FRAMELITE: DECISION-ORIENTED AESTHETIC CROPPING WITH A COMPACT VISION–LANGUAGE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Aesthetic image cropping improves composition by balancing subjects, context, and image boundaries. Large vision–language models bring semantic knowledge to this task, but their scale limits lightweight photography assistance. We propose FrameLite, a compact cropping model built on Qwen3.5-0.8B. Our approach is motivated by the observation that a small model can generate better crops than its first answer suggests. We develop three-stage training to turn this potential into stronger direct predictions. Spatial and composition training establishes localization and framing preferences. Quality-Guided Online Supervision (QOS) then learns the highest-quality complete answers among the current policy's generated candidates. Greedy-Referenced Optimization (GRO) further refines the policy using positive and negative quality differences relative to its current greedy answer. Both online stages use annotation-derived feedback, while inference returns one coordinate answer without search or an external scorer. Experiments show that FrameLite outperforms the compared large-VLM croppers across public benchmarks with a substantially smaller backbone. Controlled studies confirm that selecting high-quality training answers improves the final model beyond online sampling alone.

1 INTRODUCTION

Aesthetic image cropping selects a visually coherent region by adjusting image boundaries, removing distractions, and balancing subjects with their surroundings. Beyond locating salient objects, it determines how much context to retain, where to leave space, and which visual relationships to preserve. Multiple windows can therefore form reasonable compositions (Jia et al., 2022). Progress spans view evaluation and proposal learning (Chen et al., 2017b; Wei et al., 2018), grid anchors (Zeng et al., 2019; 2022), and composition-aware or spatial–semantic modeling (Hong et al., 2021; Zhang et al., 2022; Su et al., 2024). Feature extrapolation, professional-composition transfer, and global assessment provide further ways to learn framing (Liu et al., 2023; Hong et al., 2024; Zhang et al., 2026; Yang et al., 2026). These advances support automatic reframing and motivate stronger semantic reasoning about subject placement and photographic composition.

Vision–language models (VLMs) connect cropping with instruction following, aesthetic knowledge, and contextual examples (Zhong et al., 2022; Sheng et al., 2025; Du et al., 2026; Lee et al., 2025). Current methods also incorporate compositional reasoning, multimodal guidance, and capture-time assistance (Dong et al., 2026; You et al., 2026; Li et al., 2026). Their deployment requirements can nevertheless be demanding for everyday photography. Cropper uses an external Gemini service with retrieval and iterative refinement (Lee et al., 2025); InstructCrop's released implementation adopts LLaVA-1.6-7B (Sheng et al., 2025; Sheng, n.d.); Venus-Q and CROP also use 7B-class backbones (Du et al., 2026; Dong et al., 2026). These systems offer valuable explanations and interactions. A person composing a photograph on a phone may need one immediate, actionable suggestion (Li et al., 2025). Large model weights place pressure on device memory, while cloud access and repeated generation add dependencies and response costs. This motivates a compact VLM that can make a strong framing decision through a short, self-contained prediction path. The practical challenge is to express semantic judgments as precise boundaries within this compact model, while keeping the interaction as simple as a single actionable framing suggestion for the photographer.

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.

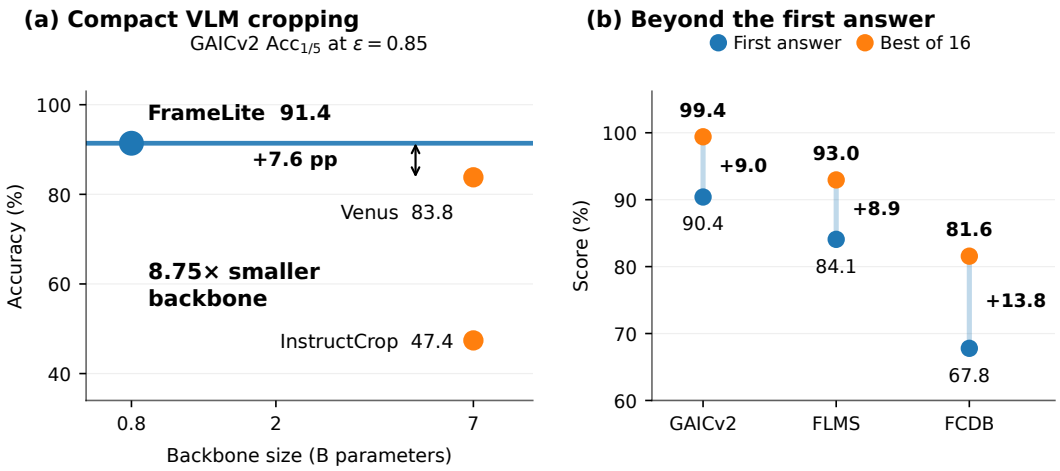

Figure 1: **Better first answers with a compact model.** (a) GAICv2 $\mathrm{Acc}_{1/5}$ at $\epsilon = 0.85$ versus nominal VLM backbone size; the horizontal line marks FrameLite. (b) First-answer and best-of-16 results from the same early supervised Qwen model. Best of 16 selects the highest-scoring sampled crop using human reference annotations. Scores are percentages of the GAICv2 $\mathrm{Acc}_{1/5}^{0.85}$ acceptance rate, FLMS maximum-reference IoU, and FCDB single-reference IoU.

Our preliminary experiments reveal a useful opportunity: a small model trained with grounding and composition supervision can generate better crops than its first answer suggests. Choosing the best of sixteen samples using human references substantially improves quality across the evaluated benchmarks (Figure 1b). This gap motivates target selection: the model must learn which of its plausible compositions best reflects the annotated preference. Test-time search would pass exploration costs to the user, while fixed targets leave the model's current alternatives unexamined. This raises a central question: *how can a compact model learn from its own alternatives to make a stronger first-answer cropping decision for the same input photograph?*

We propose FrameLite, a direct aesthetic cropping model built on **Qwen3.5-0.8B** (Qwen Team, 2026). Spatial and composition training connects localization, candidate-quality learning, and complete-frame supervision. **Quality-Guided Online Supervision (QOS)** then learns the highest-reward answer among the policy's greedy and sampled crops, adapting reward-selected learning (Dong et al., 2023) to its evolving outputs. We further propose **Greedy-Referenced Optimization (GRO)**, our crop-specific use of self-critical optimization (Rennie et al., 2017), to learn the quality differences between sampled crops and the current greedy answer. Training thus moves from learning what to frame, to finding better answers, to refining their likelihoods. Deployment retains the same compact model and a single coordinate answer, with no candidate selection.

Across public cropping benchmarks, FrameLite delivers leading performance among the compared VLM croppers with a fraction of their nominal backbone size (Figure 1a). With the same number of generated training candidates, learning the best answers yields a stronger final model than learning random ones, even after the same subsequent GRO training. These results connect compact models with accurate direct framing for lightweight photography assistance.

Our contributions connect compact crop prediction with progressive decision learning:

- **A compact aesthetic cropping model.** We propose FrameLite, a sub-billion-parameter vision–language model that combines spatial understanding and composition learning to predict a complete crop directly from the input image.
- **Quality-Guided Online Supervision (QOS).** We learn refreshed, high-quality complete answers from the current policy, connecting annotated preferences to its generated alternatives.
- **Progressive refinement with GRO.** We couple winner imitation with signed, greedy-referenced quality feedback to refine the policy within the same compact architecture.

2 Related Work

2.1 Aesthetic Cropping and Photographic Assistance

Learning-based cropping models compositional preferences through candidate scoring, ranking, and direct prediction. VFN, VEN, and VPN learn view evaluation and proposals with pairwise supervision (Chen et al., 2017b; Wei et al., 2018), while grid anchors reduce the candidate space (Zeng et al., 2019; 2022). Attention and subject-aware models guide crop generation and refinement (Wang & Shen, 2017; Yang et al., 2023). ASM-Net, CACNet, HCIC, and S^2CNet model salient-region placement, composition rules, human-relative structure, and object relationships (Tu et al., 2020; Hong et al., 2021; Zhang et al., 2022; Su et al., 2024). CGS and TransView capture relations among candidates and across crop boundaries (Li et al., 2020; Pan et al., 2021); set prediction accommodates diverse compositions (Jia et al., 2022), and spatial-aware features with rank consistency extend comparative supervision to unlabeled images (Wang et al., 2023). Feature extrapolation, outpainting supervision, professional-composition retrieval, and global feature alignment provide complementary sources of visual context, candidate comparisons, and photographic composition knowledge (Liu et al., 2023; Hong et al., 2024; Zhang et al., 2026; Yang et al., 2026).

Vision–language methods connect photographic quality with semantic intent. VILA learns aesthetic representations from image–comment pairs (Ke et al., 2023), while AesExpert and PhotoEye support aesthetic instruction following and photographic critique (Huang et al., 2024; Qi et al., 2025). ClipCrop uses CLIP representations for text- or image-conditioned cropping (Radford et al., 2021; Zhong et al., 2022); InstructCrop and Venus learn cropping instructions and rationale-supported prediction (Sheng et al., 2025; Du et al., 2026). Cropper combines retrieved examples with iterative refinement (Lee et al., 2025), whereas CROP aligns compositional reasoning with expert preferences (Dong et al., 2026). Beyond crop prediction, smart point-and-shoot photography offers camera-pose guidance (Li et al., 2025), and PhotoFramer and ShutterMuse provide multimodal composition instructions and capture-time assistance (You et al., 2026; Li et al., 2026).

2.2 Reward-Guided Learning and Policy Optimization

Reward-guided learning uses generated outputs to construct training supervision. RAFT samples candidate responses, selects high-reward examples, and fine-tunes the model on the selected outputs (Dong et al., 2023). Cropping-specific reinforcement learning includes A2-RL and Fast A3RL, which learn sequential window adjustments under aesthetic feedback (Li et al., 2018; 2019). For sequence generation, self-critical sequence training (SCST) uses the model's own inference output as a reward baseline and updates sampled sequences according to their relative rewards (Rennie et al., 2017). These approaches provide distinct ways to exploit outcome quality, through selected-output imitation or reward-weighted updates to the learned policy.

3 Method

FrameLite uses three-stage decision learning to locate important content, assess alternative framings, and commit to one answer. **Spatial and composition training** first learns from annotated regions, crop preferences, and complete reference frames (Figure 2). **Quality-Guided Online Supervision (QOS)** then selects a high-quality answer from the current policy's greedy and sampled crops (Figure 3(a)). **Greedy-Referenced Optimization (GRO)** uses each new sample's quality relative to the current greedy answer (Figure 3(b)). All three stages improve the same compact policy for direct crop prediction from a single greedy coordinate answer at inference.

3.1 Formulation and Model

A crop's quality depends on the joint placement of its boundaries. We represent it as a normalized box in $\mathcal{B} = \{(x_1, y_1, x_2, y_2) : 0 \le x_1 < x_2 \le 1,\ 0 \le y_1 < y_2 \le 1\}$. Given image I and fixed prompt p, the policy generates a coordinate sequence y:

$$\pi_\theta(y \mid I, p) = \prod_{t=1}^{T(y)} \pi_\theta(y_t \mid I, p, y_{<t}). \tag{1}$$

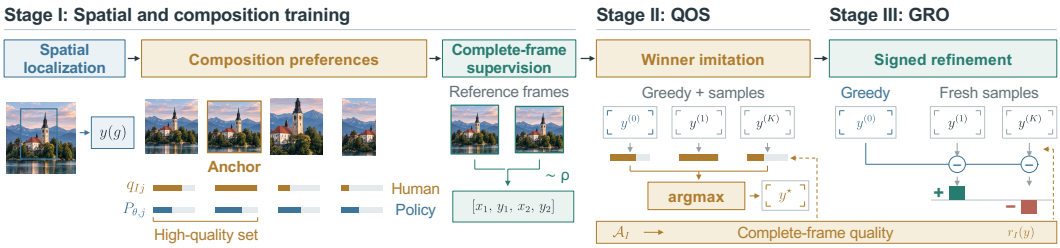

Figure 2: **Three-stage training of FrameLite.** For annotated crop j of image I, q_{Ij} and $P_{\theta,j}$ are human and policy preference probabilities; ρ gives reference-sampling weights. $y(g)$ serializes region g as normalized corner coordinates (x_1, y_1, x_2, y_2). QOS imitates the best answer $y^\star$ among greedy $y^{(0)}$ and K samples $y^{(1:K)}$; GRO refines fresh samples relative to greedy. Both use quality $r_I(y)$ defined by human crop annotations $\mathcal{A}_I$ associated with each training image.

A parser $\mathcal{P}$ maps y to a complete box or invalid symbol $\perp$. We use Qwen3.5-0.8B (Qwen Team, 2026) with a frozen visual encoder. Stage I adapts its visual merger and language LoRA (Hu et al., 2022); subsequent stages update language adapters. The training distribution $\mathcal{D}$ contains images and crop annotations $\mathcal{A}_I$, with the chosen example weights. Annotations provide supervision and quality feedback, without being supplied to the prediction model.

3.2 Stage I: Spatial and Composition Training

Spatial localization. A subject's position and extent organize the surrounding composition. We therefore first teach the policy to locate important content through referring-expression localization (Yu et al., 2016). For an expression e and annotated region g, let $y(g)$ serialize its coordinates. The loss averages negative log likelihood over valid target tokens in a training batch:

$$\mathcal{L}_{\text{loc}} = -\frac{1}{|\mathcal{T}_{\text{loc}}|} \sum_{(i,t)\in\mathcal{T}_{\text{loc}}} \log \pi_\theta(y(g_i)_t \mid I_i, e_i, y(g_i)_{<t}), \tag{2}$$

where $\mathcal{T}_{\text{loc}}$ indexes those tokens. The region target in Figure 2 teaches spatial grounding; it does not prescribe a tight subject box as the final aesthetic crop. The remaining training steps learn how to place the final frame around the located subject and its surrounding context.

Composition preferences. Several windows may preserve the same subject while differing in context, balance, and empty space; learning one reference leaves their relative quality unused. Given rated annotations $\mathcal{A}_I = \{(g_j, u_j)\}_{j=1}^{M_I}$, we therefore align the policy's preferences over complete answers with human scores. Let $z_{\theta,j} = \log \pi_\theta(y(g_j) \mid I, p)$ and define the two preference distributions over the annotated candidate frames of the same input image:

$$q_{Ij} = \frac{\exp(\widetilde{u}_j/\tau)}{\sum_k \exp(\widetilde{u}_k/\tau)}, \qquad P_{\theta,j} = \frac{\exp(z_{\theta,j})}{\sum_k \exp(z_{\theta,k})}, \tag{3}$$

where $\widetilde{u}_j$ is a normalized annotation score and τ controls the concentration of the target distribution. The model probabilities are normalized over the full annotated candidate list so that the relative likelihoods of alternative framings can be compared directly.

A good target should provide a concrete answer without discarding the fact that several answers may be acceptable. We select an anchor g_a as a quality-weighted IoU medoid within the high-quality subset $\mathcal{H}_I$, and combine three complementary signals:

$$\mathcal{L}_{\text{comp}} = \mathbb{E}_I \left[\underbrace{-z_{\theta,a}}_{\text{anchor}} \underbrace{-\sum_j q_{Ij} \log P_{\theta,j}}_{\text{relative quality}} \underbrace{-\log\left(\sum_{j\in\mathcal{H}_I} P_{\theta,j}\right)}_{\text{high-quality set}} \right]. \tag{4}$$

The anchor supplies a concrete answer, the listwise term aligns quality preferences, and the set term favors multiple acceptable compositions without fixing their probability allocation. Together, these signals adapt complete-answer learning to the ambiguity of aesthetic framing.

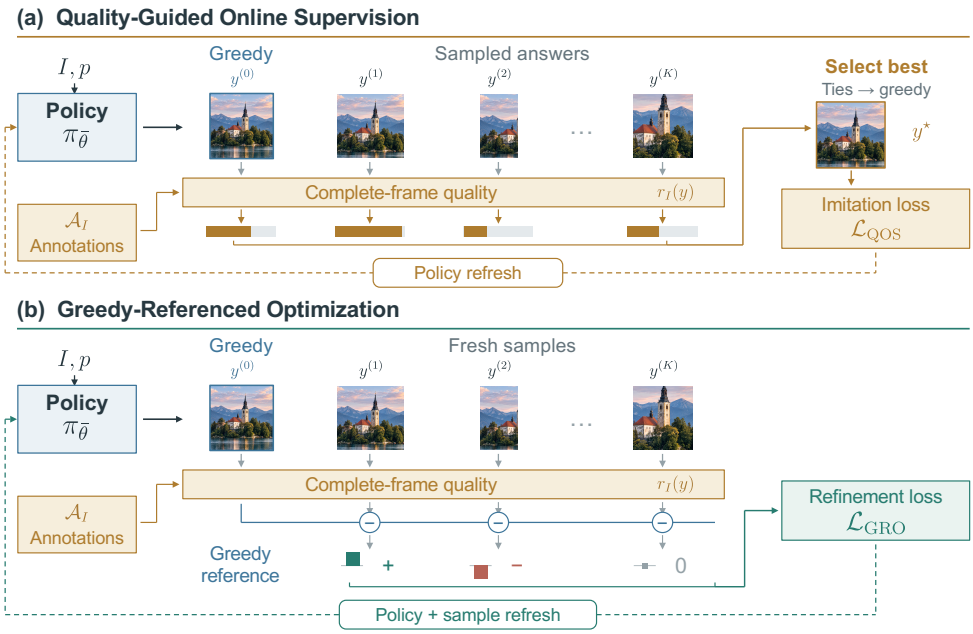

Figure 3: **Online supervision and refinement.** Given image I and prompt p, policy $\pi_{\bar{\theta}}$ with current parameters $\bar{\theta}$ generates greedy answer $y^{(0)}$ and independent samples $y^{(j)}$, $j = 1, \ldots, K$. Annotations $\mathcal{A}_I$ define quality $r_I(y)$. (a) $\mathcal{L}_{\mathrm{QOS}}$ increases the likelihood of the best answer $y^\star$, favoring greedy on ties. (b) $\mathcal{L}_{\mathrm{GRO}}$ weights sampled answers' log-likelihoods by signed quality differences from greedy. Dashed loops refresh the policy and candidates after each update.

Complete-frame supervision. Relative preferences alone specify which annotated crops should rank higher; direct prediction also requires assigning likelihood to full coordinate answers. We next consolidate this behavior using the available reference frames and their fixed annotation-derived weights $\rho_{Ij} \geq 0$, with $\sum_j \rho_{Ij} = 1$ for each training image:

$$\mathcal{L}_{\mathrm{ann}} = -\mathbb{E}_{(I,\mathcal{A}_I)\sim\mathcal{D}} \left[\sum_j \rho_{Ij} \log \pi_\theta(y(g_j) \mid I, p) \right]. \tag{5}$$

A single reference receives unit weight; multiple references can express quality or acceptance preferences. Sampling complete references under these weights preserves alternative framings, instead of creating a new box by averaging their boundaries. This supervision also accommodates images with a final crop but no candidate ranking. Spatial localization, preference learning, and full-answer supervision are successive parts of Stage I, yielding the initial policy $\pi_{\theta_{\mathrm{init}}}$.

3.3 Stage II: Quality-Guided Online Supervision (QOS)

Annotated frames do not reveal which alternatives a policy generates when making its own decisions. We propose Quality-Guided Online Supervision (QOS) to bridge this gap by selecting a complete answer from the current policy under an annotation-defined quality criterion. Refreshed during training, the target connects the model's current outputs to human framing preferences.

To compare complete frames, let $R(I, b; \mathcal{A}_I)$ evaluate a crop using its training annotations. A single accepted frame supports endpoint agreement; several acceptable frames support best-reference matching; and rated alternatives support quality-weighted agreement. We assign a sequence reward after its complete coordinates have been generated:

$$r_I(y) = \begin{cases} R(I, \mathcal{P}(y); \mathcal{A}_I), & \mathcal{P}(y) \in \mathcal{B}, \\ r_{\mathrm{invalid}}, & \mathcal{P}(y) = \bot, \end{cases} \tag{6}$$

where r_{invalid} is below any valid reward. The same fixed criterion guides both online stages, so target selection and policy refinement optimize a consistent notion of complete-frame quality.

At the start of an update, let $\bar{\theta}$ denote the current parameters. The policy generates its greedy answer and K independent alternatives, then selects the highest-reward complete answer:

$$y^{(0)} = \text{Greedy}(\pi_{\bar{\theta}}, I, p), \qquad y^{(j)} \sim \pi_{\bar{\theta}}(\cdot \mid I, p),$$
$$j^\star = \arg\max_{j \in \{0,\dots,K\}} r_I(y^{(j)}), \qquad y^\star = y^{(j^\star)}. \tag{7}$$

Including the greedy answer gives the selection a natural fallback: $r_I(y^\star) \geq r_I(y^{(0)})$ within each generated set. Ties favor the greedy answer. Thus, exploration supplies alternatives while the quality criterion decides whether any is a better target than the current first response. This relation concerns target selection, not a guarantee that every parameter update improves greedy quality.

QOS then increases the likelihood of the selected answer:

$$\mathcal{L}_{\text{QOS}}(\theta; \bar{\theta}) = -\mathbb{E}_{(I,\mathcal{A}_I)\sim\mathcal{D}} \mathbb{E}_{y^{(1:K)}\sim\pi_{\bar{\theta}}} \left[\log \pi_\theta(y^\star \mid I, p)\right]. \tag{8}$$

Candidates and selection are held fixed during differentiation, with the update evaluated at $\theta = \bar{\theta}$. The updated policy supplies the next batch of candidates. This use of reward-selected learning (Dong et al., 2023) couples target discovery to the policy's actual decisions, while human annotations keep selection tied to composition quality. Each image contributes one selected complete answer; QOS does not explicitly encode how much the unselected answers improve on or fall below the current prediction for the same input image at the current stage of policy learning.

3.4 Stage III: Greedy-Referenced Optimization (GRO)

Greedy-Referenced Optimization (GRO) extends winner imitation to feedback on all sampled frames, evaluating each complete framing decision relative to the answer the policy would currently return.

A crop's absolute score does not indicate improvement over the current decision. The greedy answer provides an image-specific, policy-dependent reference. For fresh samples from the current policy, we define a signed quality advantage for each sampled coordinate answer:

$$A_j = r_I(y^{(j)}) - r_I(y^{(0)}), \qquad j = 1, \dots, K. \tag{9}$$

A positive advantage identifies a sampled frame that is better under the training criterion; a negative advantage identifies an inferior alternative. Because the reference is recomputed from the current policy, the comparison follows its evolving framing ability rather than a fixed historical answer.

We use the self-critical sequence-training estimator (Rennie et al., 2017) to optimize the quality of complete coordinate answers under this image-specific reference:

$$\mathcal{L}_{\text{GRO}}(\theta; \bar{\theta}) = -\mathbb{E}_{(I,\mathcal{A}_I)\sim\mathcal{D}} \mathbb{E}_{y^{(1:K)}\sim\pi_{\bar{\theta}}} \left[\frac{1}{K} \sum_{j=1}^{K} \text{sg}(A_j) \log \pi_\theta(y^{(j)} \mid I, p)\right], \tag{10}$$

where sg stops gradients through the reward difference and the update is evaluated at $\theta = \bar{\theta}$. Unlike a target-selection rule, this objective retains the sign and magnitude of each sampled answer's relative quality. Its loss terms encourage higher-quality sequences and discourage inferior ones. The reward is assigned to the joint framing decision, while the full sequence probability propagates that feedback through the coordinate answer. The greedy output serves as the reference and receives no separate imitation loss for its coordinate sequence in this stage.

QOS learns high-quality answers already reachable by the policy; GRO then uses fresh alternatives to refine their relative likelihoods. The transition transfers parameters, not a fixed candidate pool. A shared quality function connects target discovery with refinement of the first answer.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate on GAICv2 (Zeng et al., 2022), FLMS (Fang et al., 2014), and FCDB (Chen et al., 2017a), with 500, 500, and 345 test images, respectively. These benchmarks offer complementary annotation formats: GAICv2 rates multiple candidate windows, FLMS provides several reference crops, and FCDB specifies one reference crop per image. Spatial training uses RefCOCO and RefCOCO+ (Yu et al., 2016); crop training uses GAIC, CPC (Wei et al., 2018), and official FCDB training images. FLMS is excluded from crop training, allowing us to examine whether the learned framing preferences transfer beyond the supervised data sources.

Table 1: **Comparison with representative cropping methods.** Acc denotes GAICv2 $\mathrm{Acc}_{1/5}$. BDE penalizes invalid predictions with 1. Bold/underline mark the best/second-best results for each metric across the displayed methods evaluated under the same benchmark protocol.

Method	Venue	GAICv2		FLMS		FCDB	
		$\mathrm{Acc}^{.85}\uparrow$	$\mathrm{Acc}^{.90}\uparrow$	IoU↑	BDE↓	IoU↑	BDE↓
Traditional methods							
VFN	MM'17	0.5500	0.2640	0.7976	0.0500	0.6433	0.0948
VEN	CVPR'18	0.0340	0.0120	0.7075	0.0619	0.5863	0.1057
VPN	CVPR'18	0.0380	0.0080	0.8234	0.0399	0.6574	0.0878
GAICD	TPAMI'22	0.8300	<u>0.7640</u>	0.8025	0.0470	0.6759	0.0819
CACNet	CVPR'21	0.5100	0.2360	0.8401	0.0368	<u>0.6967</u>	<u>0.0751</u>
VLM-based methods							
Cropper	CVPR'25	0.6900	0.6460	0.7845	0.0829	0.6519	0.1089
InstructCrop	MM'25	0.4740	0.1680	0.8145	0.0422	0.6848	0.0774
Venus	CVPR'26	<u>0.8380</u>	0.7200	<u>0.8638</u>	<u>0.0304</u>	0.6836	0.0794
FrameLite	Ours	**0.9140**	**0.8340**	**0.8679**	**0.0297**	**0.7124**	**0.0704**

Evaluation protocol. We consider automatic free-form cropping: each method returns one crop without receiving a reference aspect ratio. On GAICv2, we report $\mathrm{Acc}_{1/5}$ at IoU thresholds 0.85 and 0.90. A prediction is accepted if its IoU with any top-five rated reference, including cutoff ties, reaches the threshold. FLMS measures the maximum IoU between that single prediction and its human references; FCDB uses the unique reference. We also report boundary displacement error (BDE), the mean absolute deviation of the four normalized coordinates. FLMS BDE uses the same reference selected by maximum IoU, so overlap and boundary error describe agreement with one matched frame. Scores are averaged over the full test set. Acceptance evaluates agreement with highly rated candidates, while IoU and BDE quantify retained-area overlap and boundary placement.

Implementation details. FrameLite uses Qwen3.5-0.8B with a frozen visual encoder. Stage I adapts the visual merger and language LoRA; QOS and GRO update only the language adapters. Both online stages sample four alternatives to the current greedy answer and share an annotation-defined reward. We use AdamW (Loshchilov & Hutter, 2019) with an image batch size of 32. The final model receives one pass of QOS followed by one pass of GRO; their learning-rate schedules use base rates of 2×10^{-6} and 10^{-6}, respectively. The online adapters use LoRA rank 64 and $\alpha = 128$, and preprocessing preserves image aspect ratios. All benchmarks use the same final checkpoint, fixed crop prompt, and greedy coordinate decoding, with no candidate search or reranking.

4.2 Comparison with Existing Methods

We compare with VFN (Chen et al., 2017b), VEN and VPN (Wei et al., 2018), GAICD (Zeng et al., 2019; 2022), CACNet (Hong et al., 2021), Cropper (Lee et al., 2025), InstructCrop (Sheng et al., 2025), and Venus (Du et al., 2026). These methods span candidate scoring, composition-aware prediction, and retrieval-augmented or direct VLM cropping.

Quantitative results. The same FrameLite model leads the displayed methods in acceptance, overlap, and boundary accuracy (Table 1). GAICv2 acceptance improves by 7.6 and 7.0 percentage points over the strongest listed competitors. Relative to Venus on FLMS, BDE decreases by 2.4%, alongside higher IoU. The BDE ranking also holds on the common valid-image sets, indicating closer boundary placement beyond the effect of invalid predictions.

Qualitative results. In Figure 4(a), FrameLite preserves the curved roofline while reducing foreground rocks. In (b), it balances the mountain range and its reflection, removing excess sky and open water. In (c), it retains the pedestrian's full body and feet, which several competing crops truncate. These examples illustrate content-dependent framing: reducing peripheral regions for architecture and landscapes while retaining the spatial extent needed for an intact human subject.

4.3 Ablation Studies

We examine Stage-I initialization and the gains from QOS and GRO, then compare matched objectives to assess quality-based target selection and relative-quality refinement.

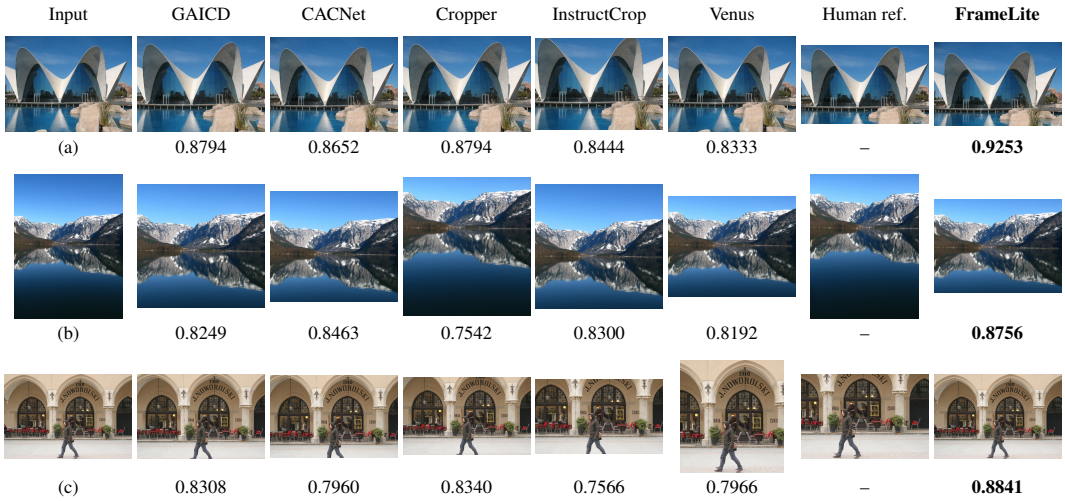

Figure 4: **Qualitative comparison.** Scores below predictions are IoU. For multi-reference evaluation, the displayed human reference may differ from each prediction's matched reference.

Table 2: **Contributions of the three training stages.** The upper block compares online paths from full Stage I; the lower block varies Stage-I components, followed by matched complete-frame supervision, QOS, and GRO. Each online stage uses one pass. Bold is determined before rounding.

Training configuration	GAICv2		FLMS		FCDB	
	$\text{Acc}^{\cdot 85}\uparrow$	$\text{Acc}^{\cdot 90}\uparrow$	IoU↑	BDE↓	IoU↑	BDE↓
Stage I	0.9040	0.7840	0.8512	0.0338	0.7013	0.0745
Stage I → GRO	0.9060	**0.8340**	0.8635	0.0302	0.7110	0.0704
Stage I → QOS	**0.9140**	0.8320	0.8673	0.0299	0.7069	0.0722
Stage I → QOS → GRO	**0.9140**	**0.8340**	**0.8679**	**0.0297**	**0.7124**	**0.0704**
Localization omitted	0.8800	0.8120	0.8553	0.0320	0.7096	0.0717
Anchor supervision only	0.9060	0.8300	0.8627	0.0307	0.7055	0.0729

Spatial and composition training. One variant omits localization; the other retains it but uses only anchor supervision for composition learning, jointly removing the listwise and high-quality-set terms. Both receive matched complete-frame supervision, QOS, and GRO. Localization omission lowers GAICv2 acceptance by 3.4/2.2 percentage points; anchor-only supervision lowers it by 0.8/0.4 points and reduces FCDB IoU by 0.0069. Both variants underperform the full model on all six final metrics (Table 2). These final gaps remain despite later training, supporting spatial grounding and multi-candidate preferences. Stagewise results appear in Appendix B.1.

Contributions of QOS and GRO. Starting from the full Stage-I model, either online stage improves direct prediction without changing the architecture or decoding (Table 2). QOS lowers FLMS BDE from 0.0338 to 0.0299, improving boundary placement alongside overlap. QOS alone performs better on FLMS, whereas GRO alone gives stronger FCDB agreement. Applying QOS followed by GRO achieves the best IoU on both datasets. These gains motivate comparing winner imitation with relative-quality refinement under matched training conditions.

Quality-based target selection. To separate exploration from target selection, we replace the QOS winner with a uniform random candidate while preserving the sampling budget and online refreshing, then apply the same GRO schedule. QOS yields lower BDE at the final endpoint: 0.0297 versus 0.0334 on FLMS, and 0.0704 versus 0.0762 on FCDB (Table 3a). Its benefit extends beyond crossing acceptance thresholds to more accurate boundary placement. QOS also improves all six final metrics over fixed-reference supervision under the same GRO refinement.

Fixed references express human preferences, while random targets follow the current policy without favoring better framing. QOS combines both through quality-based selection. Random supervision produces only 24 distinct FLMS boxes across 500 images, all sharing the same left boundary and the same right boundary. GRO increases this count to 81 but does not recover QOS's quality. This

Table 3: **Target selection and policy refinement.** (a) Models trained on different supervision targets undergo the same GRO refinement; random selection and QOS use equal candidate budgets. (b) Different update rules refine a common QOS model. Image exposure and learning-rate schedules are matched within each block. Bold marks the best values before rounding.

Training objective	GAICv2		FLMS		FCDB	
	$\text{Acc}^{.85}\uparrow$	$\text{Acc}^{.90}\uparrow$	IoU↑	BDE↓	IoU↑	BDE↓
(a) Target selection before policy refinement						
Reference targets → GRO	0.8920	0.8300	0.8580	0.0324	0.7109	0.0719
Random targets → GRO	0.8560	0.8200	0.8554	0.0334	0.6997	0.0762
QOS → GRO	**0.9140**	**0.8340**	**0.8679**	**0.0297**	**0.7124**	**0.0704**
(b) Policy refinement after QOS						
Continue QOS	0.9100	0.8260	0.8657	0.0301	0.7067	0.0721
Positive-only GRO	0.9100	0.8120	0.8661	**0.0297**	0.7085	0.0709
Full GRO	**0.9140**	**0.8340**	**0.8679**	0.0297	**0.7124**	**0.0704**

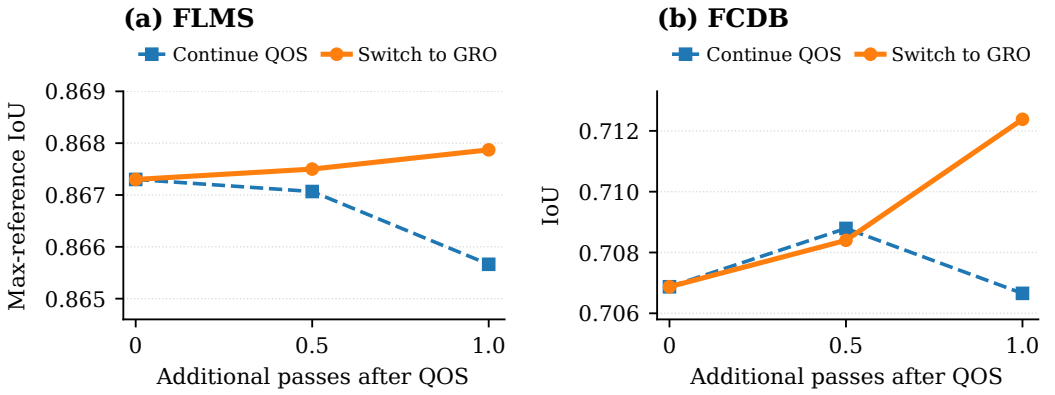

Figure 5: **Refining the QOS policy.** Both curves start from the same QOS model with matched image exposure and learning-rate schedules. Markers show initial, half-pass, and one-pass evaluations; lines connect measured greedy-output scores. The two panels use different vertical ranges.

suggests that preferring better generated answers shapes the candidates available for later policy refinement under the same annotation-defined quality criterion.

Relative-quality refinement. From the same QOS model, GRO increases FLMS IoU by about 0.0006 and FCDB IoU by 0.0055, while reducing BDE on both datasets. Continued QOS instead lowers final FLMS quality and loses its intermediate FCDB gain (Figure 5). These results support switching from winner imitation to signed quality feedback after QOS has learned to favor high-quality complete answers among its own generated cropping alternatives.

Clipping advantages to $\max(A_j, 0)$ removes feedback from worse answers. Full GRO gives higher acceptance and IoU, and lower FCDB BDE, while positive-only GRO has marginally lower FLMS BDE before rounding (Table 3b). Because clipping also changes update magnitude, these results favor the complete rule without establishing negative feedback as independently necessary.

5 Conclusion

We propose FrameLite, a compact vision–language model for direct aesthetic cropping. Spatial and composition training builds its initial capabilities, QOS identifies high-quality answers to imitate, and GRO refines their quality relative to the current greedy prediction. Experiments show strong cropping performance with a sub-billion-parameter backbone. Controlled comparisons support quality selection beyond online sampling alone and the subsequent use of relative-quality feedback. This training process strengthens direct decisions without adding search or scoring at deployment. Personalized preferences and broader photographic intentions offer promising extensions for adapting crop decisions to individual users and diverse capture contexts.

AI USE STATEMENT

In this work, we used generative AI tools for assist with translation. We have not used generative AI tools for things like generating data and supporting qualitative and thematic data analysis, and the rest of the required disclosure tasks like mathematical proofs are not applicable to this work. Additionally, we used generative AI tools for polishing our language. We have reviewed all AI-assisted work. We checked LLM-generated research ideas for potential plagiarism through a manual literature survey. LLM-generated code was verified and tested for correctness by all of the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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 *WACV*, 2017a. 6, 13

Yi-Ling Chen, Jan Klopp, Min Sun, Shao-Yi Chien, and Kwan-Liu Ma. Learning to Compose with Professional Photographs on the Web. In *ACM MM*, 2017b. 1, 3, 7

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment. *Transactions on Machine Learning Research*, 2023. 2, 3, 6, 14

Zhitong Dong, Chao Li, Jie Yu, and Hao Chen. CROP: Expert-Aligned Image Cropping via Compositional Reasoning and Optimizing Preference. arXiv preprint arXiv:2605.12545, 2026. 1, 3

Tianxiang Du, Hulingxiao He, and Yuxin Peng. Venus: Benchmarking and Empowering Multimodal Large Language Models for Aesthetic Guidance and Cropping. In *CVPR*, 2026. 1, 3, 7

Chen Fang, Zhe Lin, Radomír Měch, and Xiaohui Shen. Automatic Image Cropping Using Visual Composition, Boundary Simplicity and Content Preservation Models. In *ACM MM*, 2014. 6, 13

Chaoyi Hong, Shuaiyuan Du, Ke Xian, Hao Lu, Zhiguo Cao, and Weicai Zhong. Composing Photos Like a Photographer. In *CVPR*, 2021. 1, 3, 7

James Hong, Lu Yuan, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. Learning Subject-Aware Cropping by Outpainting Professional Photos. In *AAAI*, 2024. 1, 3

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*, 2022. 4, 13

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 *ACM MM*, 2024. 3

Gengyun Jia, Huaibo Huang, Chaoyou Fu, and Ran He. Rethinking Image Cropping: Exploring Diverse Compositions from Global Views. In *CVPR*, 2022. 1, 3

Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. VILA: Learning Image Aesthetics from User Comments with Vision-Language Pretraining. In *CVPR*, 2023. 3

Seung Hyun Lee, Jijun Jiang, Yiran Xu, Zhuofang Li, Junjie Ke, Yinxiao Li, Junfeng He, Steven Hickson, Katie Datsenko, Sangpil Kim, Ming-Hsuan Yang, Irfan Essa, and Feng Yang. Cropper: Vision-Language Model for Image Cropping through In-Context Learning. In *CVPR*, 2025. 1, 3, 7

Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. A2-RL: Aesthetics Aware Reinforcement Learning for Image Cropping. In *CVPR*, 2018. 3

Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. Fast A3RL: Aesthetics-Aware Adversarial Reinforcement Learning for Image Cropping. *IEEE Transactions on Image Processing*, 2019. 3

Debang Li, Junge Zhang, Kaiqi Huang, and Ming-Hsuan Yang. Composing Good Shots by Exploiting Mutual Relations. In *CVPR*, 2020. 3

Jiawan Li, Fei Zhou, Zhipeng Zhong, Jiongzhi Lin, and Guoping Qiu. Towards Smart Point-and-Shoot Photography. In *CVPR*, 2025. 1, 3

Jiayu Li, Yixiao Fang, Tianyu Hu, Wei Cheng, Ping Huang, Zheheng Fan, Gang Yu, and Xingjun Ma. ShutterMuse: Capture-Time Photography Guidance with MLLMs. arXiv preprint arXiv:2606.25763, 2026. 1, 3

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 *ICCV*, 2023. 1, 3

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019. 7, 13

Zhiyu Pan, Zhiguo Cao, Kewei Wang, Hao Lu, and Weicai Zhong. TransView: Inside, Outside, and Across the Cropping View Boundaries. In *ICCV*, 2021. 3

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 *CVPR*, 2025. 3

Qwen Team. Qwen3.5: Towards Native Multimodal Agents. Official model release, 2026. 2, 4, 13

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models from Natural Language Supervision. In *ICML*, 2021. 3

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-Critical Sequence Training for Image Captioning. In *CVPR*, 2017. 2, 3, 6, 14

Xiangfei Sheng. InstructCrop: Official Implementation. GitHub repository, n.d. 1

Xiangfei Sheng, Pangu Xie, Weidong Zou, Pengfei Chen, Tong Zhu, and Leida Li. InstructCrop: Teaching Multimodal Large Language Models to Crop Aesthetic Images. In *ACM MM*, 2025. 1, 3, 7

Yukun Su, Yiwen Cao, Jingliang Deng, Fengyun Rao, and Qingyao Wu. Spatial-Semantic Collaborative Cropping for User Generated Content. In *AAAI*, 2024. 1, 3

Yi Tu, Li Niu, Weijie Zhao, Dawei Cheng, and Liqing Zhang. Image Cropping with Composition and Saliency Aware Aesthetic Score Map. In *AAAI*, 2020. 3

Chao Wang, Li Niu, Bo Zhang, and Liqing Zhang. Image Cropping with Spatial-Aware Feature and Rank Consistency. In *CVPR*, 2023. 3

Wenguan Wang and Jianbing Shen. Deep Cropping via Attention Box Prediction and Aesthetics Assessment. In *ICCV*, 2017. 3

Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomír Mech, Minh Hoai, and Dimitris Samaras. Good View Hunting: Learning Photo Composition from Dense View Pairs. In *CVPR*, 2018. 1, 3, 6, 7, 13

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*, 2023. 3

Haotian Yang, Zhile Yang, Kin-Man Lam, Patrick Le Callet, and Xin Sun. Global Attention-Fused Image Cropping with Attention-Guided and Global-Aligned Crop Evaluator. arXiv preprint arXiv:2608.04821, 2026. 1, 3

Zhiyuan You, Ke Wang, He Zhang, Xin Cai, Jinjin Gu, Tianfan Xue, Chao Dong, and Zhoutong Zhang. PhotoFramer: Multi-modal Image Composition Instruction. In *CVPR*, 2026. 1, 3

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling Context in Referring Expressions. In *ECCV*, 2016. 4, 6, 13

Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Reliable and Efficient Image Cropping: A Grid Anchor Based Approach. In *CVPR*, 2019. 1, 3, 7

Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Grid Anchor Based Image Cropping: A New Benchmark and an Efficient Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 3, 6, 7, 13

Bo Zhang, Li Niu, Xing Zhao, and Liqing Zhang. Human-Centric Image Cropping with Partition-Aware and Content-Preserving Features. In *ECCV*, 2022. 1, 3

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 *AAAI*, 2026. 1, 3

Zhihang Zhong, Mingxi Cheng, Zhirong Wu, Yuhui Yuan, Yinqiang Zheng, Ji Li, Han Hu, Stephen Lin, Yoichi Sato, and Imari Sato. ClipCrop: Conditioned Cropping Driven by Vision-Language Model. arXiv preprint arXiv:2211.11492, 2022. 1, 3

A Method and Training Details

A.1 Training Data and Optimization

FrameLite uses Qwen3.5-0.8B (Qwen Team, 2026) with a frozen visual encoder. Stage I successively applies spatial localization, composition preferences, and complete-frame supervision, updating the visual merger and language LoRA (Hu et al., 2022). Quality-Guided Online Supervision (QOS) and Greedy-Referenced Optimization (GRO) subsequently update only language LoRA.

Localization uses 240,815 RefCOCO/RefCOCO+ expression–box annotations over 16,994 images (Yu et al., 2016). Composition learning uses 2,636 GAIC (Zeng et al., 2022) and 8,638 CPC (Wei et al., 2018) images, with equal weight per image. Complete-frame supervision and the online stages use 2,536 GAIC, 8,538 CPC, and 1,369 FCDB training images (Chen et al., 2017a), with source weights $(\lambda_G, \lambda_C, \lambda_F) = (0.25, 0.25, 0.50)$. FLMS (Fang et al., 2014) is used only for evaluation. For these three weighted stages, an image from source d receives $w_d = N\lambda_d/N_d$, where N_d is its source size and $N = \sum_d N_d$. The batch objective is $B^{-1}\sum_{i=1}^{B} w_{d_i}\ell(I_i)$; thus λ_d specifies the source contribution rather than a per-image multiplier independent of source size.

Table 4: Training schedule. A pass traverses the corresponding training set. Composition learning and QOS use four-pass and two-pass cosine decay schedules, respectively, and retain the model after one pass.

Objective	Training data	Passes	Base learning rate
Spatial localization	RefCOCO / RefCOCO+	1	5×10^{-5}
Composition preferences	GAIC / CPC	1	3×10^{-5}
Complete-frame supervision	GAIC / CPC / FCDB	2	3×10^{-5}
QOS	GAIC / CPC / FCDB	1	2×10^{-6}
GRO	GAIC / CPC / FCDB	1	10^{-6}

QOS and GRO use separate AdamW optimizers (Loshchilov & Hutter, 2019), 5% warmup, an image batch size of 32, and LoRA rank 64 with $\alpha = 128$. LoRA dropout is 0.05 during composition learning and zero during online training. Table 4 summarizes the learning rates and training durations.

A.2 Reference Frames and Composition Preferences

Preference distribution and anchor. For M_I annotated candidates, scores are min–max normalized within each image; all-equal scores are mapped to zero. We use temperature $\tau = 0.15$ for $q_{Ij} = \mathrm{softmax}(\widetilde{u}/\tau)_j$. The high-quality set $\mathcal{H}_I$ contains the top $\lceil 0.2M_I \rceil$ candidates and all cutoff ties. Its quality-weighted IoU medoid defines the anchor g_a. In particular, the medoid uses the annotation preference distribution:

$$\bar{q}_{Ij} = \frac{q_{Ij}}{\sum_{k\in\mathcal{H}_I} q_{Ik}}, \qquad a = \arg\max_{i\in\mathcal{H}_I} \sum_{j\in\mathcal{H}_I} \bar{q}_{Ij}\,\mathrm{IoU}(g_i, g_j). \tag{11}$$

The sum includes self-overlap, and ties are resolved deterministically. The normalization of $\bar{q}$ defines the reference weights for the medoid; the model probabilities remain normalized over the full candidate list. When all annotations have equal scores, q is uniform and $\mathcal{H}_I$ contains every candidate, making the acceptable-set term zero.

The anchor, listwise, and acceptable-set losses have equal coefficients. The model distribution $P_{\theta,j}$ is normalized over all annotated candidates, so the set term increases the total probability assigned to $\mathcal{H}_I$ without requiring a single preferred frame.

Complete-frame targets. The weights ρ_{Ij} sum to one within each image. GAIC assigns uniform weight to its top-five rated references and all cutoff ties. CPC applies a temperature-0.15 softmax to per-image min–max-normalized sums of annotation scores; FCDB uses its single reference. Sampling a complete frame under these weights preserves alternative valid compositions without averaging their coordinates. These within-image weights are distinct from the dataset weights in Section A.

A.3 Rewards and Online Learning

Let $\mathcal{G}_I^+$ contain the top-five GAIC references and all cutoff ties, and let $q_I(b) = \max_{g \in \mathcal{G}_I^+} \mathrm{IoU}(b, g)$. QOS and GRO share the following rewards for valid predictions:

$$R_G(I, b) = \tfrac{1}{2} q_I(b) + \tfrac{1}{4} \mathbb{I}[q_I(b) \geq 0.85] + \tfrac{1}{4} \mathbb{I}[q_I(b) \geq 0.90],$$
$$R_C(I, b) = \sum_j \rho_{Ij} \, \mathrm{IoU}(b, g_j), \qquad R_F(I, b) = \mathrm{IoU}(b, g_I). \tag{12}$$

GAIC combines overlap with threshold acceptance, CPC measures quality-weighted agreement, and FCDB measures single-reference agreement. Invalid answers receive -1. Only training annotations supply rewards; they are not inputs to the prediction model.

Each online stage generates the greedy answer and $K = 4$ independent samples at temperature 1, without top-k or nucleus truncation. QOS applies reward-selected supervision (Dong et al., 2023) to the highest-reward complete answer, favoring greedy on a tie. GRO uses the self-critical estimator (Rennie et al., 2017) and minimizes $-K^{-1} \sum_j \mathrm{sg}(A_j) \log \pi_\theta(y^{(j)} \mid I, p)$, where $A_j = r_I(y^{(j)}) - r_I(y^{(0)})$. Positive advantages encourage higher sequence likelihood, while negative advantages discourage inferior answers. Greedy provides the reward baseline and receives no additional imitation loss.

Candidates are refreshed after each update; rewards and selection are fixed during differentiation. Localization averages token losses, whereas complete crop answers use summed token log probabilities. Localization targets have a fixed length of 32 tokens including EOS, so averaging token losses per expression is equivalent to averaging over all valid target tokens. Online likelihoods include EOS only when emitted and use no length normalization. GRO uses neither advantage standardization nor an additional KL or supervised loss.

A.4 Crop Instruction and Coordinate Representation

Composition learning, complete-frame supervision, both online stages, and final prediction share the following crop instruction:

```
Select one aesthetically pleasing crop for this image.  Return only
the requested short structured output.
```

The image precedes this text in the model's chat template, with thinking mode disabled. The output convention is a single JSON object with a `crop` field containing (x_1, y_1, x_2, y_2). For an original image of width W and height H, the predicted pixel coordinates are (Wx_1, Hy_1, Wx_2, Hy_2). We use the official image processor with a pixel budget of 65,536–589,824. Inputs preserve their aspect ratios, and normalized coordinates refer to the original image canvas. The parser requires four finite values with $0 \leq x_1 < x_2 \leq 1$ and $0 \leq y_1 < y_2 \leq 1$; malformed outputs are rejected without coordinate repair or an additional prediction.

Human crop targets are quantized to increments of 0.001, with no requirement to pad trailing zeros. Online training uses the actual generated coordinate sequence, and inference does not impose a fixed number of decimal places. Greedy decoding produces one crop with a 32-token generation limit. A sequence is judged by its parsed content: reaching the token limit is not itself a geometric failure if a complete, valid answer has already been produced.

B Additional Ablation Studies

We examine whether spatial and composition training benefits the final crop policy after complete-frame supervision and both online stages, and clarify the controls used to distinguish the online objectives.

B.1 Contributions of Spatial and Composition Training

Design. *Localization omitted* removes the additional referring-expression training and starts composition learning from the pretrained backbone. *Anchor supervision only* retains localization but uses only complete-answer NLL for the same quality-weighted IoU medoid, jointly removing the listwise preference and high-quality-set probability terms. Each variant continues from its own parameters through complete-frame supervision, QOS, and GRO. The subsequent training images, source weights, image exposure, learning-rate schedules, and trainable modules match the full method (Section A). These comparisons assess the contributions remaining after later training can compensate for the modified initialization.

Table 2 reports the final results of these variants alongside the online-stage ablations. Figure 6 complements that comparison with measurements throughout training.

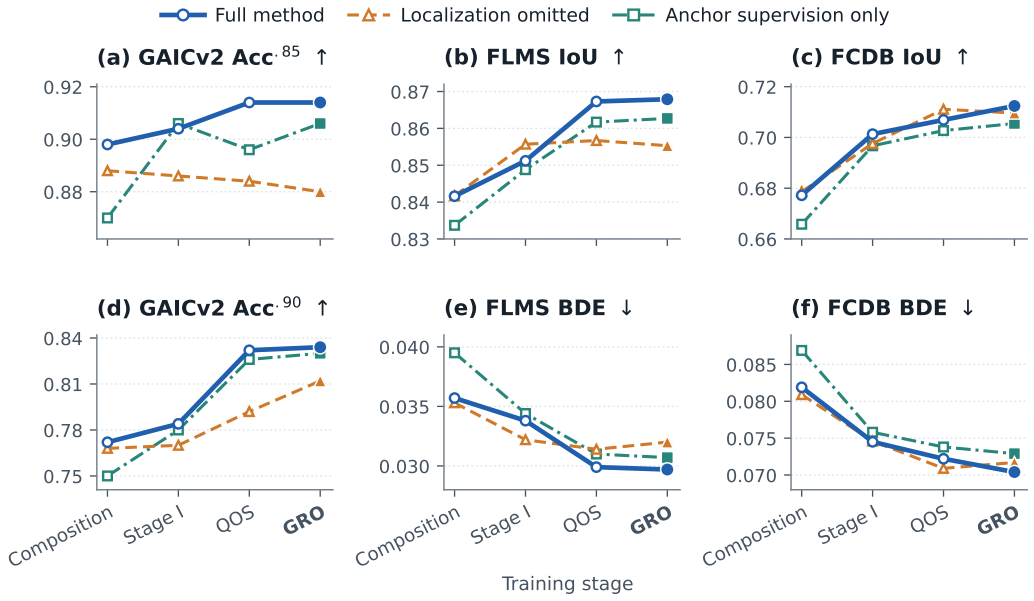

Figure 6: **How Stage-I ablations develop through training.** Markers show measurements after composition learning, at the end of Stage I, after QOS, and after GRO. Lines connect these measured points; filled markers identify the final models. Intermediate rankings vary, while the full method achieves the strongest final result on all six metrics. Upward/downward arrows indicate the preferred metric direction.

Intermediate recovery. Figure 6 shows how these differences develop. Localization improves GAIC acceptance at every evaluated stage, while FLMS/FCDB intermediate rankings vary. Their local advantages without localization disappear at the final model. For anchor-only supervision, complete-frame training substantially reduces the initial deficits; Stage-I GAIC $\text{Acc}^{\cdot 85}$ even becomes slightly higher. QOS and GRO then improve the variant's overlap and boundary accuracy, but retain a final gap. The multi-candidate benefit therefore survives subsequent complete-frame supervision and online refinement. The comparison evaluates the two preference terms jointly.

B.2 Controls for the Online Objectives

The random-target control retains QOS's greedy-plus-four candidate budget, online regeneration, source weighting, and optimization schedule, but selects a target uniformly before inspecting rewards. Repeated and invalid candidates remain eligible. Reference-target supervision instead uses fixed annotation-derived targets. The main-paper comparison after the same GRO refinement supports quality-guided selection among the current policy's answers.

Continued QOS and GRO start from the same model under matched image exposure and learning-rate schedules, testing refinement beyond winner imitation. Both use a fresh AdamW optimizer with learning rate 10^{-6}, 5% warmup, and a complete one-pass cosine schedule. Positive-only GRO replaces A_j with $\max(A_j, 0)$ while retaining all K sequence-probability calculations and the denominator K. The results favor the complete signed update in acceptance and IoU; FLMS BDE is similar. Because clipping also changes update magnitude, this control assesses the full objective without isolating negative feedback.

B.3 Number of Sampled Candidates

We compare four and eight sampled answers in QOS, retaining the greedy answer in both cases. Both settings start from the same Stage-I model and use one training pass with the same learning-rate schedule and image exposure. Table 5 shows that doubling the sampled set does not improve acceptance or IoU at the QOS endpoint; FCDB BDE is slightly lower with eight samples. This comparison supports the four-sample setting used by FrameLite, while showing that a larger candidate set alone does not guarantee a better learned policy. It evaluates candidate selection during QOS, before GRO refinement.

Table 5: **Sampled candidate count in QOS.** K excludes the greedy answer. Results are measured after one QOS pass from a common Stage-I initialization.

Sampled answers	GAICv2		FLMS		FCDB	
	$\text{Acc}^{\cdot 85}\uparrow$	$\text{Acc}^{\cdot 90}\uparrow$	IoU↑	BDE↓	IoU↑	BDE↓
$K = 4$	**0.9140**	**0.8320**	**0.8673**	**0.0299**	**0.7069**	0.0722
$K = 8$	0.9080	0.8120	0.8624	0.0303	0.7059	**0.0717**

B.4 Selection and Signed Feedback During Training

We examine the selected targets and sampled answers throughout the QOS and GRO passes. In QOS, a sampled answer replaces the greedy target for 57.7% of GAIC images, 65.8% of CPC images, and 79.5% of FCDB images. Thus, quality-guided selection frequently supplies a target different from the policy's greedy prediction.

Figure 7 shows the signs of the sampled advantages during GRO. Across the three sources, 30.9% are positive, 31.5% are zero, and 37.7% are negative. Of the 18,754 negative-advantage answers, 18,700 are valid crops: negative feedback predominantly concerns geometrically valid framings with lower reward than the greedy baseline, rather than malformed outputs. This distribution illustrates the distinction between signed refinement and winner imitation. The plotted frequencies describe encountered answers; they do not measure gradient magnitudes or the causal effect of removing an update term.

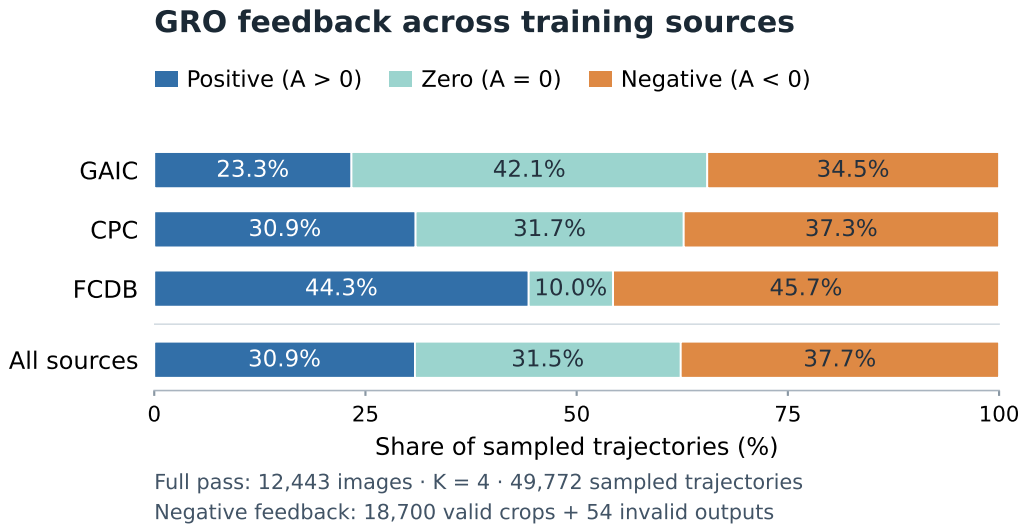

Figure 7: **Composition of feedback during GRO.** Each source is normalized by its sampled trajectories: 10,144 for GAIC, 34,152 for CPC, and 5,476 for FCDB. The last row pools these counts without source-loss reweighting. Positive and negative advantages compare each sampled crop with the same image's greedy prediction.

C User Study

We conduct a blind preference study on 40 images randomly sampled from GAICv2, FLMS, and FCDB. All 43 volunteers evaluate all 40 images. For each image, they view the crops produced by FrameLite, Venus, Cropper, InstructCrop, GAICD, and CACNet, with method identities concealed, and select the single crop they consider best. We summarize the resulting preferences as percentages in Figure 8. FrameLite receives 43.3% of the preferences, followed by Venus (21.3%), Cropper (12.9%), InstructCrop (11.2%), GAICD (7.5%), and CACNet (3.8%). FrameLite receives the largest share of preferences, providing subjective support for its predicted compositions.

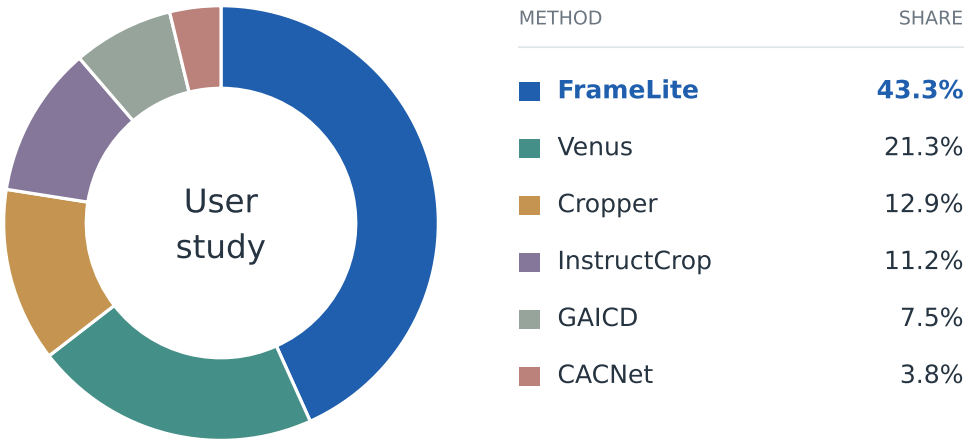

Figure 8: **Human preferences in the blind user study.** Volunteers select one preferred crop per image from six methods. Each sector represents a method's share of the preferences; percentages are rounded to one decimal place.

D Additional Qualitative Results

Figures 9 and 10 present sixteen additional comparisons from GAICv2, FCDB, and FLMS, using images distinct from Figure 4. They examine subject extent, fine structures, and scene context; column order and evaluation follow Figure 4.

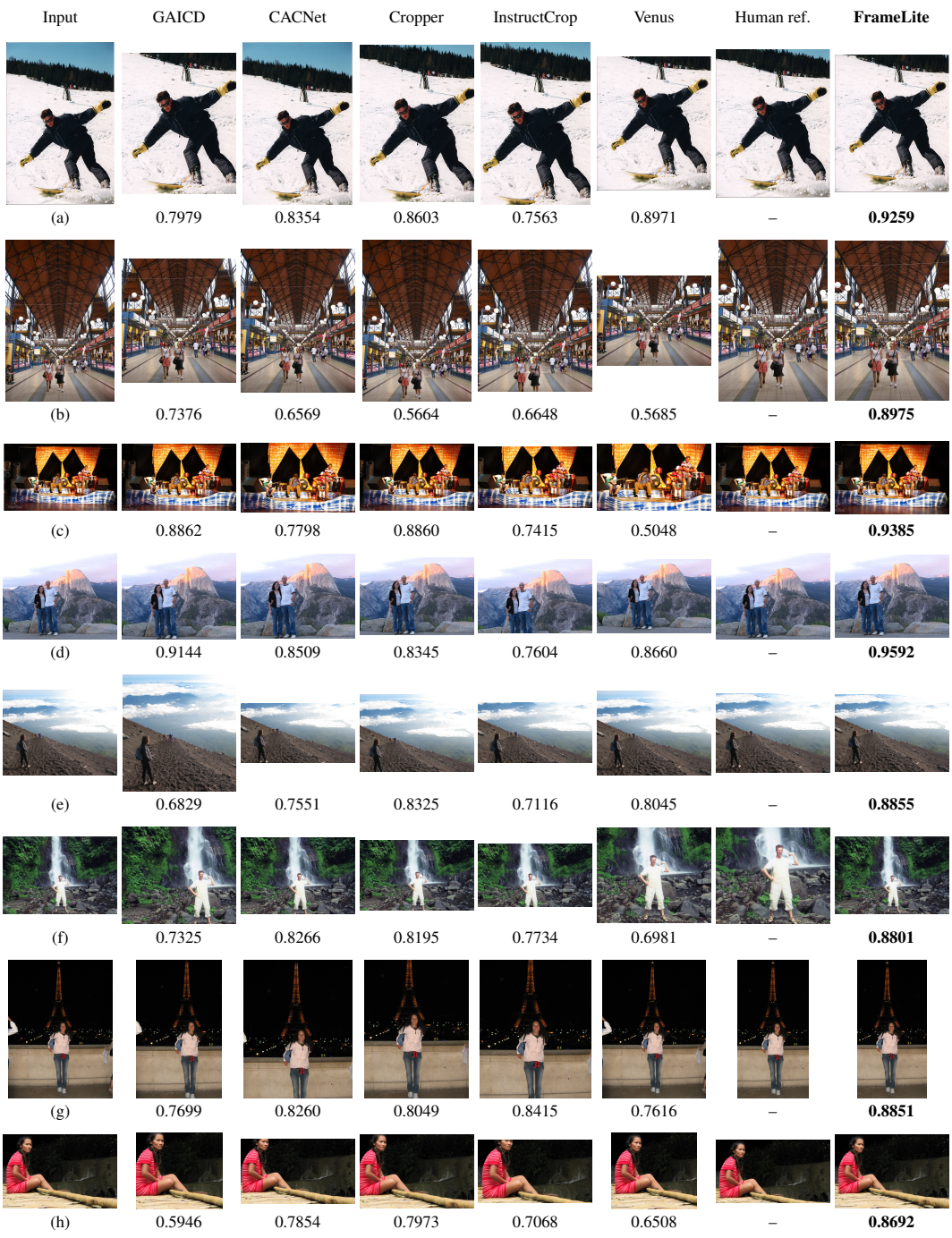

Figure 9: **Subject extent and crop boundaries.** Rows show (a) a snowboarder, (b) market pedestrians, (c) a band, (d) a mountain portrait, (e) a volcano hiker, (f) a waterfall portrait, (g) a tower portrait, and (h) a seated portrait. The comparisons illustrate visible body parts and the context surrounding each subject. Scores below predictions are IoU.



Figure 10: **Scene context and fine structures.** Rows show (a) a fountain, (b) an alley, (c) the Statue of Liberty, (d) a hibiscus, (e) ducks, (f) coastal rock arches, (g) coastal fields, and (h) a bridge at sunset. These examples compare fine structures, groups of subjects, and the allocation of foreground and background. Scores below predictions are IoU.