the visible evidence, prioritize the visible evidence.
   - Do not mention numerical values, lua field names, or technical editing operations.
6. Natural User Expression:
   - Write from the perspective of a real user giving an image editing request.
   - Use natural everyday English.
   - Avoid technical parameters, numerical values, lua field names, and professional editing jargon.
7. Reference Wording Diversity:
   - If sample-specific Reference Wording Guidance is provided, follow that guidance for this output.
   - Integrate the reference relationship naturally into the visual change description.
   - Keep the concrete transferable style characteristic clear wherever the reference relationship appears.
   - Vary how the Reference Image is introduced across different samples.

Output Requirements

- Output one refined English user instruction.
- The instruction must describe a local image editing intent, not a global image adjustment.
- Clearly identify the local subject, object, or region that should be edited.
- Keep the content mainly aligned with the Draft User Instruction.
- Ground the visual changes in concrete transferable style characteristics from the Reference Image.
- Avoid broad reference mentions that do not specify the actual transferable characteristic being followed.
- Do not turn the instruction into a generic "copy the reference" request.
- Do not mention file names such as reference.jpg, finish.jpg, or finish.lua.
- Do not output analysis, explanations, markdown, or JSON.
- Word limit: within <word_number> words.

The provided information is as follows:

Figure S8: Prompt template for reference-grounded local user-instruction refinement.

Content Summary Prompt

You are a computer-vision data annotator. Generate an extremely concise object-relation summary of the image for machine understanding.

Core Objective

1. Identify and describe only the main objects in the image and their directly visible relationships.
2. Describe the main objects concisely, accurately, and objectively, including only visible states or actions.
3. Compress the output to no more than 30 English words while preserving essential information.

Strict Rules

- Output language: English only.
- Output format: one paragraph containing exactly one sentence.
- Do not mention emotions or mental states.
- Do not use specific ages. For people, use only man, woman, child, or older adult; use person or people when the type is unclear.
- Do not use inference or hedging words such as seems, appears, or likely.
- Describe only the objects themselves and their directly visible physical or spatial relationships.
- Ignore clothing, background, environment, and decorative objects unless they are main objects or directly participate in an object relationship.
- Ignore actions unrelated to the main object relationships, such as gestures or poses.
- The output must contain 1–30 English words. Any output exceeding 30 words is invalid.

Output-Form Rules

A. When people are present

When a person is a main object or participates in a main relationship:

- Begin with the number of people, using only one, two, three, four, or a group of.
- Follow the number with the most specific visible person type: man, woman, child, or older adult.
- If all people have the same type, use a form such as two men.
- If the types are mixed or unclear, begin with two people or three people, and specify types only for people involved in the main relationship.

B. When no main person is present

- Do not write No people are visible... or any other expression meaning that nobody is present.
- Begin directly with the most important object.
- Begin with only one, two, three, four, or a group of.
- Use concrete nouns such as car, building, road, bowl, plate, tree, mountain, or river.
- Prefer directly visible spatial relationships such as on, in, under, beside, behind, above, or near.

Relationship Verbs and Quantity Limits

- Use at most 2 relationship verbs in the entire sentence.
- Relationship verbs may only be selected from: holding, touching, supporting, containing, covering, resting, and reflecting.
- Mention at most 4 objects in total, including people.
- Necessary grammatical auxiliaries such as is and are are allowed.
- Prefer physical or spatial relationships.
- Do not mention colors, lighting, style, aesthetic terms, or narrative language.
- To satisfy the 30-word limit, remove secondary objects and retain only the core objects and their main relationship.

Figure S9: Prompt template for content summarization in semantic retrieval.

Style Summary Prompt

1. Task Role

You are a professional photographic visual director. Based on the provided single image, directly analyze and define its current final visual style state, extracting its aesthetic characteristics and
visual language.

This task is applied to the original image.

2. Task Boundary

• This task contains no concepts of before-and-after comparison, post-processing changes, or an original state.
• Treat the image as a complete visual work that already exists in its current form.
• Your responsibility is to answer: "What visual style does this image itself present?"
• Discard the judgment of whether the image has been processed. Describe the image directly from the perspectives of professional photography and post-production language, abstracting its
color science, tonal structure, light-and-shadow texture, and overall atmosphere.

3. Core Analysis Process

3.1 Image Structure Recognition

• Identify the subject region, meaning the high-frequency detail-dense area.
• Identify the environment region, meaning the low-frequency light-and-shadow-bearing area.
• Use this step only to support style understanding; do not explicitly describe objects in the result.

3.2 Visual Style Decomposition

Describe the image's current established visual state through the following three dimensions:

• Color Science
• Tone Mapping
• Texture

3.3 State-Based Description Principle

• Answer only "what it is like."
• Do not answer "what it has experienced" or "what it could have been."

4. Analysis Dimensions and Strict Constraints

4.1 Overall Style Definition

• Give the image the most fitting professional style name, such as:
• low-saturation cool-toned documentary style;
• cinematic teal-and-orange grade;
• soft-haze natural-light style;
• high-contrast dark-tone style.
• Describe the overall emotion and visual character created by the style, such as:
• calm;
• restrained;
• contemplative;
• dreamlike;
• tense;
• impactful.

4.2 Color and Tone

Absolute State Description

Describe the hue tendency and saturation state of the current:

• shadow regions;
• midtone regions;
• highlight regions.

Light-and-Shadow Form

Describe the current image's:

• contrast ratio;
• tonal layering;
• transition pattern.

Descriptions may include:

• hard-toned;
• high-contrast;
• concentrated grayscale;
• softly diffused;
• highlight blooming.

5. Key Constraints

5.1 Non-Comparative Description

Strictly prohibit any comparative, change-related, or action-oriented wording.

Do not use:

• "compared with...";
• "originally...";
• "increased...";
• "decreased...";
• "darkened...";
• "emphasized...";
• "preserved...";
• "removed...".

Use only static adjectives and state judgments, such as:

• rich;
• restrained;
• cool-biased;
• low-saturation;
• high-contrast.

5.2 Object-Free Description

Do not mention any specifically identifiable object, such as:

• people;
• animals;
• plants;
• buildings;
• roads;
• lights.

Use abstract photographic terms, such as:

• subject region;
• foreground;
• midground;
• background;
• high-frequency texture region;
• environmental light-and-shadow layer.

6. Output Requirements

• The values of all JSON fields must be in English.
• Only English letters, numbers, punctuation, and spaces are allowed.
• Return only one valid JSON object.
• Do not include any additional explanatory text.

```
json
{
"style_title": "core style name",
"style_summary": "one-sentence summary of the image's current aesthetic state; no comparative words or specific object terms (within 100 characters)",
"technical_details": {
"color": "professional description of the current color distribution and color relationships (absolute state)",
"lighting": "professional description of the current light-and-shadow structure and tonal form (absolute state)",
"texture": "professional description of the current image texture and detail presentation"
},
"visual_tags": ["tag 1", "tag 2", "tag 3", "tag 4"]
}
```

Descriptions must be precise and should use professional terminology from photography and post-production.

Figure S10: Prompt template for grading-style summarization in reference reranking.

Prompt for Parameter-Grounded State-Change Captioning

Task Description
You are a professional image editing analysis expert. Please carefully observe the provided original image and edited image, and analyze and describe the visual effects and changes produced by each parameter in combination with the given configuration parameters.

Input Content
1. Original Image: [Original image]
2. Edited Image: [Edited image]
3. Configuration Parameters: [Lightroom Classic Lua format configuration file (config.lua)]
4. Parameter Reference: [Detailed descriptions of parameters that may be involved]
   - Reference format: ParameterName -> Professional Term -> Detailed Description
   - Provides mapping between LRC parameters and color grading terminology with functional explanations

Output Requirements

1. Parameter Naming Convention
- When first mentioning a technical parameter in the description, you must use the format: Professional Term (Parameter Name)
- For example: "Exposure (Exposure2012)", "Contrast (Contrast2012)", "Vibrance (Vibrance)"
- **Mandatory Requirements:**
  - Outside the parentheses must be the Professional Term (from the second column of the parameter reference table)
  - Inside the parentheses must be the Parameter Name (from the first column of the parameter reference table)
  - The professional term must strictly come from the second column of the parameter reference table (ParameterName -> Professional Term -> Detailed Description)
  - Do not use the parameter name as the professional term without consulting the table
  - Do not reverse the positions by writing "Parameter Name (Professional Term)" - this is incorrect
  - For any parameter, you must first locate it in the reference table, then use its corresponding professional term
- **Mandatory rules for subsequent mentions of parameters:**
  - You must ONLY use the professional term that appeared outside the parentheses in the first mention
  - You must NEVER use the parameter name that appeared inside the parentheses in the first mention
  - Regardless of the parameter name format (containing numbers, uppercase letters, etc.), it must not appear again in subsequent mentions

2. Analysis Content Requirements
Please provide detailed analysis for each configuration parameter, including but not limited to:
- The specific value or setting of the parameter
- The direct visual effect the parameter produces on the image
- The impact range of the parameter on specific areas or the overall image
- Synergistic effects with other parameters (if any)

3. Expression Diversity Requirements
- Use rich visual descriptive vocabulary and avoid repetitive expressions
- Avoid using the same description style for all parameters; make each parameter's description vary in sentence structure, level of detail, and order of information presentation
- Employ various sentence structures:
  - Some parameters can describe the visual effect first, then introduce the parameter name and value
  - Some parameters can use causal sentence patterns, emphasizing the relationship between parameters and effects
  - Some parameters can incorporate contrast sentences, explaining changes or differences from other approaches
  - Some parameters can be naturally integrated into the overall description without separate emphasis
  - For parameters with non-default values in the configuration file, provide detailed explanations, focusing on the specific visual effects that parameter value produces on the original image
- Vary sentence length and complexity, combining long and short sentences to avoid monotonous syntax
- Flexibly combine professional terminology with colloquial explanations
- Use rhetorical devices such as analogy and contrast to enhance expression
- Flexibly vary the order of information presentation; don't use the same narrative sequence for all parameters

4. Output Format
Please organize your analysis in flowing paragraphs, presented in a written narrative style, similar to writing detailed caption text for an image. Do not use bullet points, tables, or structured list formats; instead, naturally integrate the effects of each parameter into complete descriptions within coherent paragraphs.

The output should include the following content, but expressed in continuous paragraph form:
- An overview of the overall changes
- Visual effects, impact range, and technical explanation of each parameter
- Synergistic effects between parameters
- Summary of the overall editing effect

Ensure that when first mentioning a parameter, use the "Professional Term (Parameter Name)" format. Descriptions should be vivid, natural, and expressive.

Example Output (For Reference)

This image, after meticulous adjustment in Lightroom Classic, presents a bright and transparent visual quality without losing depth. The scene is enveloped in a soft golden warmth with distinct light-dark contrast, particularly luminous in the highlight areas. Details and textures are clearly discernible, yet the overall color palette is surprisingly restrained—not boldly vibrant, but subtly elegant, achieving a delicate balance between warmth and restraint.

Originally dim areas have become transparent and bright, thanks to the Exposure (Exposure2012) setting of 0.59—measured in EV units and applied globally, with particularly significant brightening of the midtones. With Contrast (Contrast2012) set to 35, the image's sense of depth is enhanced accordingly. The luminous effect in the highlight areas comes from the Highlights (Highlights2012) value of 72, which precisely affects the brightest 20% of the image.

The Temp Offset (IncrementalTemperature) has been shifted toward warm tones by 16 Kelvin units. This adjustment, based on custom white balance, infuses the scene with golden warmth, making white areas and skin tones appear more inviting. However, the colors haven't become intense with the warm tones; instead, they present an elegant quality—Vibrance (Vibrance) has been reduced to -42. Unlike crude overall desaturation, Vibrance intelligently avoids already saturated colors, focusing on reducing the vividness of unsaturated colors and skin tones, preventing overflow while maintaining restraint.

The enhancement of edge contours and textural quality stems from the moderate Clarity (Clarity2012) value of 18. It shapes details through mid-frequency contrast enhancement rather than traditional high-frequency sharpening, thus avoiding harshness. Parameters such as Shadows (Shadows2012), Whites (Whites2012), and Blacks (Blacks2012) have also been fine-tuned, playing roles in preserving shadow detail and tonal balance.

From an overall perspective, the combination of exposure and contrast avoids the grayish appearance that pure brightening might cause, while warm tones paired with reduced saturation strike a balance between warmth and restraint. The significant brightening of highlights combined with moderate clarity enhancement work together, making the scene both bright and transparent while retaining delicate texture. These finely coordinated parameter adjustments ultimately transform an ordinary photograph into a work with warm tones, rich layers, and abundant detail.

Important Notes
- Output Language: All descriptions and analysis must be in English
- Please base your analysis on actually observed image changes, ensuring accurate and objective descriptions
- **Core principles for parameter naming:**
  - First mention: Must look up the professional term in the table, use "Professional Term (Parameter Name)" format, with professional term outside parentheses and parameter name inside
  - Subsequent mentions: Only use the professional term (word outside parentheses), the parameter name (word inside parentheses) must not appear again
  - Under no circumstances should you make up terms intuitively; you must strictly use the corresponding terms from the parameter reference table
- While maintaining professionalism, keep descriptions vivid and accessible
- If certain parameter effects are not obvious or not enabled, this should also be noted
- Avoid using rigid, templated language
- Key: Each parameter's description should be naturally varied, with varying levels of detail, some elaborate and some simple. Avoid mechanically using the same narrative structure and level of detail for every parameter

Figure S11: Prompt template for parameter-grounded state-change captioning.

Expert Reasoning: Multi-Perspective Intent Analysis

Execute an expert-level image adjustment analysis using the provided before/after images, user requirements, reference image(s), and configuration parameters. Your task is to synthesize the textual intent with visual style characteristics from the reference, producing a comprehensive understanding of the desired editing outcome. Structure your analysis as follows:

Reverse-Engineering Intent Analysis:

- Describe Original Image Elements: Identify and articulate the primary visual elements and compositional features in the before-editing image (e.g., subject matter, key objects, spatial arrangement, dominant textures, environmental context). Focus purely on content description without mentioning color, tone, or stylistic attributes. These elements serve as the foundational carriers of style.
- Identify Original Image Baseline: Analyze the before-editing image to establish its inherent style characteristics (e.g., natural lighting, muted colors, balanced contrast, documentary tone). This baseline reveals what needed transformation.
- Deconstruct Reference Image Style: Systematically extract visual style attributes from the reference image(s), including color palette (specific hue shifts, saturation levels), tonal structure (shadow depth, highlight treatment, midtone character), texture rendering, lighting mood, and compositional emphasis.
- Parse User Instruction: Break down the textual directive into explicit editing requirements (e.g., "increase warmth," "add drama," "match reference tone") and identify which specific aspects of the reference the user wants to emulate.
- Map Reference Borrowings: Cross-reference the user's stated borrowing intentions with the reference image analysis. Verify that requested attributes (e.g., "use the golden tones from reference") are actually present in the reference and articulate their technical manifestation.
- Synthesize Editing Intent: Combine the original image's starting point, the reference's target attributes, and the user's explicit directives to reconstruct the editor's complete decision-making process, including implicit goals not directly stated (e.g., maintaining subject recognition while shifting mood).
- Classify Target Aesthetic: Define the final desired style category based on the convergence of all evidence (e.g., cinematic, editorial, vintage analog, commercial product).
- Isolate Three Defining Attributes: Identify the most critical style characteristics that bridge the gap from original to target, grounded in both reference analysis and editing outcomes.

Output Mandates:

- Use ONLY English throughout
- Maintain an expert tone and avoid using expressions such as "according to the configuration file" or "by observing the results".
- Output the result concisely (<220 words) using the exact format: Original Image Elements -> Original Image Baseline -> Reference Image Style Analysis -> User Instruction Decomposition -> Reference Borrowing Mapping -> Synthesized Editing Intent -> Final Desired Style -> Key Characteristics of That Style.

Output example:

Original Image Elements: Grand mosque with intricate domes and minarets as central subject, surrounded by palm trees, reflecting in serene water body. Twilight time period with visible sky. Symmetrical composition emphasizing architectural details and water reflections.

Original Image Baseline: Naturally lit scene with neutral color temperature, moderate contrast, documentary-style tonal distribution, minimal atmospheric haze, evenly exposed across all zones.

Reference Image Style Analysis: Golden-hour illumination aesthetic featuring saturated warm amber tones, crushed shadow tones with retained detail, elevated highlight luminance creating glow effect, blue-gold complementary color contrast in sky-structure relationship, enhanced atmospheric depth through controlled haze layering.

User Instruction Decomposition: Explicit directive to "enhance warmth and drama while matching the reference's golden-hour feel." Implicit requirement to preserve architectural detail legibility while elevating visual impact.

Reference Borrowing Mapping: Requested golden-hour warmth maps to reference's warm amber bias in highlights and midtones. "Drama" correlates to reference's shadow depth and blue-gold color separation. Atmospheric quality mirrors reference's haze rendering technique.

Synthesized Editing Intent: Transform neutral documentary baseline into cinematic architectural portrayal by applying reference's warm color grading, tonal contrast elevation, and atmospheric depth enhancement while maintaining structural clarity and compositional integrity.

Final Desired Style: Cinematic Architectural Portraiture.

Key Characteristics of That Style:

1. Saturated Golden-Hue Dominance: Warm amber tones dominate illuminated surfaces, contrasting with cool twilight blues.
2. Dynamic Range Control: Shadows retain texture while highlights glow without clipping, preserving depth.
3. Enhanced Atmospheric Perspective: Controlled haze layering amplifies subject-background separation and reflective symmetry.

Input Image Order:

1. First image: Original image (before editing)
2. Second image: Edited image (after editing)
3. Third image: Reference image (style target)

The provided user requirements and configuration parameters are as follows:

Figure S12: Prompt template for structured interpretation of grading intent.

Expert Reasoning: Technical Adjustment Strategy

Act as an image editing process analyst performing reverse-engineering of editing decisions. Given the original image, completed editing result, reference image, user instruction, professional intent analysis, and configuration parameters, reconstruct the editor's technical reasoning process during the editing operation.

Input Image Order:

1. Original image (before editing - the starting point for all edits)
2. Edited image (after editing - use this to understand what was achieved)
3. Reference image (style target - provides visual guidance)

Additional Provided Information:

4. User's instruction (textual directive referencing which aspects of the reference to apply)
5. Configuration parameters (e.g., Lightroom preset/LUT values)
6. Professional analysis of the user's intent statement

Your Task:

Reconstruct the editor's thought process during editing by reverse-engineering from the final result. Generate a single-paragraph technical analysis (about 120 words, no headers) that articulates:

1. Decision rationale for technical adjustments:
   - What specific aspects the user instruction (analyzed through professional interpretation) requests to be modified
   - Which reference image characteristics address those user requirements and how they provide technical guidance
   - How the combination of user intent and reference analysis determines the specific technical operations to be employed (e.g., "will apply global luminance curve reshaping," "will use selective chroma suppression")
   - The intended visual changes these operations will achieve (e.g., "to darken shadows," "to warm colors," "to soften background")
2. Predicted visual outcomes:
   - What specific effects the parameter adjustments will produce on the image
   - How these effects fulfill the user's requirements while incorporating the reference's demonstrated style attributes
   - The final visual character expected from the integrated application of user intent and reference guidance

Structure Requirements:

- Write as a continuous narrative in Future Tense (describing what will be done and what effects will result)
- DO NOT reference the edited image directly (e.g., avoid "the output shows," "the result exhibits")
- Instead, frame as predictions: "will produce," "will achieve," "will exhibit," "this will result in"
- Avoid mentioning specific numerical values
- Maintain expert tone without phrases like "according to the configuration file"
- Vary sentence structures: Mix simple, compound, and complex sentences; alternate between active constructions and passive voice where appropriate; use diverse transition words and clause arrangements to avoid monotonous patterns

Special Notes:

1. All outputs must be strictly in English, prohibiting the use of any other languages.
2. Please strictly enforce the word limit (about 120 words).
3. Mandatory: Use Future Tense throughout - describe intended actions and predicted outcomes, not completed results.
4. Prioritize linguistic variety: Avoid repetitive sentence openings, vary clause length, and employ different grammatical structures to create a natural, professional flow.

Example:

"To achieve the user's request for warmer, atmospheric tones, the reference's golden-hour color palette will guide a graduated temperature shift that pushes midtones toward amber while maintaining architectural detail clarity through selective sharpening. By adopting the reference's layered depth approach, localized haze adjustments will enhance subject-background separation via controlled shadow compression. The tonal curve will be reshaped to mirror the reference's luminance distribution, lifting shadow detail without sacrificing highlight integrity. Expected outcomes include visibly enriched golden highlights contrasting with cooler twilight zones, improved atmospheric perspective, and enhanced color harmony that reflects the reference's cinematic character while fulfilling the user's creative intent."

The provided information is as follows:

Figure S13: Prompt template for mapping grading intent to parameter decisions.

Expert Reasoning: Final Trajectory Integration

You are an expert in image editing and are assigning image editing tasks to your apprentice. You need to organize your language based on the available information to clearly explain the requirements to the apprentice.

IMPORTANT - Role Consistency:
You ARE the expert speaking directly to your apprentice. Use first-person perspective ("I think", "my approach", "here's how I see it") when sharing your analytical thinking. NEVER refer to yourself in third person (avoid phrases like "based on expert analysis", "the expert suggests", "according to the analysis").

Available Information:
- Original user intent statement
- Task analysis and decomposition (inc. the analysis of the original image, the explicit and implicit needs of the users, the style type that the users ultimately hope to achieve, the key features of that style)
- Your analytical thinking on adjustment strategies
- Original image (first image input)
- Reference image(s) (second image input)

Core Principle:
Your task is NOT to list information, but to demonstrate reasoning. Show how you use reference image analysis to decode what the user's instruction really means, revealing both stated and unstated intentions. The reference analysis is a tool for intent discovery, not a separate description to include.

Guidelines:
1. Conversational & Colloquial Style:
   Use casual, spoken language as if explaining face-to-face to your apprentice. Employ colloquial expressions like "pretty much", "what they're really after", "that'll give us". Use contractions (we're, they're, that'll, it's) and informal connectors (so, now, but). Vary your transitional phrases throughout - don't repeat the same expressions; mix it up naturally. Avoid formal academic language, bullet points, or structured formatting - make it flow like natural speech.
2. Vary Opening Phrases:
   IMPORTANT: Every response MUST start with the word "Alright" as the very first word. After "Alright", vary the continuation to maintain freshness and natural variety. Use diverse casual phrases like "Alright, so here's what we're working with", "Alright, looking at this image", "Alright, the goal here is", "Alright, what we've got is", "Alright, here's the situation", "Alright, what we need to do here is", or similar natural conversational continuations that fit the specific image and task.
3. First-Person Strategic Thinking:
   Share YOUR analytical thinking naturally using conversational first-person expressions like "my approach is to", "I want to", "the way I see it", "here's my thinking". Talk through YOUR technical rationale as if thinking aloud with the apprentice, not lecturing them.
4. Intent-Driven Narrative Flow:
   Don't simply list information elements. Instead, build a reasoning narrative that shows how you deduce the user's true intent by connecting the dots: Start from what you observe in the original image, use the reference image analysis to decode what the user's instruction is really pointing at (which specific visual attributes they want borrowed), reveal both explicit and implicit needs through this cross-analysis, conclude with the target style and key characteristics, and naturally transition into YOUR strategic thinking on achieving it. The narrative should feel like detective work—piecing together clues from multiple sources to uncover the complete picture.
5. Decode Instructions Through Reference Analysis:
   The user's textual instruction is NOT a standalone command—it's a pointer to specific features in the reference image. Your job is to decode this relationship: When the user says "add warmth," you must analyze the reference to identify WHICH warmth (e.g., "those golden amber tones in the highlights"), and then reason about WHY they want it (explicit: warmer colors; implicit: more drama, cinematic feel). The reference image analysis exists solely to unlock the hidden meaning behind user instructions. Never describe the reference in isolation—always tie its characteristics back to what the user is actually trying to achieve.
6. Seamless Integration:
   The output must NOT be a simple concatenation of "task analysis" section followed by "technical thinking" section. Instead, weave together the task context and your strategic thinking into a cohesive, natural narrative where analytical observations flow smoothly into technical decisions. The narrative should feel like a continuous conversation, not separate modules pasted together.
7. Content Restriction:
   The output content must be based on the information provided. Absolutely prohibit mentioning any numerical values related to adjustment operations (e.g., avoid "+15 exposure", "saturation 40", "contrast -10", "2800K color temperature"). Describe adjustments qualitatively using terms like "boost", "lift", "darken", "enhance", "warm tones", "deeper shadows" instead of specific numbers.

Output Requirements:
- Language: All outputs must be strictly in English, prohibiting the use of any other languages
- Length: Strictly enforce the 150-word limit
- Structure: Output as one single paragraph that seamlessly weaves together all required elements: original image analysis, reference image style characteristics, both explicit and implicit user needs (as professionally deconstructed from their intent), target aesthetic style, key style features, AND your strategic technical thinking on how to achieve the outcome. The narrative should flow naturally from observation to strategy, creating one cohesive expert explanation rather than separate description and plan sections.
- No Numerical Values: Absolutely prohibit mentioning any specific adjustment numbers or parameter values (e.g., "+15", "40%", "2800K", "-10"). Use qualitative descriptions only.

Examples:

Alright, so here's what we're working with—a grand mosque at twilight, pretty cool and blue overall with just a hint of golden light. The user's asking to "apply the golden-hour warmth from the reference to make it more majestic," which tells me they're pointing at something specific in that reference image. When I look at it, I can see what caught their eye: those saturated amber tones that transform ordinary architecture into something cinematic, that blue-gold contrast creating visual tension, and shadows deep enough to add drama but still holding detail. Now, on the surface they're asking for warmth, but reading between the lines—wanting "majestic" feel—they're really after the emotional impact that comes from the whole golden-hour treatment: expanded tonal range for depth, heightened saturation for presence, and atmospheric layering for that epic quality. That points us toward Enhanced Architectural Twilight style—saturated golden-hue dominance, controlled dynamic range, enhanced atmospheric perspective. My thinking is to use those reference amber tones as a roadmap for temperature shifts, mirror that luminance curve to lift shadows without losing highlight structure, and build atmospheric depth through localized adjustments—that's how we'll bridge the gap from documentary to cinematic.

The provided information is as follows:

Figure S14: Prompt template for integrating intent analysis and parameter decisions into a reasoning trajectory.

Direct Visual-Quality Reward Evaluation

You are an evaluator of rendered-image quality for photographic color grading. Assess only the intrinsic quality of the rendered output relative to the source image, using the listed quality criteria. Do not use the reference image, the user instruction, or the parameter configuration; those govern a separate task-alignment assessment.

EVALUATION PRINCIPLES

1. Compare the source image with the rendered output to judge whether each listed criterion's property is preserved or damaged. Judge every criterion directly from the two supplied images.
2. Evaluate only the intrinsic rendered-outcome quality and editing damage. Do not judge whether the output matches the reference or fulfills the instruction; those are out of scope here.
3. A global change in brightness, contrast, saturation, color temperature, or style is not itself a defect and must not block a clean score. Only a concrete, visible consequence for the specific criterion counts as a defect. Likewise, strong stylization, mood, or attractiveness cannot excuse a visible defect.
4. Score each criterion independently. One visible root cause may affect multiple criteria; describe it once per affected criterion rather than inventing unrelated justifications.

QUALITY CRITERIA CATALOG

Score every criterion in the following catalog. Each entry gives its name and the property to inspect.

{rubric_catalog}

SCORING CRITERIA

For each criterion, assign one of the following scores:

100: The property is visibly preserved from the source, with a concrete supporting observation.
70: A trace or mild defect is visible only on close inspection, or the property cannot be conclusively judged from the supplied images.
35: A defect is clearly visible under normal viewing and materially affects the property.
0: A severe or critical defect makes the property unacceptable.
null: The criterion has no applicable subject in this image pair.

Distinguish an inapplicable criterion (null), an unverifiable one (70, with the limitation stated), and a visibly unmet one (35 or 0). Do not use null for an observable defect or merely insufficient evidence.

OUTPUT FORMAT

Return only one valid JSON object containing:

- criteria: an array with one entry per catalog item, each containing:
  - criterion_id: the catalog identifier.
  - score: 100, 70, 35, 0, or null.
  - evidence: a concise observation in the form "Source: ...; Rendered: ...; Comparison: ...", or a brief statement of the inapplicability or evidential limitation when score is null or 70.

Score every listed criterion. Do not compute an aggregate quality score; that is calculated separately from these per-criterion scores.

INPUTS

Quality criteria catalog:
{rubric_catalog}

Source image:
{source_image}

Rendered output:
{rendered_output}

Score every listed criterion under the system rubric and return the specified JSON object.

Figure S15: Prompt template for the fourteen-criterion visual-quality reward.

Direct Parameter-Aware Instruction-Following Evaluation

You are an evaluator of photographic color grading. Assess instruction fulfillment using the source image, reference image, user instruction, rendered output, and parameter configuration actually used to produce the output.

EVALUATION PRINCIPLES

1. Evaluate against the user instruction. The reference image provides stylistic and tonal context, while the instruction determines which attributes to adopt and which adjustments to prioritize. Do not require the output to reproduce the reference in full or introduce requirements absent from the instruction.
2. Compare the source image with the rendered output to determine whether the direction, degree, and spatial scope of the adjustments satisfy the instruction. For reference-related requirements, also assess the relevant attributes in the reference image.
3. Use the parameter configuration only as auxiliary evidence of adjustment direction and scope. Base fulfillment judgments on visible outcomes. Do not infer success solely from parameter names, signs, or magnitudes, or assume that a visual effect requires a unique parameter configuration.
4. Evaluate only explicitly requested attributes. Unrequested aesthetic quality, naturalness, noise, artifacts, and similar issues belong to a separate visual-quality assessment and must not affect this score.
5. Respect the requested degree and spatial scope. For direction-only requests, do not impose an unstated target magnitude. Do not reject a local adjustment solely because the global appearance changes differently. An unchanged attribute may satisfy the instruction when preservation is requested or the source already meets the target.

EVALUATION DIMENSIONS

Assign a score directly to each of the following dimensions:

1. color: Explicit requirements concerning hue, warmth or coolness, color casts, saturation, palette, color preservation, and reference colors.
2. light_shadow: Explicit requirements concerning brightness, exposure, highlights, shadows, illumination, and luminance preservation.
3. technical: Explicit requirements concerning contrast, clarity, sharpness, texture, dehazing, noise reduction, and artifact correction.
4. high_level: Explicit requirements concerning overall style, mood, period appearance, and composite visual characteristics.

Assess every dimension against the complete original instruction. Do not infer specific color or lighting targets solely from broad descriptions such as "cinematic" or "serene"; these requirements may instead be evaluated under high_level.

SCORING CRITERIA

For each applicable dimension, assign one of the following scores:

100: The requirements are fully satisfied, with the requested adjustment direction, degree, and spatial scope.
70: The main requirements are satisfied, with limited remaining discrepancies.
35: The requirements are only partially satisfied, with substantial discrepancies remaining.
0: The requirements are not satisfied, or the output visibly contradicts them.
null: The dimension is inapplicable, or the available evidence is insufficient to assess fulfillment.

A dimension is inapplicable when the instruction contains no requirement for it. A requirement is unverifiable when it is present but the evidence is insufficient. Neither case should be confused with a visibly unmet requirement. Distinguish these cases in the brief explanation.

When a dimension contains multiple requirements, assess their fulfillment collectively. Do not let success on one requirement obscure other explicit requirements.

OVERALL SCORE

Compute instruction_score as the equal-weight mean of the applicable dimensions with numeric scores, divided by 100. The resulting score must lie between 0 and 1.

Exclude null entries from the denominator.

If all dimensions are null, return instruction_score=0 and valid=false, and explain that no scoreable requirements or sufficient evidence are available. Otherwise, return valid=true.

Calculate only instruction fulfillment. Do not calculate visual quality, format validity, or the final training reward.

OUTPUT FORMAT

Return only one valid JSON object containing:

• color: 100, 70, 35, 0, or null.
• light_shadow: 100, 70, 35, 0, or null.
• technical: 100, 70, 35, 0, or null.
• high_level: 100, 70, 35, 0, or null.
• instruction_score: The normalized overall score.
• valid: Whether at least one dimension can be scored.
• reason: A concise explanation of the main fulfilled requirements, discrepancies, and relevant evidential limitations. Do not provide an extended reasoning trace.

INPUTS

Original user instruction:
{user_instruction}

Source image:
{source_image}

Reference image:
{reference_image}

Rendered output:
{rendered_output}

Parameter configuration actually used to render the output:
{parameter_configuration}

Parameter definitions and spatial scope (optional):
{parameter_context}

Figure S16: Prompt template for parameter-aware instruction-following reward evaluation.

Online Derived Single Instruction Prompt

[ROLE]
You are an image-editing intent generation expert.

[INPUTS]
- Original Image: the image before the current edit.
- Edited Result Image: the image after the current edit.
- Reference Image: a source of transferable color characteristics.

[TASK]
Infer the visible color change from the Original Image to the Edited Result Image. Then write one natural English user request that would reproduce this change while grounding the target colors in the Reference Image.

[CONTENT SCOPE]
Describe only color-related edits:
- hue;
- saturation;
- palette;
- color temperature;
- color distribution.

Identify the most relevant visible region or regions. For a global change, describe the foreground and background separately when both are visibly affected. For a local change, name the specific visible subject or region. Preserve the actual color direction supported by the Original Image and the Edited Result Image.

[REFERENCE IMAGE USE]
Use concrete color characteristics from the Reference Image when appropriate. Transfer color properties only. Do not copy or request the transfer of objects, people, composition, layout, or background content.

[HARD CONSTRAINTS]
- Do not request changes to brightness, lighting, exposure, contrast, sharpness, clarity, texture, detail, composition, geometry, or object/content.
- Do not invent objects, regions, or color changes that are unsupported by the images.
- If a detail is uncertain, omit it instead of guessing.
- Do not mention technical parameters, software controls, files, analysis, or image-processing operations.
- Write a normal user instruction rather than a technical description.

[OUTPUT FORMAT]
Output exactly one plain-text English instruction.
Do not output JSON, analysis, markdown, labels, quotation marks, file names, or any additional text.
Keep the instruction within <word_number> words.

Figure S17: Prompt template for image-conditioned instruction generation during dynamic data derivation.



Instruction: Make my photo's colors and contrast sharp like the reference image.



Analysis, planning and decision making

<think> Alright, this one's a rescue mission. Our original shot of Dubai is pretty flat and hazy, but the user wants it to look "sharp like the reference." Looking at that reference, it's clear they're not just asking for a bit more contrast; they're after that hyper-real, graphic style. See how the reference has that intensely saturated blue sky, crushed blacks, and super vibrant colors? That's the goal: a Hyper-Real Urban Architectural look. My approach is to first cut through the haze with some dehaze to reveal the building details. Then, I'll aggressively boost the contrast and clarity to create that sharp, graphic quality. I'm going to shift the color palette, transforming the muted tones into that signature teal-and-orange scheme, and then use local adjustments to make the sky and temple pop. That'll give us the dramatic, high-impact result they're really after. </think>

First-step editing (global)

<think> Alright, let's break down this task. The original photo is a high-contrast coastal scene at dusk, with a dark foreground and a cool, blue sky. The user wants to transform this into a warm, inviting sunset, making it feel like a nostalgic memory. </think> <tool_call> CompatibleVersion = 251854848, ProcessVersion = "15.4", WhiteBalance = "Custom", Temperature = 5590, Tint = 10, Contrast2012 = 53, Highlights2012 = -41, Shadows2012 = 89, Whites2012 = 46, Blacks2012 = -40, Texture = 16, Clarity2012 = 34, Dehaze = 2, Vibrance = 53, Saturation = -30, PostCropVignetteAmount = -8, PostCropVignetteMidpoint = 50, PostCropVignetteFeather = 50, PostCropVignetteStyle = 1, Sharpness = 40, SharpenRadius = 1, SharpenDetail = 25, ParametricShadowSplit = 25, ParametricMidtoneSplit = 50, ParametricHighlightSplit = 75, HueAdjustmentRed = 49, HueAdjustmentOrange = -11, HueAdjustmentYellow = -1, HueAdjustmentAqua = -4, HueAdjustmentBlue = -15, SaturationAdjustmentRed = 41, SaturationAdjustmentOrange = 80, SaturationAdjustmentYellow = 75, SaturationAdjustmentAqua = 33, SaturationAdjustmentBlue = 1, SaturationAdjustmentPurple = -65, SaturationAdjustmentMagenta = -62, LuminanceAdjustmentRed = 23, LuminanceAdjustmentOrange = 42, LuminanceAdjustmentYellow = 1, LuminanceAdjustmentAqua = 19, LuminanceAdjustmentBlue = 15, SplitToningShadowHue = 241, SplitToningShadowSaturation = 14, SplitToningHighlightHue = 348, SplitToningHighlightSaturation = 2, ColorGradeMidtoneSat = 10, ColorGradeBlending = 50, ColorGradeGlobalHue = 34, ColorGradeGlobalSat = 5, ToneCurveName2012 = "Custom"… </tool_call>



Editing Tools Sandbox

Figure S18: Intermediate reasoning, predicted parameter configurations, and rendered outputs of JarvisAura.

System Prompt for Reference-Guided Color Transfer Evaluation (SC / PQ)

You are an expert evaluator of reference-guided photographic color transfer.

You will receive exactly three images in this fixed order:
1. SOURCE: the original photograph whose content must be preserved.
2. REFERENCE: a different photograph that defines the desired transferable color and tonal style. Its subject, objects, composition, and geometry may be unrelated to SOURCE.
3. EDITED: the candidate result produced by applying REFERENCE's color and tonal style to SOURCE.

There is no ground-truth image. Judge EDITED only from SOURCE and REFERENCE. Produce two independent scores from 0 to 10; half-point increments are allowed.

SC — Style Consistency:
Evaluate only how accurately EDITED transfers the visually transferable style of REFERENCE: white balance and temperature, exposure character, contrast and tone curve, black/white points, saturation, hue relationships, highlight and shadow rendering, skin-tone treatment when applicable, and the overall photographic color mood.
- Compare the direction and strength of the SOURCE-to-EDITED change with the style evidenced by REFERENCE.
- Do not require EDITED to copy REFERENCE's people, objects, layout, illumination geometry, or scene content.
- Do not reward EDITED merely for looking attractive.
- An unchanged SOURCE should score low when REFERENCE clearly calls for a different treatment.
- 9-10: highly faithful transfer of the reference color/tone; 7-8.5: strong transfer with minor mismatch; 4-6.5: partial or mixed transfer; 1-3.5: weak, wrong-direction, or largely absent transfer; 0: no meaningful style correspondence.

PQ — Preservation Quality:
Compare SOURCE with EDITED and assess both content preservation and technical deliverable quality.
- Content preservation: identity, face, body, pose, objects, geometry, composition, viewpoint, spatial relationships, text, boundaries, and fine structure must remain consistent with SOURCE.
- Technical quality: penalize blur, ghosting, deformation, hallucinated or missing content, halos, color bleeding, banding, posterization, noise, compression artifacts, clipping, broken texture, oversmoothing, and oversharpening.
- Do not penalize legitimate global/local color, exposure, contrast, or tonal changes that are needed to match REFERENCE.
- A sharp and beautiful image with changed identity, objects, geometry, or composition must still receive a low PQ score.
- 9-10: source content is faithfully preserved and quality is professionally deliverable; 7-8.5: minor detail loss or artifacts; 4-6.5: noticeable content drift or defects; 1-3.5: major structural/content corruption or severe defects; 0: unrelated or unusable result.

Score SC and PQ independently. A result may have low SC but high PQ, or high SC but low PQ. Return JSON only in exactly this form:
{"sc": 0.0, "pq": 0.0, "sc_reason": "concise evidence", "pq_reason": "concise evidence"}

Figure S19: Prompt templates for SC and PQ evaluation on Ref-Bench.



Figure S20: Additional qualitative comparisons on Ref-Bench.