DISCRETE ANNOTATION, CONTINUOUS PREFERENCE: RETHINKING SUPERVISION FOR ACCURATE AND GENERALIZABLE AESTHETIC IMAGE CROPPING

Anonymous authors
Paper under double-blind review

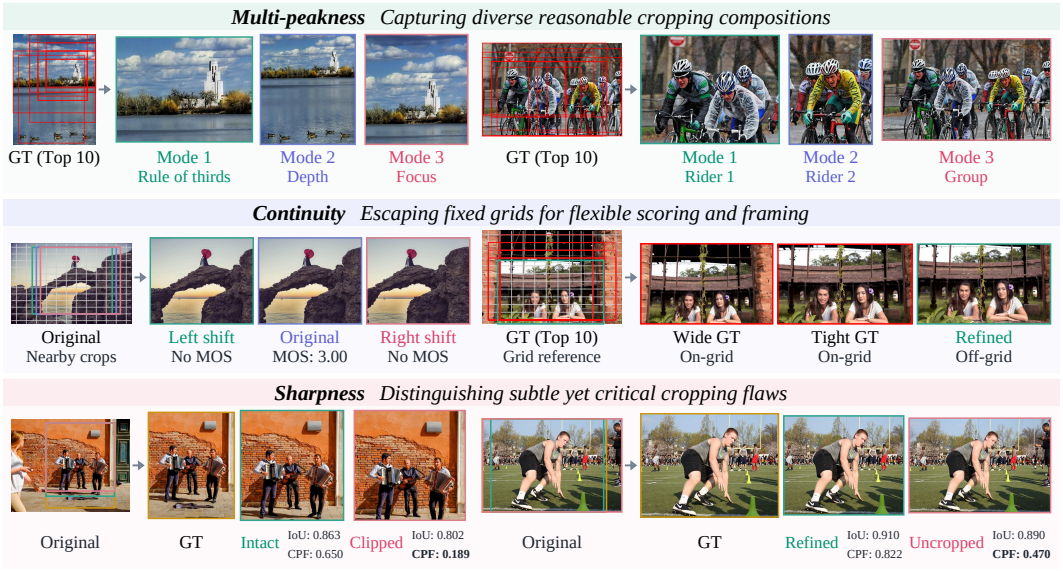

Figure 1: **Illustration of three key properties of our Continuous Preference Field (CPF)**: (1) multi-peakness, (2) continuity, and (3) sharpness. (Zoom-in for best view)

ABSTRACT

Aesthetic image cropping aims to identify the optimal crop of an image in terms of aesthetics and composition. While supervision based on annotated data is fundamental, the field has been hindered by a long-standing problem: existing datasets suffer from (1) human subjectivity and (2) rigid discreteness confined to fixed sampling grids. These flawed annotations not only limit the accuracy and generalization of trained models but also severely distort fair evaluation. To overcome this, we propose to model human cropping preference as a **multi-peaked, continuous, and sharp** field over the crop space. We introduce the **Continuous Preference Field** (CPF), which recovers a dense preference landscape from discrete annotations through (1) peak clustering, (2) off-lattice refinement, (3) negative shaping, and (4) field assembly. Based on this, we train **CPIC**, a VLM-based cropping model optimized via GRPO with the CPF reward, which overcomes template collapse, achieving state-of-the-art performance and exceptional out-of-domain generalization. Finally, to resolve the long-standing benchmark evaluation crisis, we introduce **CPICD**, a comprehensive recalibration of existing ground-truth boxes. By leveraging the CPF to correct grid-bound artifacts across mainstream benchmarks, CPICD establishes a rigorous and reliable foundation for future cropping research. Extensive experiments and user studies demonstrate the superiority of our CPF, CPIC, and CPICD. Code will be released soon.

1 INTRODUCTION

Aesthetic image cropping is a classic computer vision task that seeks to crop the sub-image from an image with the best aesthetics and composition. It serves critical roles in real-world applications such as automated photo-editing and intelligent camera view recommendation. Over the past decade,

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.

Table 1: Cropping datasets. Each restricts the coordinate space to make annotation affordable, and each records the taste of a small pool of people.

Dataset	Candidates	Subjects
CPC (Wei et al., 2018)	4×6 views	6 workers
FLMS (Fang et al., 2014)	10 boxes	10 experts
FCDB (Chen et al., 2017a)	1 box	7 workers
GAIC (Zeng et al., 2019)	12×12 grid	19 raters

Table 2: Distinct coordinate values and distinct boxes produced over the 200 GAIC test images: at most 800 and 200 respectively.

Method	Output	Values	Boxes
HCIC (Zhang et al., 2022)	anchors	18	93
GAIC (Zeng et al., 2019)	anchors	19	105
Venus (Du et al., 2026)	tokens	43	66
CPIC (ours)	tokens	**157**	**199**

solutions have evolved from traditional networks scoring pools of proposals (Yan et al., 2013; Chen et al., 2017b; Wei et al., 2018; Zeng et al., 2019; 2020; Zhang et al., 2022), to models directly regressing continuous coordinates (Hong et al., 2021; Yang et al., 2023; Liu et al., 2023; Zhang et al., 2026a), and most recently to vision-language models (VLMs) decoding bounding boxes as discrete text tokens (Lee et al., 2025; Sheng et al., 2025; Du et al., 2026). By leveraging powerful visual pretraining knowledge and advanced reasoning capabilities, VLMs have significantly advanced scene understanding. Nevertheless, despite these architectural leaps, the vast majority of state-of-the-art approaches share a core dependency: to accurately align with human aesthetic preference, they fundamentally rely on fitting the specific data bounding boxes that human annotators scored highest.

However, a critical examination of existing datasets reveals two major flaws. First, they are **(1) subjective**. As summarized in Table 1, mainstream datasets record the highly idiosyncratic preferences of a tiny pool of raters (e.g., relying on just 6 to 19 subjects), causing models to exhibit **poor generalization** across different benchmarks. Second, they are artificially **(2) discrete**. To make annotation affordable, these datasets strictly confine candidate boxes to sparse grids or predefined aspect ratios. Training on these restricted grids forces models into template collapse, resulting in **poor spatial accuracy**. As demonstrated in Table 2, Venus (Du et al., 2026), for instance, produces only 43 distinct coordinate values across 800 test slots.

How, then, can we break free from these deeply flawed dataset priors? As long as models are trained to strictly mimic the limited, discrete bounding boxes provided by annotators, they will never uncover the true continuous nature of aesthetics. To solve this, we must shift our perspective from discrete fitting to modeling the global aesthetic landscape.

Insight 1. Aesthetic cropping preference is a multi-peaked, continuous, and sharp distribution.

As illustrated in Figure 1, true human preference differs drastically from rigid grid annotations. (1) It is *multi-peaked*: a single image frequently supports multiple disjoint, valid framings. For instance, both a tight portrait crop and a wider rule-of-thirds composition can be equally pleasing. (2) It is *continuous*: shifting a good crop by a few pixels still yields a good crop, demanding sub-grid spatial continuity rather than artificial score drops between grid points. (3) It is *sharp*: however, once a critical visual boundary is crossed (e.g., clipping a subject's face), the aesthetic score plummets steeply—a strict penalty that standard metric smoothing (like IoU) fails to reflect.

If true preference exhibits these complex geometric properties, how can we bridge the gap between this continuous ideal and our discrete dataset reality?

Insight 2. Discrete annotations can provide reliable anchors to recover continuous landscape.

Crucially, we do not need to discard decades of existing datasets. While individual grid boxes are discrete and noisy, collectively, they roughly sample the core tendencies of human preference. We can construct a continuous approximation to human preference by using them as starting points.

Guided by these insights, we introduce the **Continuous Preference Field (CPF)**, constructed via a four-stage pipeline. **(1) Peak clustering** identifies the multiple valid framings within an image to guarantee the *multi-peaked* property. **(2) Off-lattice refinement** shifts these targets off the discrete grid to their precise continuous local optima, ensuring *continuity*. **(3) Negative shaping** injects hard boundary penalties (e.g., truncated subjects) to enforce *sharpness*. **(4) Field assembly** merges them into a dense, continuous reward function.

With this dense preference landscape established, we introduce **CPIC**, fundamentally changing the VLM cropping paradigm. Apart from only using standard instruction tuning to regress discrete tokens, CPIC leverages Group Relative Policy Optimization (GRPO) (Shao et al., 2024) guided by the CPF reward. This allows CPIC to explore and optimize in the continuous space, completely overcoming template collapse and achieving unmatched out-of-domain generalization.

Finally, the same discrete flaws plaguing training data also compromise testing. Treating current validation ground truths as flawed anchors, we apply our CPF to introduce **CPICD**. This recalibration of benchmark annotations purges historical grid artifacts, providing the field with its first rigorous continuous evaluation standard. In summary, our contributions are:

- We construct **Continuous Preference Field** (CPF), which models aesthetic cropping preference as a multi-peaked, continuous, and sharp distribution.
- We propose **CPIC**, a VLM-based cropping model trained with CPF as a GRPO reward, which overcomes template collapse to achieve state-of-the-art accuracy.
- We introduce **CPICD**, recalibrating mainstream ground-truth boxes to resolve the long-standing benchmark evaluation crisis.
- Extensive experiments and user studies validate the superiority of our framework.

2 Related Work

2.1 Aesthetic image cropping

Aesthetic image cropping is the task of identifying and extracting the most visually appealing (Liu et al., 2024b) and optimally composed sub-region from a given source photograph. Early learned methods approach aesthetic cropping by scoring candidate proposals, either ranking professional crops (Chen et al., 2017b), regressing quality changes (Yan et al., 2013), scoring predefined views (Wei et al., 2018), or evaluating dense candidates generated from grid anchors (Zeng et al., 2019; 2020; Zhang et al., 2022). To bypass exhaustive pooling, a second family regresses bounding box coordinates directly, conditioned on various image features (Hong et al., 2021; Yang et al., 2023; Liu et al., 2023; Zhang et al., 2026a). The current state of the art formulates cropping autoregressively, utilizing VLMs to decode bounding boxes as discrete text tokens via prompting, instruction tuning, or pretraining (Lee et al., 2025; Sheng et al., 2025; Du et al., 2026). Because discrete coordinate vocabularies seamlessly encode dataset annotation lattices, these models are highly susceptible to reproducing structural discretization artifacts. In contrast, our method constructs a continuous optimization signal directly from the underlying structure of the original annotations themselves.

2.2 Cropping datasets

Existing cropping datasets often impose strict constraints that artificially limit the valid coordinate space. ICDB initiated the domain with professional crops (Yan et al., 2013), while FCDB retains a single crop preferred over the original image (Chen et al., 2017a), and FLMS collects unordered boxes from crowd workers (Fang et al., 2014). To evaluate denser candidates, CPC restricts views to four aspect ratios (Wei et al., 2018), and GAIC strictly confines approximately ninety scored candidates per image to a 12×12 anchor grid (Zeng et al., 2019; 2020). Other datasets supply subject-centric ranked pairs (Yang et al., 2023) or rely on image-level aesthetic scores for auxiliary supervision without bounding boxes (Murray et al., 2012). To avoid subjective human labeling, GenCrop synthesizes training pairs by outpainting professional photographs (Hong et al., 2024). Rather than discarding human judgments, our approach builds on existing discrete annotations and enables crop evaluation beyond annotated candidates and refinement of existing ground-truth boxes.

3 Methodology

3.1 Preliminary: Rethinking the Supervision Signal

Given a source image I, aesthetic image cropping predicts an admissible bounding box $b = (x_1, y_1, x_2, y_2)$. Despite its conceptual simplicity, formulating a reliable supervision signal is persistently hindered by two structural flaws in existing benchmarks.

First, aesthetic preference is inherently **subjective**. As the last column of Table 1 shows, every benchmark rests on a small pool of just 6 to 19 annotators, meaning the mean opinion score (MOS)

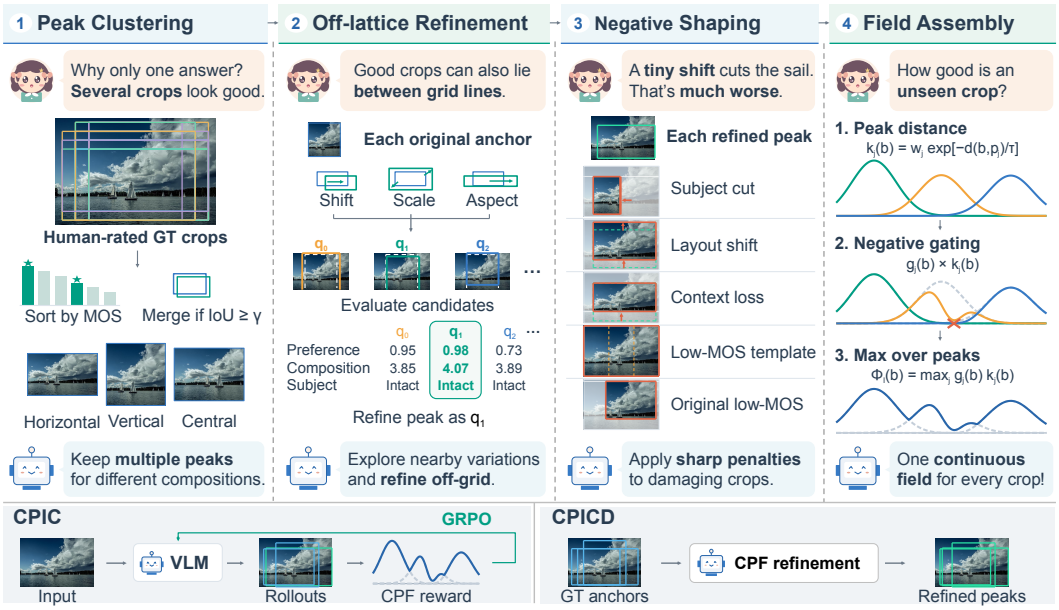

Figure 2: Overview of CPF construction and its applications. (1) Peak clustering identifies distinct preference modes from human-scored crops. (2) Off-lattice refinement adjusts each peak using preference, composition, and subject-preservation scores. (3) Negative shaping constructs five types of undesirable crops. (4) Field assembly combines peak-distance and negative gating into CPF. CPF serves as the reward for GRPO training of CPIC and guides annotation recalibration for CPICD.

contains error noise and cannot be absolutely trusted. Additionally, this causes a lack of generalizability from one dataset to another, as the annotators of different datasets may prefer differently.

Second, bounding box evaluation is treated as **discrete**. The candidate-space column of Table 1 shows that canonical datasets restrict candidate crops to coarse anchor grids or sparse predefined aspect ratios. This discreteness systematically excludes the unconstrained coordinate space, forcing models to artificially approximate fluid geometric operations. Table 2 shows the severe cost of this discreteness. Across two hundred images, published croppers (Zhang et al., 2022; Zeng et al., 2019; Du et al., 2026) emit only 18 to 43 distinct coordinate values, repeating whole boxes across unrelated photographs; the continuous coordinates they cannot produce are simply the ones their supervision never contained, which restricts the performance from reaching its optimal upper bound.

3.2 CPF: Constructing the Continuous Preference Field

To bridge the gap between discrete, noisy labels and the human-preferred aesthetic landscape, we introduce the Continuous Preference Field (CPF). Our goal is to explicitly instantiate the three insights identified in the Introduction—multi-peakedness, continuity, and sharpness—into a dense, evaluable scalar function. The four-stage construction is illustrated in Figure 2.

Stage 1: Peak Clustering. To model aesthetic preference, we must capture its inherently **multi-peaked** distribution. We observe that existing human annotations naturally manifest this property: high-scoring crop candidates concentrate densely around specific compositional modes while remaining sparse elsewhere. This spatial concentration directly embodies the multi-peakedness we seek, which we recover using the clustering procedure in Algorithm 1. Furthermore, by assigning dynamic weights via the dataset's annotation noise floor T, we derive both the number of peaks and their relative weights directly from each specific image's native score spread. Employing a fixed weight ladder would incorrectly impose an identical rank-1 to rank-2 gap across all images, acting as a rigid hyper-parameter that erases inherent aesthetic subjectivity.

Algorithm 1 Peak Clustering

Require: candidates C with scores s; γ, Δ, K, T
1: sort C by s desc.; $s_{\max} \leftarrow \max s$; $P \leftarrow \emptyset$
2: **for** $c_j \in C$ **while** $|P| < K$ **do**
3: **if** $s_j < s_{\max} - \Delta$ **and** $P \neq \emptyset$ **then continue**
4: **if** $\mathrm{IoU}(c_j, p) < \gamma$ for all $p \in P$ **then** $P \leftarrow P \cup \{c_j\}$
5: **end for**
6: **if** $P = \emptyset$ **then** $P \leftarrow \{\arg\max_j s_j\}$
7: $w_j \leftarrow \exp\big(-(s_{\max} - s_j)/T\big)$ for each $p_j \in P$
8: **return** peaks P with weights W

Stage 2: Off-lattice Refinement. The peaks extracted in Stage 1 still reside on the quantized annotation grid. To restore **continuity**, we expand the local search space beyond the discrete lattice. We refine each clustered peak by selecting the best candidate from the original anchor and its continuously perturbed neighbors, combining learned preference with rule-based regularization to reduce subjective annotation bias. Around an anchor center (c_x, c_y) with width w and height h, deterministic perturbations are seeded via the image key, peak index, and anchor rank:

$$\begin{aligned} c'_x &= c_x + \mathcal{N}(0, \sigma_{\text{shift}} w), \qquad & w' &= w \exp(s + a/2), \\ c'_y &= c_y + \mathcal{N}(0, \sigma_{\text{shift}} h), \qquad & h' &= h \exp(s - a/2), \end{aligned} \tag{1}$$

where $s \sim \mathcal{N}(0, \sigma_{\text{scale}})$ and $a \sim \mathcal{N}(0, \sigma_{\text{aspect}})$. We refine targets using the mean score $z = (z_{\text{pref}} + z_{\text{comp}} + z_{\text{subj}})/3$. The score fuses:

- **(1) Human preference (z_{pref}):** Averages the standardized anchor MOS and pre-trained GAIC score (Zeng et al., 2019; 2020): $z_{\text{pref}} = (z_{\text{MOS}} + z_{\text{GAIC}})/2$. Because the GAIC scorer learns generalized rules from dataset annotations, it effectively smooths out the sharp noise and discreteness of raw human labels.
- **(2) Composition score (z_{comp}):** Provided by SAMP-Net (Zhang et al., 2021), this robustly evaluates fine-grained sub-grid geometric alignment.
- **(3) Subject preservation (z_{subj}):** Computes the explicit area retention percentage of YOLOv8 (Jocher et al., 2023)-detected subject bounding boxes.

Composition and subject-preservation criteria regularize the learned preference score with explicit visual constraints, reducing reliance on subjective annotations and biases inherited by one scorer.

Stage 3: Negative Shaping. Spatial distance alone implies a smooth decay, which contradicts the fact that crossing critical visual boundaries severely plummets perceived aesthetics. To enforce this **sharpness**, we explicitly model "valleys" representing non-preference. We surround each continuous peak with structured negatives encompassing five classes:

- **(1) Truncated subject:** We intentionally push the box edges inward to clip the bounding box of YOLOv8 (Jocher et al., 2023; Hong et al., 2024)-detected main subjects.
- **(2) Broken composition:** We shift the box to deliberately move detected subjects away from their original ideal placements (e.g., rule-of-thirds or center) to awkward positions near the frame edges, and shift detected horizon lines away from rule-of-thirds/center axes.
- **(3) Discarded value:** We tighten the bounding box edges from all directions without cutting the subject, blindly discarding meaningful contextual background information.
- **(4) Templated crops:** We include the uncropped full image and highly repetitive boxes lying on discrete grid points; if these templates do not possess a natively high MOS in the dataset, they are strictly marked as negatives.
- **(5) Original low-MOS:** We dynamically mine original human-annotated boxes that highly overlap with the refined peak but possess a demonstrably low MOS.

By explicitly synthesizing these five boundary conditions, we inject necessary sharpness into the continuous space, ensuring the field rigorously models the severe aesthetic penalties humans perceive when fundamental compositional rules are violated.

Stage 4: Field Assembly. Finally, we unify the peaks, their dynamic weights, and the structured negatives into a continuous reward field. To measure geometric discrepancy without introducing additional weighting hyperparameters, we combine four complementary terms for overlap, boundary displacement, scale, and shape:

$$d(b, b') = \frac{1}{4}\left[(1 - \text{IoU}(b, b')) + \text{BDE}(b, b') + \frac{|\log(A/A')|}{1 + |\log(A/A')|} + \frac{|\log(r/r')|}{1 + |\log(r/r')|}\right], \tag{2}$$

where A and r denote the area and aspect ratio, respectively, and BDE is computed on normalized coordinates. All four terms are thus bounded to comparable ranges and contribute equally, avoiding manually tuned distance weights.

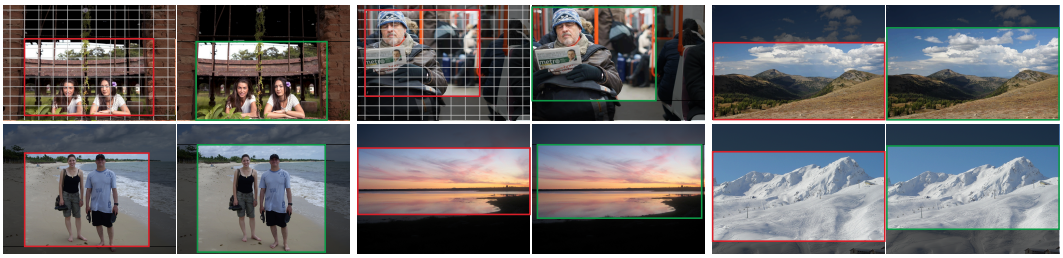

Figure 3: Comparison of original benchmark annotations and CPICD recalibrated boxes. Recalibration improves unbalanced compositions with truncated subjects or overly narrow crops.

For each compositional peak p_j, its refined location induces a smoothly decaying similarity score $\exp(-d(b, p_j)/\tau)$, which attracts a query crop b toward the corresponding high-preference region. Spatial proximity alone, however, cannot capture the sharp preference drop near undesirable crops. We therefore compare the distance to the positive peak against that to its nearest structured negative and define a soft negative gate:

$$g_j(b) = \sigma\left(\left[\min_{n \in N_j} d(b, n) - d(b, p_j)\right]/\tau\right), \tag{3}$$

where $\sigma(\cdot)$ denotes the sigmoid function. The gate remains high when b is closer to the preferred peak, but rapidly decreases as it approaches an undesirable region. The shared temperature τ controls both the locality of each peak and the sharpness of this transition.

Finally, since an image may support multiple valid framings, a query crop only needs to match one preference mode. We therefore take the maximum over all peak responses, weighted by their relative importance w_j:

$$\Phi_I(b) = \max_j \left[w_j \exp\big(-d(b, p_j)/\tau\big) g_j(b)\right]. \tag{4}$$

Through this assembly, $\Phi_I(b)$ forms our **multi-peaked, continuous, and sharp** preference field.

3.3 CPIC: Leveraging CPF as Supervision Signal

With our continuous and highly distinguishable reward landscape established, we introduce CPIC. We first employ a cold-start supervised fine-tuning stage (DeepSeek-AI, 2025) using the refined continuous peaks as cross-entropy targets, explicitly shifting the VLM's output support off the original discrete annotation lattice. Following this cold start, we apply group-relative policy optimization (GRPO) (Shao et al., 2024), sampling a group of crops $o_1 \dots o_G$ per image q to optimize:

$$J(\theta) = \mathbb{E}_{q,\,\{o_i\}\sim\pi_{\text{old}}}\left[\frac{1}{G}\sum_{i=1}^{G}\left(\min\big(\rho_i A_i,\ \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)A_i\big) - \beta\,\mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\right)\right], \tag{5}$$

where $\rho_i = \frac{\pi_\theta(o_i|q)}{\pi_{\text{old}}(o_i|q)}$, and the group-relative advantage $A_i = \frac{R_i - \text{mean}(R)}{\text{std}(R)}$ normalizes the reward $R_i = \Phi_I(b_i)$, the CPF score of the box decoded from rollout o_i. Specifically, our CPF allows good boxes and bad boxes to have a large degree of distinguishability. Because of this, the reward maintains a healthy gap and does not collapse, thereby driving the model's predicted boxes to further optimize toward better ones.

3.4 CPICD: Recalibration of Benchmark Ground Truth with CPF

As established in Section 3.1, the quantization errors that hinder model training equally distort benchmark evaluations, penalizing models for predicting plausible off-grid crops instead of matching flawed grids. To resolve this long-standing evaluation crisis, we deploy our CPF refinement machinery (Stage 2) to audit and recalibrate the standard test sets. By treating each historical benchmark annotation as a flawed initial anchor, applying local continuous spatial perturbations, and selecting the refined crop using the same rule-regularized preference score, we quantitatively correct the grid-bound artifacts (yielding the precise green boxes in Figure 3). This recalibration yields **CPICD**, establishing an alternative set of continuous ground truth for aesthetic cropping evaluation.

Table 3: Comparison on the four benchmarks. Best per column in **bold**, second underlined.

Method	GAIC			FLMS			FCDB			CPICD	
	ACC_{10} ↑	SRCC↑	LAION↑	IoU↑	BDE↓	LAION↑	IoU↑	BDE↓	LAION↑	IoU↑	BDE↓
Traditional networks											
VFN (MM'17)	0.725	0.425	**5.286**	0.763	0.058	5.035	0.630	0.100	4.913	0.701	0.076
VEN (CVPR'18)	0.295	0.482	5.239	0.708	0.062	5.041	0.586	0.106	4.910	0.631	0.087
VPN (CVPR'18)	0.405	0.495	5.271	0.825	0.040	5.078	0.658	0.088	4.933	0.726	0.067
GAIC (CVPR'19)	**0.975**	0.557	5.263	0.803	0.047	5.062	0.676	0.082	4.931	0.775	0.054
CACNet (CVPR'21)	0.800	0.457	5.259	0.840	0.037	5.067	0.697	0.075	4.932	0.765	0.057
HCIC (ECCV'22)	0.440	0.144	5.200	0.705	0.071	5.017	0.560	0.119	4.878	0.636	0.093
UNIC (ICCV'23)	0.845	0.544	5.279	0.839	0.040	5.065	0.681	0.080	4.931	0.767	0.058
Zero-shot VLMs											
Qwen3-VL-2B	0.290	0.285	5.137	0.804	0.049	5.020	0.566	0.119	4.863	0.649	0.094
Qwen3-VL-4B	0.215	0.261	5.029	0.696	0.063	4.905	0.507	0.130	4.779	0.550	0.113
Qwen3-VL-8B	0.250	0.295	5.141	0.736	0.056	5.019	0.572	0.107	4.863	0.628	0.086
Qwen2.5-VL-7B	0.185	0.097	4.920	0.656	0.077	4.871	0.390	0.171	4.638	0.494	0.131
LLaVA-1.5-7B	0.065	0.114	4.603	0.428	0.134	4.570	0.365	0.178	4.504	0.356	0.169
InternVL3.5-2B	0.385	0.321	5.282	0.817	0.047	5.041	0.634	0.102	4.930	0.704	0.081
VLM-based cropping methods											
InstructCrop (MM'25)	0.630	0.326	5.238	0.761	0.055	5.040	0.652	0.087	4.910	0.709	0.069
Venus (CVPR'26)	0.945	0.541	5.270	**0.856**	**0.032**	5.076	0.683	0.079	4.933	0.763	0.057
CPIC (ours)	0.905	**0.590**	5.283	0.842	0.037	**5.082**	**0.714**	**0.070**	**4.934**	**0.789**	**0.049**

Table 4: User study on cropping quality (%): the fraction of trials in which each method's crop is chosen as the best one.

GAIC	CACNet	UNIC	InstructCrop	Venus	**CPIC (ours)**
32.7	37.2	42.0	22.8	32.5	**46.3**

Table 5: User study on ground truth recalibration (%).

Original GT	**CPICD (ours)**
35.5	**64.5**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets. We train our model on the CPC (Wei et al., 2018) and GAIC (Zeng et al., 2019; 2020) datasets, optimized via our CPF. Following prior work (Zeng et al., 2020; Zhang et al., 2026a), we evaluate on the GAIC test split, FLMS (Fang et al., 2014), and FCDB (Chen et al., 2017a).

Metrics. We comprehensively consider multiple metrics. Following standard work (Zeng et al., 2020; Zhang et al., 2026a), we report the Spearman rank correlation coefficient (SRCC) and ACC_{10} on the GAIC dataset. For FLMS and FCDB, we report the Intersection over Union (IoU) and the boundary displacement error (BDE) (Chen et al., 2017b). We further introduce the no-reference LAION-Aesthetics score (Schuhmann et al., 2022) to measure the aesthetic quality of the crop box.

Implementation details. We train our CPIC model based on Qwen3-VL-2B-Instruct (Bai et al., 2025a) using LoRA (Hu et al., 2022) with a rank of 64 and the AdamW optimizer (Loshchilov & Hutter, 2019). In the SFT stage, we train with a learning rate of 1×10^{-4} and batch size of 16 for 2500 steps on a single RTX 4090 GPU. In the GRPO (Shao et al., 2024) stage, we train with a learning rate of 3×10^{-5} and batch size of 16 for 1200 steps on 4 RTX 4090 GPUs. We sample $G = 8$ crops per image and set $\beta = 0.04$ and $\epsilon = 0.2$.

4.2 COMPARISON WITH THE STATE-OF-THE-ART METHODS

Compared methods. We compare against three families. *Traditional networks* that score or regress boxes without a large VLM: VPN (Wei et al., 2018), GAIC (Zeng et al., 2019; 2020), CACNet (Hong et al., 2021), HCIC (Zhang et al., 2022) and UNIC (Liu et al., 2023). *Zero-shot VLMs* prompted to emit a crop box without any cropping-specific training: Qwen3-VL at 2B, 4B and 8B (Bai et al., 2025a), Qwen2.5-VL-7B (Bai et al., 2025b), LLaVA-1.5-7B (Liu et al., 2024a) and InternVL3.5-2B (Wang et al., 2025). *VLM-based cropping methods* that fine-tune a vision-language model for the task: InstructCrop (Sheng et al., 2025), Venus (Du et al., 2026) and our CPIC.

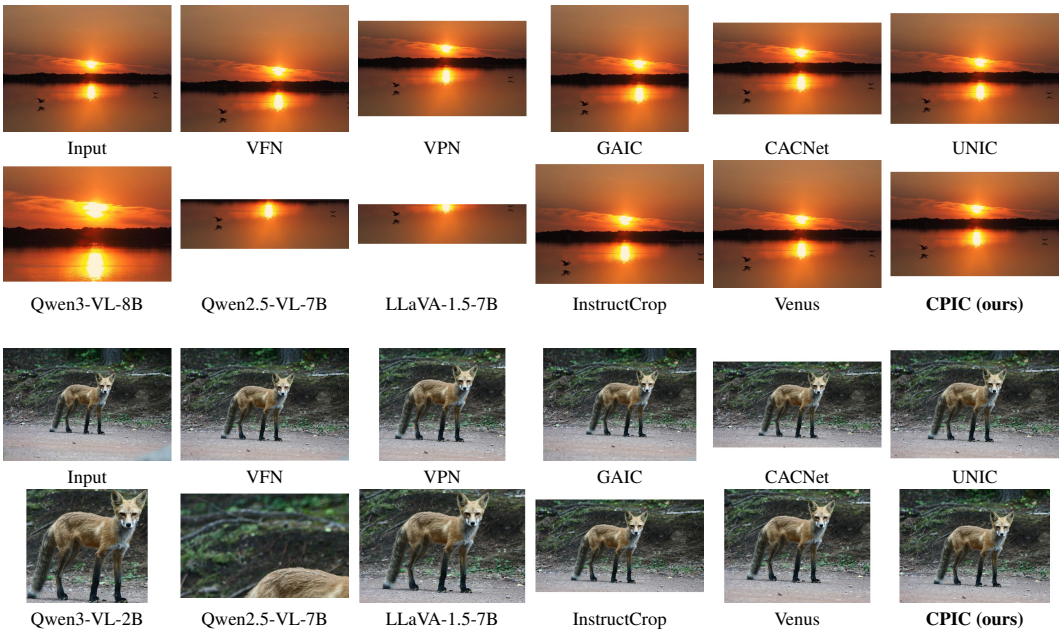

Figure 4: Qualitative results on in-domain (GAIC) and out-of-domain (FCDB) images.

Quantitative comparison. As detailed in Table 3, CPIC demonstrates the strongest generalization across all three method families and under both reference-based and no-reference metrics, ranking first or second on ten of the eleven columns. On the GAIC benchmark, GAIC and Venus achieve the top ACC_{10} scores due to their grid priors, as this metric inherently rewards reproducing a lattice-bound annotation. Nevertheless, CPIC yields higher absolute crop quality on the same dataset, outperforming both grid-bound methods on the SRCC and LAION metrics. Crucially, CPIC maintains dominant performance on the FLMS and FCDB benchmarks. Because neither FLMS nor FCDB appears in our training data, these test sets are entirely out of domain for us. Succeeding in these unseen environments provides the strongest evidence of our method's robust generalization.

Qualitative comparison. Figure 4 demonstrates our method's aesthetic cropping capabilities compared to baselines. In the first example, CPIC achieves a horizontal, centered composition that perfectly balances the two flying birds on the left and right. Other methods fail to capture this harmony: GAIC, InstructCrop, and Venus leave excessive empty sky and water; VPN and CACNet yield overly thin wide strips. In the second example, CPIC keeps the main subject (the fox) centered while preserving an appropriate margin of context to maintain a harmonious visual balance. Conversely, baseline models either retain distracting empty ground (UNIC, VFN, InstructCrop), crop too tightly around the fox without sufficient breathing room.

User Study. We conduct two users studies with implementation details provided in Appendix Sec. D. In the first study evaluating cropping quality (Table 4), participants are presented with crops from six methods and instructed to select their top 1 to 3 aesthetically pleasing choices per image. Our CPIC achieves the highest overall selection rate, demonstrating its superior alignment with human visual preference. In the second study evaluating ground-truth recalibration (Table 5), participants perform a two-alternative choice between the original benchmark annotation and our recalibrated CPICD. The results show a significant preference for CPICD, verifying that our continuous refinement can correct historical annotation flaws and yield better aesthetic optima.

4.3 Ablation Studies

To systematically validate our design choices, we conduct extensive ablation studies in Table 6.

Effect of the continuous preference field. We first establish that substituting cross-entropy (CE) with GRPO is not inherently responsible for the success. As shown in **(a)** and **(b)**, applying GRPO with a naive IoU reward against discrete ground truths severely limits spatial exploration, capping output diversity at only 92 distinct boxes. While shifting CE targets to our refined continuous peaks **(c)** expands diversity, the accuracy remains limited. Even when applying GRPO with an IoU reward

Table 6: Ablation studies. Boxes counts the distinct crops the model emits over the 200 GAIC test images (at most 200), the diversity measure of Table 2.

	Configuration	CPF ingredient			GAIC	FLMS		FCDB	Boxes↑
		multi-peak	continuity	sharpness	SRCC↑	IoU↑	LAION↑	IoU↑	
	Supervision strategies								
(a)	CE on annotated GT				0.523	0.798	5.060	0.642	56
(b)	+ GRPO w/ IoU				0.564	0.822	5.074	0.671	92
(c)	CE on refined peaks		✓		0.538	0.827	5.076	0.704	154
(d)	+ GRPO w/ IoU		✓		0.559	0.835	**5.082**	0.701	199
	+ GRPO with CPF								
(e)	CPF w/ single peak		✓	✓	**0.619**	0.829	5.079	0.706	199
(f)	CPF w/o refinement	✓		✓	0.517	0.829	5.065	0.672	88
(g)	CPF w/o negatives	✓	✓		0.574	0.839	5.076	0.708	**200**
(h)	**CPF (ours)**	✓	✓	✓	0.590	**0.842**	**5.082**	**0.714**	199

on these refined continuous targets **(d)**, the model fails to capture true aesthetic boundaries, yielding sub-optimal results. The substantial, comprehensive improvements across all reference-based and no-reference metrics materialize only when the policy is guided by our CPF **(h)**.

Effect of multi-peak supervision. Row **(e)** restricts the CPF to a single peak. Because a single photograph frequently supports several disjoint valid framings, collapsing the field to one exclusive target discards valid alternative crops and forces the model to mimic a single subjective choice. Consequently, while **(e)** slightly overfits to the GAIC training domain, it artificially penalizes good alternative predictions, suffering noticeable drops in out-of-domain generalization on the FLMS and FCDB benchmarks. By incorporating multi-peak supervision, our full model **(h)** better aligns with diverse human consensus, yielding a robust balance across all domains.

Effect of off-lattice continuity. Row **(f)** completely ablates the off-lattice refinement, forcing the CPF positive anchor to revert to the old lattice-bound annotated box. This explicitly destroys the continuity of the target space. As a result, the model's spatial precision collapses, and its ability to break the discrete token constraint is crippled, causing output diversity to plummet back to a mere 88 distinct boxes. Taking the target mathematically off the annotation grid in **(h)** provides an accurate continuous optimum, allowing the model to freely explore sub-grid coordinates (reaching 199 distinct boxes) and significantly boosting overlap metrics.

Effect of sharp structured negatives. Row **(g)** ablates the five structured negative classes. Without these negatives, the preference reward decays smoothly and indiscriminately with geometric distance, diluting the learning signal. It fails to aggressively penalize critical aesthetic violations, such as slightly cutting into a subject's face, because the IoU shift is minor. Explicitly gating the field with structured negatives in **(h)** ensures that preference falls off sharply into a "valley" when visual boundaries are crossed. This sharpness penalty is crucial for fine-grained aesthetic discrimination, maximizing both the no-reference aesthetic quality (LAION) and bounding box accuracy.

Due to page limits, key validity justifications are deferred to Appendix Sec. C, which we kindly hope will be taken into account for a proper interpretation of our method and results.

5 Conclusion

In this work, we rethink and overcome the long-standing limitations of discrete, grid-bound annotations in aesthetic image cropping. We reframe human preference as a multi-peaked, continuous, and sharp distribution, introducing the Continuous Preference Field (CPF) to transform coarse dataset labels into a dense reward landscape through peak clustering, off-lattice refinement, structured negative shaping, and field assembly. By utilizing the CPF as a GRPO reward, we train CPIC, a VLM-based cropping model optimized directly for this continuous space. CPIC decisively escapes the discrete annotation lattice and achieves state-of-the-art accuracy alongside exceptional out-of-domain generalization. Furthermore, we apply the CPF machinery to audit existing mainstream validation sets, introducing **CPICD** to recalibrate flawed ground-truth boxes. By correcting the historical grid-bound artifacts that have long distorted fair model comparison, CPICD establishes a rigorous, continuous, and reliable foundation for future aesthetic cropping research.

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

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

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.

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.

Zheng Chen, Kai Liu, Jingkai Wang, Xianglong Yan, Jianze Li, Ziqing Zhang, Jue Gong, Jiatong Li, Lei Sun, Xiaoyang Liu, Radu Timofte, Yulun Zhang, et al. The fourth challenge on image super-resolution ($\times$4) at NTIRE 2026: Benchmark results and method overview. In *CVPRW*, 2026.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

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.

Chaoyi Hong, Shuaiyuan Du, Ke Xian, Hao Lu, Zhiguo Cao, and Weicai Zhong. Composing photos like a photographer. In *CVPR*, 2021.

James Hong, Lu Yuan, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. Learning subject-aware cropping by outpainting professional photos. In *AAAI*, 2024.

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.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8. `https://github.com/ultralytics/ultralytics`, 2023.

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.

Xiaojian Lin, Hanhui Li, Yuhao Cheng, Yiqiang Yan, and Xiaodan Liang. GDrag: Towards general-purpose interactive editing with anti-ambiguity point diffusion. In *ICLR*, 2025.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024a.

Kai Liu, Ziqing Zhang, Wenbo Li, Renjing Pei, Fenglong Song, Xiaohong Liu, Linghe Kong, and Yulun Zhang. Dog-IQA: Standard-guided zero-shot MLLM for mix-grained image quality assessment. *arXiv preprint arXiv:2410.02505*, 2024b.

Kai Liu, Haoyang Yue, Zeli Lin, Zheng Chen, Jingkai Wang, Jue Gong, Jiatong Li, Xianglong Yan, Libo Zhu, Jianze Li, Ziqing Zhang, Zihan Zhou, Xiaoyang Liu, Radu Timofte, Yulun Zhang, et al. The first challenge on remote sensing infrared image super-resolution at NTIRE 2026: Benchmark results and method overview. In *CVPRW*, 2026a.

Xiao Liu, Kai Liu, Naiyang Guan, Hongliang Lu, Zhixin Wang, Zhikai Chen, Renjing Pei, and Yulun Zhang. Accelerating rectified flow models via trajectory-aware caching. *arXiv preprint arXiv:2605.16789*, 2026b.

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.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W. Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Xiangfei Sheng, Pangu Xie, Weidong Zou, Pengfei Chen, Tong Zhu, and Leida Li. InstructCrop: Teaching multimodal large language models to crop aesthetic images. In *ACM MM*, 2025.

Xinrui Shi, Kai Liu, Ziqing Zhang, Jianze Li, Anqi Li, and Yulun Zhang. DRScaffold: Boosting dense-scene reasoning in lightweight vision language models. *arXiv preprint arXiv:2605.26038*, 2026.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. LGM: Large multi-view gaussian model for high-resolution 3D content creation. In *ECCV*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomír Měch, Minh Hoai, and Dimitris Samaras. Good view hunting: Learning photo composition from dense view pairs. In *CVPR*, 2018.

Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang. Learning the change for automatic image cropping. In *CVPR*, 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. *CVM*, 2023.

Rui Yang, Wei Zhou, Dingyong Gou, Xiaohui Cui, Cong Li, Yinyin Gong, Yipo Huang, and Jiliang Zhao. COMEX: A composition-grounded benchmark and learning framework for explainable aesthetic image cropping. *arXiv preprint arXiv:2608.07570*, 2026.

Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Reliable and efficient image cropping: A grid anchor based approach. In *CVPR*, 2019.

Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Grid anchor based image cropping: A new benchmark and an efficient model. *TPAMI*, 2020.

Bo Zhang, Li Niu, and Liqing Zhang. Image composition assessment with saliency-augmented multi-pattern pooling. In *BMVC*, 2021.

Bo Zhang, Li Niu, Xing Zhao, and Liqing Zhang. Human-centric image cropping with partition-aware and content-preserving features. In *ECCV*, 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 *AAAI*, 2026a.

Lingzhi Zhang, Tarmily Wen, and Jianbo Shi. Deep image blending. In *WACV*, 2020.

Ziqing Zhang, Kai Liu, Zheng Chen, Xi Li, Yucong Chen, Bingnan Duan, Linghe Kong, and Yulun Zhang. InfVSR: Toward consistency-driven streaming generative video super-resolution. In *ICML*, 2026b.

APPENDIX

This supplementary material provides additional implementation details, extended ablations, and discussions to support the main manuscript. It is organized as follows:

- Sec. A summarizes implementation hyperparameters and inference details.
- Sec. B demonstrates CPIC's zero-shot interpretable and interactive cropping capabilities.
- Sec. C provides additional justifications and ablation studies.
- Sec. D details the protocols and participant demographics for the human evaluations.
- Sec. E presents additional qualitative visualizations.

A IMPLEMENTATION DETAILS

Hyperparameters. We summarize the construction and training settings below.

- **Peak clustering.** We retain at most $K = 4$ peaks, with a candidate MOS tolerance $\Delta = 0.75$ and initial clustering threshold $\gamma = 0.72$. Final peak consolidation uses an IoU threshold of 0.70. Peak weights use $T = 0.35$ in Algorithm 1.
- **Local perturbations.** We sample 40 local perturbations per peak. The three $(\sigma_{\text{shift}}, \sigma_{\text{scale}}, \sigma_{\text{aspect}})$ settings are $(0.012, 0.022, 0.025)$, $(0.026, 0.045, 0.050)$, and $(0.050, 0.075, 0.075)$, sampled with probabilities 0.4, 0.4, and 0.2.
- **Scoring and negatives.** The three visual scores are averaged equally, with MOS and GAIC scores equally combined within the preference term. The drift penalty uses $\lambda = 0.15$, and the field uses a shared temperature $\tau = 0.10$. Negatives with IoU ≥ 0.85 to their associated refined peak are removed to avoid conflicting supervision.
- **Optimization.** LoRA uses rank 64 and alpha 32, with a warmup ratio of 0.03. SFT and GRPO use learning rates 10^{-4} and 3×10^{-5}, respectively, and an effective batch size of 16. GRPO uses $G = 8$, $\beta = 0.04$, and $\epsilon = 0.2$.
- **Training rollouts.** Crops are sampled with temperature 1.0, top-$p = 0.95$, and a maximum of 64 generated tokens per rollout.

Inference details. At inference time, CPIC predicts a single aesthetic crop for each image using a fixed instruction:

```
Please suggest the single best aesthetic crop region for this image.
Use normalized coordinates in a 0-1000 image coordinate system: left
edge x=0, right edge x=1000, top edge y=0, bottom edge y=1000. Return
only JSON in this exact format: {"bbox_2d": [x1, y1, x2, y2]}.
```

The prompt defines a normalized coordinate system in which the left and right image boundaries are x=0 and x=1000, while the top and bottom boundaries are y=0 and y=1000. Each crop is represented by four coordinates in the JSON field `"bbox_2d"`. We use greedy decoding without sampling, so inference is deterministic and does not require a random seed. All evaluations use single-precision floating-point arithmetic to ensure bit-reproducible results across hardware architectures.

B ZERO-SHOT INTERPRETABILITY AND INTERACTION

Table 7: Output-target ablation. Best **bolded**.

Output target	GAIC	FLMS	FCDB
	$\text{ACC}_{10}\uparrow$	IoU↑	IoU↑
Box only	0.905	**0.842**	0.714
Short text + box	**0.910**	0.822	**0.723**

We compared box-only output with a variant that first emits a short, image-specific statement about the subject and composition while retaining crop localization as the task objective. Table 7 shows no consistent benefit: the added text improves GAIC and FCDB but degrades FLMS, with only small changes overall. We therefore train CPIC to output only boxes, isolating CPF supervision and avoiding unnecessary generation overhead. This agrees with COMEX (Yang et al., 2026), where

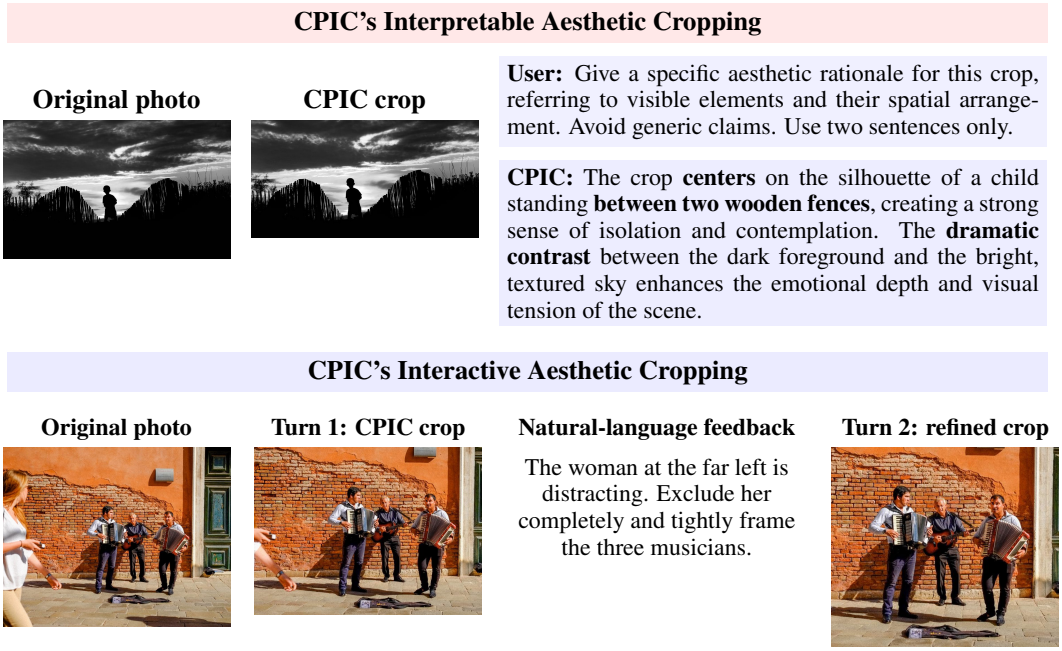

Figure 5: **Zero-shot interpretable and interactive aesthetic cropping.** Top: CPIC provides a concrete rationale grounded in the selected subject, spatial arrangement, contrast, and visual effect. Bottom: after its initial crop, CPIC follows free-form feedback to remove the distracting foreground pedestrian and tightly reframe the three musicians.

removing the semantic explanation reward changes crop IoU only marginally. Longer, explicitly grounded chains of thought (Shi et al., 2026) remain an important direction for future work.

Beyond predicting accurate boxes, a practical VLM-based cropper should be able to explain its decisions and respond to a user's intent. Following the qualitative protocol of Venus (Du et al., 2026), we therefore probe CPIC in two settings: (1) after its initial crop, we ask why the crop improves the overall aesthetics; and (2) we provide a natural-language follow-up requesting a change in subject, framing, or crop tightness. Importantly, CPIC is not trained with aesthetic-rationale annotations or multi-turn crop-refinement dialogues. These abilities are evaluated directly, without demonstrations or parameter updates, using the same checkpoint as our quantitative evaluation.

Figure 5 shows both capabilities with one representative example each. For interpretation, CPIC grounds its explanation in the selected subject, spatial arrangement, tonal contrast, and their visual effect. For interaction, its predicted coordinates respond coherently to free-form feedback by removing an unwanted foreground subject and reframing the requested musicians. Thus, CPF-based crop optimization preserves the underlying model's language capabilities while coupling them to continuous spatial decisions, enabling interpretable and personalized refinement in a zero-shot manner.

C Additional Justification

Due to the main-paper page limit, we defer several important discussions and additional justifications to this section. It contains validity analyses that are necessary for a proper assessment of several central claims in this paper, rather than merely additional implementation details. We therefore ask that these results be considered jointly with the main-paper experiments when evaluating the soundness of the proposed framework.

Justification for continuity. Continuous crop optimization requires a coordinate space dense enough to express precise boundary adjustments. A fixed anchor set limits the attainable accuracy even when the best candidate is selected perfectly. Table 8 quantifies this limit by selecting the

Table 8: Oracle IoU on three benchmarks.

Coordinates	GAIC	FLMS	FCDB
12×12 anchors	**1.0000**	0.9099	0.8273
0–1000 integers	0.9994	**0.9997**	**0.9989**

highest-IoU representable box for each test image under the benchmark's ground-truth matching protocol. GAIC's 12×12 anchor template reaches an oracle IoU of 1 on GAIC because its annotations use the same template, but only 0.9099 and 0.8273 on FLMS and FCDB. To remain consistent with the model's pretrained coordinate format, we represent crop boundaries as integers in $[0, 1000]$. Although this representation is not strictly continuous, **its quantization error is negligible**: the near-unit oracle IoU across all three datasets shows that it substantially raises the attainable accuracy beyond the original grid anchors on FLMS and FCDB. Therefore, it is CPF that provides supervision beyond the annotation grid that enables the model to move away from fixed crop templates; we do not claim that the model itself produces strictly continuous coordinates.

Justification for off-lattice refinement. Table 9 compares (a) CPF without refinement, (b) directly using the learned preference score as the reward without refinement, and (c–e) progressively adding aesthetic, composition, and subject-preservation scores for local refinement. The checkmarks indicate scores used for refinement. Relative to the SFT baseline in Table 6(a), direct optimization of the learned preference score yields modest gains; its GAIC and FLMS improvements are smaller than those of the IoU reward in Table 6(b), while FCDB IoU is equal. Our refinement uses human annotations to determine preference modes and anchor the local search, while learned preference, regularized by composition and subject-preservation criteria, distinguishes among nearby off-lattice candidates. **This retains human preference as the foundation and uses visual criteria to improve local targets**, a structure that direct score maximization does not explicitly enforce. Adding composition and subject-preservation regularization in (c–e) improves all three metrics in successive steps. We do not claim to recover the exact human preference landscape; rather, we construct a MOS-grounded approximation and regularize its off-lattice extension to improve alignment with human preference.

Table 9: Refinement-score ablation.

ID	Config.	GAIC	Comp.	Subj.	GAIC SRCC↑	FLMS IoU↑	FCDB IoU↑
(a)	None	–	–	–	0.517	0.829	0.672
(b)	Direct	–	–	–	0.526	0.812	0.671
(c)	Refine	✓	–	–	0.563	0.827	0.704
(d)	Refine	✓	✓	–	0.578	0.828	0.709
(e)	Refine	✓	✓	✓	**0.590**	**0.842**	**0.714**

Justification for sharpness and structured negatives. High geometric overlap does not necessarily imply a good composition. An uncropped image or a fixed, oversized template can retain substantial overlap with the ground-truth crop while leaving distracting content or excessive margins. Conversely, a small boundary shift can cut through a subject's head or another visually critical region while barely changing IoU. A preference field that only rewards proximity to positive crops cannot reliably distinguish these failures from acceptable local variations. Our structured negatives explicitly cover such undesirable crops and shape the field to assign lower scores to them, preserving tolerance around good compositions while penalizing visually damaging changes. Figure 6 demonstrates this distinction: CPF penalizes distracting content, excessive margins, and subject truncation even when IoU remains high. These examples provide compelling qualitative evidence that CPF distinguishes good compositions from perceptually flawed crops in a manner more consistent with human visual preference than geometric overlap alone.

Justification for CPICD. CPICD is intended to provide better reference annotations for future cropping research, rather than to establish CPIC's superiority through evaluation against its own refinement criterion. Our performance claims are already supported by results on the three established benchmarks, GAIC, FLMS, and FCDB, using their original annotations (Table 3). The value of CPICD is instead assessed through human preference: participants favor its recalibrated crops over the original ground truth in Table 5, providing evidence beyond agreement with the construction score. By retaining human annotations as anchors while relaxing their discrete sampling constraints and incorporating composition and subject preservation, CPICD offers a more principled construction of reference crops. Our aim is to make annotations better aligned with human satisfaction available as a shared resource for subsequent research.

D User study details

We conducted a web-based questionnaire with a total of 1,200 human preference evaluations on cropping quality and ground-truth recalibration (Figures 7 and 8). Twenty participants were instructed to select the most aesthetically pleasing compositions based on their first impressions, with no strictly correct answers. The study size is consistent with prior computer vision user studies (Tang

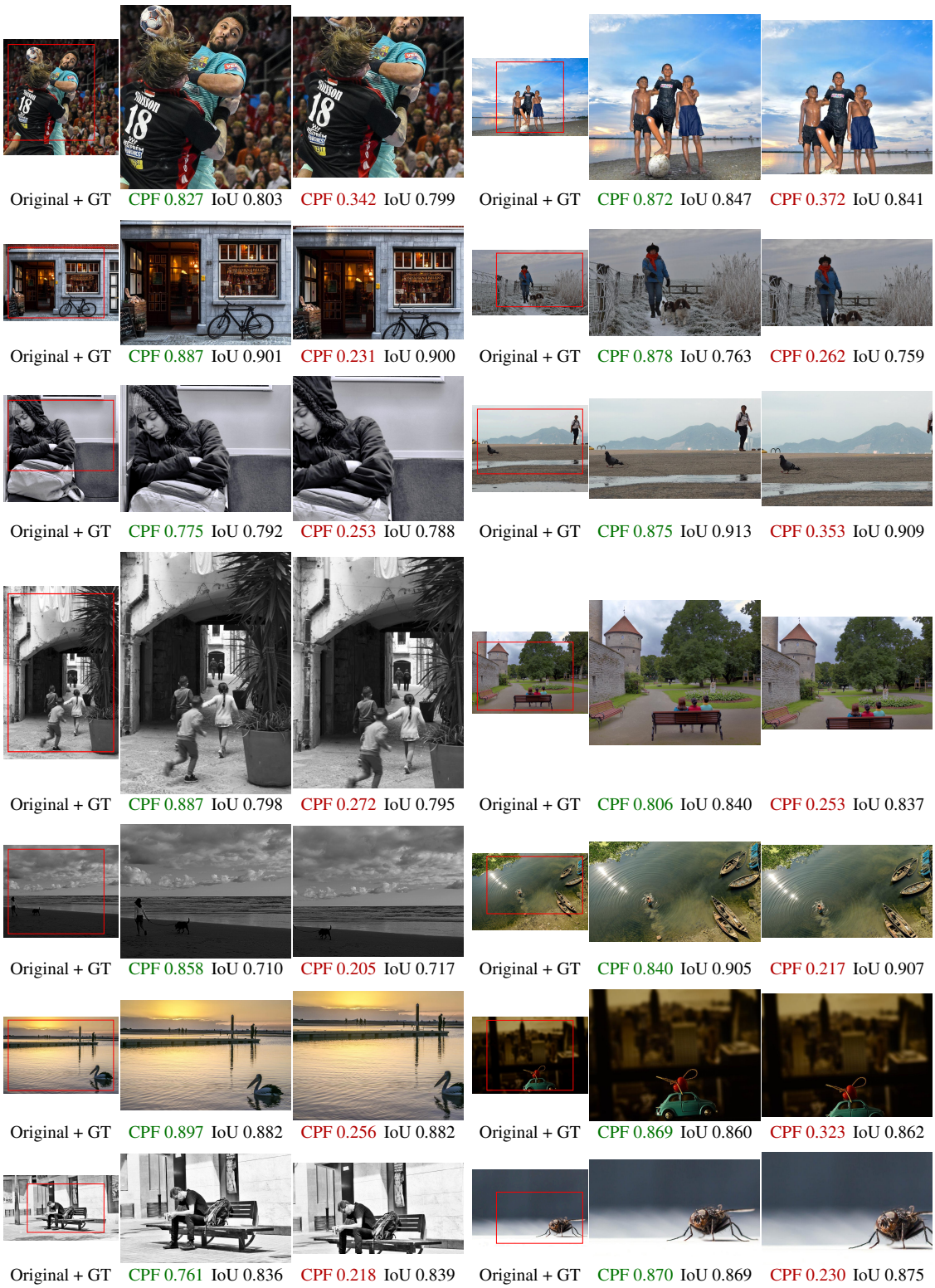

Figure 6: CPF scores and IoU for pairs of crops. CPF distinguishes coherent compositions from visually flawed alternatives that IoU fails to adequately penalize, providing qualitative evidence of its closer alignment with human visual preference.

et al., 2024; Zhang et al., 2020; Lin et al., 2025; Zhang et al., 2026b), while the balanced sampling across datasets and repeated evaluations from multiple participants provide sufficient human judgments for statistically reliable comparisons (Chen et al., 2026; Liu et al., 2026a;b). The evaluation comprised two sub-studies, each utilizing 30 test images evenly drawn from GAIC, FLMS, and FCDB (10 per dataset), yielding a total of 1,200 human preference evaluations. For the recalibration study (Table 5), users were shown the source image and chose the single best crop between the original benchmark annotation and our CPICD box. For the cropping quality study (Table 4), they viewed six anonymized crops (GAIC, CACNet, UNIC, InstructCrop, Venus, and CPIC) and selected their top 1 to 3 choices. To ensure unbiased results, all method names were completely hidden (labeled only A–F), and both the question order and option layouts were randomly shuffled for every participant.

E ADDITIONAL VISUALIZATIONS

To further demonstrate the aesthetic superiority of our method, Figures 9–12 provide additional qualitative comparisons against traditional networks, zero-shot VLMs, and VLM-based croppers using test images from GAIC, FCDB and FLMS. Furthermore, to illustrate the superior quality of our recalibrated benchmark, Figure 13 visualizes supplementary examples (Table 5). These visual pairs highlight instances across all three datasets where our CPICD recalibrated boxes (green) visibly excel the original annotations (red).

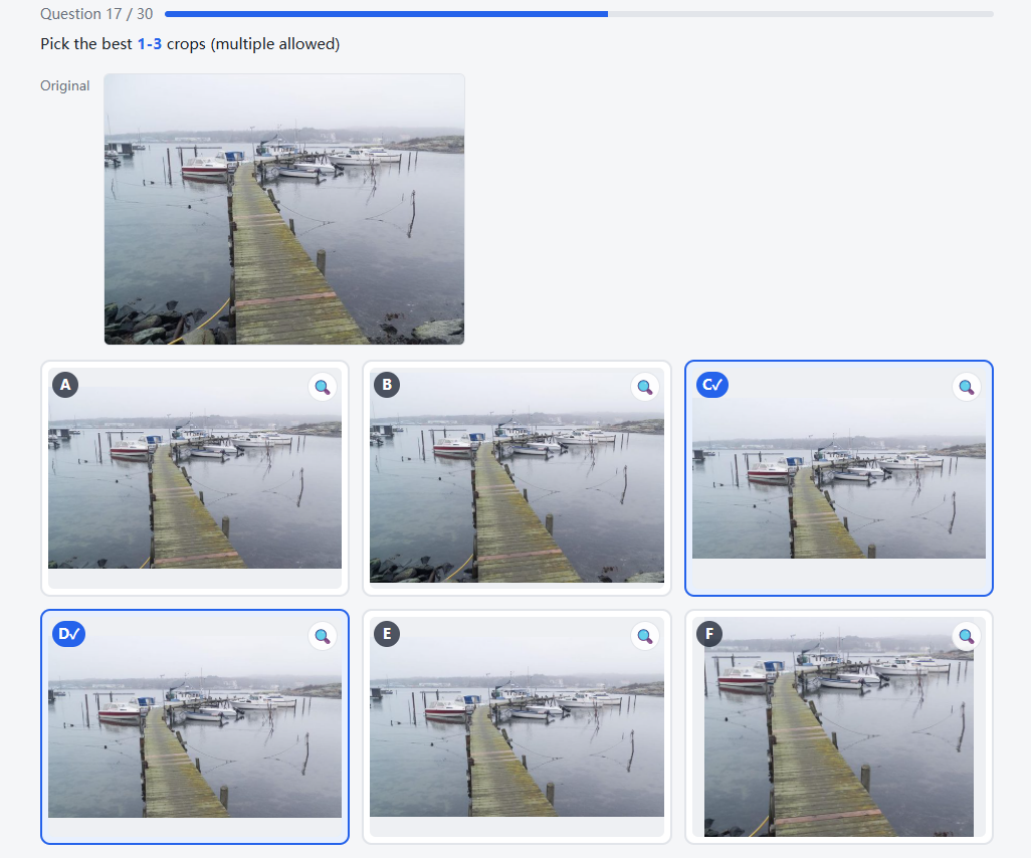

Figure 7: User study interface for cropping quality.

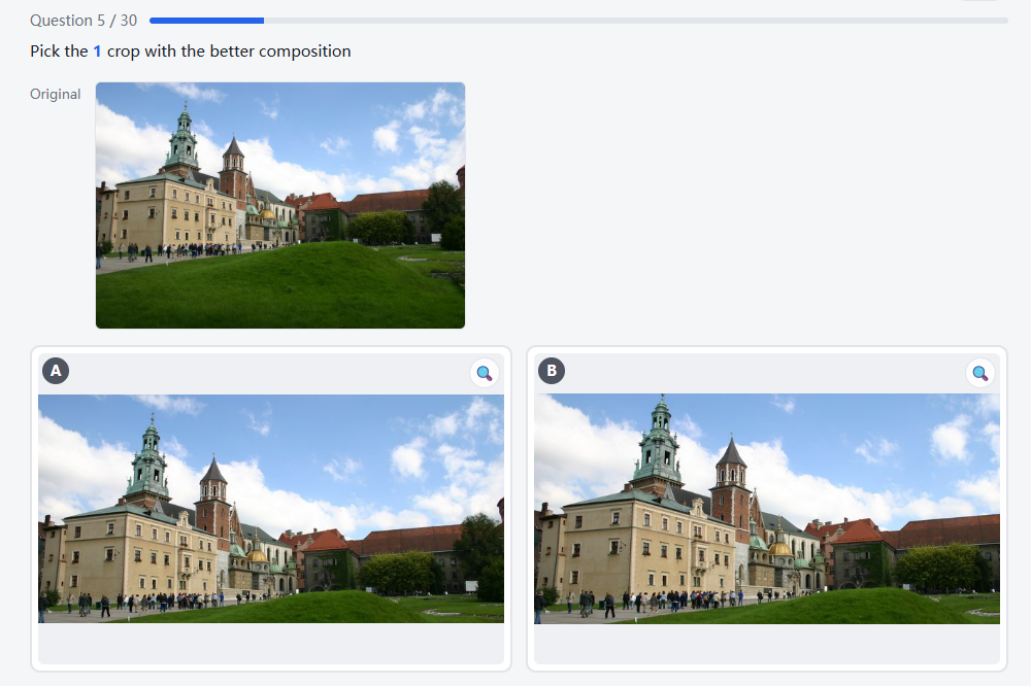

Figure 8: User study interface for ground-truth recalibration.

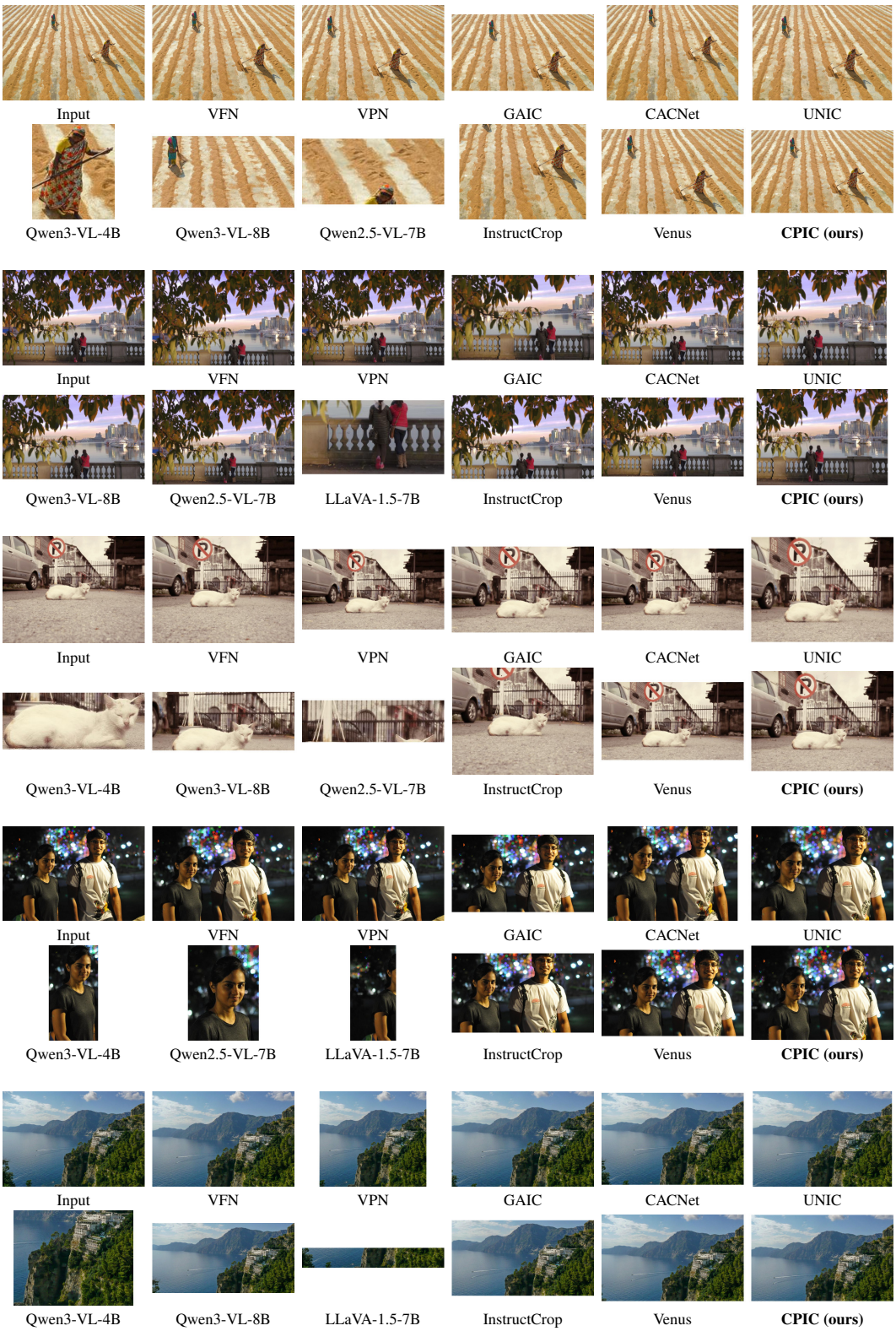

Figure 9: Additional qualitative comparison (1/4).

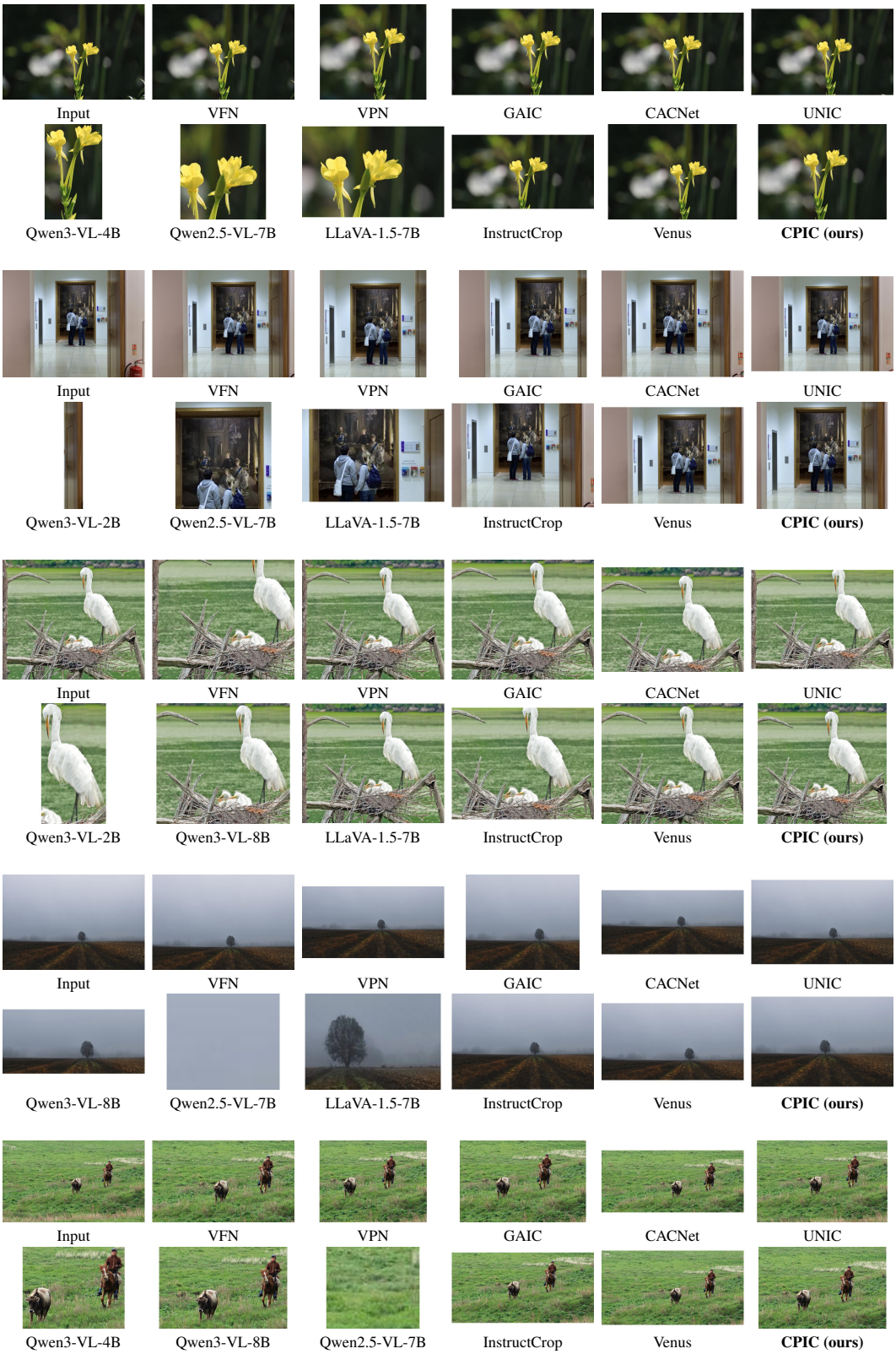

Figure 10: Additional qualitative comparison (2/4).

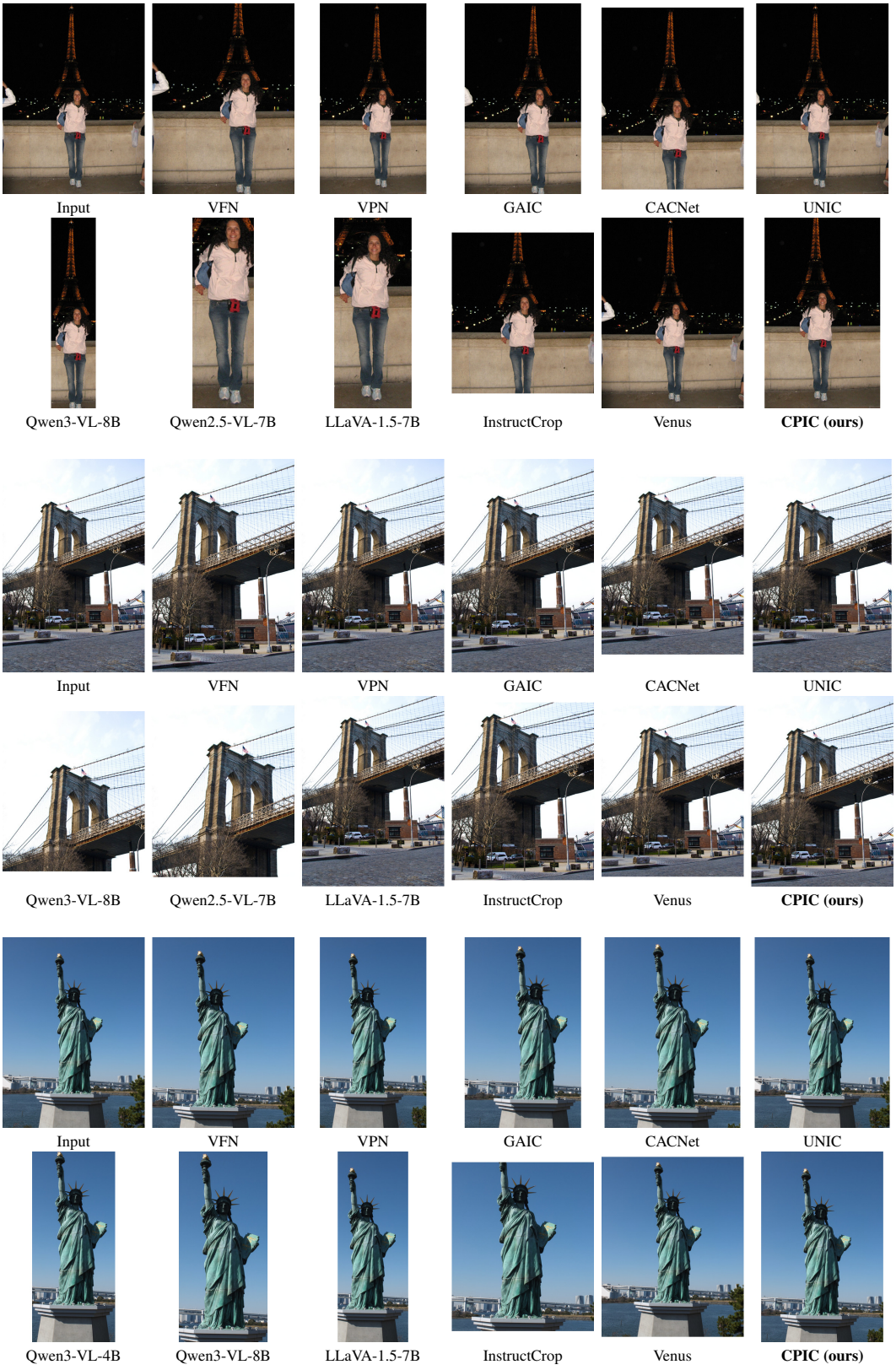

Figure 11: Additional qualitative comparison (3/4).

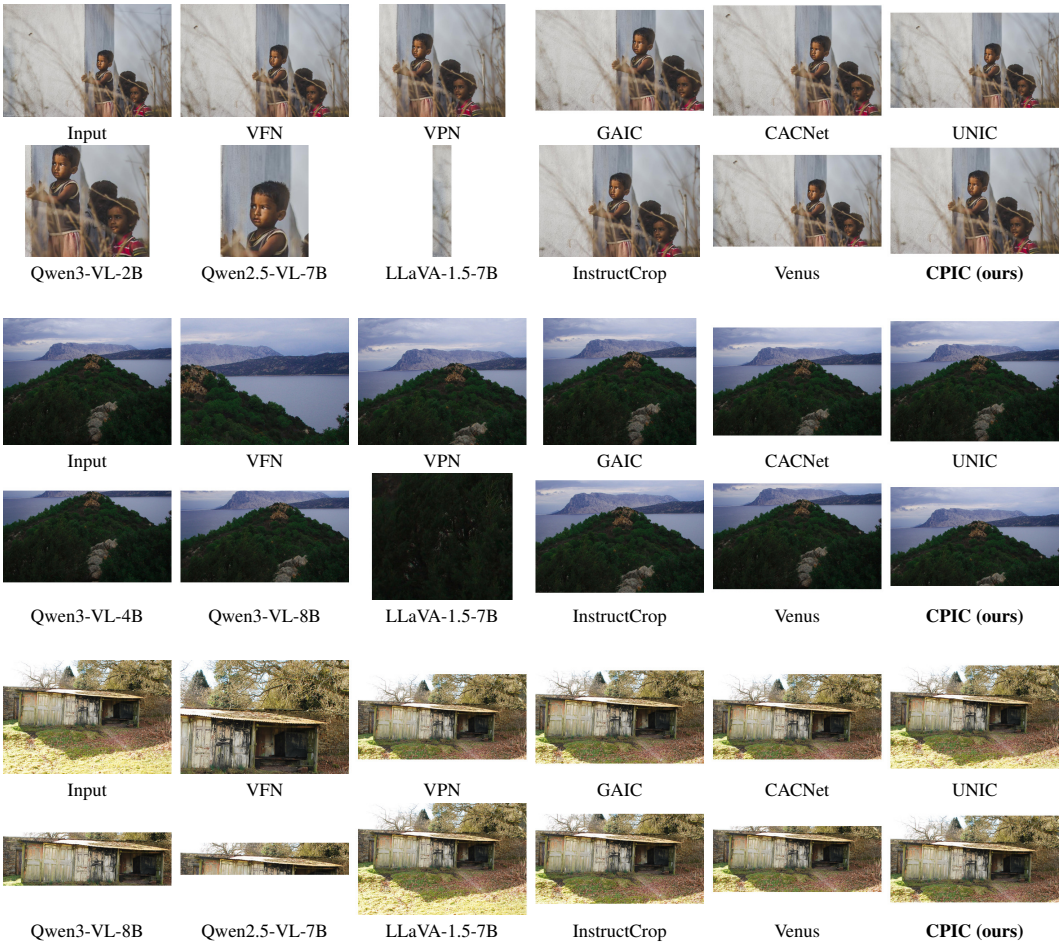

Figure 12: Additional qualitative comparison (4/4).



Figure 13: Original ground truth (red) vs. our recalibrated boxes (green).