SHUTTERMUSE: CAPTURE-TIME PHOTOGRAPHY GUIDANCE WITH MLLMS

Anonymous authors
Paper under double-blind review

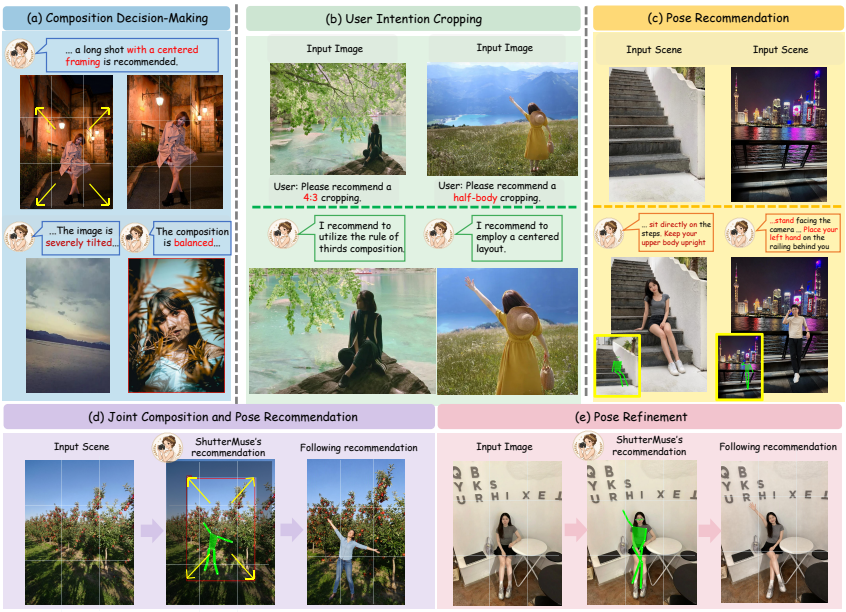

Figure 1: **Showcases of ShutterMuse.** (a) Composition decision-making with `refine`, `reject`, and `keep` outcomes. (b) Instruction-guided framing under diverse user intentions. (c) Scene-conditioned pose recommendation. (d) Joint composition and pose recommendation. (e) Refinement of an existing person's pose. The keypoint poses are rendered by GPT-Image-2.

ABSTRACT

Computational photography has largely focused on post-capture optimization, while many photographic failures could be prevented at capture time. Capture-time assistance requires more than aesthetic scoring or crop prediction: it must determine whether the current view should be preserved, refined, or abandoned, and guide how a subject should pose within the scene. We formulate this problem as *capture-time photography guidance* and introduce **ShutterBench**, a benchmark evaluating both sides of the camera. Photographer-side guidance evaluates composition decisions (`keep`/`refine`/`reject`) and spatial refinement, while subject-side guidance evaluates scene-conditioned pose recommendation. Our evaluation reveals complementary limitations: general-purpose MLLMs make reasonable composition decisions but struggle with precise spatial refinement, whereas specialized cropping models localize effective crops but cannot reason over the broader decision space; MLLMs also generate valid poses, but their aesthetic quality remains limited. To enable unified learning, we construct **ShutterData**, comprising 130K samples with textual rationales and structured composition and pose supervision, and develop **ShutterMuse**, a unified MLLM trained with supervised and reinforcement fine-tuning. ShutterMuse achieves the strongest overall photographer-side performance on ShutterBench and competitive subject-side performance at substantially lower inference cost, demonstrating its potential to shift

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.

computational photography from post-hoc optimization toward actionable assistance before capture.

1 INTRODUCTION

Computational photography has traditionally focused on post-capture optimization, yet many consequential decisions must be made before the shutter is pressed. A capture-time assistant should determine whether the current framing should be preserved, refined, or abandoned, and, for portraits, how the subject should pose within the scene. We formulate this problem as **capture-time photography guidance**. Existing aesthetic cropping benchmarks (Yan et al., 2013; Hong et al., 2024; Yang et al., 2023; Zeng et al., 2019; Wei et al., 2018; Chen et al., 2017; Fang et al., 2014; Zhang et al., 2022) mainly evaluate post-hoc crop prediction, typically assuming that cropping is necessary and sufficient. They therefore cannot fully evaluate capture-time guidance when the current view should be preserved or rejected, or when subject-side pose recommendations are needed.

MLLMs provide a natural foundation for this broader problem. Multimodal large language models (MLLMs) combine visual understanding and instruction following (Alayrac et al., 2022; Li et al., 2023; Dai et al., 2023; Zhu et al., 2024; Bai et al., 2025; 2023; Liu et al., 2023a; Team et al., 2023; Wang et al., 2025) and have demonstrated capabilities in visual aesthetics and photographic composition (Huang et al., 2024b; Qi et al., 2025; Wu et al., 2023; Liu et al., 2025; Cao et al., 2025). However, their ability to translate photographic judgment into an appropriate, spatially precise intervention before capture remains underexplored.

To study this capability, we introduce **ShutterBench**, covering two complementary tasks: **photographer-side guidance** evaluates `keep`/`refine`/`reject` decisions and framing boxes when refinement is needed; **subject-side guidance** evaluates scene-conditioned pose recommendations. Our evaluation reveals complementary limitations: general-purpose MLLMs (Wang et al., 2025; Bai et al., 2025; Team et al., 2023; Achiam et al., 2023; Moonshot AI, 2025) can often make composition decisions but lack precise refinement localization, whereas specialized cropping models (Sheng et al., 2025; Du et al., 2026; Hong et al., 2021; Liu et al., 2023b) localize crops effectively but do not natively support `keep` or `reject`. For subject-side guidance, MLLMs can produce valid structured poses but struggle to recommend aesthetically pleasing ones. These findings highlight the gap between photographic judgment and actionable spatial guidance.

To address these limitations, we construct **ShutterData**, comprising 130K atomic-task samples with textual rationales and structured composition and pose annotations. Photographer-side annotations are scaled through an expert-seeded, MLLM-verified self-distillation pipeline; subject-side examples pair person-free scenes with expert-verified pose keypoints, visibility states, and rationales. Based on ShutterData, we propose **ShutterMuse** (Fig. 1), a unified MLLM trained with supervised fine-tuning followed by photographer-side reinforcement fine-tuning. Experiments on ShutterBench show that ShutterMuse achieves the best overall photographer-side performance among evaluated baselines and competitive subject-side pose recommendations with substantially lower inference cost than image-editing methods, demonstrating the potential of MLLMs as interactive assistants during image capture.

Our main contributions are summarized as follows:

- We formulate **capture-time photography guidance**, covering both photographer-side composition decisions and refinement and subject-side pose recommendation.
- We introduce **ShutterBench** to evaluate these capabilities and identify complementary limitations of existing models.
- We construct **ShutterData**, a large-scale dataset pairing textual rationales with structured composition and pose annotations.
- We propose **ShutterMuse**, achieving the best overall photographer-side performance among evaluated baselines and efficient, competitive subject-side pose recommendations.

2 Related Work

Aesthetic Cropping Benchmarks. Aesthetic cropping benchmarks evaluate composition refinement through reference crop annotations. FCDB (Chen et al., 2017) and FLMS (Fang et al., 2014) provide expert cropping references, while SACD (Zhang et al., 2022) incorporates subject-aware composition annotations. These resources primarily assess where to crop, rather than whether cropping is needed or sufficient. ShutterBench complements this setting with explicit `keep`/`refine`/`reject` decisions and a separate scene-conditioned pose recommendation task.

Composition Recommendation. Cropping methods typically rank candidate regions (Su et al., 2024; Zeng et al., 2019; Zhang et al., 2026) or directly regress crop coordinates (Hong et al., 2021; Huang et al., 2024a; Pan et al., 2021). Recent MLLM-based methods extend crop prediction with aesthetic reasoning and instruction following: InstructCrop (Sheng et al., 2025) learns explanatory, intention-aware crop suggestions, and Venus (Du et al., 2026) uses two-stage training for aesthetic composition guidance. Our focus extends beyond crop localization to deciding whether the current view should be retained, refined, or rejected, while supporting subject-side guidance within the same model.

Pose Recommendation for Photography. Photography-oriented pose guidance has been explored through reference selection and aesthetic optimization. Collaborative composition models (Zhang et al., 2015) and InspireMePosing (Sheng et al., 2018) recommend poses or portrait references using scene and aesthetic features. InstaPose (Ji & Hsiao, 2025) retrieves scene-relevant poses, reranks them for diversity, and provides textual adjustment guidance. Photo Stand-Out (Wang et al., 2020) optimizes a virtual character's pose for aesthetic compatibility with a user's pose. More recently, Before the Shutter (Jiang & Chen, 2026) studies portrait planning in 3D scenes, including pose, camera, lighting, and exposure. Building on this broader goal of photography assistance, we study direct generation of structured pose recommendations from a single scene image by MLLMs, with composition and pose guidance assessed separately in ShutterBench.

3 Dataset and Benchmark

ShutterData provides 130K examples for the two atomic guidance tasks: 100K for photographer-side guidance and 30K for subject-side guidance. ShutterBench evaluates the two tasks separately. Training and benchmark distributions are detailed in Appendix A.3.

3.1 ShutterData

Photographer-side guidance. We construct a 12K-image seed set annotated by 10 trained annotators with cross-review, assigning each image a `refine`, `keep`, or `reject` label, a textual rationale, and a framing box for `refine` cases. To scale this supervision, we use an *expert-seeded, MLLM-verified self-distillation pipeline* (EMDP; Fig. 2): a model trained on the seed set generates pseudo annotations for unlabeled images, an MLLM verifier checks rationale correctness and rationale–box consistency, and verified samples support iterative retraining. A fixed expert validation set monitors weak composition patterns, while an independent expert test set assesses pipeline reliability. Seed annotation and rationale construction details are provided in Appendix A.1.

Subject-side guidance. Each subject-side training example pairs a person-free scene with target keypoints, visibility labels, and a rationale explaining the pose's suitability for the scene. We extract the target pose from an original portrait and remove the person to construct the input scene, discarding samples that fail an artifact check. Professional annotators inspect the extracted skeleton overlaid on this scene against the original portrait, rejecting poses that are inconsistent with the reference or aesthetically unsuitable for the scene. For retained samples, they correct visibility labels when needed without changing keypoint coordinates. An MLLM generates recommendation rationales from the original portrait, followed by human review for consistency with the scene and pose. The construction pipeline (Fig. 8), tool details, annotation definitions, and verification procedures are provided in Appendix A.2. ShutterMuse also supports joint composition and pose recommendation and refinement of an existing person's pose through dedicated training data (Appendix A.4).

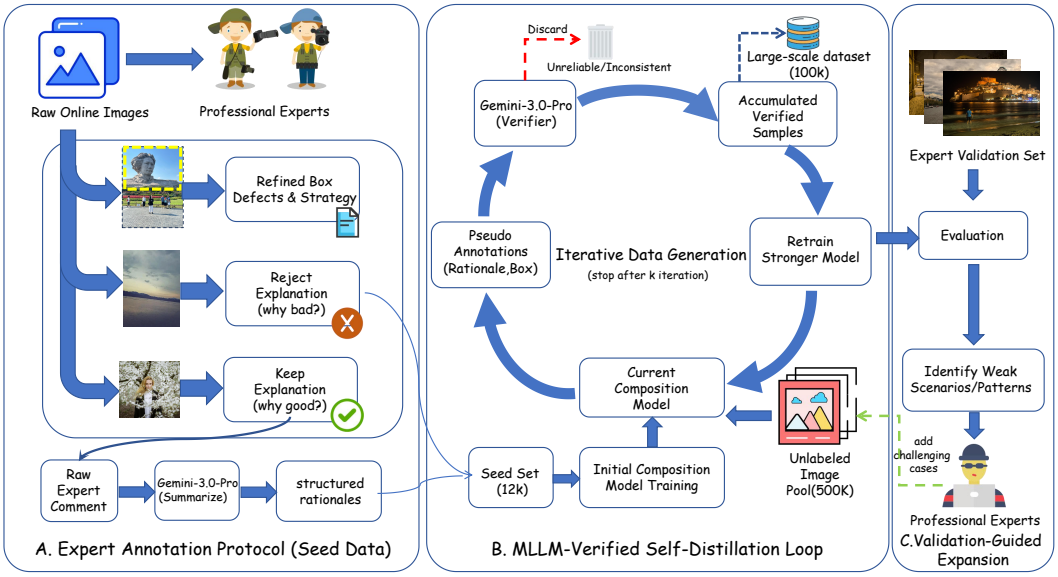

Figure 2: Overview of the expert-seeded, MLLM-verified self-distillation pipeline for photographer-side data construction. Expert seed annotations are structured by an MLLM, expanded through pseudo-labeling and verification, and monitored with a held-out validation set.

3.2 SHUTTERBENCH

ShutterBench evaluates the two atomic guidance tasks; joint composition and pose recommendation and refinement of an existing person's pose are shown only qualitatively in Fig. 4(b,c).

The photographer-side benchmark contains 567 examples annotated following the EMDP expert-seed protocol: 337 `refine`, 96 `keep`, and 134 `reject`, with 3–5 expert reference boxes per `refine` example. The subject-side benchmark contains 552 examples reserved for evaluation, with no overlap with the training set, plus 100 real empty-scene images, totaling 652 examples. All benchmark examples are excluded from both SFT and RFT.

Photographer-side evaluation metrics. Following prior work (Fang et al., 2014; Zhang et al., 2022), we compute **IoU**, normalized **boundary displacement error (BDE)**, and **refinement success rate (R@0.7)** over all ground-truth `refine` examples, including those predicted as `keep`. For a `refine` prediction, we evaluate the predicted framing box; for a `keep` prediction, we treat the full image as the predicted box and include it in all three metrics. IoU takes the maximum overlap with the reference boxes, and BDE takes the minimum error. Predictions of `reject` or `refine` with an invalid or malformed box are localization failures, assigned IoU 0 and BDE 1. R@0.7 is the percentage of all ground-truth `refine` examples whose evaluated box achieves IoU > 0.7, including `keep` predictions whose full-image box meets this threshold.

Reject success rate (RSR) and **keep success rate (KSR)** are the classification accuracies on ground-truth `reject` and `keep` examples, respectively. **Macro-F1** averages the F1 scores of the three decision classes; unparseable or out-of-set decisions are incorrect.

MLLM-Score assesses composition quality beyond geometric overlap. For ground-truth non-reject examples, an MLLM judge evaluates the predicted refinement box or, for a `keep` prediction, the full image, assessing whether it preserves annotated strengths and addresses annotated weaknesses; a `reject` prediction or an invalid frame receives 0. For ground-truth `reject` examples, the score is 1 for a correct `reject` prediction and 0 otherwise. Scores are averaged over the entire photographer-side benchmark. The judging prompt is provided in Appendix B.3.

Subject-side evaluation metrics. Since multiple poses can suit the same scene, we evaluate pose quality beyond exact keypoint matching. An MLLM judge scores the predicted skeleton overlaid on

the input scene along three dimensions: **physical plausibility**, **scene interaction**, and **pose aesthetics**. **Mean** averages these three scores; judging prompts appear in Appendix B.4.

4 SHUTTERMUSE

ShutterMuse is a unified MLLM for capture-time photography guidance, built on Qwen3-VL-8B (Bai et al., 2025). A central challenge is that photographer-side guidance couples a discrete intervention decision with a continuous spatial action that is meaningful only when refinement is warranted. Subject-side guidance further requires structured generation of scene-conditioned human poses. We train ShutterMuse in two stages. Supervised fine-tuning (SFT) first learns a unified representation of composition and pose guidance, followed by photographer-side reinforcement fine-tuning (RFT) that directly optimizes the coupled decision–localization objective.

Unified Structured Guidance. We represent capture-time guidance through structured outputs that make semantic decisions and spatial interventions jointly learnable. Responses contain a rationale and structured annotations: `composition_xy` encodes `reject` as an empty value, `keep` as $[0, 0, 1, 1]$, and `refine` as another normalized framing box; pose outputs contain 17 normalized keypoints and visibility labels. Detailed output schemas are provided in Appendix E.1.

Supervised Fine-Tuning. SFT uses 130K atomic-task examples plus 10K joint-guidance and pose-refinement examples (Appendix A.4). SFT teaches the model to jointly produce photographic rationales and structured interventions across photographer- and subject-side tasks. We apply next-token prediction loss only to target responses (loss detailed in Appendix E).

Reinforcement Fine-Tuning. Applying a localization reward indiscriminately would incorrectly encourage cropping even for views that should be kept or rejected. We therefore formulate photographer-side RFT as decision-conditioned spatial optimization and apply GRPO to 20K photographer-side samples from ShutterData.

The decision reward is $R_{\text{dec}} = \mathbb{I}[\hat{c} = c^\star]$, where $c^\star \in \{\texttt{reject}, \texttt{keep}, \texttt{refine}\}$ is the reference decision and $\hat{c}$ is parsed from `composition_xy` as described above.

Multiple framings can be photographically reasonable, while a crop with substantial reference overlap may still remove visually important content. We therefore augment reference-box agreement with a salient-content preservation constraint. Let b be the predicted box and $\mathcal{B}^\star = \{b_k^\star\}_{k=1}^K$ the reference boxes. BiRefNet (Zheng et al., 2024) extracts a binary salient-object mask $M \in \{0, 1\}^{H \times W}$ from the input image. The fraction of salient content retained by b is

$$\text{Cov}(b, M) = \frac{\sum_{u,v} M(u, v)\mathbf{1}_b(u, v)}{\sum_{u,v} M(u, v) + \epsilon}, \tag{1}$$

where $\mathbf{1}_b(u, v)$ indicates whether pixel (u, v) lies inside b, and ϵ is a small numerical constant. The localization reward is

$$R_{\text{loc}} = \mathbb{I}[c^\star = \hat{c} = \texttt{refine}] \underbrace{\max_k \text{IoU}(b, b_k^\star)}_{\text{reference agreement}} \underbrace{\min\left(1, \frac{\text{Cov}(b, M)}{\tau_m}\right)}_{\text{content preservation}}, \tag{2}$$

where τ_m is the target saliency-coverage threshold. IoU encourages alignment with the reference composition, while the coverage factor downweights crops that retain insufficient salient content and saturates at 1 once the threshold is reached. Localization reward is enabled only when both the reference and predicted decisions are `refine`.

The reward combines decision correctness and composition localization:

$$R_{\text{photo}} = \lambda_{\text{dec}} R_{\text{dec}} + \lambda_{\text{loc}} R_{\text{loc}}, \tag{3}$$

where λ_{dec} and λ_{loc} weight the two components.

GRPO optimizes sampled responses using group-relative rewards; the complete objective and notation are provided in Appendix E.3.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Implementation Details. We run EMDP for three rounds and train ShutterMuse with SFT followed by photographer-side RFT. Training hyperparameters and inference settings are provided in Appendix E.4.

Evaluation Protocol. We follow the evaluation metrics defined in Section 3.2, using Qwen3.8-Max (Qwen Team, 2026b) as the primary MLLM judge. Individual rubric scores take values in $\{0, 0.5, 1\}$, with aggregation following the task-specific definitions. Results with alternative judges are reported in Appendix B.1. For the subject-side benchmark, we additionally report the average number of output tokens and wall-clock inference time per recommendation.

5.2 QUANTITATIVE ANALYSIS

Table 1: Quantitative results on the photographer-side guidance subset of ShutterBench. $R@0.7$ denotes the refinement success rate, RSR and KSR denote reject success rate and keep success rate, respectively. Macro-F1 is the macro-averaged F1 score over the three decision categories. A dash indicates that the corresponding metric is not applicable to the method. Best and second-best results are in **bold** and underlined, respectively. Tied best results are all bolded.

Method	IoU%(↑)	BDE(↓)	$R@0.7$%(↑)	RSR%(↑)	KSR%(↑)	Macro-F1%(↑)	MLLM-Score(↑)
Open-source General MLLMs							
InternVL3.5-8B Wang et al. (2025)	42.86	0.127	8.61	0.00	22.92	29.09	0.18
Kimi-K2.6 Moonshot AI (2025)	65.60	0.092	38.28	4.48	96.88	28.79	0.47
Qwen3-VL-8B-Instruct Bai et al. (2025)	54.44	0.107	16.91	2.24	40.62	29.27	0.35
Qwen3-VL-32B-Instruct Bai et al. (2025)	64.33	0.099	37.39	7.46	96.88	20.14	0.43
Qwen3-VL-235B-A22B-Instruct Bai et al. (2025)	63.97	0.094	35.61	10.45	**100.00**	25.77	0.48
Qwen3.5-9B Team (2026)	61.94	0.094	30.86	5.97	77.08	31.37	0.46
Qwen3.6-27B Qwen Team (2026a)	53.96	0.084	31.45	50.00	66.67	45.24	0.48
Proprietary General MLLMs							
Gemini-3.0-Flash Google (2025b)	62.46	0.077	36.80	**74.63**	70.83	68.99	0.60
Gemini-3.0-Pro Google (2025a)	64.08	0.070	47.77	72.39	81.25	63.39	0.63
Gemini-3.1-Pro Google (2026a)	65.42	0.069	49.85	72.39	80.21	64.31	0.65
Gemini-3.5-Flash Google (2026b)	66.05	0.075	46.29	57.46	86.46	57.39	0.63
GPT-5.4 OpenAI (2026a)	60.03	0.089	35.91	46.27	67.71	48.89	0.48
GPT-5.5 OpenAI (2026b)	69.74	0.078	55.49	5.22	45.83	33.75	0.48
Specialized Aesthetic Cropping Models							
CACNet Hong et al. (2021)	67.57	0.080	51.93	0.00	0.00	24.85	0.37
UNIC Liu et al. (2023b)	64.35	0.081	36.20	0.00	0.00	24.85	0.29
InstructCrop Sheng et al. (2025)	69.53	0.072	56.97	0.00	0.00	24.85	0.36
Venus Du et al. (2026)	69.90	0.075	59.05	0.00	2.08	26.24	0.44
ShutterMuse (Ours)	**74.72**	**0.053**	**68.25**	68.66	79.17	**70.10**	**0.69**

We compare against general-purpose MLLMs and specialized cropping models for photographer-side guidance, and direct keypoint-generation and image-editing models for subject-side guidance. For image-editing baselines, we extract keypoints and evaluate the same skeleton visualization format. Baseline prompts are provided in Appendix C.

Photographer-side guidance results. Table 1 shows that ShutterMuse achieves the best overall balance between composition accuracy and decision making on ShutterBench. In particular, it obtains the highest IoU, lowest BDE, best refinement success rate and the highest Macro-F1 scores. By contrast, specialized models such as Venus achieve competitive crop quality but do not natively support keep/reject decisions. General-purpose MLLMs are better at three-way decision making, but their crop predictions are less accurate.

We further conduct evaluation on the widely used FLMS benchmark to assess out-of-domain generalization. As shown in Table 2, ShutterMuse achieves an IoU of 87.61% and a BDE of 0.027. Together, the results on ShutterBench and FLMS demonstrate that ShutterMuse combines flexible capture-time decision making with state-of-the-art crop-localization accuracy.

Subject-side guidance results. Table 3 shows that ShutterMuse achieves the highest mean pose-quality score among the evaluated VLMs (0.50), followed by Gemini-3.1-Pro (0.43). It also outperforms GPT-5.5 on all three pose-quality criteria. Its mean score approaches GPT-Image-2 (0.53) but remains below Nano-Banana-Pro (0.60). The advantage of image-editing models may reflect stronger anatomical and scene-interaction priors acquired through broad-scale pretraining. ShutterMuse requires 4.96 seconds per recommendation, compared with 102.61 seconds for GPT-Image-2 and 55.16 seconds for Nano-Banana-Pro. This combination of competitive pose quality and substantially lower measured latency supports its use for interactive capture-time guidance.

Table 2: FLMS results. Best and second-best results are **bold** and underlined.

Method	IoU% ↑	BDE ↓
Gemini-3.0-Pro Google (2025a)	77.13	0.049
Gemini-3.5-Flash Google (2026b)	77.69	0.048
GPT-5.5 OpenAI (2026b)	64.99	0.049
InstructCrop Sheng et al. (2025)	80.98	0.043
CACNet Hong et al. (2021)	84.04	0.037
Venus Du et al. (2026)	86.63	0.030
ShutterMuse (Ours)	**87.61**	**0.027**

Table 3: Quantitative results on the subject-side guidance subset of ShutterBench. Mean averages Plausibility, Interaction, and Aesthetics. Time and # Tokens denote average response time and output token count per pose. Best and second-best results are marked in **bold** and underlined, respectively, separately within the image-editing and vision-language model groups.

Method	Plausibility ↑	Interaction ↑	Aesthetics ↑	Mean ↑	Time (s) ↓	# Tokens ↓
Image-Editing Models						
Nano-Banana-Pro Google (2025c)	**0.87**	**0.48**	**0.45**	**0.60**	**55.16**	**1370**
GPT-Image-2 OpenAI (2026c)	0.81	0.40	0.38	0.53	102.61	1427
Vision-Language Models						
Qwen3-VL-8B-Instruct Bai et al. (2025)	0.26	0.17	0.06	0.16	5.17	284.60
Qwen3-VL-235B-A22B-Instruct Bai et al. (2025)	0.66	0.12	0.05	0.28	11.37	366.56
Kimi-K2.6 Moonshot AI (2025)	**0.81**	0.12	0.08	0.34	5.99	**249.47**
Gemini-3.1-Pro Google (2026a)	0.78	0.22	0.29	0.43	20.85	439.46
GPT-5.5 OpenAI (2026b)	0.54	0.21	0.25	0.33	19.15	1052
ShutterMuse (Ours)	0.80	**0.33**	**0.36**	**0.50**	**4.96**	412



Figure 3: Qualitative comparisons of different models on photographer-side guidance.

5.3 Qualitative Analysis

Figure 3 illustrates composition decisions and instruction-conditioned framing; Venus still crops the illustrated `keep`/`reject` cases. Figure 4(a) compares pose recommendations with image-editing baselines, while panels (b,c) demonstrate joint composition and pose recommendation and existing-pose refinement.

5.4 Ablation and Diagnostic Analyses

Effect of Reinforcement Fine-Tuning and Reward Design. Table 4 shows that RFT improves IoU from 72.39% to 74.72% and raises RSR and KSR by 11.20 and 11.46 percentage points over SFT, respectively. These gains support both framing accuracy and capture-time decisions about whether to adjust the view. Removing R_{dec} substantially lowers RSR and KSR, while ablating either the reference-box IoU or saliency-coverage factor reduces localization performance. The full reward achieves the best results across all four reported metrics.

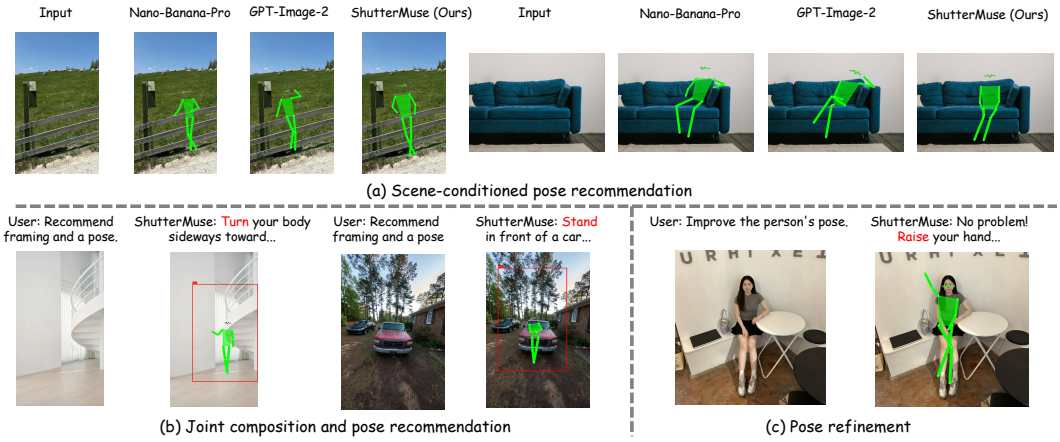

Figure 4: Pose guidance examples: (a) comparison with image-editing baselines; (b) joint composition and pose recommendation; (c) refinement of an existing person's pose. Red boxes indicate recommended framing; green skeletons indicate recommended poses.

Table 4: Photographer-side ablations of reinforcement fine-tuning and reward components. Best results are in bold.

Method	IoU% ↑	RSR% ↑	KSR% ↑	MLLM-Score ↑
ShutterMuse-SFT	72.39	57.46	67.71	0.60
ShutterMuse-RL w/o R_{dec}	74.10	56.72	68.75	0.65
ShutterMuse-RL w/o mask term	73.48	60.45	71.88	0.63
ShutterMuse-RL w/o IoU term	73.76	59.70	72.92	0.63
ShutterMuse-RL (Ours)	**74.72**	**68.66**	**79.17**	**0.69**

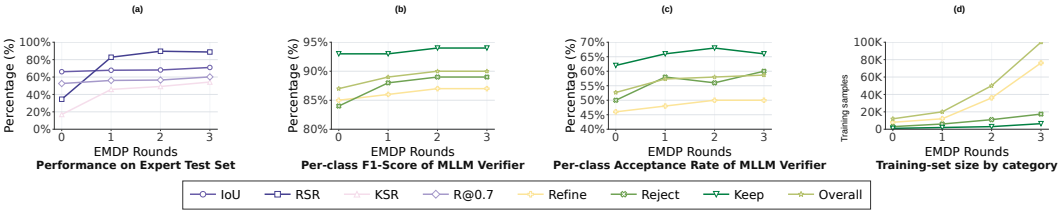

Figure 5: Reliability analysis of EMDP across three rounds. (a) Performance on the expert test set. (b) MLLM Verifier F1-score. (c) Acceptance rate. (d) Training-set size.

Reliability Analysis of EMDP. We evaluate three EMDP rounds on 450 held-out expert-annotated samples excluded from training and model selection (Appendix A.1). We also assess the verifier against expert judgments on 100 randomly sampled examples per category per round. Figure 5(a) shows overall improvements from the seed model to the final round across all four metrics, despite a slight RSR decline in the last round. Verifier F1 remains at least 84% across categories and rounds (Fig. 5(b)), indicating strong agreement with experts on the audited samples. After three rounds, the training set reaches 100K samples (Fig. 5(d)). These results support the utility of EMDP data.

Generalization to Real-Empty Scenes. To examine whether pose recommendation extends to inputs without subject-removal artifacts, we compare the subject-erased and real-empty subsets of ShutterBench (Appendix B.2, Table 7). Performance remains comparable on real-empty scenes, with a mean score of 0.52 versus 0.49 on subject-erased scenes. This supports generalization to scenes captured without subjects, but the unpaired comparison cannot isolate the effect of removal artifacts.

6 User Study

Agreement with expert ratings. Six photography experts independently rate 100 evaluation instances from each subset of ShutterBench using the same three-level scoring criteria as the MLLM judge. Photographer-side instances are sampled with approximately balanced coverage of `refine`, `keep`, and `reject`. For subject-side guidance, each expert gives one overall score considering physical plausibility, scene interaction, and pose aesthetics. We average the six expert ratings for each instance and compute Spearman's correlation with the corresponding MLLM score, obtaining 0.67 for photographer-side and 0.75 for subject-side guidance.

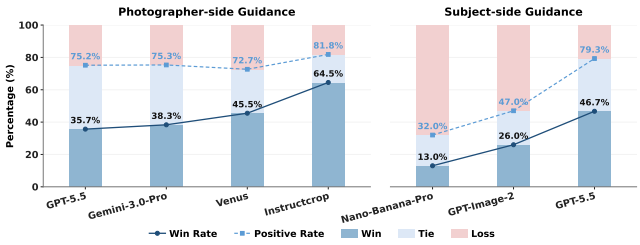

Figure 6: Pairwise human preferences for ShutterMuse against photographer-side (left) and subject-side (right) baselines. Bars show wins, ties, and losses; curves show win rate and positive rate (win + tie).

Pairwise human preference. Using the same samples and experts, we additionally assess pairwise preferences between ShutterMuse and competing methods (Fig. 6). Win rate denotes comparisons favoring ShutterMuse; positive rate includes both wins and ties. On the photographer side, ShutterMuse wins more often than it loses against every compared baseline, although ties account for a substantial fraction of comparisons. On the subject side, it achieves a 46.7% win rate and a 79.3% positive rate against GPT-5.5. Human preferences favor ShutterMuse over GPT-5.5 but favor Nano-Banana-Pro and GPT-Image-2 over ShutterMuse, consistent with the relative pose-quality scores in Table 3.

7 Conclusion

We introduced **ShutterBench**, a benchmark for evaluating MLLMs in capture-time photography guidance, covering both photographer-side composition refinement and subject-side pose recommendation. We constructed **ShutterData**, a large-scale dataset with structured annotations and textual rationales, and proposed **ShutterMuse**, a unified MLLM trained with supervised fine-tuning and reinforcement fine-tuning. Experiments on ShutterBench show that ShutterMuse achieves the best overall photographer-side performance among evaluated baselines, especially in balancing refinement accuracy with *keep*/*reject* decisions, and provides competitive subject-side pose recommendations with substantially lower inference cost. This work suggests that MLLMs can serve as practical interactive assistants for photography during capture.

AI USE STATEMENT

Generative AI tools were used for dataset construction, automated evaluation, and illustrative pose rendering, as described in the corresponding sections of this paper. An AI writing assistant also assisted with manuscript editing, consistency checks, and LaTeX formatting. Human verification of dataset annotations and comparisons with expert ratings are described in Section 3 and Section 6, respectively.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

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.

Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, et al. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. *arXiv preprint arXiv:2507.14533*, 2025.

Yi-Ling Chen, Tzu-Wei Huang, Kai-Han Chang, Yu-Chen Tsai, Hwann-Tzong Chen, and Bing-Yu Chen. Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pp. 226–234. IEEE, 2017.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

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

Chen Fang, Zhe Lin, Radomir Mech, and Xiaohui Shen. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1105–1108, 2014.

Google. Gemini 3. https://blog.google/products-and-platforms/products/gemini/gemini-3/, 2025a. Accessed: 2026-06-14.

Google. Gemini 3 Flash. https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/, 2025b. Accessed: 2026-06-14.

Google. Nano Banana Pro. https://blog.google/innovation-and-ai/products/nano-banana-pro/, 2025c. Accessed: 2026-06-14.

Google. Gemini 3.1 Pro. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro, 2026a. Accessed: 2026-06-14.

Google. Gemini 3.5 Flash. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/#gemini-3-5-flash, 2026b. Accessed: 2026-06-14.

Chaoyi Hong, Shuaiyuan Du, Ke Xian, Hao Lu, Zhiguo Cao, and Weicai Zhong. Composing photos like a photographer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7057–7066, 2021.

James Hong, Lu Yuan, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. Learning subject-aware cropping by outpainting professional photos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2175–2183, 2024.

Yipo Huang, Leida Li, Pengfei Chen, Haoning Wu, Weisi Lin, and Guangming Shi. Multi-modality multi-attribute contrastive pre-training for image aesthetics computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):1205–1218, 2024a.

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

Yu Ji and I-Han Hsiao. InstaPose: Scene-aware pose recommendation via vision transformers and diversity-optimized reranking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 2747–2753, 2025. URL https://openaccess.thecvf.com/content/ICCV2025W/SAUAFG/html/Ji_InstaPose_Scene-Aware_Pose_Recommendation_via_Vision_Transformers_and_Diversity-Optimized_Reranking_ICCVW_2025_paper.html.

Ruixiang Jiang and Chang Wen Chen. Before the shutter: Aesthetic and actionable portrait photography planning in 3D scenes. *arXiv preprint arXiv:2605.30318*, 2026. URL https://arxiv.org/abs/2605.30318.

Glenn Jocher, Jing Qiu, Mengyu Liu, Shuai Lyu, Fatih Cagatay Akyon, and Muhammet Esat Kalfaoglu. Ultralytics yolo26: Unified real-time end-to-end vision models, 2026. URL https://arxiv.org/abs/2606.03748.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

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

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

Xiaoyu Liu, Ming Liu, Junyi Li, Shuai Liu, Xiaotao Wang, Lei Lei, and Wangmeng Zuo. Beyond image borders: Learning feature extrapolation for unbounded image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13023–13032, 2023b.

Yuti Liu, Shice Liu, Junyuan Gao, Peng-tao Jiang, Hao Zhang, Jinwei Chen, and Bo Li. Advancing comprehensive aesthetic insight with multi-scale text-guided self-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 5748–5756, 2025.

Moonshot AI. Kimi K2.6: Moonshot AI Technical Blog. https://www.kimi.com/blog/kimi-k2-6, 2025. Accessed: 2026-06-14.

OpenAI. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026a. Accessed: 2026-06-14.

OpenAI. Introducing GPT-5.5. https://openai.com/index/introducing-gpt-5-5/, 2026b. Accessed: 2026-06-14.

OpenAI. Introducing ChatGPT Images 2.0. https://openai.com/index/introducing-chatgpt-images-2-0/, 2026c. Accessed: 2026-06-14.

Zhiyu Pan, Ke Xian, Hao Lu, and Zhiguo Cao. Robust image cropping by filtering composition irrelevant factors. In *International Conference on Image and Graphics*, pp. 277–289. Springer, 2021.

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

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27b dense model, April 2026a. URL https://qwen.ai/blog?id=qwen3.6-27b.

Qwen Team. Qwen3.8. https://qwen.ai/blog?id=qwen3.8, 2026b. Qwen Blog, accessed 2026-08-03.

Bin Sheng, Yuxi Jin, Ping Li, Wenxiao Wang, Hongbo Fu, and Enhua Wu. InspireMePosing: Learn pose and composition from portrait examples. In *Pacific Graphics Short Papers*, pp. 33–35. The Eurographics Association, 2018. doi: 10.2312/pg.20181274.

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

Yukun Su, Yiwen Cao, Jingliang Deng, Fengyun Rao, and Qingyao Wu. Spatial-semantic collaborative cropping for user generated content. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4988–4997, 2024.

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.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Yujia Wang, Sifan Hou, Bing Ning, and Wei Liang. Photo stand-out: Photography with virtual character. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 781–788, 2020. doi: 10.1145/3394171.3413957.

Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomir Mech, Minh Hoai, and Dimitris Samaras. Good view hunting: Learning photo composition from dense view pairs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5437–5446, 2018.

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

Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang. Learning the change for automatic image cropping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 971–978, 2013.

Guo-Ye Yang, Wen-Yang Zhou, Yun Cai, Song-Hai Zhang, and Fang-Lue Zhang. Focusing on your subject: Deep subject-aware image composition recommendation networks. *Computational Visual Media*, 9(1):87–107, 2023.

Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Reliable and efficient image cropping: A grid anchor based approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5949–5957, 2019.

Bo Zhang, Li Niu, Xing Zhao, and Liqing Zhang. Human-centric image cropping with partition-aware and content-preserving features. In *European Conference on Computer Vision*, pp. 181–197. Springer, 2022.

Ke Zhang, Tianyu Ding, Jiachen Jiang, Tianyi Chen, Ilya Zharkov, Vishal M Patel, and Luming Liang. Procrop: Learning aesthetic image cropping from professional compositions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 12600–12608, 2026.

Yanhao Zhang, Qingming Huang, Lei Qin, Sicheng Zhao, Xiusheng Lu, Xiaoshuai Sun, and Hongxun Yao. Strategy for aesthetic photography recommendation via collaborative composition model. *IET Computer Vision*, 9(5):691–698, 2015. doi: 10.1049/iet-cvi.2014.0276.

Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *arXiv preprint arXiv:2401.03407*, 2024.

Deyao Zhu, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pp. 18378–18394, 2024.

APPENDIX CONTENTS

A Data Statistics and Annotation Protocol 15
A.1 Photographer-side Annotation Guidelines 15
A.2 Subject-side Annotation Guidelines 15
A.3 Dataset and Benchmark Statistics 17
A.4 Training Data for Extended Guidance Tasks 17

B MLLM-based Evaluation Details 18
B.1 Evaluation with Alternative MLLM Judges 18
B.2 Generalization to Real-Empty Scenes 19
B.3 Photographer-side MLLM-Score Prompt 20
B.4 Subject-side MLLM-Score Prompt 21

C Baseline Prompt Templates 23
C.1 Photographer-side Prompt 23
C.2 Subject-side Prompt 24

D Failure Case Analysis for Subject-side Guidance 25

E Structured Outputs and Training Objectives 26
E.1 Output Schemas 26
E.2 Supervised Fine-Tuning Objective 27
E.3 Reinforcement Fine-Tuning Objective 27
E.4 Implementation Details 27

A Data Statistics and Annotation Protocol

A.1 Photographer-side Annotation Guidelines

Seed annotation and rationale construction. We collect seed images from multiple online platforms. Ten trained annotators assign labels with cross-review; ambiguous cases and those with low agreement are re-annotated, yielding a 12K-image seed set. Each image receives a decision label and free-form comments. For `refine`, annotators provide a target framing box and describe the compositional weaknesses and recommended adjustment. For `keep`, they describe the strengths of the current framing; for `reject`, they identify defects that cropping cannot resolve. An MLLM consolidates these comments into structured rationales, paired with the decision labels and framing boxes where applicable. Figure 7 illustrates the three decision categories and their annotation criteria.

Refine. An image is labeled `refine` when its subject or visual content is usable, but cropping can improve the framing. Typical weaknesses include insufficient subject prominence, an unbalanced visual center, a subject that occupies too little of the frame, and distracting background elements. Annotators identify these weaknesses and provide a framing box that improves composition while preserving the main subject and important context.

Keep. An image is labeled `keep` when its current framing is compositionally appropriate and further cropping would remove useful context or reduce composition quality. Annotators describe strengths such as clear subject placement, visual balance, effective background context, and coherent scene layout, explaining why the view should be preserved.

Reject. An image is labeled `reject` when defects such as severe blur, occlusion, poor exposure, a missing subject, or tilted framing cannot be adequately corrected by cropping. Annotators identify these defects and explain why a framing adjustment is insufficient.

EMDP reliability evaluation protocol. Before training any composition model, we reserve 450 expert-annotated samples as a fixed held-out expert test set. These samples are excluded from the initial seed training set, the EMDP expansion process, SFT, RFT, verifier calibration, and hyperparameter selection. To assess verifier agreement with experts, we randomly sample 100 examples per category at each round and compare the verifier's decisions with expert judgments.

A.2 Subject-side Annotation Guidelines

Scene–pose pair construction. Figure 8 summarizes the construction pipeline. Given an original portrait, we use Nano-Banana-Pro (Google, 2025c) to remove the person while preserving the background layout and contextual information. Qwen3-VL-27B (Bai et al., 2025) checks the resulting scene for removal artifacts, and samples that fail this check are discarded. In parallel, a YOLO-based pose estimator (Jocher et al., 2026) extracts target keypoints from the original portrait. The person-free image serves as the training input, paired with the extracted pose and its visibility labels.

Pose representation and verification. For subject-side annotations, we follow the COCO-17 human keypoint order: nose, left eye, right eye, left ear, right ear, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle, and right ankle. The left/right labels are defined from the subject's anatomical perspective.

Each subject-side instance is stored as a structured annotation containing the task type, a natural-language rationale, normalized keypoint coordinates, and visibility labels. Specifically, `keypoints_xyn` contains 17 normalized (x, y) coordinates following the COCO-17 order, and `visibility` contains the corresponding visibility state for each keypoint. We use a three-level visibility scheme: `1` indicates that the keypoint is inside the image and visible; `0` indicates that the keypoint is inside the image but not visible, e.g., due to occlusion; and `-1` indicates that the keypoint is outside the image boundary.

Professional annotators overlay the extracted skeleton on the person-free scene and compare it with the original portrait. They discard samples whose poses differ substantially from the reference or are

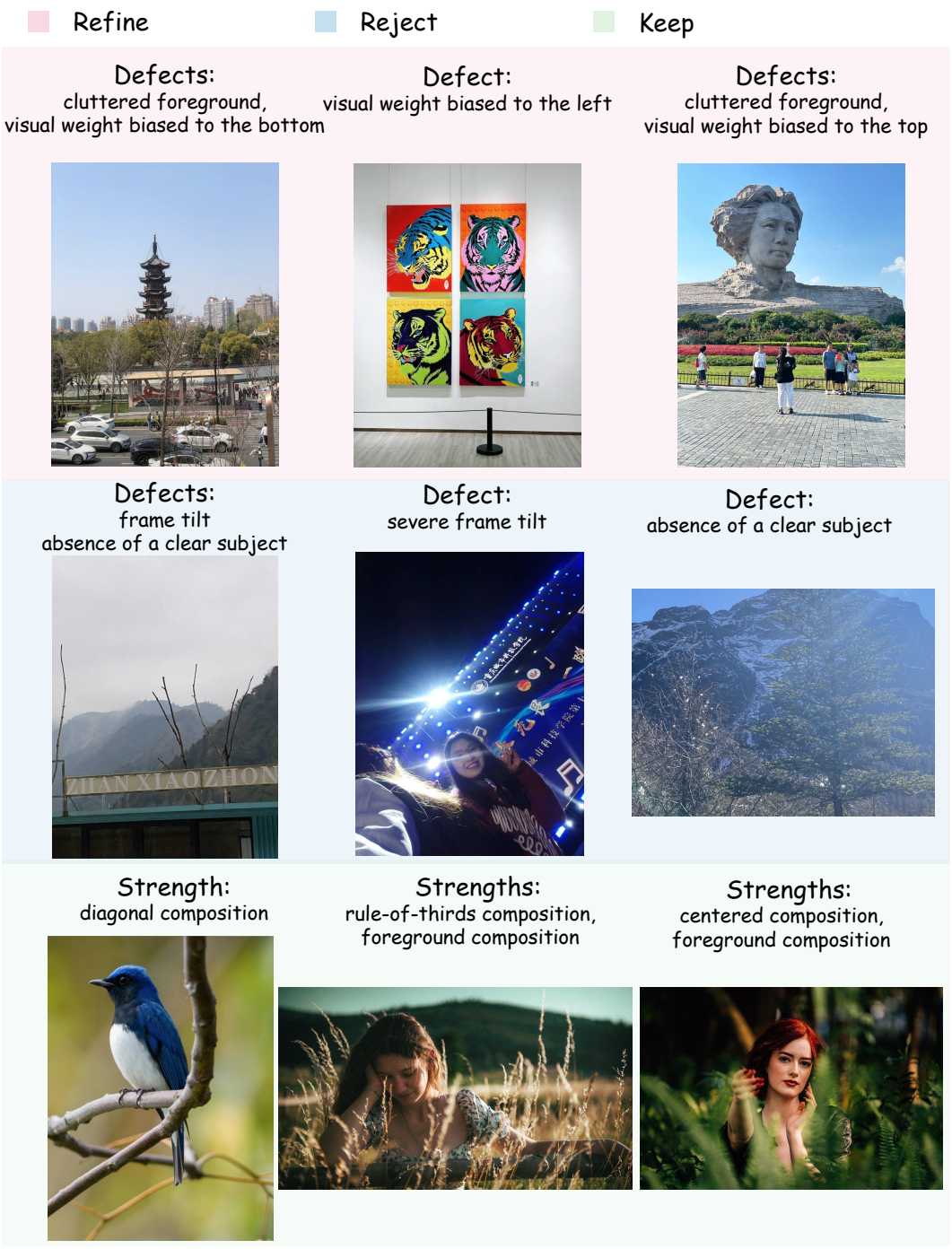

Figure 7: Examples illustrating the annotation criteria. Rows show `refine` cases, where cropping improves composition; `reject` cases, where cropping cannot resolve the defects; and `keep` cases, where the current framing is already appropriate.

aesthetically unsuitable for the scene. For retained samples, annotators correct inaccurate visibility labels while leaving the keypoint coordinates unchanged. Visibility describes the recommended subject pose, not the person-free input scene.

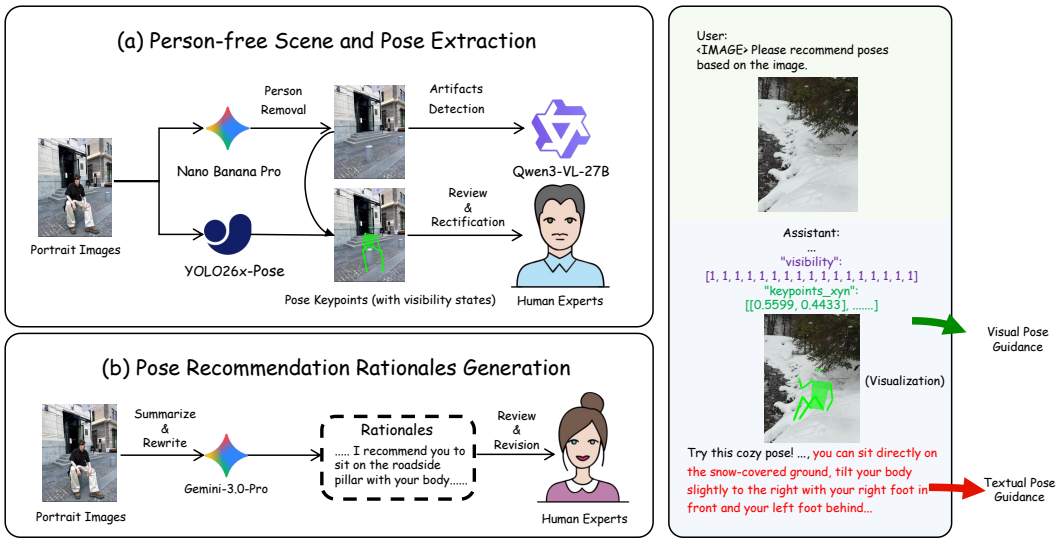

Figure 8: Subject-side data construction. Portrait images are converted into person-free scenes and paired with verified pose keypoints, visibility labels, and textual rationales.

Rationale generation and review. Gemini-3.0-Pro analyzes the scene and subject pose in the original portrait and expresses this analysis as a recommendation rationale. Professional annotators review and revise each rationale for accuracy, scene relevance, and consistency with the verified pose. Five professional photographers further assess whether the described pose cues are sufficient to recover the visualized pose. Rationales that do not consistently describe the intended pose are revised or removed from the final annotations.

A.3 Dataset and Benchmark Statistics

Figure 9 summarizes ShutterData and ShutterBench. The 130K atomic-task training examples comprise 100K photographer-side and 30K subject-side examples. The figure reports composition aspect ratios, decision categories, and scene and pose categories. ShutterBench evaluates the two atomic tasks separately, and its examples are excluded from both SFT and RFT.

A.4 Training Data for Extended Guidance Tasks

In addition to the 130K atomic-task examples in ShutterData, we construct 5K examples for joint composition and pose recommendation and 5K for refinement of an existing person's pose. All 10K additional examples are included in SFT for the model reported in the main comparison tables, bringing the total SFT training set to 140K examples. ShutterBench evaluates only the two atomic tasks; the extended capabilities are illustrated qualitatively in Fig. 4(b,c).

Joint composition and pose recommendation. Following the outpainting-based crop supervision of Hong et al. (2024), we derive compatible framing and pose targets from high-quality portraits. We extract the subject's keypoints and remove the person. GPT-Image-2 then outpaints the person-free image to create an expanded scene, which serves as the input. The original image region defines the target framing box, and the extracted keypoints are mapped into the expanded image coordinates to define the target pose.

Let (x_0, y_0, x_1, y_1) denote the normalized target framing box in the expanded image, and let (u_i, v_i) be keypoint i normalized with respect to the original portrait. Its coordinates in the expanded image are

$$u'_i = x_0 + u_i(x_1 - x_0), \qquad v'_i = y_0 + v_i(y_1 - y_0). \tag{4}$$

This mapping preserves the target pose's placement within the recommended framing box. Visibility labels are inherited from the original portrait and manually rechecked for consistency with the

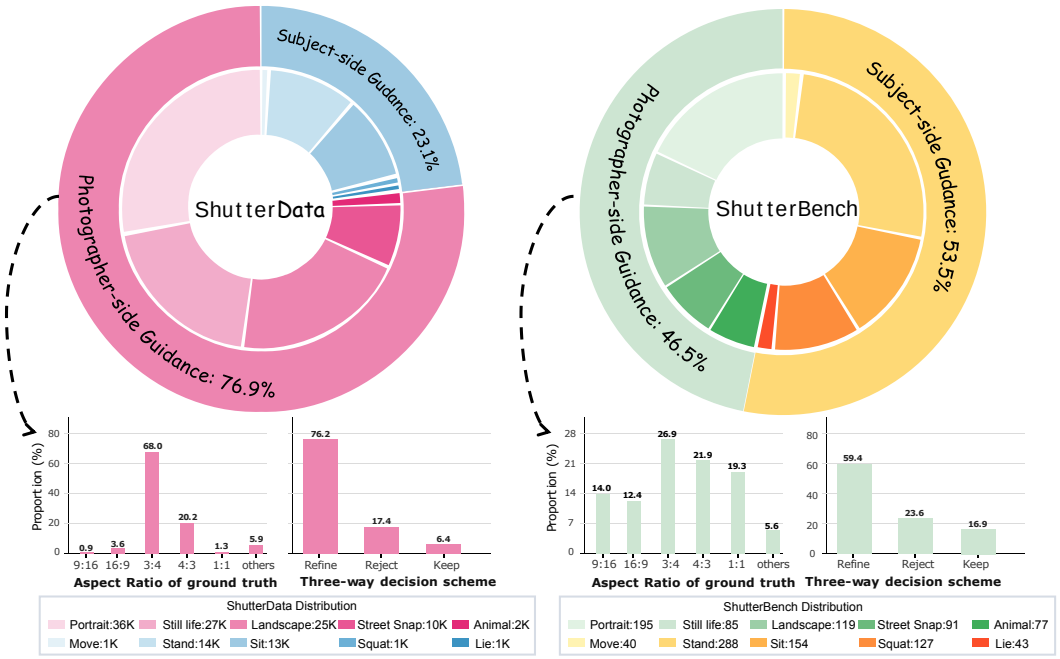

Figure 9: Distributions of the 130K atomic-task training examples in ShutterData and the evaluation examples in ShutterBench.

recommended pose and framing. They describe the recommended final pose, rather than visibility in the person-free input. For this joint task, -1 denotes a keypoint outside the recommended framing box, rather than outside the full input image; keypoint coordinates remain normalized to the full input image.

Refinement of an existing person's pose. For each high-quality portrait, we extract the subject's keypoints as the target pose. We then use GPT-Image-2 to replace the original pose with an ordinary, less visually interesting one while preserving the scene and image dimensions. The edited image serves as the input, and the original keypoints specify the desired pose. Because the edit does not constrain the subject's position, the recommendation may adjust both body configuration and placement. All target coordinates remain normalized to the full original image.

Recommendation rationales. For both extended tasks, Gemini-3.0-Pro generates the recommendation rationale stored in the `reason` field. The rationale explains the recommended pose and its relationship to the scene, while the structured fields specify the framing and/or keypoints. Output schemas are provided in Appendix E.1.

B MLLM-BASED EVALUATION DETAILS

B.1 EVALUATION WITH ALTERNATIVE MLLM JUDGES

We assess sensitivity to the MLLM judge by reevaluating the same model outputs with Gemini-3.5-Flash Google (2026b), alongside the primary judge, Qwen3.8-Max. Evaluation samples, prompts, and scoring protocols remain fixed. We report scores and rankings within the set of methods evaluated by both judges.

On photographer-side guidance (Table 5), both judges produce the same ranking of the eight compared methods. ShutterMuse ranks first, with scores of 0.671 under Gemini-3.5-Flash and 0.690 under Qwen3.8-Max. Gemini-3.1-Pro ranks second (0.650 and 0.646), followed by Gemini-3.5-Flash (0.623 and 0.631). The remaining methods also retain their relative ordering.

Table 5: Comparison of MLLM-based evaluation results under different judges on the photographer-side guidance subset of ShutterBench. Scores are rounded to three decimal places, while ranks are computed using the unrounded scores among the compared methods under each judge. The best and second-best scores for each judge are highlighted in **bold** and underlined, respectively.

MLLM Judge	Method	MLLM-Score ↑	Rank ↓
Gemini-3.5-Flash Google (2026b)	Gemini-3.1-Pro Google (2026a)	0.650	2
	Gemini-3.5-Flash Google (2026b)	0.623	3
	GPT-5.5 OpenAI (2026b)	0.502	4
	Kimi-K2.6 Moonshot AI (2025)	0.499	5
	CACNet Hong et al. (2021)	0.398	7
	InstructCrop Sheng et al. (2025)	0.338	8
	Venus Du et al. (2026)	0.405	6
	ShutterMuse (Ours)	**0.671**	1
Qwen3.8-Max Qwen Team (2026b)	Gemini-3.1-Pro Google (2026a)	0.646	2
	Gemini-3.5-Flash Google (2026b)	0.631	3
	GPT-5.5 OpenAI (2026b)	0.478	4
	Kimi-K2.6 Moonshot AI (2025)	0.467	5
	CACNet Hong et al. (2021)	0.368	7
	InstructCrop Sheng et al. (2025)	0.359	8
	Venus Du et al. (2026)	0.444	6
	ShutterMuse (Ours)	**0.690**	1

On subject-side guidance (Table 6), the four reevaluated methods have the same ordering under both judges: Nano-Banana-Pro, GPT-Image-2, ShutterMuse, and GPT-5.5. ShutterMuse receives a mean score of 0.50 under both judges, compared with 0.39 and 0.33 for GPT-5.5. These results preserve the main findings for the reevaluated methods: ShutterMuse leads the photographer-side comparison and outperforms GPT-5.5 in direct pose recommendation, while image-editing models achieve higher subject-side scores. Agreement between these two judges supports the stability of these comparisons despite differences in absolute scores.

Table 6: Comparison of MLLM-based evaluation results under different judges on the subject-side guidance subset of ShutterBench. Mean denotes the average score across Plausibility, Interaction, and Aesthetics. Rank is determined by the mean score under each judge. The best and second-best scores for each judge are highlighted in **bold** and underlined, respectively.

MLLM Judge	Method	Plausibility ↑	Interaction ↑	Aesthetics ↑	Mean ↑	Rank ↓
Gemini-3.5-Flash Google (2026b)	Nano-Banana-Pro Google (2025c)	**0.81**	**0.51**	**0.44**	**0.58**	**1**
	GPT-Image-2 OpenAI (2026c)	0.77	0.44	**0.44**	0.55	2
	GPT-5.5 OpenAI (2026b)	0.60	0.27	0.29	0.39	4
	ShutterMuse (Ours)	0.74	0.39	0.36	0.50	3
Qwen3.8-Max Qwen Team (2026b)	Nano-Banana-Pro Google (2025c)	**0.87**	**0.48**	**0.45**	**0.60**	**1**
	GPT-Image-2 OpenAI (2026c)	0.81	0.40	0.38	0.53	2
	GPT-5.5 OpenAI (2026b)	0.54	0.21	0.25	0.33	4
	ShutterMuse (Ours)	0.80	0.33	0.36	0.50	3

B.2 Generalization to Real-Empty Scenes

Table 7 reports ShutterMuse's performance on the 552 subject-erased scenes and 100 real-empty scenes in the subject-side subset of ShutterBench. Both subsets use the same evaluation protocol. Mean pose quality is comparable across the two subsets (0.49 and 0.52, respectively), supporting generalization to scenes captured without subjects. Since the subsets are unpaired, this comparison does not isolate the effect of subject-removal artifacts.

Table 7: Performance on the subject-erased and real-empty subsets of ShutterBench.

Evaluation Set	#Samples	Plausibility ↑	Interaction ↑	Aesthetics ↑	Mean ↑
Subject-erased	552	0.80	0.33	0.35	0.49
Real-empty	100	0.81	0.36	0.40	0.52

B.3 PHOTOGRAPHER-SIDE MLLM-SCORE PROMPT

The following prompt implements the composition-quality assessment described in Section 3.2. For ground-truth non-`reject` examples with a valid framing prediction, the judge receives the original image with a red box indicating the evaluated region, together with the annotated compositional strengths and weaknesses. The box is the predicted refinement for `refine` and the full image for `keep`. The judge scores only the indicated region. Invalid frames and `reject` predictions on these examples receive 0. Ground-truth `reject` examples are scored directly by decision correctness, without invoking this prompt.

Photographer-side MLLM-Score Prompt

You are a strict photography composition reviewer. You will see a single image containing the original image with the model-predicted composition frame overlaid as a red box. Please evaluate the quality of the composition corresponding to the red box. Output only one JSON object and do not output Markdown.
Reasons why the original image is good: `<origin_good>`
Reasons why the original image is poor: `<origin_bad>`
Three-level scoring rules. You only need to score the model-predicted composition frame. The score must be one of `0`, `0.5`, or `1`.

- **Score 0: severe composition problem.** Assign a score of 0 if the red box contains any of the following problems:
  1. The foreground or scene is cluttered, e.g., crowded tourists at a scenic spot, making the image look casual and preventing the subject from standing out.
  2. For portrait composition, check whether the red box cuts through important joints of the main subject, such as the ankle, knee, or neck. If the boundary is merely close to a joint, it should not be considered as cutting the joint. Only when the red boundary passes through the joint should it be regarded as a severe composition problem. Cutting background pedestrians is not considered a severe composition error.
  3. The crop severely violates the original photographic intention. For example, if the original image is intended to capture both the person and the scenery, but the crop contains only the person or only the scenery; or if the original image is intended as a group photo, but the crop contains only a few individuals.
- **Score 0.5: partially fixes the original problems.** Assign a score of 0.5 if the red box fixes only part of the problems described in the "poor reasons" of the original image. For example, if the left-biased visual center is corrected but the low visual center remains unresolved, the score should be 0.5. Each listed problem should be checked individually.
- **Score 1: fixes all original problems and improves aesthetics.** Assign a score of 1 if the red box fixes all original problems, introduces no obvious common composition errors, and uses an effective composition technique to improve visual balance, rhythm, layering, or spatial relationships.

Composition knowledge. The following common photography composition techniques should be used as references during evaluation:

- **Centered composition:** placing the subject near the image center, suitable for symmetric scenes or cases where the subject should be emphasized.
- **Rule-of-thirds composition:** placing the subject along the thirds lines or near the intersections of a 3×3 grid to create visual tension and dynamics.

- **Diagonal composition:** creating diagonal structures in the image to enhance depth and motion.
- **Symmetric composition:** arranging the subject to form symmetry, producing a stable and solemn visual effect.
- **Leading-line composition:** using environmental lines, such as roads, railings, or rivers, to guide the viewer's attention toward the subject.
- **Negative-space composition:** using a plain background or repeated elements to occupy approximately 80% of the image, while the subject occupies only a small region.
- **Frame-within-a-frame composition:** using mirrors, doors, windows, caves, branches, or other natural frames to enclose the subject.
- **Triangular composition:** using the subject's shape or pose to form a triangle, thereby improving visual stability.
- **Foreground composition:** using a blurred foreground object to increase layering and spatial depth.
- **Long shot:** in portrait photography, a composition where the full body of the subject is included.
- **Medium shot:** in portrait photography, a composition where the subject is included from the mid-calf or mid-thigh upward.
- **Close-up shot:** in portrait photography, a composition where the subject is included from the waist upward, or where the face is shown in close-up.

The following common composition errors should be carefully checked:

- **Incomplete subject:** important parts of the key subject are cropped out, e.g., the top of a person's head is cut off. In group photos, all main subjects should be included. For landscape or architectural composition, this does not apply as long as the important parts remain inside the frame; for example, excluding the base of a statue or the bottom of a mountain is not necessarily considered subject incompleteness.
- **Distracting elements:** irrelevant pedestrians, clutter, or objects in the background distract attention from the subject.
- **Tilted image:** the image is noticeably tilted, such as a slanted ground plane or tilted person. A severely tilted image may indicate that no recommendable composition is available.
- **Insufficient visual space:** when the subject's body or gaze has a clear direction, insufficient space is left in that direction. This is common in rule-of-thirds compositions.
- **Unclear subject:** the image is too cluttered to identify a clear subject, or the potential subject is strongly distracted by other visual elements.

The output format must strictly be:

```
{"score":0 or 0.5 or 1,
 "level":"L0 or L0.5 or L1",
 "reason":"brief explanation",
 "fixed_origin_bad":0 or 1,
 "used_composition_technique":"optional, brief description"}
```

B.4 SUBJECT-SIDE MLLM-SCORE PROMPT

We assess subject-side pose recommendations along three dimensions: **physical plausibility**, **scene interaction**, and **pose aesthetics**. Each dimension uses a dedicated prompt. The judge receives a single visualization of the predicted skeleton overlaid on the input scene, following Section 3.2. Three-level rubric scores are mapped to $\{0, 0.5, 1\}$, with higher values indicating better quality; Mean averages the three dimension scores. The prompts are reproduced below.

Subject-side MLLM-Score Prompt: Physical Plausibility

You are a strict but reasonable evaluator for human pose recommendations. This round evaluates only **physical plausibility**.

You will receive only one image: the model-generated pose visualization, shown as a skeleton image. In the skeleton image, red body parts indicate that the model believes these parts are occluded by objects. There are three types of poses: full-body poses, three-quarter-body poses, and upper-body poses. A full-body pose displays 17 keypoints. For simplicity, the head and the torso are not connected in the visualization. A three-quarter-body pose does not display the left ankle or the right ankle. An upper-body pose does not display the lower body.

Scoring rules:

- **score=0:** The model output severely violates physical constraints, and the pose cannot be imitated by a human. **Note that the separation between the head and the torso is not a severe physical violation; it is caused by the visualization style. Do not assign score 0 for this reason.**
- **score=0.5:** The pose slightly violates physical plausibility. For example, some red body parts are not actually occluded by objects, or the pose contains minor penetration or floating artifacts. However, the recommended pose can still be understood and imitated.
- **score=1.0:** The body scale is natural with respect to the environment, the contact between the person and the scene is credible, there is no obvious floating or penetration, and the pose can be reasonably imitated by a human.

Notes:

1. The evaluation does not need to be overly strict; if the pose can be understood and imitated, it is acceptable.
2. Slight floating is normal for standing poses, because the skeleton visualization only shows keypoints down to the ankles. This should not be regarded as floating.
3. Crossed legs do not necessarily indicate an error, because the two legs may be positioned at different depths.
4. When the person is sitting, floating feet can be physically reasonable.
5. **Do not mistakenly judge upper-body poses or three-quarter-body poses as incomplete bodies or floating bodies.**

Required output format:

```
{"score": 0 or 0.5 or 1.0, "reason": "brief explanation"}
```

Subject-side MLLM-Score Prompt: Scene Interaction

You are a strict but reasonable evaluator for human pose recommendations. This round evaluates only **interaction with the scene**.

You will receive only one image: the model-generated pose visualization, shown as a skeleton image. There are three types of poses: full-body poses, three-quarter-body poses, and upper-body poses. A full-body pose displays 17 keypoints. A three-quarter-body pose does not display the left ankle or the right ankle. An upper-body pose does not display the lower body.

Scoring rules:

- **score=0:** Weak interaction. The person does not interact with any object or scene element. Simply standing or squatting on the ground does not count as interaction.
- **score=0.5:** Moderate interaction. The pose shows simple interaction, such as sitting on a chair, leaning against a wall, sitting on steps, or leaning on a railing.
- **score=1.0:** Strong interaction. The pose clearly interacts with distinctive objects or scene elements, including interaction through body movement or gaze. Examples include interacting with small animals, flowers, signs, scenery, or other scene-specific objects.

Required output format:

```
{"score": 0 or 0.5 or 1.0, "reason": "brief explanation"}
```

Subject-side MLLM-Score Prompt: Pose Aesthetics

You are a strict but reasonable evaluator for human pose recommendations. This round evaluates only **pose aesthetics**.

You will receive only one image: the model-generated pose visualization, shown as a skeleton image. There are three types of poses: full-body poses, three-quarter-body poses, and upper-body poses. A full-body pose displays 17 keypoints. A three-quarter-body pose does not display the left ankle or the right ankle. An upper-body pose does not display the lower body.
Scoring rules:

- **score=0:** The pose is an ordinary static pose, with neither dynamic movement nor meaningful pose details, such as standing upright or sitting straight.
- **score=0.5:** The pose contains certain details or dynamics, such as standing with crossed legs, raising one leg, crossing the legs while seated, walking, looking back at the camera, or waving.
- **score=1.0:** The pose contains rich action details, has high visual interest and expressive tension, and is well adapted to the environment.

Notes: Only evaluate the pose itself. Do not consider irrelevant factors such as the person's appearance, clothing, or image composition.
Required output format:

```
{"score": 0 or 0.5 or 1.0, "reason": "brief explanation"}
```

C Baseline Prompt Templates

This section documents the prompts used for baseline evaluation. Specialized cropping models use official checkpoints and default inference settings, with task-specific prompts from their original papers where applicable. General-purpose MLLMs use a unified photographer-side prompt. Subject-side evaluation includes direct keypoint generation with vision-language models and image editing with GPT-Image-2 and Nano-Banana-Pro. We provide the photographer-side templates, the shared direct-generation prompt for all VLM baselines, and the shared image-editing prompt below.

C.1 Photographer-side Prompt

Specialized Cropping Baselines. For specialized cropping baselines, including InstructCrop and Venus, we use their official checkpoints and default settings. We follow the prompt templates provided in their original papers.

Venus Prompt

Please provide the bounding box coordinate of the most visually balanced and aesthetically pleasing composition area.

InstructCrop Prompt

Please suggest the best cropping area in this image, and explain the reason.

General MLLM Baselines. For general MLLM baselines, we use the following unified prompt. The prompt asks the model to first determine whether the original image has good composition, improvable composition, or extremely poor composition that is difficult to salvage. It then requires the model to output either a normalized bounding box with the specified aspect ratio or the special token `<non>` when no aesthetically valid crop exists.

Unified Photographer-side Prompt for General MLLM Baselines

Role. You are a senior photography composition mentor.

Input. You will receive one image, i.e., the original image. Its composition may fall into one of the following three cases, and you need to determine which case it belongs to:

1. The original image has excellent composition.
2. The original image has room for compositional improvement.
3. The original image has extremely poor composition and is difficult to salvage.

Task. Based on the composition quality of the original image, generate a logically coherent, natural, and concise analysis. Do not use a list format. The output should be concise and accurate and should not introduce subjective assumptions or unsupported inferences. Keep the analysis within 100 characters.
The reasoning logic should be as follows.

Case 1: The original image has excellent composition. If the original image already has excellent composition, first identify the photographic subject. If the subject includes both a person and other elements, such as a person–scenery photo, also specify the scenery included in the subject. Then explain why the composition is aesthetically pleasing, explicitly state that no modification is needed and that preserving the original image is the best choice. Finally, output the bounding box of the entire image.

Case 2: The original image has room for compositional improvement. If the original image has room for improvement, first determine whether the subject is not prominent. If so, directly state that the subject in the original image is not prominent. Otherwise, identify the photographic subject. If the image is intended as a person–scenery photo, explicitly state that the photographic intention is to capture both the person and the scenery.
Then describe the strengths and weaknesses of the original composition. Finally, if the photographic subject is a person, first examine the original image and determine an appropriate shot scale. A close-up shot refers to a half-body portrait, a long shot refers to a full-body portrait, and a medium shot refers to a portrait cropped from around the knees upward. After specifying the recommended shot scale, explain how to modify the original image to improve its composition, and provide the cropping bounding box.

Case 3: The original image has extremely poor composition and is difficult to salvage. If the original image has extremely poor composition and is difficult to salvage, first determine whether the subject is not prominent. If so, directly state that the subject in the original image is not prominent. Otherwise, identify the photographic subject. If the image is intended as a person–scenery photo, explicitly state that the photographic intention is to capture both the person and the scenery.
Then point out the defects in the image. If the reference defects include "cluttered background" or "cluttered foreground", specify what causes the clutter, for example, "the crowded pedestrians at the bottom cause foreground clutter". Finally, explicitly state that "there is no composition scheme in this image that satisfies aesthetic standards", and output the special token `<non>`.

Bounding box format. If a bounding box needs to be returned, it must follow the aspect ratio specified by `<target_ratio>` and be returned in the format `[x1,y1,x2,y2]`, where `(x1,y1)` is the top-left corner and `(x2,y2)` is the bottom-right corner. All coordinates must be normalized to the range $[0.0, 1.0]$.

C.2 Subject-side Prompt

Specialized cropping models such as InstructCrop and Venus predict framing boxes and do not directly support pose recommendation. As described in the main experiments, direct keypoint-generation baselines include Qwen3-VL-8B-Instruct, Qwen3-VL-235B-A22B-Instruct, Kimi-K2.6, Gemini-3.1-Pro, and GPT-5.5. All five VLMs use the same prompt and output schema shown below. Image-editing baselines comprise GPT-Image-2 and Nano-Banana-Pro, which share a separate editing prompt.

C.2.1 Direct Keypoint Generation with VLMs

Each direct-generation VLM predicts a structured pose from the input scene. Its response contains an `instance_info` list with one instance and the fields `task_type`, `reason`, `keypoints_xyn`, and `visibility`. The 17 keypoints follow the COCO-17 order and are normalized to the full image. Visibility labels are 1 for visible keypoints, 0 for occluded keypoints, and -1 for keypoints outside the frame. Predictions are rendered in the same skeleton format as ShutterMuse and evaluated using the same subject-side criteria. The prompt below is shared by all direct keypoint-generation VLM baselines; the keypoint-name list is expanded for clarity.

Shared Direct Keypoint-Generation Prompt for VLM Baselines

```
You are a portrait photography pose analysis expert. Based on the
    image, recommend one aesthetically pleasing, natural, physically
    plausible human pose that fits the scene and can be used as a
    photography guide. Provide the relative coordinates of 17 human
    keypoints and whether each keypoint is visible in the image in
    JSON format. The 17 keypoints are, in order: nose, left eye, right
     eye, left ear, right ear, left shoulder, right shoulder, left
    elbow, right elbow, left wrist, right wrist, left hip, right hip,
    left knee, right knee, left ankle, right ankle.

Output requirements:
1. Return exactly one JSON object and no markdown or additional text.
2. Use exactly this schema:
{"instance_info":[{"task_type":"pose","reason":"A concise, actionable
    description of the recommended pose.","keypoints_xyn":[[x,y
    ],...],"visibility":[v,...]}]}
3. keypoints_xyn must contain exactly 17 [x,y] pairs in the order
    listed above. Both x and y must be normalized floating-point
    coordinates in [0,1], relative to the full input image.
4. visibility must contain exactly 17 integers. Use 1 when the
    keypoint is visible in the recommended final pose and 0 when it is
     occluded, -1 when it is outside the frame.
```

C.2.2 Image-Editing Baselines

GPT-Image-2 and Nano-Banana-Pro use a shared prompt to insert a human mesh representing a recommended pose into the input scene. To reduce the influence of rendering style, texture, lighting, and editing artifacts, we extract keypoints from the generated subjects and render them in the same skeleton format used for ShutterMuse and direct keypoint-generation baselines. The MLLM judge evaluates these standardized scene–skeleton visualizations rather than the edited images.

Subject-side Editing Prompt for GPT-Image-2 and Nano-Banana-Pro

Please edit this image by adding a human mesh into the scene for photography pose recommendation. The pose of the human mesh must be highly compatible with the current environment, visually aesthetic and natural, and physically plausible.

Hard requirements:

1. The pose should be stable, without floating or penetration. The limb proportions should be natural, and the joint directions should be reasonable.
2. The pose should be aesthetically pleasing and useful as a photography pose reference. Avoid stiff, distorted, or unnatural movements.
3. The relationship between the human mesh and the environment should be reasonable, e.g., standing, sitting, or leaning on a physically supportive location, with natural body orientation and gaze direction.
4. Keep the original scene and composition largely unchanged, and do not add any additional people.
5. Output a single edited image.

D Failure Case Analysis for Subject-side Guidance

We examine apparent body–scene contact mismatches in pose visualizations and distinguish limitations of the skeleton representation from errors in the recommended pose.

Figure 10 illustrates the limitations of representing a human body using sparse keypoints and connecting lines. Skeletons encode joint positions but omit body volume and contact surfaces. In a

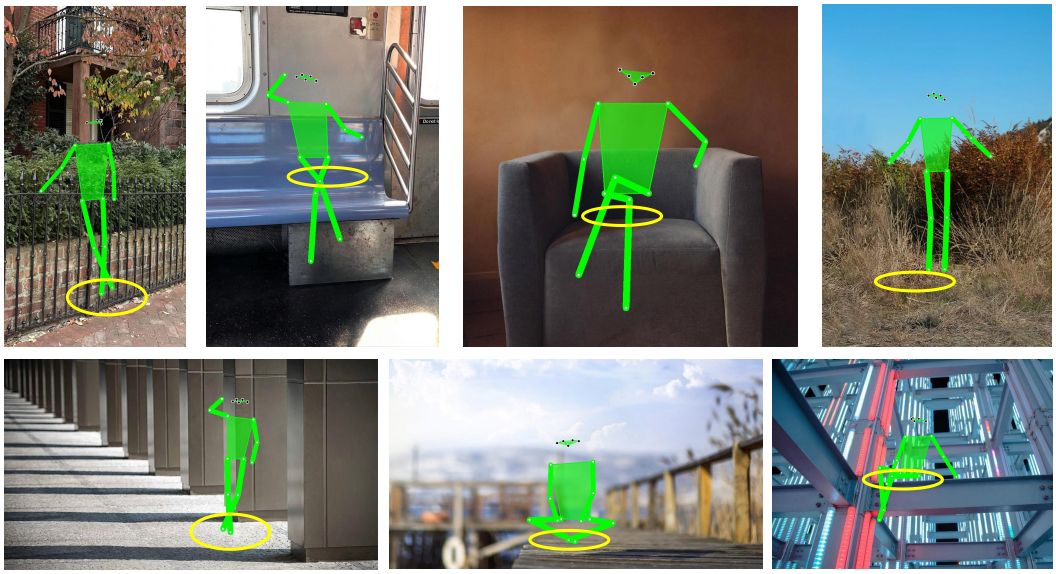

Figure 10: Examples of apparent contact mismatches in subject-side skeleton visualizations. Yellow circles highlight gaps near the feet or between the body and supporting surfaces. A sparse skeleton does not represent body thickness or the full contact surface of a person.

seated pose, the hips and thighs occupy space around the corresponding joints, so a visible gap between the skeleton and a seat does not necessarily imply that the body would lack support. Likewise, COCO-17 represents ankles but not soles or full foot–ground contact regions, which can make a standing pose appear to float.

These representational limitations can explain some apparent contact gaps, but they do not rule out pose errors. Sparse 2D keypoints alone cannot reliably distinguish omitted body volume from incorrect positioning, limiting the precision of contact-related guidance. Incorporating body shape, surface geometry, or explicit contact representations could improve the assessment and generation of physically supported poses.

E Structured Outputs and Training Objectives

E.1 Output Schemas

Each training example consists of an input image $\mathbf{x}$, a text prompt $\mathbf{p}$, and a target response $\mathbf{y}$. The response is a JSON object whose schema depends on the guidance task, implementing the unified structured guidance described in Section 4.

For photographer-side guidance, the fields are `task_type`, `reason`, and `composition_xy`, with `task_type` set to `composition`. The `composition_xy` field encodes both the decision and framing: an empty value denotes `reject`, $[0, 0, 1, 1]$ denotes `keep`, and another valid normalized box $[x_1, y_1, x_2, y_2] \in [0, 1]^4$ denotes `refine`.

For subject-side guidance, the fields are `task_type`, `reason`, `keypoints_xyn`, and `visibility`, with `task_type` set to `pose`. The `keypoints_xyn` field stores 17 normalized keypoint coordinates in COCO-17 order. The corresponding visibility vector uses 1 for visible keypoints, 0 for occluded keypoints within the image, and -1 for keypoints outside the frame.

The COCO-17 keypoint order and annotation conventions are detailed in Appendix A.2.

Extended-task output schemas. For joint composition and pose recommendation, the JSON response contains the fields `task_type`, `reason`, `keypoints_xyn`, and `composition_xy`. The field `task_type` is set to `composition and pose`. The field `keypoints_xyn` stores

the normalized coordinates of 17 human keypoints, while `composition_xy` specifies the recommended framing box. Both are normalized with respect to the full input image. For pose refinement, the JSON response contains `task_type`, `reason`, and `keypoints_xyn`, with `task_type` set to `pose_refine`. The keypoints specify the recommended pose for the person in the input image. For both tasks, `reason` provides a recommendation rationale generated by Gemini-3.0-Pro.

E.2 Supervised Fine-Tuning Objective

Let $q = (\mathbf{x}, \mathbf{p})$ denote the image-prompt input and $\mathbf{y}^\star = (y_1^\star, \ldots, y_L^\star)$ denote the target JSON response. We optimize the response-only next-token prediction loss:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(q,\mathbf{y}^\star)\sim\mathcal{D}_{\mathrm{SFT}}}\left[\frac{1}{L}\sum_{t=1}^{L}\log \pi_\theta\left(y_t^\star \mid q, y_{<t}^\star\right)\right]. \tag{5}$$

E.3 Reinforcement Fine-Tuning Objective

Let $q = (\mathbf{x}, \mathbf{p})$ denote an image-prompt input. As described in Section 4, GRPO is applied only to photographer-side guidance using R_{photo}, which combines decision correctness with localization quality conditioned on both decisions being `refine`.

For each input q, we sample G responses $\{\mathbf{y}_i\}_{i=1}^{G}$ from the old policy $\pi_{\theta_{\mathrm{old}}}$ and compute $r_i = R_{\mathrm{photo}}(q, \mathbf{y}_i)$. The group-relative advantage is

$$A_i = \frac{r_i - \bar{r}}{s_r + \epsilon}, \tag{6}$$

where $\bar{r}$ and s_r are the mean and standard deviation of the rewards within the group, and ϵ is a small numerical constant. For token $y_{i,t}$, we define

$$\rho_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t} \mid q, \mathbf{y}_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(y_{i,t} \mid q, \mathbf{y}_{i,<t})}. \tag{7}$$

Let L_i be the length of response $\mathbf{y}_i$. The GRPO loss is

$$\mathcal{L}_{\mathrm{GRPO}} = -\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{L_i}\sum_{t=1}^{L_i}\left(\min[\rho_{i,t}A_i,\ \mathrm{clip}(\rho_{i,t}, 1-\epsilon_c, 1+\epsilon_c)A_i] - \beta D_{i,t}^{\mathrm{KL}}\right)\right], \tag{8}$$

where

$$D_{i,t}^{\mathrm{KL}} = D_{\mathrm{KL}}(\pi_\theta(\cdot \mid q, \mathbf{y}_{i,<t}) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid q, \mathbf{y}_{i,<t})), \tag{9}$$

Here, π_{ref} is the frozen SFT reference policy, ϵ_c is the clipping parameter, and the KL divergence is computed at each decoding step under the corresponding conditioning context.

E.4 Implementation Details

The EMDP pipeline is conducted for three rounds. SFT is performed on eight A800 GPUs using the AdamW optimizer with a learning rate of 1×10^{-4} and an effective batch size of 64 for 5 epochs. During inference, we employ vLLM (Kwon et al., 2023). The RFT stage is initialized from the SFT model and uses the SFT model as the reference policy. For GRPO, we use an effective batch size of 64 and sample 32 rollouts per input for group-wise reward normalization and policy optimization. We train for 1 epoch with a learning rate of 1×10^{-6}, a weight decay of 0.1, and a KL regularization coefficient of $\beta = 0.01$. For the saliency-coverage factor in R_{loc}, we set the coverage threshold to $\tau_m = 0.9$. We set the reward weights to $\lambda_{\mathrm{dec}} = \lambda_{\mathrm{loc}} = 1$, assigning equal weight to decision correctness and composition localization.