MATCHVLM: SYNERGIZING VISION-LANGUAGE MODELS AND LOCAL FEATURE MATCHING FOR CROSS-VIEW POINT CORRESPONDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-View Point Correspondence (**CVPC**) aims to identify the same physical point across different viewpoints, providing precise spatial guidance for embodied interaction. Despite recent progress, vision-language models (VLMs) can confuse similar object instances under viewpoint changes, revealing the difficulty of inferring fine-grained correspondences from global visual context alone. To address this limitation, we propose **MatchVLM**, a collaborative framework that integrates feature matching with a compact VLM, leveraging local correspondence cues to support precise cross-view point prediction. We develop this capability through two training stages. *Reference-Oriented Spatial Perception Learning* jointly supervises reference reliability assessment and answer prediction through supervised fine-tuning. *Distance-Aware Point Refinement* employs Group Relative Policy Optimization (GRPO) with a point-level distance reward that directly reflects prediction error, providing spatial feedback beyond token-level supervision. A curriculum within this stage emphasizes cases that remain challenging after supervised fine-tuning. We further introduce **CVPC-Bench**, comprising 500 manually annotated cross-view point pairs across eight outdoor and two indoor scenes. Experiments demonstrate that MatchVLM-3B achieves the highest overall Score on CrossPoint-Bench among the evaluated models and outperforms both CroPond-3B and CroPond-7B on CVPC-Bench under zero-shot evaluation, supporting its effectiveness in cross-view point correspondence and generalization to diverse scenes.

1 INTRODUCTION

Cross-View Point Correspondence (**CVPC**) aims to establish precise point-level spatial correspondence across different viewpoints (Wang et al., 2025). Such correspondence is important for embodied applications, including robotic manipulation (Ye et al., 2025), navigation (Qiao et al., 2025), and multi-agent collaboration (Wang et al., 2025). As vision-language models (VLMs) are increasingly adopted in these settings, improving their ability to establish accurate cross-view correspondence becomes important for reliable embodied interaction.

To advance this capability, CroPond (Wang et al., 2025) constructs a large-scale CVPC dataset, CrossPoint-378K, and adapts general-purpose VLMs to CVPC through supervised fine-tuning (SFT). While this approach substantially improves CVPC performance, the resulting models can still confuse visually similar instances under viewpoint changes. As illustrated in Figure 1(a), CroPond predicts a point on the wrong lamp, despite identifying the correct object category. This failure suggests that task-specific SFT may leave VLMs reliant on semantic similarity without sufficiently grounding their predictions in cross-view spatial relationships. Target-point supervision specifies where to predict but does not explicitly reveal how surrounding regions correspond across views. Providing local correspondence cues can complement this supervision with visual evidence that helps distinguish similar instances and locate the same physical point. This motivates providing VLMs with additional correspondence cues, so that fine-grained cross-view relationships can be inferred from explicit local visual evidence rather than global semantic similarity alone.

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.

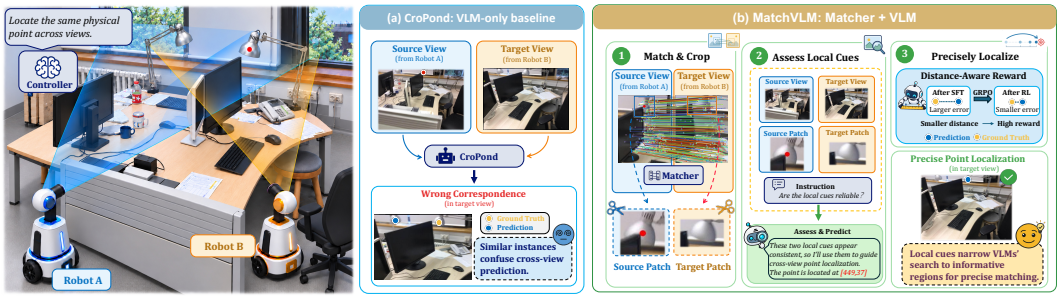

Figure 1: **Comparison between the VLM-only baseline CroPond and our MatchVLM for cross-view point correspondence.** (a) CroPond directly outputs the predicted target point from global views but can confuse semantically similar instances under viewpoint changes. (b) MatchVLM provides local matching cues, jointly learns reference reliability assessment and basic cross-view spatial perception through SFT, and improves localization precision through GRPO training with a distance-based reward. Please zoom in for a clearer view.

Feature matching models provide a natural source of such cues. They are explicitly designed to establish dense or semi-dense correspondences across views. Models such as LoFTR (Sun et al., 2021) and Efficient LoFTR (Wang et al., 2024) can identify multiple correspondences within the neighborhood of a query point, revealing how the surrounding regions align across views. These matches provide spatial context beyond a single annotated point and can be converted into paired local references for the VLM.

Motivated by these observations, we propose **MatchVLM**, a collaborative framework for CVPC, as shown in Figure 1(b). Given a source query point, MatchVLM extracts its local neighborhood and uses a lightweight matcher to identify candidate correspondences in the target view. These correspondences are clustered into a compact target patch and paired with the source patch, providing the VLM with more focused evidence. Meanwhile, since matching results may be imperfect, MatchVLM needs to assess their reliability and selectively exploit useful correspondence evidence for target-point prediction. To enable MatchVLM to exploit reliable matcher cues, reject misleading ones, and produce accurate cross-view point predictions, we adopt a two-stage training strategy. (1) In *Reference-Oriented Spatial Perception Learning*, we jointly supervise reference reliability assessment and final answer prediction through supervised fine-tuning (SFT), establishing selective reference use and basic cross-view spatial perception. (2) In *Distance-Aware Point Refinement*, we further train the model using Group Relative Policy Optimization (GRPO) with a point-level distance reward. By mapping each valid answer to its corresponding coordinates, this reward directly reflects point localization error and provides spatial distance information not explicitly encoded in the token-level cross-entropy objective used during SFT. In addition, we incorporate curriculum learning to focus optimization on challenging cases that remain unresolved after SFT.

Experimental results on CrossPoint-Bench (Wang et al., 2025) demonstrate that MatchVLM-3B outperforms prior VLM baselines, surpassing CroPond-7B with only $\sim 0.4\times$ the model scale. To further examine its generalization to open outdoor environments and unseen indoor scenes, we introduce **CVPC-Bench**, a new benchmark comprising 500 cross-view point pairs meticulously annotated by hand across eight outdoor and two indoor scenes. Without any training or adaptation on this benchmark, MatchVLM-3B still outperforms both CroPond baselines, demonstrating strong zero-shot generalization. Our contributions are summarized as follows:

- **MatchVLM**, an expert-guided collaborative framework for cross-view point correspondence that couples a lightweight matcher with a 3B-parameter VLM, achieving better overall performance than the 7B-parameter CroPond baseline.
- A two-stage training paradigm comprising ***Reference-Oriented Spatial Perception Learning*** and ***Distance-Aware Point Refinement.*** The first stage jointly supervises reference reliability assessment and final answer prediction through SFT, while the second uses GRPO with a point-level distance reward and curriculum learning to develop spatial distance awareness and focus optimization on challenging cases.

- **CVPC-Bench**, a new benchmark comprising 500 cross-view point pairs meticulously annotated by hand across eight outdoor and two indoor scenes. Extensive experiments on CrossPoint-Bench and CVPC-Bench demonstrate that MatchVLM achieves state-of-the-art performance and strong zero-shot generalization.

2 Related Work

Spatial understanding and cross-view correspondence in VLMs. Spatial understanding connects language instructions with visual locations and relationships. RoboPoint (Yuan et al., 2024) learns spatial affordance point prediction from synthetic instruction data, while RoboRefer (Zhou et al., 2025) combines depth-aware perception and reinforcement fine-tuning for spatial referring and multi-step reasoning. Beyond instruction-conditioned localization, cross-view point correspondence requires identifying the same physical point across viewpoints. CroPond (Wang et al., 2025) formalizes CVPC, introduces CrossPoint-Bench for evaluating grounding, visibility, and correspondence, and constructs CrossPoint-378K to develop these capabilities through SFT.

Precise correspondence remains challenging among semantically similar objects, as category recognition does not establish physical identity and final-point supervision leaves local cross-view relationships implicit. MatchVLM provides local correspondence references and supervises their reliability assessment to support selective use of spatial evidence for point prediction.

3 Problem Formulation

From direct answer prediction to reference-assessed correspondence. Following (Wang et al., 2025), Cross-View Point Correspondence (CVPC) aims to identify the same physical point across different viewpoints and involves single-view grounding and two cross-view tasks: visibility assessment and point correspondence. Grounding localizes an instructed target within one image, while the cross-view tasks determine whether a source query point is visible in the target image and identify its corresponding location, respectively.

For query-conditioned cross-view tasks, the input consists of a source image I_s, a target image I_t, a source query point p_s, and an instruction T. Letting $x = (I_s, I_t, p_s, T)$, the original formulation directly predicts the task answer:

$$a = \mathcal{F}_\theta(x). \tag{1}$$

Depending on T, a specifies target coordinates, a candidate-point selection, or a binary judgement about correspondence or visibility. When the queried location is visible in I_t, its ground-truth correspondence p_t^* and p_s are projections of the same physical 3D point.

We reformulate these same tasks by introducing local-reference assessment before answer prediction. Given paired local patches $\mathcal{R} = (I_s^L, I_t^L)$ as candidate correspondence evidence, MatchVLM first predicts a reliability decision z and then generates the original task answer a:

$$\pi_\theta(z, a \mid x, \mathcal{R}) = \pi_\theta(z \mid x, \mathcal{R})\pi_\theta(a \mid x, \mathcal{R}, z). \tag{2}$$

Here, z indicates whether the local references provide reliable evidence for the queried correspondence. This reformulation preserves the original answer space while making reference assessment explicit before prediction. Single-view tasks and cross-view tasks without a designated query point retain their original direct-answer formulation.

4 Methodology

As shown in Figure 2, we build MatchVLM on Qwen2.5-VL-3B (Bai et al., 2025) and train it using CrossPoint-378K (Wang et al., 2025), which contains diverse single-view and cross-view spatial tasks. While target-point supervision specifies the desired location, it does not explicitly describe how the surrounding regions correspond across views. To provide this spatial evidence, a frozen matching expert constructs paired local patches as auxiliary references for the query-conditioned cross-view tasks defined above: visibility assessment and point correspondence.

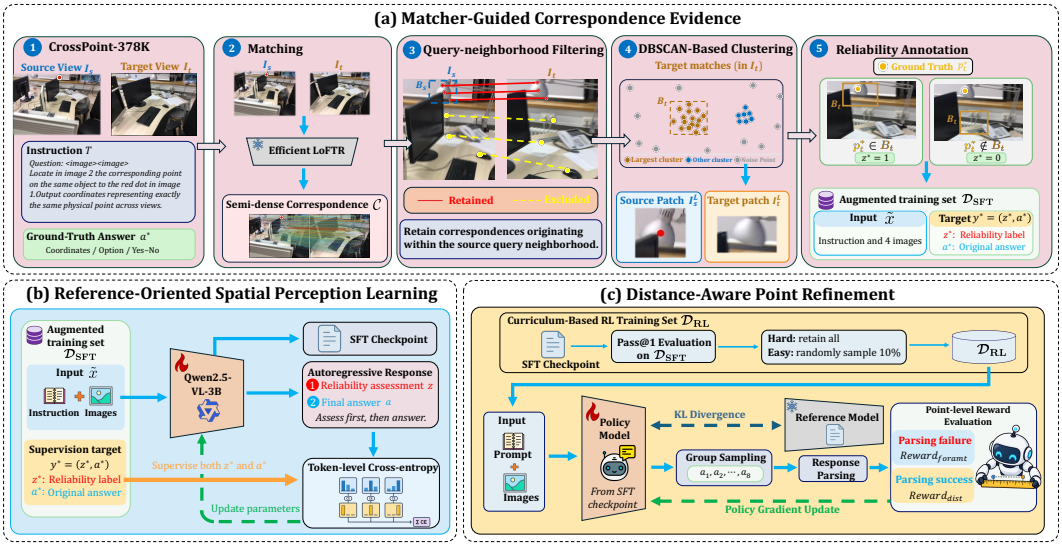

Figure 2: **Data construction and training stages of MatchVLM.** (a) *Matcher-Guided Correspondence Evidence*. A frozen matcher provides local correspondences to construct paired image patches, which augment the training data alongside reference reliability labels. (b) *Reference-Oriented Spatial Perception Learning*. SFT jointly supervises reference reliability assessment and final answer prediction, encouraging selective reference use and establishing basic cross-view spatial perception. (c) *Distance-Aware Point Refinement*. GRPO further refines point prediction with a point-level distance reward that provides continuous feedback on prediction error, while curriculum learning prioritizes examples whose final-answer strings do not exactly match the annotations after SFT.

We adopt two training stages to learn selective reference use and further improve localization precision. In *Reference-Oriented Spatial Perception Learning*, SFT jointly supervises reference reliability assessment and answer prediction for reference-augmented examples, while retaining all other examples with their original inputs and answer supervision. In *Distance-Aware Point Refinement*, GRPO uses a distance-based reward to optimize single-view grounding and point correspondence, with the latter restricted to coordinate prediction and candidate-point selection.

4.1 Matcher-Guided Correspondence Evidence

We construct local references for visibility assessment and point correspondence, both of which involve a designated source query point. Single-view examples retain their original inputs without reference augmentation. We use Efficient LoFTR (Wang et al., 2024) as a frozen matching expert $\mathcal{M}$ for its balance between matching accuracy and computational efficiency. Given the source and target views, it establishes semi-dense correspondences:

$$\mathcal{C} = \mathcal{M}(I_s, I_t) = \{(u_i, v_i)\}_{i=1}^{N}, \tag{3}$$

where u_i and v_i denote matched source and target locations. To obtain evidence relevant to p_s, we retain correspondences whose source locations fall within a neighborhood B_s centered on the query point:

$$\mathcal{C}_{p_s} = \{(u_i, v_i) \in \mathcal{C} \mid u_i \in B_s\}. \tag{4}$$

If $\mathcal{C}_{p_s}$ is empty, no target matches are available for reference construction. We therefore retain the example in its original form for SFT without local references or reliability supervision. Otherwise, we construct the target reference from the retained matches as follows.

To obtain a spatially coherent target reference while reducing the influence of scattered matches, we apply DBSCAN (Ester et al., 1996) to the target endpoints in $\mathcal{C}_{p_s}$ and define the candidate region B_t as the bounding box of the largest non-noise cluster. If no valid cluster is found, we use all retained target endpoints to construct B_t.

Paired local patches I_s^L and I_t^L are obtained by cropping B_s from I_s and B_t from I_t, respectively. With $\mathcal{R} = (I_s^L, I_t^L)$, the augmented input is

$$\tilde{x} = (x, \mathcal{R}), \tag{5}$$

where x denotes the original task input. For the cross-view tasks considered here, $x = (I_s, I_t, p_s, T)$, with p_s visually marked in I_s. For examples without local references, we set $\tilde{x} = x$. Global views provide context, while local patches support reliability assessment and task answering.

4.2 Reference-Oriented Spatial Perception Learning

Matcher-derived references are not always reliable, and using them indiscriminately may mislead answer prediction. We therefore jointly supervise reference assessment and task answering to encourage selective reference use and develop foundational spatial perception.

Construction of reference-augmented training data. For reference-augmented examples, let v^* indicate whether the queried location is visible in the target view. When visible, its ground-truth correspondence p_t^* is obtained from dataset metadata. Given the target reference region B_t, we define

$$z^* = \begin{cases} \mathbb{I}[p_t^* \in B_t], & v^* = 1, \\ 0, & v^* = 0, \end{cases} \tag{6}$$

where $\mathbb{I}[\cdot]$ is the indicator function. Invisible locations receive an unreliable label without requiring a target coordinate. The target response is $y^* = (z^*, a^*)$, placing reference assessment before the original task answer a^*. All other examples, including those without usable matches, retain their original inputs and answer-only targets $y^* = a^*$.

Joint optimization in SFT. Both groups form $\mathcal{D}_{\mathrm{SFT}}$ and share an autoregressive objective:

$$\mathcal{L}_{\mathrm{SFT}} = -\mathbb{E}_{(\tilde{x}, y^*) \sim \mathcal{D}_{\mathrm{SFT}}} \left[\sum_{j=1}^{|y^*|} \log \pi_\theta \left(y_j^* \mid \tilde{x}, y_{<j}^* \right) \right], \tag{7}$$

where $\tilde{x}$ denotes the model input, with local references when available, and π_θ is the VLM's conditional token distribution. This objective retains supervision across all tasks and conditions answer prediction on reference assessment for augmented examples.

4.3 Distance-Aware Point Refinement

Point-level distance reward. SFT establishes reference assessment and foundational spatial perception through annotated responses. However, token-level cross-entropy measures agreement with the target sequence rather than the spatial distance between predicted and ground-truth points. Smaller coordinate errors therefore do not necessarily incur lower training losses. We use Group Relative Policy Optimization (GRPO) with a point-level distance reward to provide explicit spatial feedback for localization refinement. This stage includes single-view grounding and point correspondence with coordinate or candidate-point answers. Binary judgements, including visibility assessment, and other tasks without point-localization targets are excluded.

Let p^* denote the ground-truth localization point: the annotated point in the input image for single-view grounding, or p_t^* in the target view for point correspondence. For each example, $\Phi(a)$ converts the response a into a predicted point. Coordinate responses are parsed directly, while corresponding-point selections are mapped to their candidate coordinates provided by the example. The localization error is

$$d(a) = \|\Phi(a) - p^*\|_2 . \tag{8}$$

Responses that cannot be parsed into valid coordinates receive a format penalty:

$$Reward_{\mathrm{format}}(a) = -1. \tag{9}$$

A valid but incorrect candidate selection is evaluated by its localization error rather than treated as a parsing failure. For successfully parsed responses, the distance reward is

$$Reward_{\mathrm{dist}}(a) = \frac{d_{\mathrm{th}} - d(a)}{d_{\mathrm{th}} + d(a)}, \tag{10}$$

where $d_{\text{th}} = 20$ pixels specifies the error at which the reward equals zero. The final reward is

$$Reward(a) = \begin{cases} Reward_{\text{format}}(a), & \text{if parsing fails,} \\ Reward_{\text{dist}}(a), & \text{otherwise.} \end{cases} \tag{11}$$

The distance reward equals 1 for an exact prediction and decreases continuously toward -1 as the error increases, distinguishing near misses from distant errors.

GRPO with curriculum learning. To focus optimization on cases that remain challenging after SFT, we evaluate the SFT model on the eligible point-localization training subset using Pass@1. An example is considered easy if its extracted final-answer string exactly matches the ground-truth answer string; otherwise, it is considered hard. We retain all hard examples and randomly sample 10% of the easy examples:

$$\mathcal{D}_{\text{RL}} = \mathcal{D}_{\text{hard}} \cup \mathcal{D}_{\text{easy}}^{10\%}. \tag{12}$$

For each input $\tilde{x}$ from $\mathcal{D}_{\text{RL}}$, we sample $G = 8$ responses from π_{old} and standardize their rewards within the group to obtain response-level advantages A_i. These advantages are shared across response tokens in the clipped GRPO objective (Shao et al., 2024):

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}[\min(\rho_{i,j}A_i,\ \text{clip}(\rho_{i,j}, 1-\delta, 1+\delta)A_i)] + \beta\mathcal{L}_{\text{KL}}, \tag{13}$$

where $\rho_{i,j} = \pi_\theta(a_{i,j} \mid \tilde{x}, a_{i,<j}) / \pi_{\text{old}}(a_{i,j} \mid \tilde{x}, a_{i,<j})$ is the token-level probability ratio and δ is the clipping parameter.

5 Experiments

5.1 Datasets

CrossPoint-378K and CrossPoint-Bench. CrossPoint-378K (Wang et al., 2025) contains approximately 378K visual question–answer pairs. Its training tasks extend beyond those evaluated in CrossPoint-Bench to include spatial understanding, cross-view depth variation, and spatial transformation. All task types are retained for SFT, while GRPO is applied only to point-localization tasks. CrossPoint-Bench (Wang et al., 2025) comprises 1,000 examples across fine-grained grounding, visibility reasoning, correspondence judgement, and correspondence pointing, each covering objects and semantic parts.

CVPC-Bench. To evaluate zero-shot generalization to open outdoor environments and additional indoor scenes, we construct CVPC-Bench using images from ETH3D (Schops et al., 2017). Originally developed for multi-view stereo reconstruction, ETH3D provides images of real-world indoor and outdoor environments captured from diverse viewpoints, covering both natural and man-made scenes. This variety makes it a suitable source for evaluating cross-view correspondence beyond a single environment type. Building on these images, we manually annotate 500 cross-view point pairs across eight outdoor and two indoor scenes.

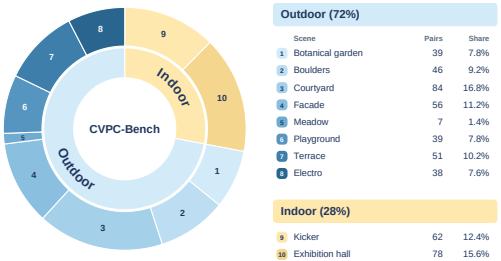

Figure 3: **Overview of CVPC-Bench.** The benchmark includes indoor and outdoor environments, each comprising multiple distinct scenes.

As illustrated in Figure 3, CVPC-Bench covers multiple distinct scenes within each environment category. The outdoor subset spans natural settings, such as botanical garden, boulders, and meadow, as well as built environments, including courtyard, facade, playground, and terrace. The indoor subset complements this coverage with kicker and exhibition hall. Together, these scenes broaden the range of visual contexts for assessing whether learned correspondence capabilities transfer to new environments. We use CVPC-Bench exclusively for evaluation, without additional training or adaptation on its data.

Table 1: **Performance of various models on CrossPoint-Bench.** The best and second-best results in each column, excluding human performance, are highlighted in **bold** and underline respectively.

Model	Score	Fine-grained Ground			Visibility Reason			Correspondence-Judge			Correspondence-Point		
		object	part	sum	object	part	sum	object	part	sum	object	part	sum
Human Evaluation													
Human	91.75	83.57	76.37	79.51	96.15	89.66	92.73	98.49	96.63	97.44	95.87	91.18	93.63
Proprietary Models													
Claude-Sonnet-4	23.10	17.14	7.69	11.80	66.35	50.86	58.18	41.67	37.04	38.46	7.02	3.17	5.18
Gemini-2.5-Pro	37.10	50.00	19.78	32.92	68.27	66.38	67.27	75.00	53.70	60.26	20.25	12.22	16.41
o3	32.60	11.43	6.59	8.70	79.81	71.55	75.45	70.83	60.19	63.46	10.74	9.50	10.15
Open-Source Vision-Language Models													
Qwen2.5-VL-3B-Instruct	25.40	67.14	41.76	52.80	50.96	43.97	47.27	33.33	34.26	33.97	2.89	2.26	2.59
Qwen2.5-VL-7B-Instruct	26.80	65.71	35.16	48.45	50.00	63.79	57.27	50.00	37.04	41.03	0.00	0.00	0.00
Qwen2.5-VL-32B-Instruct	29.10	32.86	14.29	22.36	56.73	58.62	57.73	47.92	40.74	42.95	19.83	5.88	13.17
Qwen2.5-VL-72B-Instruct	36.50	65.71	46.15	54.66	49.04	58.62	54.09	45.83	51.85	50.00	23.14	10.86	17.28
Qwen3-VL-30B-A3B-Instruct	28.10	55.71	26.37	39.13	66.35	52.59	59.09	50.00	41.67	44.23	4.55	3.62	4.10
Qwen3-VL-30B-A3B-Thinking	39.70	72.86	38.46	53.42	49.04	59.48	54.55	45.83	45.37	45.51	32.23	19.00	25.92
Qwen3-VL-235B-A22B-Instruct	35.20	70.00	52.75	60.25	59.62	56.90	58.18	79.17	54.63	62.18	7.85	4.98	6.48
Qwen3-VL-235B-A22B-Thinking	52.70	74.29	**58.24**	**65.22**	65.38	59.48	62.27	77.08	61.11	66.03	50.83	26.70	39.31
Spatial / Referring Specialist Models													
RoboPoint-v1-Vicuna-13B	14.60	7.14	7.69	7.45	33.65	41.38	37.73	37.50	25.93	29.49	0.83	1.36	1.08
RoboRefer-8B-SFT	21.00	62.86	32.97	45.96	37.50	37.07	37.27	22.92	31.48	28.85	1.65	2.26	1.94
RoboBrain2.0-7B(no-thinking)	25.40	64.29	41.76	51.55	46.15	55.17	50.91	39.58	33.33	35.26	0.41	1.36	0.86
RoboBrain2.0-7B(thinking)	26.00	72.86	40.66	54.66	46.15	52.59	49.55	29.17	41.67	37.82	0.41	1.36	0.86
Molmo-7B-D	25.80	62.86	36.26	47.83	49.04	51.72	50.45	37.50	35.19	35.90	4.13	1.81	3.02
CroPond-3B	71.60	**78.57**	43.96	59.01	79.81	65.52	72.27	91.67	70.37	76.92	**84.30**	62.44	73.87
CroPond-7B	76.60	71.43	51.65	60.25	81.73	76.72	79.09	91.67	84.26	86.54	82.64	**72.40**	**77.75**
MatchVLM-3B (Ours)	**77.90**	**78.57**	48.35	61.49	**84.62**	**79.31**	**81.82**	**97.92**	**87.04**	**90.38**	**84.30**	70.14	77.54

Table 2: **Zero-shot performance on CVPC-Bench by scene.** Mean EPE is reported in pixels (↓). Aggregate results are sample-weighted over 360 outdoor and 140 indoor image pairs.The best and second-best results are highlighted in **bold** and underline, respectively.

Model	Outdoor									Indoor			Overall
	Botanical garden	Boulders	Courtyard	Facade	Meadow	Playground	Terrace	Electro	All	Kicker	Exhibition hall	All	
CroPond-3B	17.64	8.72	17.39	30.79	**5.10**	61.87	28.22	68.25	29.88	41.65	63.93	54.07	36.65
CroPond-7B	18.02	10.51	17.95	21.96	11.69	34.64	27.62	**17.78**	20.67	35.25	53.46	45.40	27.59
MatchVLM-3B (Ours)	**13.52**	**7.11**	**16.04**	**17.50**	6.45	**24.16**	**18.34**	25.66	**16.89**	**19.22**	**50.39**	**36.59**	**22.40**

5.2 Evaluation Metrics

Overall accuracy. Following CrossPoint-Bench (Wang et al., 2025), multiple-choice predictions are evaluated by answer correctness, while point predictions are evaluated by whether they fall within the ground-truth masks. We report Score as the percentage of correct predictions over all test examples, with unparseable responses counted as incorrect.

Point prediction error. On CVPC-Bench, we report mean endpoint error (Mean EPE), the average Euclidean distance between predicted and ground-truth target points:

$$\text{Mean EPE} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{p}_{t,i} - p^{*}_{t,i}\right\|_2, \tag{14}$$

where $N = 500$. All evaluated models produce successfully parsed predictions for all 500 examples, with no samples excluded. Errors are measured in pixels in the original target-image coordinate system.

5.3 Implementation details

Please refer to Appendix A

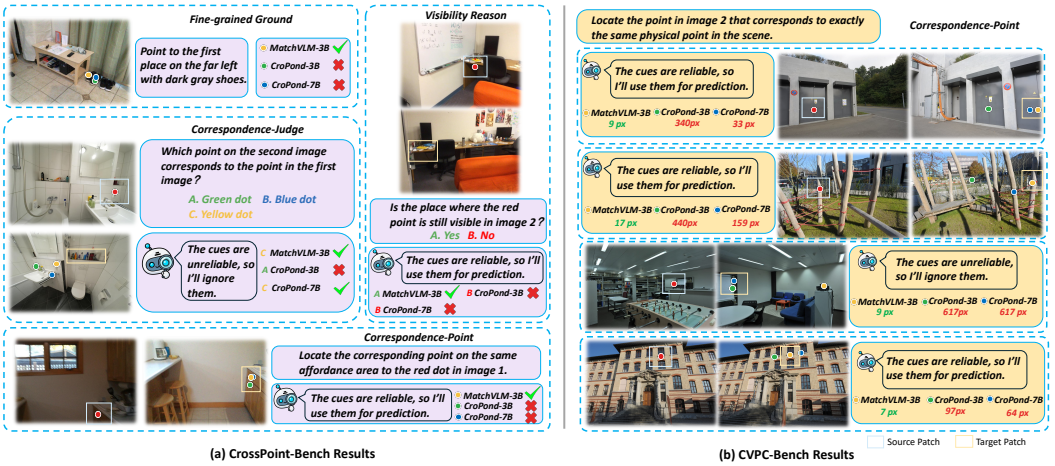

Figure 4: **Qualitative results.** Our method is compared with CroPond-3B and CroPond-7B (Wang et al., 2025) on CrossPoint-Bench and our proposed CVPC-Bench. Evaluation on CVPC-Bench is conducted on unseen outdoor scenes and additional indoor scenes in a zero-shot setting.

Table 3: **Effects of reference-augmented SFT and GRPO on CrossPoint-Bench.** Original denotes the unmodified CrossPoint-378K dataset. Reference-augmented denotes its version with local references and reliability labels. Local references are provided at inference only for models trained on the reference-augmented dataset.

SFT dataset	GRPO	Score	Fine-grained Ground	Visibility Reason	Correspondence-Judge	Correspondence-Point
Original	✗	71.60	59.01	72.27	76.92	73.87
Original	✓	75.10	60.25	79.09	85.90	74.73
Reference-augmented	✗	76.60	59.01	81.36	87.82	76.67
Reference-augmented	✓	77.90	61.49	81.82	90.38	77.54

5.4 Main Results

Quantitative results. We evaluate MatchVLM's overall CVPC performance on CrossPoint-Bench and its zero-shot generalization on CVPC-Bench. On CrossPoint-Bench, we compare against general-purpose VLMs and spatial specialist models, including CroPond-3B and CroPond-7B (Wang et al., 2025) as the most directly related baselines. As shown in Table 1, MatchVLM-3B achieves the highest overall Score among the evaluated models, outperforming both CroPond variants without scaling the VLM to 7B. This demonstrates the effectiveness of our framework for improving CVPC performance with a compact VLM. On CVPC-Bench, we further compare against both CroPond baselines on unseen outdoor scenes and additional indoor scenes, without additional training or adaptation. As reported in Table 2, our method achieves lower mean EPE on both the outdoor and indoor subsets and delivers the best results in most individual scenes. These results support its generalization across environments beyond the task-specific training data.

Qualitative results. Figure 4 presents representative comparisons with CroPond-3B and CroPond-7B on CrossPoint-Bench and CVPC-Bench. The illustrated cases highlight MatchVLM's ability to distinguish visually similar instances and predict points closer to the ground truth, including across diverse indoor and outdoor scenes from CVPC-Bench. Reliable local cues provide fine-grained visual evidence for resolving cross-view ambiguity, complementing distance-aware refinement through GRPO. The examples also show that MatchVLM can identify unreliable references while still producing correct localizations with access to the global views. These qualitative observations illustrate the benefits of synergizing a matcher with a VLM for precise cross-view point correspondence.

Table 4: **Ablation study on matcher selection.** LightGlue (Lindenberger et al., 2023) uses SuperPoint (DeTone et al., 2018) features, while OmniGlue (Jiang et al., 2024) combines SuperPoint features with DINOv2 (Oquab et al., 2023) guidance. We compare these pipelines with Efficient LoFTR (Wang et al., 2024). FPS and crop success rate are measured on CrossPoint-Bench's two-image matching tasks eligible for reference construction. FPS covers feature extraction, matching, and patch construction; crop success rate is the percentage of eligible examples yielding valid paired patches. Score after SFT is evaluated on the full benchmark. The best and second-best results are highlighted in **bold** and underline, respectively.

Matching Pipeline	FPS ↑	Crop Success (%) ↑	Score after SFT ↑
LightGlue	**30.62**	69.1	74.20
OmniGlue	0.53	**86.1**	<u>74.80</u>
Efficient LoFTR	<u>30.12</u>	<u>82.5</u>	**76.60**

5.5 Ablation Study

Component and training stage ablation. Table 3 compares original and reference-augmented SFT with and without GRPO.

(1) Reference-augmented SFT improves cross-view spatial understanding. Without GRPO, reference-augmented SFT outperforms SFT on the original dataset in overall Score, with improvements in visibility reasoning, correspondence judgement, and correspondence pointing. Fine-grained grounding remains unchanged. These results support the effectiveness of the reference-augmented training and inference scheme for cross-view tasks, although they do not isolate the contribution of reliability supervision from that of additional visual evidence.

(2) GRPO improves both SFT configurations. GRPO increases overall Score and performance across all four tasks under both SFT configurations. The reference-augmented model retains its advantage in overall Score and the three cross-view tasks after GRPO, indicating that distance-based refinement complements rather than replaces reference augmentation. Although visibility reasoning also improves, it receives no direct RL supervision.

(3) The combined configuration performs best. Reference-augmented SFT followed by GRPO achieves the highest overall Score and the best performance on all four tasks. The results support combining reference-oriented spatial learning with subsequent distance-aware refinement.

Matcher ablation. Table 4 compares reference construction efficiency and downstream performance. LightGlue achieves the highest processing speed but the lowest crop success rate, suggesting that its sparse matches limit local reference availability. OmniGlue achieves the highest crop success rate at substantially lower processing speed, without yielding the best downstream performance. In contrast, Efficient LoFTR operates at a speed close to LightGlue while providing a higher crop success rate and the best Score after SFT. These results support its favorable balance between reference availability, computational efficiency, and downstream accuracy.

More ablation results. Additional hyperparameter ablations and analyses of reference reliability and utility are provided in Appendix B.

6 Conclusion

We introduced MatchVLM, a collaborative framework that combines a lightweight visual matcher with a compact VLM for cross-view point correspondence. MatchVLM assesses matcher-derived local references to support point prediction, with reference-oriented supervised learning and distance-aware reinforcement learning developing this capability. We further constructed CVPC-Bench to evaluate zero-shot generalization across indoor and outdoor scenes. Experiments on CrossPoint-Bench and CVPC-Bench demonstrate the effectiveness and generalization of MatchVLM. Our work highlights the potential of combining local correspondence evidence with vision-language reasoning for precise cross-view spatial understanding.

7 AI USAGE DISCLOSURE

Generative AI tools were used to assist with code checking, language polishing, and suggestions on manuscript organization. They did not contribute to methodology development, experimental design, hypothesis formulation, implementation of the proposed method, or interpretation of experimental results. All AI-assisted outputs were manually reviewed and verified by the authors, who assume full responsibility for the final manuscript, including its text, claims, code, and any other AI-assisted content.

REFERENCES

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, 2025. URL https://arxiv.org/abs/2502.13923.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *2018 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 337–33712. IEEE, 2018.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pp. 226–231, 1996.

Hanwen Jiang, Arjun Karpur, Bingyi Cao, Qixing Huang, and André Araujo. Omniglue: Generalizable feature matching with foundation model guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19865–19875, 2024.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 17627–17638, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Yanyuan Qiao, Haodong Hong, Wenqi Lyu, Dong An, Siqi Zhang, Yutong Xie, Xinyu Wang, and Qi Wu. Navbench: Probing multimodal large language models for embodied navigation. *arXiv preprint arXiv:2506.01031*, 2025.

Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3260–3269, 2017.

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.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8922–8931, 2021.

Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. Efficient loftr: Semi-dense local feature matching with sparse-like speed. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21666–21675, 2024.

Yipu Wang, Yuheng Ji, Yuyang Liu, Enshen Zhou, Ziqiang Yang, Yuxuan Tian, Ziheng Qin, Yue Liu, Huajie Tan, Cheng Chi, et al. Towards cross-view point correspondence in vision-language models. *arXiv preprint arXiv:2512.04686*, 2025.

Angen Ye, Zeyu Zhang, Boyuan Wang, Xiaofeng Wang, Dapeng Zhang, and Zheng Zhu. VLA-R1: Enhancing reasoning in vision-language-action models. *arXiv preprint arXiv:2510.01623*, 2025.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721*, 2024.

Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. *arXiv preprint arXiv:2506.04308*, 2025.

A Implementation Details

Matcher and local references. We use the full configuration of Efficient LoFTR (Wang et al., 2024) with its pretrained outdoor checkpoint, keeping the matcher frozen in FP16. We retain matches within a 140×140-pixel neighborhood centered on the source query point. DBSCAN (Ester et al., 1996) is applied to the retained target endpoints with $\varepsilon = 120$ pixels and $\mathrm{min_samples} = 3$. The largest non-noise cluster defines the target crop; if no valid cluster is found, all retained target endpoints are used.

Supervised fine-tuning. We perform full-parameter fine-tuning on $4\times$ A800 GPUs. The model is trained for one epoch with a learning rate of 10^{-5}, a cosine learning-rate schedule, 100 warmup steps, and a weight decay of 0.1. We use an effective batch size of 16, BF16 precision, and a maximum sequence length of 4,096 tokens. Reference reliability assessment and the original task answer are jointly supervised in a single autoregressive response.

Reinforcement learning. We initialize the policy from the SFT checkpoint and retain a frozen copy for KL regularization with coefficient $\beta = 0.001$. Full-parameter GRPO training runs for one epoch with a learning rate of 10^{-6}. RL includes single-view grounding and point correspondence with coordinate or candidate-point answers. Binary judgements, including visibility assessment, and other non-localization tasks are excluded. Within this eligible subset, curriculum selection retains all examples that fail the final-answer string exact-match criterion and randomly samples 10% of those that satisfy it.

Training uses four GPUs with a per-device batch size of 16, one gradient accumulation step, BF16 precision, and DeepSpeed ZeRO-1. We sample $G = 8$ responses per input at temperature 1.0, with a maximum response length of 64 tokens. Following Equation 13, token-level probability ratios are clipped to $[1-\delta, 1+\delta]$ with $\delta = 0.2$.

Coordinate responses are parsed directly, while corresponding-point selections are mapped to the coordinates of the selected candidates. Both use the distance-based reward with $d_{\mathrm{th}} = 20$ pixels. Responses that cannot be parsed into valid coordinates receive a penalty of -1.

B More Ablation Results

B.1 Hyperparameter ablation.

We investigate three hyperparameters governing local reference construction, distance-based optimization, and training sample composition. Each parameter is varied while the others remain at their default values. Table 5 reports the overall Score and per-task accuracy on CrossPoint-Bench. Neighborhood sizes are evaluated after SFT, whereas reward scales and retention ratios are evaluated after GRPO.

Query-neighborhood size. A moderate neighborhood size provides effective local references for cross-view correspondence (Table 5(a)). With $L = 80$, fewer matched keypoints are retained, which can hinder target-side clustering and patch construction under large viewpoint changes. Increasing L to 200 produces larger patches that may include redundant visual information not focused on the query point. Both settings yield lower visibility reasoning and correspondence performance than $L = 140$, while grounding accuracy remains unchanged. We therefore use $L = 140$ pixels to balance sufficient matching evidence with focused local references.

Distance reward scale. The reward scale d_{th} controls sensitivity to localization error (Table 5(b)). With $d_{\mathrm{th}} = 10$, reward differences are more pronounced near the correct point but become compressed at larger errors, potentially limiting feedback for poorly localized predictions. Increasing d_{th} to 30 reduces sensitivity near zero error and yields lower correspondence performance than the intermediate setting. In contrast, $d_{\mathrm{th}} = 20$ achieves the best overall and correspondence-pointing performance. We therefore use $d_{\mathrm{th}} = 20$ pixels to balance feedback for larger errors with fine-grained refinement.

Table 5: **Hyperparameter ablation of MatchVLM.** Overall Score and per-task accuracy on CrossPoint-Bench. We vary one hyperparameter at a time. Shaded rows indicate the default configurations. The best and second-best results within each group are highlighted in **bold** and underline, respectively, with ties receiving the same formatting.

Configuration	Score	Fine-grained Ground	Visibility Reason	Correspondence-Judge	Correspondence-Point
(a) Query-neighborhood side length L					
$L = 80$ px	74.40	**59.01**	78.64	80.77	75.59
$L = 140$ px	**76.60**	**59.01**	**81.36**	**87.82**	**76.67**
$L = 200$ px	74.60	**59.01**	79.09	82.05	75.38
(b) Distance reward scale d_{th}					
GRPO, $d_{\text{th}} = 10$ px	77.20	60.87	81.36	**90.38**	76.46
GRPO, $d_{\text{th}} = 20$ px	**77.90**	**61.49**	**81.82**	**90.38**	**77.54**
GRPO, $d_{\text{th}} = 30$ px	77.40	**61.49**	**81.82**	89.74	76.67
(c) Easy-sample retention ratio ρ					
GRPO, $\rho = 0\%$	76.10	60.87	80.00	87.82	75.59
GRPO, $\rho = 10\%$	**77.90**	**61.49**	**81.82**	**90.38**	**77.54**
GRPO, $\rho = 20\%$	77.30	**61.49**	81.36	89.10	76.89

Table 6: **Reference reliability assessment on CrossPoint-Bench.** Constructing paired local references yields 608 usable examples from Correspondence-Pointing and Correspondence-Judgement. For Pointing, only ground-truth masks are available, so crop–mask intersection serves as an evaluation proxy for reliability rather than the point-containment criterion used in training. For Judgement, reliability requires the target crop to contain the correct candidate's center. Precision and recall treat reliable references as the positive class. All metrics are reported in percentages. The best and second-best results are highlighted in **bold** and underline, respectively.

Model	Precision ↑	Recall ↑	Macro-F1 ↑
CroPond-3B	86.17	55.14	52.48
MatchVLM-3B after SFT	**97.40**	92.39	88.31
MatchVLM-3B after GRPO	97.22	**93.62**	**89.30**

Easy-sample retention. The retention ratio ρ balances hard-case optimization with continued exposure to examples already solved after SFT (Table 5(c)). With $\rho = 0\%$, training focuses exclusively on hard examples and yields lower performance across all tasks than retaining a small proportion of easy examples. Increasing ρ to 20% still outperforms hard-only training but provides no further benefit over 10%, suggesting that a larger easy-sample proportion is less effective for the intended hard-case focus. The intermediate setting $\rho = 10\%$ achieves the best overall performance. We therefore retain 10% of easy examples to complement hard-case optimization.

B.2 Analyses of reference reliability and utility

We examine whether MatchVLM can assess reference reliability and whether providing local references improves its final predictions. Both analyses use the same 608 CrossPoint-Bench examples with paired local references and valid reliability labels.

Reference reliability assessment. Table 6 shows that MatchVLM-3B after SFT substantially outperforms CroPond-3B in precision, recall, and Macro-F1. The higher precision indicates fewer unreliable references being accepted, while the higher recall indicates better recognition of reliable references. These results are consistent with the intended role of reference-oriented SFT in developing reliability assessment. After GRPO, recall and Macro-F1 improve further with a slight decrease in precision, suggesting that distance-aware refinement preserves this ability.

Utility of local references. Table 7 examines the effect of reference-guided inputs on MatchVLM-3B after SFT. The results show that reliable references improve final-task accuracy, supporting the value of local correspondence evidence. The unreliable-reference subset also exhibits a positive

Table 7: **Effects of reliable and unreliable references.** The same MatchVLM-3B checkpoint after SFT is evaluated under two conditions. *w/o References* uses the two global images and the original task question. *w/ References* additionally provides two local patches and prompts reliability assessment before answering. Groups are defined by ground-truth reference reliability. We report final-task accuracy (%) and changes in percentage points (pp), computed from unrounded scores.

Reference quality	Samples	w/o References	w/ References	Δ (pp)
Reliable	486	71.60	80.45	+8.85
Unreliable	122	63.11	68.85	+5.74
Overall	608	69.90	78.13	+8.22

Table 8: **Task performance and inference cost.** We report overall Score on CrossPoint-Bench, Mean EPE on CVPC-Bench, inference latency, and peak GPU memory. MatchVLM-3B costs include reference construction and VLM inference. The best and second-best results are highlighted in **bold** and underline, respectively.

Model	CrossPoint-Bench Score ↑	CVPC-Bench Mean EPE (px) ↓	Latency (ms) ↓	Peak GPU memory (GiB) ↓
CroPond-3B	71.60	36.65	**197.6**	**7.405**
CroPond-7B	76.60	<u>27.59</u>	253.0	15.867
MatchVLM-3B (Ours)	**77.90**	**22.40**	<u>222.3</u>	<u>8.441</u>

aggregate change rather than an overall performance drop. Reliable references account for most of the net improvement across all examples. These results support the effectiveness of our reference-guided prediction scheme after SFT.

C Analysis of Inference Efficiency

We evaluate inference latency and peak GPU memory to assess the computational cost of MatchVLM-3B relative to CroPond-3B and CroPond-7B.

Experimental setup. Measurements are conducted on the same GPU with batch size 1 using 20 CrossPoint-Bench examples covering all four tasks. Each example is measured three times. The VLMs use BF16 with Transformers SDPA and KV caching, generating up to 768 tokens with early stopping at the end-of-sequence token. Efficient LoFTR uses its reparameterized full configuration in FP16 and processes the 15 two-image examples. MatchVLM-3B is profiled across reference construction and VLM inference. Peak GPU memory includes model weights and inference tensors, excluding unused allocator-reserved memory and transient model-loading overhead.

Results. Table 8 shows that while MatchVLM-3B introduces a modest latency and memory overhead relative to the base CroPond-3B, it remains significantly more lightweight than CroPond-7B. Crucially, by outperforming CroPond-7B in overall task accuracy at a substantially lower computational cost, MatchVLM-3B achieves a strictly superior accuracy-efficiency trade-off. These results demonstrate that explicit matcher–VLM collaboration provides a more effective scaling paradigm for high-fidelity correspondence than naively increasing VLM parameter count.