SEE IT BEFORE YOU EDIT: A LARGE-SCALE DATASET FOR OPERATION-AWARE IMAGE EDITING ANTICIPATION

Anonymous authors
Paper under double-blind review

ABSTRACT

With the development of multi-modal large language models, many fields have begun to introduce related agents, among which photographer agents are one example. Existing agents are typically trained on holistic parameter data, outputting Lightroom editing parameters for images, yet lacking the ability to anticipate the results of their operations. This paper introduces SIBYE, a tripartite dataset comprising about 190,000 image pairs and operation parameters, corresponding to cropping, global editing, and masked local editing in photographic post-processing. The cropping subset provides about 47,000 compositional variations, generated using normalized cropping ratios along the top, bottom, left, and right edges. The editing subset includes about 100,000 global adjustments based on 10 basic operations, 24 HSL color grading operations, and 6 color calibration operations in Lightroom. The masking subset contains about 43,000 linear and radial gradient masks, with 10 basic operations applied per mask type to achieve local editing. Each data entry is a quadruple (original image, result image, operation type, operation parameters), enabling the model to learn the forward mapping from operations to visual outcomes and the reverse inference from results to operations. We design tasks of varying difficulty, including single-operation classification, operation parameter prediction, multi-operation parameter prediction, among others—and establish benchmark experiments to validate the dataset's effectiveness and challenges. This dataset will provide critical support for research areas such as interpretable image editing, operation-aware generative models, and intelligent photography assistants. The dataset can be downloaded at `https://doi.org/10.5281/zenodo.19369384`.

1 INTRODUCTION

In modern photography, post-processing is an extremely critical step. By adjusting overall or regional exposure, color, contrast, and tone, it is possible to produce expressive, high-quality images. Today, many photographer agents have been introduced to assist people in photo post-processing. Previous work Dutt et al. (2025); Chen et al. (2025a); Lin et al. (2025) typically uses editing datasets built on Lightroom Bychkovsky et al. (2011); Liang et al. (2021). They often directly train agents to output complete processing parameters based on input images and instructions, or incorporate auxiliary mechanisms like chain-of-thought to enhance agent capabilities. However, they overlook a crucial aspect: these agents lack perception of the actual effects of these parameters. While such agents may know that increasing exposure brightens an image, they cannot foresee the specific results and differences between an exposure increase of +0.3 versus +0.5 as shown in Figure1.

To address this issue, we introduce SIBYE, a large-scale dataset designed to support the development of operation-aware photo editing models. SIBYE is built upon the widely used MIT-Adobe FiveK dataset and comprises 190,000 structured quadruples (original image, edited image, operation type, operation parameters). We used scripts to apply operations in Photoshop's Camera Raw Filter to obtain editing results that closely resemble real-world outcomes. SIBYE is organized into three distinct editing modalities:

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.

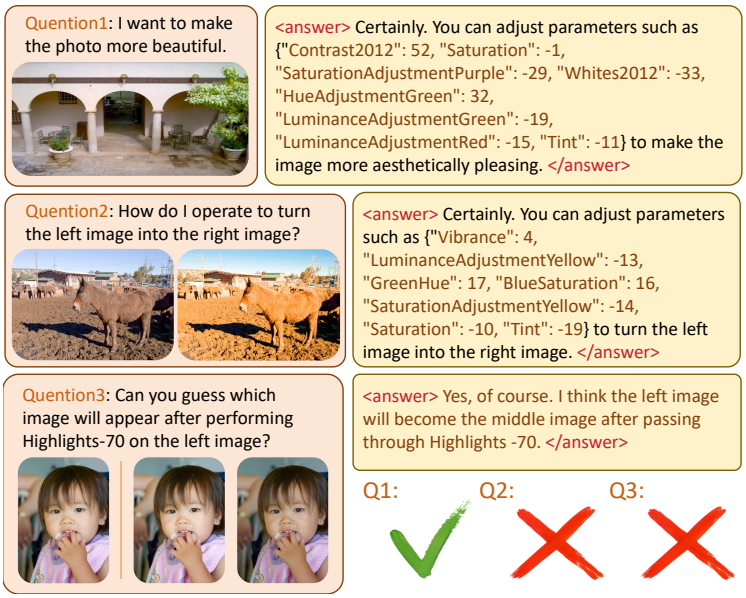


Figure 1: Existing photographer agents can provide image beautification suggestions, but they cannot perceive the actual effects of specific parameters, and thus cannot distinguish which parameters lead to which results.

- Cropping: about 47,000 samples, with cropping ratios specified as thousandths of the original width and height, enabling precise simulation of compositional variation.
- Global editing: about 100,000 samples, covering 10 basic Lightroom operations, 24 HSL channel adjustments, and 6 color calibration operations, representing the most widely used global editing space.
- Mask-based editing: about 43,000 samples (22,000 linear gradients, 21,000 radial gradients), each with a masked region and 10 internal basic operations, supporting localized editing.

Beyond its scale, SIBYE introduces a multi-level difficulty hierarchy to facilitate progressive evaluation. We segment alignment tasks of varying difficulty based on the magnitude of change between the edited image and the original, as well as the prediction precision required from the agent. This enables a gradual enhancement of the agent's perception of post-processing operations, from easy to hard.

The contributions of our work are as follows:

- We present SIBYE, a large-scale dataset of 190,000 structured operation-aware editing samples, covering crop, global edit, and mask-based edit modalities with fine-grained parameter annotations.
- We propose a multi-level difficulty taxonomy across all three editing modalities, enabling systematic evaluation of model capability from coarse to fine-grained perception and operation prediction.
- We release the dataset with pre-defined training/validation splits and difficulty labels to support research in interpretable image editing, operation-aware generation, and intelligent photography agents.

By providing a structured, scalable, and difficulty-aware resource, SIBYE aims to serve as a foundational dataset for advancing the perceptual and operational capabilities of next-generation photo editing models. The dataset, code, and benchmarks are publicly available.

2 Related Work

Photo Editing. Existing automated workflows have been proposed to simplify manual editing tasks. Zero-order and first-order optimization methods Yu et al. (2021); Tseng et al. (2022) represent early attempts, but they are constrained by limited parameter prediction capabilities and a reliance on pre-trained proxy models. Reinforcement Learning (RL)-based methods Ke et al. (2022); Ouyang et al. (2023); Wu et al. (2024) attempt to mimic human workflows and offer a degree of transparency, yet they fail to capture artistic perspective and lack deeper user interaction. Diffusion models Zhang et al. (2023); Xiao et al. (2025) dominate in high-fidelity image synthesis, but they depend on static prompts and lack multi-round reasoning or flexible language-alignment capabilities, limiting open-ended editing. Furthermore, recent unified image editing models have achieved breakthroughs in both understanding and generation. Notable examples include closed-source models (e.g., GPT-5 Singh et al. (2025) and Gemini-3 Team et al. (2023)) and open-source models (e.g., Janus-Pro Chen et al. (2025b), QLIP Zhao et al. (2025), and VARGPT-v1.1 Zhuang et al. (2025)).

Image Editing Datasets. Unlike AIGC-based methods, image editing datasets are typically built on tools such as Photoshop and Lightroom. They usually provide original images, editing results, and editing sequences, or directly include DNG files that embed editing operations. MIT-Adobe FiveKBychkovsky et al. (2011) is one of the most classic paired datasets in the field of image editing. Jointly released by MIT and Adobe, it contains 5,000 high-quality raw photos covering diverse scenes such as indoors, outdoors, portraits, and landscapes. For each photo, both the RAW data and the edited results independently produced by five professional retouchers are provided, meaning that a single input has five different expert-style outputs. PPR10KLiang et al. (2021) comprises approximately 10,000 high-resolution portrait images along with their corresponding professional editing results. Its scale and portrait-specific focus significantly surpass traditional general-purpose editing datasets, making it particularly suitable for tasks such as portrait enhancement, intelligent editing, portrait beautification, personalized editing, and editing parameter prediction. MMart DatasetLin et al. (2025) is a large-scale multi-modal photo editing dataset constructed as part of the JarvisArtLin et al. (2025) work. It contains 55,000 high-quality samples covering mainstream photography scenarios such as portraits, landscapes, and still lifes. Each sample is accompanied by the original image, the edited result, natural language instructions, Lightroom professional operation sequences, and artist chain-of-thought annotations. This dataset introduces for the first time a chain-of-thought annotation paradigm for professional editing, built upon a global-local-parameter multi-granularity labeling system. However, the operation sequences in these datasets are holistic; an agent cannot perceive the specific impact of each individual operation from the results of dozens of operations, nor can it know the exact differences between parameter values.

Reinforcement Fine-tuning. Rule-based reinforcement fine-tuning, as demonstrated by OpenAI's o1 Jaech et al. (2024) and Deepseek-R1 Guo et al. (2025), has shown remarkable performance in tasks such as mathematical reasoning Cai et al. (2024); Yang et al. (2025) and code generation Yang et al. (2025); Zhang et al. (2025; 2024). Subsequent research has extended this approach to multi-modal models by designing task-specific reward functions for visual perception tasks. Our work provides training data specifically for the reinforcement fine-tuning of photographer agents.

Image Editing Agent. MonetGPT Dutt et al. (2025) divides the editing process into three stages: lighting, saturation and white balance, and selective color edits, corresponding to six basic lighting operations, four basic color operations, and the HSL panel in Lightroom. The decision process can be adjusted between stages, offering greater interpretability and controllability compared to outputting all operations at once. However, the agent's perception of atomic operations remains insufficiently precise. PhotoArtAgent Chen et al. (2025a) formulates photo post-processing as an agent-driven multi-step decision problem, gradually completing image enhancement and aesthetic optimization through a pipeline of "image perception - problem analysis - editing planning $\rightarrow$ operation execution". At each step, the agent can observe the outcome of the current operation sequence. JarvisArt Lin et al. (2025) is capable of understanding user intentions and mimicking the reasoning process of professional artists. It undergoes chain-of-thought supervised fine-tuning to establish basic reasoning and tool-use skills, followed by Group Relative Policy Optimization Refinement (GRPO-R) to further improve its decision-making capabilities and tool proficiency. JarvisArt, nevertheless, still operates on full operation sequences.

Table 1: The structure of the SIBYE dataset, including cropping, global editing, mask-based Editing, and their subcategories.

All	**Total**	**Crop**	**Global Editing**		**Mask-based Editing**	
			Single	**Multi**	**Line**	**Circle**
Count	193953	47386	70000	33281	22104	21182
Train	**Total**	**Crop**	**Global Editing**		**Mask-based Editing**	
			Single	**Multi**	**Line**	**Circle**
Count	182276	44994	65219	32096	20101	19866
Val	**Total**	**Crop**	**Global Editing**		**Mask-based Editing**	
			Single	**Multi**	**Line**	**Circle**
Count	5593	477	2097	829	1107	1083

Therefore, we propose SIBYE, designed as pre-training data for the photographer agent to enhance the agent's fine-grained perception of Lightroom operation parameters and its anticipation of their effects, thereby laying the foundation for subsequent image beautification functions.

3 DATASET

We begin by outlining the overall structure and data sources of SIBYE (See It Before You Edit) in Section 3.1. Subsequently, we introduce the three dedicated generation pipelines of SIBYE: the crop dataset (Section 3.2), the global edit dataset (Section 3.3), and the mask-based edit dataset (Section 3.4). Finally, we present the multi-level difficulty taxonomy, which stratifies SIBYE into easy, medium, and hard tiers according to task-specific criteria (Section 3.5).

3.1 OVERVIEW

We first outline the overall construction pipeline of SIBYE, a large-scale dataset designed to support research on operation-aware photo editing. The dataset is built upon the widely used MIT-Adobe FiveK datasetBychkovsky et al. (2011), and employs the edited results from Expert E in Adobe FiveK as the original images of SIBYE. Our decision to build SIBYE upon the Adobe FiveK dataset is motivated by three primary considerations: (1) Adobe FiveK is originally designed for image-editing tasks, which makes it more pertinent to our research; (2) the dataset allows others to publish new datasets derived from it; and (3) its images are relatively small in size, enabling more efficient processing. Each data entry in SIBYE consists of a quadruple (original image, editing image, operation type, operation parameters) and encompasses three distinct generation pipelines corresponding to the core editing modalities in professional photography: cropping, global editing, and mask-based editing. Among these, the cropping data is processed using the PIL library, while the global editing and mask-based editing use Photoshop's Camera Raw Filter as the processing tool (the operations in Camera Raw Filter are interoperable with those in Lightroom). The cropping subset includes about 47,000 samples, with cropping precision specified as one-thousandth of the original dimensions. The global editing subset contains about 100,000 samples, covering 40 distinct operations: 10 basic Lightroom adjustments (Exposure2012, Contrast2012, Highlights2012, Shadows2012, Whites2012, Blacks2012, Temperature, Tint, Vibrance, and Saturation), 24 HSL channel operations, and 6 color calibration operations, with parameter values aligned with those in Lightroom. The mask-based editing subset comprises about 43,000 samples (22,000 linear gradients and 21,000 radial gradients), each combining a masked region with 10 basic operations. In total, SIBYE provides about 190,000 structured editing samples. Additionally, we design multi-level difficulty tasks for SIBYE to facilitate progressive training.

3.2 THE CROPPING DATASET

We construct the cropping dataset to enable the model to learn compositional adjustments through precise cropping operations. For each original image, we generate 10 cropped results, targeting a

total of 50,000 samples. After validity checking of the generated crops, we obtain 49,852 valid samples.

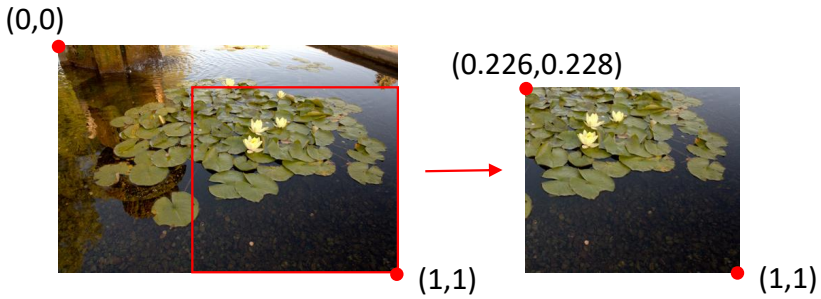


Figure 2: The lack of result awareness for specific image post-processing parameters in existing agents.

Cropping pattern sampling. We define 15 cropping patterns, each corresponding to a specific combination of cropped sides: top, bottom, left, right, top-left, top-right, bottom-left, bottom-right, top-bottom, left-right, only left, only right, only top, only bottom, and all four sides. These patterns are uniformly sampled during generation.

Parameter sampling. For each selected pattern, we sample cropping ratios for the relevant sides using a triangular distribution centered at zero. Specifically, the raw value X is drawn from a triangular distribution over the interval $[-1, 1]$ with mode 0. As shown in Eq.1 The final cropping ratio is taken as the absolute value $|X|$, which yields a distribution where small crops (ratios close to 0) are more common than large ones. This reflects real-world editing behavior, where subtle compositional refinements are more frequent than aggressive clipping.

$$X \sim Triangular(-1, 1, 0), \quad Y = |X| \tag{1}$$

Geometric constraints. To ensure that the cropped results are visually reasonable and retain meaningful content, we impose three constraints: (1) The resulting width and height must be strictly positive. (2) The aspect ratio of the cropped area must lie within $[0.5, 2]$, avoiding extremely wide or tall outputs that are rarely used in practice. (3) The cropped area must be at least 5% of the original image area.

Additional filtering. We further apply a filtering step to exclude samples where the combined cropping effect is negligible. Specifically, the sum of the cropping parameters for a given sample must exceed 0.05. This ensures that each sample undergoes a meaningful crop.

After applying the above pipeline, the final cropping dataset contains 47,386 samples.

3.3 The Global Editing Dataset

The global editing dataset aims to provide comprehensive coverage of non-destructive tone and color adjustments commonly used in professional photo editing. We construct this dataset through two complementary strategies: (1) exhaustive single-operation generation covering the entire parameter space of each operation, and (2) random multi-operation combinations that simulate real-world editing workflows where multiple adjustments are applied sequentially. After quality filtering, the dataset contains a total of 103,281 valid samples.

Operation space definition. We analyze the MMart Dataset published by JarvisArt and count the most frequent operations. Finally, we select 40 of the most common and effective global editing operations, categorized as follows: 1. Ten basic operations (Exposure2012, Contrast2012, Highlights2012, Shadows2012, Whites2012, Blacks2012, Temperature, Tint, Vibrance, Saturation). 2. Twenty-four HSL operations: Hue, Saturation, and Luminance adjustments for eight color channels (Red, Orange, Yellow, Green, Aqua, Blue, Purple, Magenta). 3. Six color calibration operations: Hue and Saturation adjustments for the Red, Green, and Blue primaries. All operation values are aligned with Adobe Lightroom.

Operation value ranges. To establish realistic parameter bounds, we compute the minimum and maximum values for each operation in the MMart Dataset, round them, and use the resulting intervals as the range for random parameter sampling, ensuring that the data remains close to real-world scenarios. For Exposure2012, which uses a continuous floating-point scale, we adopt a step size of 0.5. We employ relative color temperature adjustments from –100 to 100 (step 10) for single-operation tasks, and absolute color temperature from 2,000 to 16,000 (step 10) for multi-operation tasks, reflecting two perceptual difficulty levels—the latter being more prevalent in existing editing datasets. The other operations likewise use a step size of 10. This step-based sampling ensures comprehensive coverage of the feasible parameter space without being so fine-grained that the resulting edits lack visually noticeable differences.

Single-operation. We take the first 2,000 of the 5,000 original images. For each original image, we apply all 40 operations, and for each operation we generate multiple copies according to its range and step size.

Multi-operation. We take another 2,000 original images that are not used in single-operation generation. For each source image, we generate 20 editing sequences. Each editing sequence is generated as follows: 1. Sample the number of operations k from [2, 40] using a triangular distribution, such that small operation sets (2–5 operations) are more frequent than large combinations. 2. For each selected operation, sample its parameter value using a triangular distribution based on the minimum and maximum values obtained from MMart Dataset. For Exposure2012, the parameter is rounded to one decimal place; for other integer operations, the value is rounded to an integer.

Quality filtering. To ensure visual plausibility and avoid trivial or extreme edits, we apply a filtering step based on the Mean Squared Error (MSE) between the original and the edited image: 1. Samples with $MSE \leq 20$ are discarded because the visual difference is imperceptible and would not provide meaningful training signals. 2. Samples with $MSE > 10000$ are discarded because the edits are too aggressive (e.g., Exposure2012 +4) and unlikely to occur in realistic scenarios. After applying this filtering, the global editing dataset contains 103,281 valid samples, comprising 70,000 single-operation samples and 33,281 multi-operation samples.



Figure 3: Global editing. Above is the single-operation editing: {"Temperature": 70}; below is the multi-operation editing: {"HueAdjustmentBlue": -2, "HueAdjustmentRed": -4, "SaturationAdjustmentRed": -10, "Highlights2012": 19, "HueAdjustmentYellow": 24, "BlueHue": 7, "HueAdjustmentAqua": 12, "Vibrance": 7, "Exposure2012": 1.4, "RedSaturation": 7, "BlueSaturation": 26, "LuminanceAdjustmentOrange": 18, "LuminanceAdjustmentYellow": -19, "SaturationAdjustmentGreen": 6, "HueAdjustmentPurple": -19, "SaturationAdjustmentYellow": -5, "Contrast2012": 52, "Saturation": -1, "SaturationAdjustmentPurple": -29, "Whites2012": -33, "HueAdjustmentGreen": 32, "LuminanceAdjustmentGreen": -19, "LuminanceAdjustmentRed": -15, "Tint": -11}

3.4 The Mask Editing Dataset

The mask-based editing dataset aims to enable models to learn local adjustments via gradient masks. We use all 5,000 images to construct this dataset. For each image, we generate 5 linear mask samples and 5 radial mask samples, yielding an initial target of 50,000 samples.

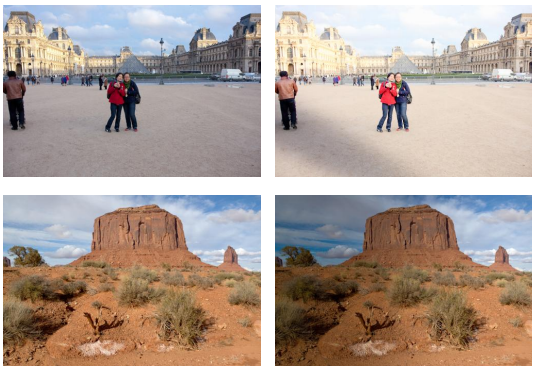


Figure 4: Mask editing. Above is the cicle-mask editing: {"Exposure2012": 0.13, "Highlights2012": 0.2, "Shadows2012": 0.48, "Whites2012": -0.04, "Temperature": 0., "Top": 0.093, "Left": -0.485, "Bottom": 1.36, "Right": 0.646, "Angle": 26., "MaskInverted": "true", "Midpoint": 98, "Feather": 62}; below is the line-mask editing: {"Exposure2012": -0.37, "Contrast2012": -0.76, "Saturation": 0.52, "ZeroX": 0.911, "ZeroY": 0.298, "FullX": 0.234, "FullY": 0.027}

Mask types and parameter definitions. Considering reproducibility and commonality of masks, we select only linear masks and radial masks (with concretely defined parameters that yield unique results). Aligned with Lightroom parameters, we define linear masks using ZeroX, ZeroY, FullX, FullY, and radial masks using Top, Left, Bottom, Right, Angle, MaskInverted, Midpoint, Feather, with parameter ranges also corresponding to those in Lightroom. Each mask sample is associated with a set of editing operations applied only within the masked region. We restrict the operations to the 10 basic Lightroom adjustments.

Generation pipeline. For each mask, the generation process consists of the following steps:

Mask parameter sampling: Randomly generate parameters according to the defined mask parameter ranges, then detect and retain valid parameters.

Area constraint: Compute the area of the masked region (areas where mask intensity ¿ 0.5 after feathering). If the mask area is less than 5% of the total image area, discard the sample and regenerate, up to a maximum of 10 attempts.

Operation sampling: Sample a set of operations and their parameters as described in Section3.3, with the difference that the number of operations ranges from 1 to 10.

Image editing: Apply the operations to the original image, blending the edited region with the unedited background using mask weights (including feathering).

Result validation: Compute the MSE between the original and the edited image. If $MSE \leq 20$, the sample is discarded because the visual difference is imperceptible.

We apply the area constraint before image composition to avoid generating overly small local edits. The MSE filter is applied after composition to ensure that the overall change is visible. Finally, we obtain 22104 linear mask samples and 21182 radial mask samples.

3.5 Dataset Splits and Difficulty Taxonomy

We introduce a disjoint splitting strategy for SIBYE and construct subtasks of varying difficulty to progressively enhance the agent's perceptual capabilities.

Training/validation set split. To prevent data leakage and enable fair evaluation across editing modalities, we build separate training and validation sets for the three editing directions (cropping, global editing, and mask-based editing). Each validation set contains 50 original images, and we ensure that none of the original images in any validation set appears in any training set.

Difficulty levels assigned to training tasks.

Cropping dataset. We construct three difficulty levels based on the precision of the cropping values. The core idea is to evaluate the model's ability to distinguish cropping parameters at different precision levels. Given a cropping sample defined by four directional ratios (t, b, l, r), we quantize each value to a target precision level. Specifically:

Easy: Ratios are rounded to one decimal place (0.1 precision). A sample is retained for this difficulty level if $|x - round(x, 1)| \leq 0.01$ for each directional ratio x; otherwise, the sample is excluded.

Medium: Ratios are rounded to two decimal places (0.01 precision). The same filtering condition applies, with a tolerance of 0.001.

Difficult: Ratios are kept to three decimal places (0.001 precision), requiring the model to predict cropping parameters with fine-grained precision.

Global editing data and mask-based editing data. We adopt a two-dimensional difficulty stratification based on perceptual salience (measured by MSE) and operational complexity (number of operations and granularity of parameter prediction).

Perceptual salience levels. We divide samples into three levels according to the MSE between the original and edited images:

Easy: $MSE > 100$, the edited result is very obviously different from the original image.

Medium: $30 \leq MSE \leq 100$, the edited result is clearly different from the original image.

Difficult: $MSE < 30$, the difference between the edited result and the original image is not obvious and requires side-by-side comparison to be perceived.

Agent training can follow an easy-to-hard curriculum, gradually improving perceptual abilities.

4 Experiment

The SIBYE dataset encompasses the most prevalent post-processing operations in photographic editing, facilitating the construction of diverse training data for agent-centric perception enhancement. To validate the practical utility of SIBYE, we designed a series of experiments to evaluate the resulting improvements in agent capabilities.

4.1 Experimental Setup

Across all seven experiments, we employed Qwen3-VL-8B-Instruct Bai et al. (2025) as the baseline model for fine-tuning. The specific configurations are as follows:

- **Framework**: All experiments are implemented using the **ms-swift** Zhao et al. (2024) framework.
- **Fine-tuning Strategy**: We freeze all non-LLM components and apply Low-Rank Adaptation (LoRA) exclusively to the language module of the LLM. LoRA adapters are inserted into every linear layer of the LLM, with rank $r = 8$ and scaling factor $\alpha = 32$.
- **Optimization**: The model is optimized with a learning rate of 1×10^{-4} and a linear warmup ratio of 0.05.
- **Hardware and Training Dynamics**: Training is performed on two NVIDIA RTX 4090D GPUs, leveraging DeepSpeed ZeRO-2 for distributed acceleration. We use a per-device batch size of 1 and accumulate gradients over 8 steps, resulting in an effective batch size of 16. Each experiment is trained for one epoch.
- **Comparative Models**: For performance benchmarking, we include several state-of-the-art models for comparison, namely Gemini-3-Flash-Preview Team et al. (2023), GPT-5-Nano Singh et al. (2025), and JarvisArt Lin et al. (2025).

4.2 Experimental Design

We design seven experiments (Exp. A–G) that probe the agent's capability in a progressive manner, from naming a single operation to executing multi-operation edits and finally to transferring

Table 2: Overall results of Exp. A–G. Exp. A–F are evaluated on SIBYE, whereas Exp. G follows the MMart-Bench protocol of JarvisArt Lin et al. (2025). Acc.: accuracy; MSE: mean squared error between the edited image and the ground truth; MAE: mean absolute error of the regressed value; IoU: intersection over union of the cropping box; n: number of valid outputs in Exp. F; $L1$/$L2$: pixel-level distances ($\times 10^2$/$\times 10^3$). ↑/↓ mark whether higher or lower is better, and "–" means the corresponding method reports no result. Bold marks the best value in each column. †Exp. G is run on the model set of the JarvisArt paper, whose commercial baselines differ from those used in Exp. A–F; the row "Ours" is fine-tuned on SIBYE for Exp. A–F and on SIBYE+MMart for Exp. G.

	Exp. A	Exp. B	Exp. C		Exp. D		Exp. E	Exp. F		Exp. G	
Model	Acc.↑	Acc.↑	Acc.↑	MAE↓	MSE↓	MSE↓	IoU↑	MSE↓	n	$L1$↓	$L2$↓
Gemini-3-Flash	0.6800	**0.5600**	0.4189	1.4194	320.57	–	0.6862	1454.19	78	–	–
GPT-5-Nano	0.4200	0.4400	0.2706	2.5217	784.29	–	0.3461	1735.42	67	–	–
JarvisArt	0.2097	0.2013	0.1000	3.5161	700.12	–	0.0885	649.54	1	12.44	30.56
Qwen3 (Zero-shot)	0.2892	0.2136	0.2450	5.36	715.36	1156.54	0.4133	3912.66	16	10.97	23.4
Qwen3 (Ours)	**0.8997**	0.5283	**0.8792**	**0.625**	**136.21**	**405.34**	**0.9069**	**317.24**	168	**8.94**	**16.3**
Gemini-2-Flash†	–	–	–	–	–	–	–	–	–	23.07	90.99
GPT-4o†	–	–	–	–	–	–	–	–	–	22.84	92.23
Qwen3 (MMart-ft)†	–	–	–	–	–	–	–	–	–	9.75	19.3

to an external benchmark. Exp. A and Exp. B test single-operation recognition over a restricted 10-operation and the full 40-operation label space; Exp. C additionally requires regressing the exact adjustment value; Exp. D evaluates multi-operation editing by applying the predicted parameters to the original image; Exp. E tests cropping-box localization; Exp. F covers linear and radial mask editing; and Exp. G measures transfer to MMart-Bench Lin et al. (2025). The MSE thresholds, sampling strategy, per-experiment training/test sizes, and evaluation protocols of all seven settings are deferred to Appendix A.

4.3 Experimental Results

As shown in Table 2, our fine-tuned model demonstrates superior capability in distinguishing specific editing operations applied to images, particularly in terms of brightness/color adjustments, cropping, and geometric masking. Even for more common tasks such as single-edit recognition and cropping, the model trained on SIBYE data generally outperforms commercial alternatives. In the Exp. G columns of Table 2, we evaluate a more practical application scenario using the MMart-Bench benchmark; the experimental results indicate that images edited by our fine-tuned model are significantly closer to the ground truth. It is worth noting that, since our MMart fine-tuning dataset omits many adjustment parameters that have relatively minor impacts compared to the original data, and focuses solely on global editing, while we also provide only 40 operations in the prompts, even the baseline model without fine-tuning achieves better performance than JarvisArt.

5 Conclusions

This paper presents the SIBYE dataset, which is designed to provide training for image-editing agents based on Photoshop & Lightroom, aiming to enhance their perceptual capabilities for editing. SIBYE comprises about 190,000 structured quadruples, organized into three distinct editing modalities: cropping, global retouching, and mask-based retouching. Beyond its scale, SIBYE introduces a multi-level difficulty hierarchy to facilitate progressive evaluation. We partition alignment tasks of varying difficulty based on the magnitude of change between the retouched result and the original image, as well as the prediction precision required from the agent, enabling a gradual enhancement of the agent's perception of post-processing operations from easy to hard.

AI use statement

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

In this work, we used generative AI tools (commercial large language model assistants) for a limited set of tasks, reported below using the task categories of the ICLR 2027 AI Policy for Authors. We separate the tasks with required disclosure from those with recommended disclosure.

Tasks with required disclosure. (i) *Assist with translation*: generative AI tools were used to help translate the authors' original drafts into English. (ii) *Implement methods* and *clean and reformat dataset*: generative AI tools were used to help write the code that implements our experimental pipeline and the code that constructs and cleans the SIBYE dataset. All code was specified, reviewed, tested and executed by the authors, and every dataset statistic and experimental result reported in this paper was obtained by the authors from that code.

We have not used generative AI tools to generate synthetic data sets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, support qualitative and thematic data analysis, or interpret results. These tasks were carried out entirely by the authors, and the remaining tasks with required disclosure are not applicable to this work.

Tasks with recommended disclosure. Generative AI tools were used to *edit the paper to improve readability* and to *suggest a structure for the paper*: we used them to check the LaTeX source for consistency and formatting errors, and to help relocate material that exceeded the main-text page limit into the appendix.

Verification and responsibility. We have reviewed all AI-assisted work. The translated text was read and revised by the authors, and every technical claim was checked against the original draft. The AI-assisted code was inspected, tested and executed by the authors; the dataset and experimental statistics reported in the paper were recomputed from its output and verified independently. Every change to the manuscript, including the relocation of content to the appendix, was reviewed by the authors to confirm that no claim or result was altered in the process. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

See the ICLR 2027 AI Policy for Authors for further details.

ETHICS STATEMENT

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

SIBYE is built upon the MIT-Adobe FiveK dataset Bychkovsky et al. (2011), whose license permits the publication of datasets derived from it. The released version of SIBYE contains both the edited result images and their corresponding input images, and both are derived from FiveK; users of SIBYE are therefore required to comply with the original MIT-Adobe FiveK terms in addition to the terms we impose. The image files retain the EXIF metadata of the original photographs as distributed in FiveK, which includes photographer credits and copyright notices such as "All Rights Reserved"; this metadata originates from the upstream dataset, belongs to the original photographers and right holders, and is neither covered by nor extended by the terms we impose. All editing results are produced programmatically by scripts, and this work involves no human subjects, no crowdsourced or expert annotation, no user studies, and no personally identifiable information about any individual is collected or released by us. We note as a limitation that the underlying photographs are limited to the 5,000 images of FiveK and are skewed toward particular scenes and subjects, and that aesthetic judgment is inherently subjective and culturally dependent; SIBYE is intended for research on operation-aware photo editing and should not be used for automated aesthetic decision-making. The authors declare no competing interests.

REPRODUCIBILITY STATEMENT

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

We have released the pipeline behind this work to facilitate reproduction. The anonymized supplementary material contains (i) the scripts used to generate the SIBYE data, (ii) the training code for all experiments reported in this paper, and (iii) the label files used in those experiments. The dataset itself is described in detail in Sections 3.1–3.5: the generation pipeline of each of the three editing modalities, the geometric and quality filtering criteria, and the definition of the difficulty

taxonomy. Table 1 summarizes the resulting structure, sample counts and splits, and Table 2 reports the full comparison results. The MMart-Bench transfer results of Exp. G are reported in Table 3 and discussed in Appendix B. The experimental protocol behind Exp. A–G is stated in Section 4.2 and expanded in Appendix A, which specifies for each setting the data subsets, the evaluation metrics and the difficulty levels involved. All baseline models reported in Table 2 are evaluated using their publicly released weights under default inference settings. The dataset, code and benchmarks are publicly available.

REFERENCES

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

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.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

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*, 2025a.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

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.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson WH Lau. Harmonizer: Learning to perform white-box image and video harmonization. In *European conference on computer vision*, pp. 690–706. Springer, 2022.

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 Conference on Computer Vision and Pattern Recognition*, 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*, 2025.

Wenqi Ouyang, Yi Dong, Xiaoyang Kang, Peiran Ren, Xin Xu, and Xuansong Xie. Rsfnet: A white-box image retouching approach using region-specific color filters. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12160–12169, 2023.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

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.

Ethan Tseng, Yuxuan Zhang, Lars Jebe, Xuaner Zhang, Zhihao Xia, Yifei Fan, Felix Heide, and Jiawen Chen. Neural photo-finishing. *ACM Trans. Graph.*, 41(6):238–1, 2022.

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.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13294–13304, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Ke Yu, Zexian Li, Yue Peng, Chen Change Loy, and Jinwei Gu. Reconfigisp: Reconfigurable camera image processing pipeline. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4248–4257, 2021.

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.

Kechi Zhang, Ge Li, Yihong Dong, Jingjing Xu, Jun Zhang, Jing Su, Yongfei Liu, and Zhi Jin. Codedpo: Aligning code models with self generated and verified source code. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15854–15871, 2025.

Yuxiang Zhang, Shangxi Wu, Yuqi Yang, Jiangming Shu, Jinlin Xiao, Chao Kong, and Jitao Sang. o1-coder: an o1 replication for coding. *arXiv preprint arXiv:2412.00154*, 2024.

Yue Zhao, Fuzhao Xue, Scott Reed, Linxi Fan, Yuke Zhu, Jan Kautz, Zhiding Yu, Philipp Krähenbühl, and De-An Huang. Qlip: Text-aligned visual tokenization unifies auto-regressive multimodal understanding and generation. *arXiv preprint arXiv:2502.05178*, 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.

Xianwei Zhuang, Yuxin Xie, Yufan Deng, Dongchao Yang, Liming Liang, Jinghan Ru, Yuguo Yin, and Yuexian Zou. Vargpt-v1. 1: Improve visual autoregressive large unified model via iterative instruction tuning and reinforcement learning. *arXiv preprint arXiv:2504.02949*, 2025.

A Detailed Experimental Design

This appendix expands the seven evaluation settings summarized in Section 4.2. Unless stated otherwise, the sample counts below refer to the split used for that experiment, and every metric is computed against the ground-truth edit.

A.1 Exp. A: Single-Operation Recognition (Restricted Label Space)

We select samples that involve 10 basic Lightroom operations and whose magnitude of change satisfies $\mathrm{MSE} > 100$. This yields 14,931 training samples and 968 test samples. The model must name the applied operation, and accuracy over the 10 classes is the evaluation metric.

A.2 Exp. B: Single-Operation Recognition (Full Label Space)

Expanding the label space to all 40 editing operations under the same $\mathrm{MSE} > 100$ threshold gives 5,221 training samples and 1,058 test samples. The task and the metric are identical to Exp. A.

A.3 Exp. C: Single-Operation Recognition with Value Regression

Building on the setting of Exp. A, we lower the threshold to $\mathrm{MSE} > 30$. Besides identifying the operation, the model must predict the exact adjustment value; to make this well posed, the valid range and step size of every operation are provided in the prompt. The dataset contains 18,590 training samples and 1,465 test samples. We report the classification accuracy and the mean absolute error (MAE) of the predicted value.

A.4 Exp. D: Multi-Operation Execution and MSE Evaluation

We use multi-operation data with $\mathrm{MSE} > 30$ and perform balanced sampling with respect to the number of operations (2 to 40 per image), keeping 10 test samples and 500 training samples per operation count. This yields 16,235 training samples and 335 test samples. The predicted parameters are applied to the original images with Photoshop scripts, and the MSE between the resulting image and the ground truth is the metric. In addition to the in-dataset test set, we randomly draw 200 sample groups from the MMart dataset as a cross-dataset test benchmark.

A.5 Exp. E: Cropping Box Localization and IoU Evaluation

We construct a cropping-focused dataset with balanced distributions over box position, aspect ratio, and box size, containing 13,246 training samples and 278 test samples. The model outputs the cropping box coordinates, and we report the Intersection over Union (IoU) against the ground-truth box.

A.6 Exp. F: Linear and Radial Mask Editing and MSE Evaluation

We sample equal numbers of linear-mask and radial-mask samples to form the training and test sets, resulting in 38,405 training samples and 189 test samples. The predicted edits are applied to the original images with Photoshop scripts, and the MSE between the prediction and the ground truth is reported, together with the number of valid outputs.

A.7 Exp. G: Transfer to MMart-Bench

To test whether an agent trained on SIBYE transfers to an external benchmark, we select 2,309 samples from the MMart dataset that involve only the 40 global editing operations. Separately, we randomly draw 5,000 instances from the single-operation and multi-operation global editing data of SIBYE with $\mathrm{MSE} > 30$, and merge them into the training set for fine-tuning the baseline model. In a parallel setting, the same baseline is fine-tuned using only the 2,309 MMart samples. Both variants are then evaluated on MMart-Bench Lin et al. (2025).

B Detailed Results on MMart-Bench

Table 3: Scores of the MMart-Bench test results in Exp. G. "MMart-Bench-source" denotes the reference edits produced by the photographers in the benchmark. Images were rated by GPT-5.4 on a 1-to-10 scale with 0.1 precision; higher is better.

Results	Avg.↑
MMart-Bench-source	7.666
Qwen3 (Zero-shot)	7.547
Qwen3 (MMart-ft)	7.665
Qwen3 (SIBYE&MMart-ft)	7.738

Table 3 reports the MMart-Bench scores of Exp. G. The untuned model is weakest at 7.547, 0.119 below the benchmark's reference edits (7.666); fine-tuning on MMart alone closes that gap (7.665), while adding SIBYE gives the best score overall (7.738), exceeding both the MMart-only variant and the reference edits. As the margins are small on a 1-to-10 scale, this ordering is best read as a trend.