ANYPOINT: UNPACKING CROSS-VIEW UNDERSTANDING GRANULARITY IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review



Figure 1: **Motivation and overview.** (a-c) Existing protocols are global but coarse, discrete, and imprecise, or precise but local. (d) AnyPoint-Bench jointly localizes arbitrary pixels in one turn across diverse scenes and query settings. (e) Our CorrVLM-8B reveals substantial headroom beyond general-purpose VLMs.

ABSTRACT

Fine-grained cross-view understanding requires Vision-Language Models (VLMs) to localize arbitrary pixel locations across changes in viewpoint and appearance. This capability is fundamental to spatial reasoning and embodied AI. Existing benchmarks rely on discrete candidate selection or isolated pixel queries. However, discrete candidate selection limits localization precision, while isolated pixel queries yield query-dependent scores and fail to assess global understanding. To address these limitations, we introduce AnyPoint-Bench, the first benchmark to formulate fine-grained cross-view understanding as the cross-view localization of arbitrary pixels. Given an image pair, a VLM jointly localizes an arbitrary set of reference-view pixels by predicting their coordinates in the target view. This unified protocol evaluates both fine-grained localization and global understanding. AnyPoint-Bench covers diverse scenes, viewpoint and appearance changes, and query configurations. Our evaluation reveals a pronounced gap between proprietary and open-source VLMs. GPT-5.6-Sol achieves 36.07% accuracy, compared with 17.19% for Qwen3.8. To probe the potential of the VLM paradigm, we develop CorrVLM, a cross-view specialist trained with vocabulary repurposing and coarse-to-fine learning. Despite using substantially fewer image pairs than specialized visual-geometric models, CorrVLM matches their performance at moderate thresholds and surpasses GPT-5.6-Sol by 31.28 percentage points. These results demonstrate that strong fine-grained cross-view understanding is attainable within the VLM paradigm, yet remains largely unrealized in current general-purpose VLMs. AnyPoint-Bench offers a new lens for evaluating this capability and advancing fine-grained cross-view understanding in VLMs. Our benchmark, code, and models will be publicly available.

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.

1 INTRODUCTION

Cross-view understanding in Vision-Language Models (VLMs) is the ability to associate visual content across changes in viewpoint and appearance. It is a cornerstone of spatial reasoning Yang et al. (2025); Li et al. (2026) and embodied AI Huang et al. (2023); Feng et al. (2026). Recent vanilla VLMs have shown promising capabilities in 3D scene understanding Ouyang et al. (2025) and dynamic reasoning Li et al. (2025); Yin et al. (2026). However, their cross-view ability is still evaluated mainly through view-level spatial judgments, such as camera motion and relative viewpoint Yang et al. (2025); Yeh et al. (2026b); Wu et al. (2026). As shown in Fig. 1 (a), these tasks capture global cross-view relations but do not measure fine-grained localization. Recently, task-specific studies show that vanilla VLMs can learn pixel-level depth Yeh et al. (2026a) and correspondence Cai et al. (2026). However, these studies characterize models adapted to specific tasks rather than general-purpose VLMs. It therefore remains unclear to *what extent current general-purpose VLMs possess fine-grained cross-view understanding*.

As illustrated in Fig. 1 (b-c), existing fine-grained evaluation can be broadly grouped into marker-based and coordinate-based protocols. Marker-based protocols restrict predictions to predefined candidates Fu et al. (2024); Zhong et al. (2026). Although multiple markers can probe correspondences across the image pair, discrete candidate selection limits localization precision. Coordinate-based protocols instead directly predict pixel locations Xu et al. (2026); Wang et al. (2025); Wu et al. (2026), enabling precise localization. However, they typically query one pixel at a time, so each prediction evaluates only an isolated local correspondence and cannot assess global cross-view understanding. As shown in Fig. 2, our attention visualization further suggests that multi-pixel queries induce more spatially distributed cross-view attention than a single-pixel query. Moreover, covering broader regions with single-pixel queries requires repeated inference and makes evaluation sensitive to pixel sampling.

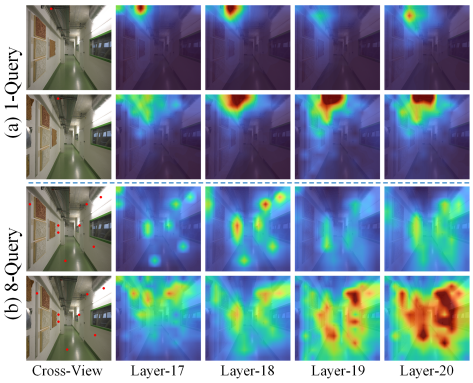


Figure 2: **Single-query versus multi-query reasoning.** Queries are provided as textual coordinates. Red markers are visualization overlays and are not part of the visual input.

To address these limitations, we introduce AnyPoint-Bench. AnyPoint-Bench formulates fine-grained cross-view understanding through an arbitrary multi-pixel query protocol. Given an image pair and an ordered set of reference-view pixels, a VLM jointly predicts their corresponding coordinates in the target view in a single response. Each prediction is explicitly associated with its source query. Direct coordinates expose localization errors, while joint predictions assess correspondence consistency across the image pair. AnyPoint-Bench supports eight query settings ranging from one to 144 pixels and covers diverse scenes with substantial viewpoint and appearance changes. We derive reliable correspondence labels from flow prediction Zhang et al. (2025) and verify them using cycle consistency and epipolar geometry Hartley et al. (2003). AnyPoint-Full contains 37K image pairs and 670K VQA pairs, while AnyPoint-4K provides 4K VQA pairs for efficient evaluation.

To quantify the fine-grained cross-view understanding of current general-purpose VLMs, we evaluate representative models across all eight query settings. GPT-5.6-Sol achieves 36.07% accuracy, while Qwen3.8 reaches 17.19%. However, the limited performance of general-purpose VLMs does not imply an inherent limitation of the VLM paradigm. To probe its potential, we develop CorrVLM, a cross-view specialist built on a standard autoregressive VLM. Unlike VLM^3 Cai et al. (2026), which directly decodes target coordinates autoregressively, CorrVLM introduces two complementary designs. First, it repurposes about 2,000 existing vocabulary entries as patch and coordinate tokens, yielding a compact spatial vocabulary that narrows the effective decoding space. Second, it progressively learns correspondence from patch-level associations to precise coordinates, forming a coarse-to-fine process that eases optimization. CorrVLM-8B achieves 67.35% accuracy, surpassing GPT-5.6-Sol by 31.28 percentage points. At a 9-pixel localization threshold, it matches RoMa Edstedt et al. (2024) while using fewer training image pairs (700K vs. 1.1M), and generalizes to query settings unseen during training. These results demonstrate that strong fine-grained

cross-view understanding is attainable within the VLM paradigm, yet remains largely unrealized in current general-purpose VLMs.

Our contributions are threefold. **1) AnyPoint-Bench**, to the best of our knowledge, the first unified benchmark that jointly evaluates pixel-level localization precision and global cross-view understanding through arbitrary multi-pixel queries in a single task. **2) A systematic evaluation** of representative VLMs across diverse scenes and eight query settings, revealing a pronounced gap between proprietary and open-source VLMs and considerable headroom. **3) CorrVLM**, a cross-view specialist with vocabulary repurposing and coarse-to-fine learning that surpasses general-purpose VLMs and matches specialized visual-geometric models with fewer training pairs.

2 Related Work

2.1 Cross-View Understanding Benchmarks

Cross-view benchmarks mainly evaluate scene-level reasoning or point-level correspondence. Recent scene-level benchmarks extend spatial evaluation from single images to multiple views and videos. VSI-Bench Yang et al. (2025) tests 3D scene understanding from egocentric videos through direction, distance, and route planning. All-Angles Yeh et al. (2026b) evaluates relative direction, manipulation, and camera pose across real-world views. MMSI-Bench Yang et al. (2026) evaluates positions, attributes, motion, and multi-step reasoning across multiple images. These benchmarks cover diverse scene-level relations, but mainly use discrete labels, scalar estimates, or short answers. Precise cross-view localization remains outside their scope.

Point-level benchmarks adopt marker- or coordinate-based protocols. BLINK Fu et al. (2024) evaluates marker-based correspondence within a general visual understanding suite, while MultiSPA Xu et al. (2026) and SpatialScore Wu et al. (2026) incorporate point correspondence into broader spatial benchmarks. CrossPoint-Bench Wang et al. (2025) and ReasonMatch-Bench Zhong et al. (2026) instead focus directly on cross-view correspondence. CrossPoint-Bench predicts a target coordinate for one query pixel. ReasonMatch-Bench binds predefined markers under wide viewpoint changes. Table 1 compares these protocols.

Table 1: Comparison of benchmarks involving fine-grained cross-view tasks. *# Scenes* denotes the number of scene types covered.

Dataset	Granularity	Multi-Query	Output	# Scenes	# Pairs	# VQAs
BLINK	Marker	✗	MCQ	5	336	344
CrossPoint	Marker, Pixel	✗	MCQ, Coord.	1	553	619
MultiSPA	Marker, Pixel	✗	MCQ, Coord.	1	600	600
ReasonMatch	Marker	✓	Bound Markers	3	2.8K	2.8K
SpatialScore	Pixel	✗	Coord.	4	229	229
AnyPoint-Full	Pixel	✓	Query-Bound Coord.	8	37K	670K
AnyPoint-4K	Pixel	✓	Query-Bound Coord.	8	3.6K	4K

Despite this progress, marker-based protocols hide localization error. Coordinate-based protocols expose localization error, but each query probes only one local point. Such isolated queries cannot assess global cross-view understanding. Their results depend on the selected query pixel. They therefore provide a limited measure of fine-grained cross-view understanding. AnyPoint-Bench instead localizes an arbitrary set of pixels with query-bound coordinates in one turn. It therefore unifies localization precision and global cross-view understanding within one task.

2.2 Pixel Correspondence

Recent progress in pixel correspondence is driven by specialized dense matchers. PDC-Net+ jointly predicts dense flow and uncertainty, allowing unreliable matches to be identified Truong et al. (2023). DKM uses learned kernel regression to establish dense correspondences for geometric estimation Edstedt et al. (2023). RoMa further improves robustness under large changes in scale, illumination, viewpoint, and texture Edstedt et al. (2024). More recently, UFM unifies optical flow and wide-baseline matching through direct dense-flow regression Zhang et al. (2025). These methods achieve strong geometric accuracy, but rely on correspondence-specific architectures and objectives.

VLM^3 Cai et al. (2026) shows that standard VLMs can learn pixel correspondence through text-based coordinate generation. The architecture supports depth, camera pose, and object-level 3D tasks. CorrVLM follows this unified VLM paradigm, but focuses on arbitrary multi-pixel localization. Under our protocol, the accuracy of VLM^3 drops sharply as the number of joint queries increases. CorrVLM combines a compact spatial vocabulary with coarse-to-fine learning. These designs improve localization under large query sets while preserving the standard VLM architecture.

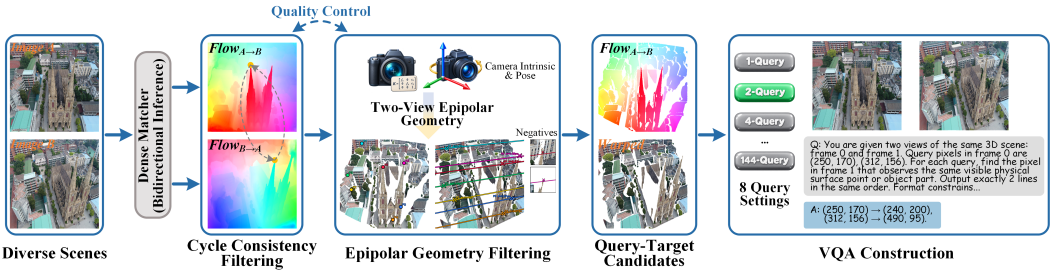


Figure 3: **Construction pipeline of AnyPoint-Bench**. We obtain bidirectional flow from diverse image pairs, and apply cycle-consistency as well as two-view epipolar-geometry filtering for quality control. The resulting query–target candidates are organized into eight query settings and converted into coordinate-based VQA instances.

3 METHODOLOGY

3.1 ANYPOINT-BENCH

Overview. Existing fine-grained cross-view benchmarks rely mainly on marker selection or single-pixel localization. Marker-based protocols restrict predictions to predefined candidates and therefore hide localization error. Single-pixel protocols preserve coordinate precision, but probe only one local match. They cannot reveal global cross-view understanding, and their scores depend strongly on the selected query. Consequently, current benchmarks provide either imprecise or query-sensitive estimates of model capability. Closing this gap calls for arbitrary multi-pixel queries across diverse scenes. The missing element is not another image corpus. It is a protocol that turns reliable visual geometry into comprehensive VLM evaluation. We therefore build AnyPoint-Bench on ZEB Shen et al. (2024), a visual matching benchmark with calibrated image pairs drawn from 12 datasets. These datasets cover diverse domains, including BlendedMVS Yao et al. (2020), ETH3D Schops et al. (2017), KITTI Geiger et al. (2012), SceneNet McCormac et al. (2017), and ICL-NUIM Handa et al. (2014). This scene diversity and two-view geometry provide a natural foundation for constructing arbitrary multi-pixel localization tasks. The resulting benchmark offers broad coverage of cross-view conditions and query settings, as summarized in Table 1.

Correspondence and Verification. As illustrated in Fig. 3, given an image pair $(\boldsymbol{\mathcal{I}_A}, \boldsymbol{\mathcal{I}_B})$, we apply UFM (Zhang et al., 2025) bidirectionally to obtain dense flow fields $\boldsymbol{f}_{A\to B}$ and $\boldsymbol{f}_{B\to A}$. For a pixel $\boldsymbol{u}$ in $\boldsymbol{\mathcal{I}_A}$, its initial target is $\boldsymbol{v} = \boldsymbol{u} + \boldsymbol{f}_{A\to B}(\boldsymbol{u})$. Cycle-consistency filtering retains this pair only when mapping $\boldsymbol{v}$ back to $\boldsymbol{\mathcal{I}_A}$ returns within τ_{cyc} pixels of $\boldsymbol{u}$. We then verify the pair using the camera intrinsics $(\boldsymbol{K}_A, \boldsymbol{K}_B)$ and relative pose $(\boldsymbol{R}, \boldsymbol{t})$. The fundamental matrix $\boldsymbol{F} = \boldsymbol{K}_B^{-\top}[\boldsymbol{t}]_\times \boldsymbol{R}\boldsymbol{K}_A^{-1}$, and epipolar distance is defined as,

$$d(\boldsymbol{u}, \boldsymbol{v}) = \frac{(\tilde{\boldsymbol{v}}^T \widehat{\boldsymbol{F}} \tilde{\boldsymbol{u}})^2}{\|\boldsymbol{F}\tilde{\boldsymbol{u}}\|_{[1]}^2 + \|\boldsymbol{F}\tilde{\boldsymbol{u}}\|_{[2]}^2 + \|\boldsymbol{F}^T\tilde{\boldsymbol{v}}\|_{[1]}^2 + \|\boldsymbol{F}^T\tilde{\boldsymbol{v}}\|_{[2]}^2}, \tag{1}$$

where $\tilde{\boldsymbol{u}}$ and $\tilde{\boldsymbol{v}}$ are homogeneous coordinates, and $[\boldsymbol{t}]_\times$ is the skew-symmetric matrix of $\boldsymbol{t}$ Hartley et al. (2003). $\|\cdot\|_{[i]}$ denotes the i-th element. We retain candidates with $d < \tau_{\text{epi}}$. The filters assign negative labels to predictions that violate bidirectional consistency or two-view geometry.

Multi-Query VQA Construction. The verified correspondences form the query-target pool. To standardize evaluation, we resize all images to 384×384 and sample ordered query sets from the centers of a uniform 12×12 grid across eight settings, with cardinalities ranging from one to 144. Each prompt provides all source coordinates, while the answer returns their target coordinates in the same order. This query-bound format avoids permutation ambiguity and supports variable query cardinalities. Direct coordinates expose localization error, while joint prediction tests global cross-view consistency within one inference turn. Complete prompt templates are provided in the Appendix. AnyPoint-Full contains 37K image pairs and 670K VQAs. AnyPoint-4K contains 3.6K pairs and 4K VQAs for efficient evaluation. Although its construction is deliberately simple, AnyPoint-Bench fills an important gap by turning established visual geometry data into a comprehensive testbed for fine-grained cross-view understanding in VLMs.

3.2 MULTI-QUERY LOCALIZATION

We consider N pixel queries in the source image $\boldsymbol{\mathcal{I}_A}$ and their corresponding locations in the target image $\boldsymbol{\mathcal{I}_B}$. Let $H_A \times W_A$ and $H_B \times W_B$ denote the image sizes, and define the coordinate domains as $\Omega_A = [0, W_A) \times [0, H_A)$ and $\Omega_B = [0, W_B) \times [0, H_B)$. The query-target correspondences are,

$$\mathcal{Q} = \{\boldsymbol{Q}_i = (\boldsymbol{u}_i, \boldsymbol{v}_i) \mid i = 1, \ldots, N\}, \boldsymbol{u}_i \in \Omega_A, \boldsymbol{v}_i \in \Omega_B, \tag{2}$$

where $\boldsymbol{u}_i = (x_i^A, y_i^A)$, $\boldsymbol{v}_i = (x_i^B, y_i^B)$. $\boldsymbol{u}_i$ is an arbitrary query coordinate and $\boldsymbol{v}_i$ is its corresponding target location. We write the input queries as an ordered sequence $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N)$, whose order also determines the output order. Unlike single-query formulations that process one pixel at a time, our formulation allows N to vary across samples. A single model can therefore handle different query cardinalities. Our objective is to predict all target locations conditioned on the two images and the complete query sequence,

$$\hat{\boldsymbol{V}} = (\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_N) = \mathcal{F}_\theta(\boldsymbol{\mathcal{I}_A}, \boldsymbol{\mathcal{I}_B}, \boldsymbol{U}), \tag{3}$$

where $\mathcal{F}_\theta$ denotes a VLM parameterized by θ.

3.3 CORRVLM

VLM^3 Cai et al. (2026) learns pixel query localization through direct autoregressive coordinate prediction. This single-stage paradigm requires a VLM to establish cross-view associations and resolve precise target locations simultaneously. Such coupling becomes harder to optimize as query cardinality grows. Our ablations confirm this limitation, accuracy drops sharply when more queries are processed jointly. Inspired by coarse-to-fine matching Sun et al. (2021); Wang et al. (2024), CorrVLM first learns patch-level correspondences and then refines them into pixel coordinates. It also repurposes existing vocabulary entries as patch and coordinate tokens. This compact spatial vocabulary narrows the decoding space and simplifies multi-query optimization.

Coarse-Level Optimization. We first formulate this stage as conditional classification. Specifically, we uniformly partition $\boldsymbol{\mathcal{I}_B}$ into a grid with G_h rows and G_w columns. The grid resolution is model-specific and should be adjusted for different VLMs. This partition yields $M_p = G_h G_w$ spatial patches. To represent these patches in the autoregressive output space, we assign a dedicated token to each patch and define $\mathcal{V}_p = \{\langle p_m \rangle \mid m = 0, \ldots, M_p - 1\}$. We do not expand the original vocabulary. Instead, we reserve its final M_p entries as patch tokens. Their original lexical meanings are discarded for correspondence learning, and their associated embeddings and output weights are optimized to encode spatial regions.

During training, each ground-truth target pixel is quantized into a patch index, which serves as its classification label. To be specific, for a target pixel $\boldsymbol{v}_i = (x_i^B, y_i^B)$, we first identify its grid cell and then flatten the two-dimensional cell coordinate into a patch index,

$$r_i = \left\lfloor \frac{G_h y_i^B}{H_B} \right\rfloor, \qquad c_i = \left\lfloor \frac{G_w x_i^B}{W_B} \right\rfloor, \qquad k_i = G_w r_i + c_i, \tag{4}$$

where $r_i \in \{0, \ldots, G_h - 1\}$, $c_i \in \{0, \ldots, G_w - 1\}$, and $k_i \in \{0, \ldots, M_p - 1\}$. The supervision for query $\boldsymbol{u}_i$ is the patch token $t_i^c = \langle p_{k_i} \rangle$, and the target sequence is $\boldsymbol{T}^c = (t_1^c, \ldots, t_N^c)$. Conditioned on the paired images and $\boldsymbol{U}$, the VLM autoregressively predicts this sequence. The objective is,

$$\mathcal{L}_c(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \log p_\theta(t_i^c \mid \boldsymbol{\mathcal{I}_A}, \boldsymbol{\mathcal{I}_B}, \boldsymbol{U}, \boldsymbol{T}_{<i}^c), \tag{5}$$

where $\boldsymbol{T}_{<i}^c = (t_1^c, \ldots, t_{i-1}^c)$ contains the preceding ground-truth patch tokens under teacher forcing. This objective teaches the VLM to associate each source query with a coarse region in the target view. We denote the optimized parameters by θ_c^* and use them to initialize fine-level optimization.

Fine-Level Optimization. Starting from θ_c^*, we optimize the VLM for fine-grained localization. Directly regressing continuous coordinates is not naturally aligned with autoregressive token prediction. We therefore quantize the coordinate space and cast localization as coordinate-token prediction. Specifically, we uniformly quantize each spatial dimension into K bins and associate each bin with a dedicated token $\mathcal{V}_c = \{\langle c_m \rangle \mid m = 0, \ldots, K - 1\}$. As with the patch tokens, we do not add new

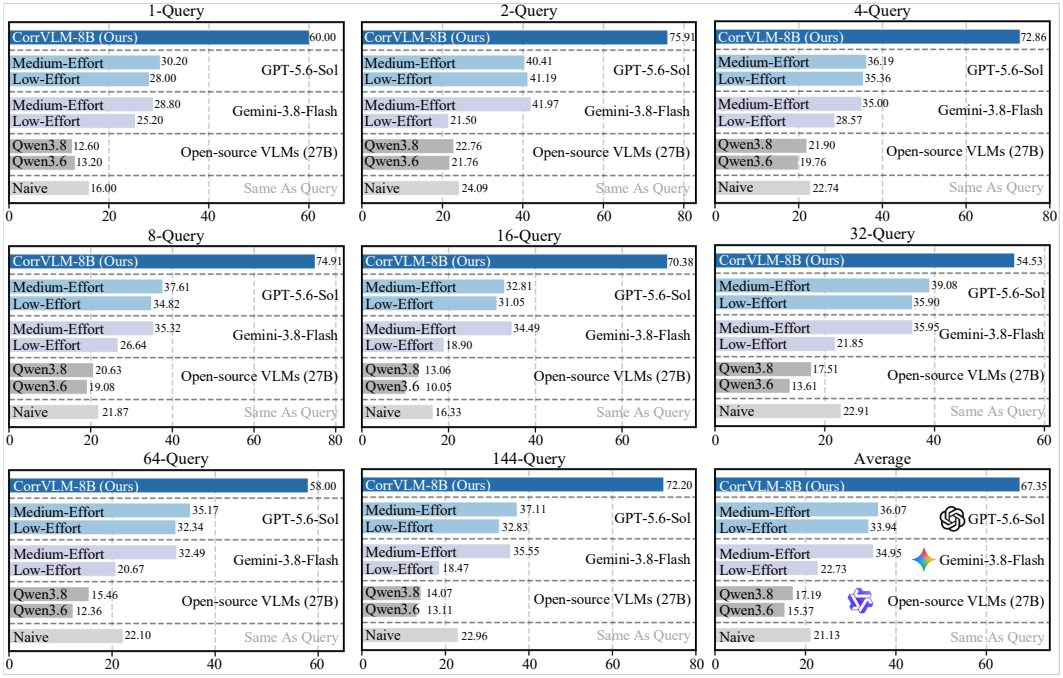


Figure 4: **General-purpose VLMs remain limited in fine-grained cross-view understanding.** We compare proprietary GPT-5.6-Sol and Gemini-3.8-Flash, open-source Qwen3.8 and Qwen3.6, and specialist CorrVLM-8B on AnyPoint-4K. PCK@9px (%) is reported across eight query cardinalities and their average. The proprietary models are evaluated under low- and medium-effort inference.

vocabulary entries. We reserve the K entries immediately preceding the patch-token range as coordinate tokens. Thus, $\mathcal{V}_c$ and $\mathcal{V}_p$ occupy disjoint ranges of the original vocabulary. This design keeps the vocabulary size unchanged and represents each two-dimensional location with only two tokens.

During fine-grained optimization, each target point $\boldsymbol{v}_i = (x_i^B, y_i^B)$ is represented by two discrete coordinate indices,

$$a_i = \left\lfloor \frac{K x_i^B}{W_B} \right\rfloor, \qquad b_i = \left\lfloor \frac{K y_i^B}{H_B} \right\rfloor, \qquad a_i, b_i \in \{0, \ldots, K-1\}. \tag{6}$$

Accordingly, the fine-level supervision for query $\boldsymbol{u}_i$ is represented by the coordinate-token pair $t_i^f = (\langle c_{a_i} \rangle, \langle c_{b_i} \rangle)$. For multiple queries, we do not supervise the VLM with an isolated sequence of target coordinates. Instead, each target is explicitly paired with its source query,

$$\boldsymbol{T}^f = \text{Serialize}\left(\boldsymbol{u}_1 \to t_1^f, \boldsymbol{u}_2 \to t_2^f, \ldots, \boldsymbol{u}_N \to t_N^f\right). \tag{7}$$

We optimize the response sequence with the standard autoregressive language-modeling objective,

$$\mathcal{L}_f(\theta) = -\frac{1}{S_f} \sum_{j=1}^{S_f} \log p_\theta\left(T_j^f \mid \boldsymbol{\mathcal{I}_A}, \boldsymbol{\mathcal{I}_B}, \boldsymbol{U}, \boldsymbol{T}_{<j}^f\right), \tag{8}$$

where S_f is the number of tokens in $\boldsymbol{T}^f$, T_j^f is its j-th token, and $\boldsymbol{T}_{<j}^f$ contains all preceding ground-truth response tokens under teacher forcing. We denote the resulting parameters by θ_f^*. Details of inference with an arbitrary number of queries are provided in the Appendix.

4 Experiments

CorrVLM Implementation. We use Qwen3-VL Bai et al. (2025) as the base model and set $G_h = G_w = 12$ and $K = 2{,}000$, yielding 144 patch tokens and 2,000 coordinate tokens. The training set comprises 700K image pairs constructed from DL3DV-10K Ling et al. (2024). We retain the

Table 2: **Dataset-wise comparison on AnyPoint-4K under the 144-Query setting.** We report PCK@9px (%, ↑) on eight real and four simulated datasets. *Same as Query* copies each source query to the same coordinate in the target image. † and ‡ denote low- and medium-effort inference.

Method	Mean	Real								Simulated			
		GL3	BLE	ETI	ETO	KIT	WEA	SEA	NIG	MUL	SCE	ICL	GTA
Naive Prediction													
SAME AS QUERY	22.96	0.05	0.53	1.16	1.63	46.11	56.90	50.11	70.82	20.59	0.91	18.07	5.92
Open-source General VLMs													
Qwen3.6-27B	13.11	0.05	0.48	0.32	0.81	22.24	37.45	36.21	38.95	8.22	0.21	11.87	3.05
Qwen3.8-27B	14.07	0.25	0.57	0.68	0.65	33.56	36.91	38.44	27.34	10.38	0.32	17.90	2.65
Proprietary General VLMs													
Gemini-3.8-Flash†	18.47	3.17	0.18	1.35	0.57	45.04	41.83	42.13	40.21	20.09	1.23	17.90	5.59
Gemini-3.8-Flash‡	35.55	21.35	8.60	25.74	30.31	65.93	58.80	55.37	51.74	41.99	20.83	25.08	17.03
GPT-5.6-Sol†	32.83	17.67	3.73	17.61	16.66	63.93	61.48	58.63	66.76	36.34	14.44	22.23	10.85
GPT-5.6-Sol‡	37.11	23.11	8.42	24.02	19.71	66.50	62.26	60.80	74.24	41.44	18.48	28.02	14.24
CorrVLM-8B	**72.20**	**39.78**	**37.76**	**67.97**	**79.89**	**94.23**	**83.52**	**91.59**	**85.12**	**84.24**	**57.96**	**66.35**	**70.09**

bidirectional UFM Zhang et al. (2025) prediction and cycle-consistency filtering for the training dataset construction, but without epipolar-geometry verification. During training, we use only 1-, 16-, and 144-query configurations. The same checkpoint is directly evaluated at arbitrary query cardinalities, including those unseen during training. Inspired by E-RayZer Zhao et al. (2026), we further adopt a scene-wise sampling curriculum. For each scene, we construct three pair pools by sampling 48, 32, and 16 frames. Training progresses from the denser 48-frame pool to the 32-frame pool and finally the sparser 16-frame pool, gradually reducing cross-view overlap and increasing matching difficulty. We perform full-parameter fine-tuning with AdamW Loshchilov & Hutter (2019), using learning rates of 2×10^{-6} and 2×10^{-5} for the vision encoder and language model, respectively. Training the 8B model takes approximately 70 hours on 196 H20e GPUs.

Evaluation Setting. We use three evaluation sets. Experiments involving proprietary VLMs use AnyPoint-4K, a fixed subset of AnyPoint-Full, for efficient evaluation. (Fig. 4 and Tab. 2). Models with public checkpoints are evaluated on AnyPoint-Full (Tabs. 3). Both AnyPoint splits serve as out-of-domain test sets for all evaluated VLMs and visual-geometric models. No model receives AnyPoint-specific training or adaptation. For in-domain evaluation, we use a held-out split from DL3DV-10K (Fig. 6). To avoid potential annotation noise in negative samples, we evaluate all predictions on queries with positive ground-truth correspondences.

4.1 Main Results

General-Purpose VLMs. As shown in Fig. 4, general-purpose VLMs remain limited in fine-grained cross-view understanding. Under medium-effort inference, GPT-5.6-Sol and Gemini-3.8-Flash achieve average PCK@9px scores of 36.07% and 34.95%, while Qwen3.8 and Qwen3.6 reach only 17.19% and 15.37%. Both open-source models underperform the *Same as Query* baseline at every query cardinality. This result shows that general visual-language capability does not directly translate into pixel-accurate cross-view localization. Increasing inference effort improves Gemini from 22.73% to 34.95%, but only raises GPT from 33.94% to 36.07%. Reasoning effort can therefore reveal latent correspondence ability, but does not resolve the underlying spatial grounding problem. CorrVLM reaches 67.35%, showing that the remaining gap is not an inherent limitation of the VLM paradigm.

The dataset-wise results in Tab. 2 reveal complementary strengths and strong dependence on viewing conditions. GPT-5.6-Sol consistently outperforms Gemini-3.8-Flash on WEA, SEA, and NIG. The largest gap appears on NIG, where GPT-5.6-Sol achieves 74.24% compared with 51.74% for Gemini-3.8-Flash. This pattern indicates stronger robustness to appearance changes caused by weather, season, and nighttime illumination. Gemini-3.8-Flash performs better on ETI and ETO and leads on three of the four simulated subsets, suggesting a relative advantage under better-lit and simulated conditions. The *Same as Query* baseline is strong on KIT, WEA, SEA, and NIG, but nearly collapses on GL3, BLE, ETI, and ETO. Its behavior indicates that the driving subsets contain many correspondences with limited image-plane displacement. GPT-5.6-Sol's advantage on WEA, SEA, and NIG therefore reflects stronger appearance robustness under relatively aligned

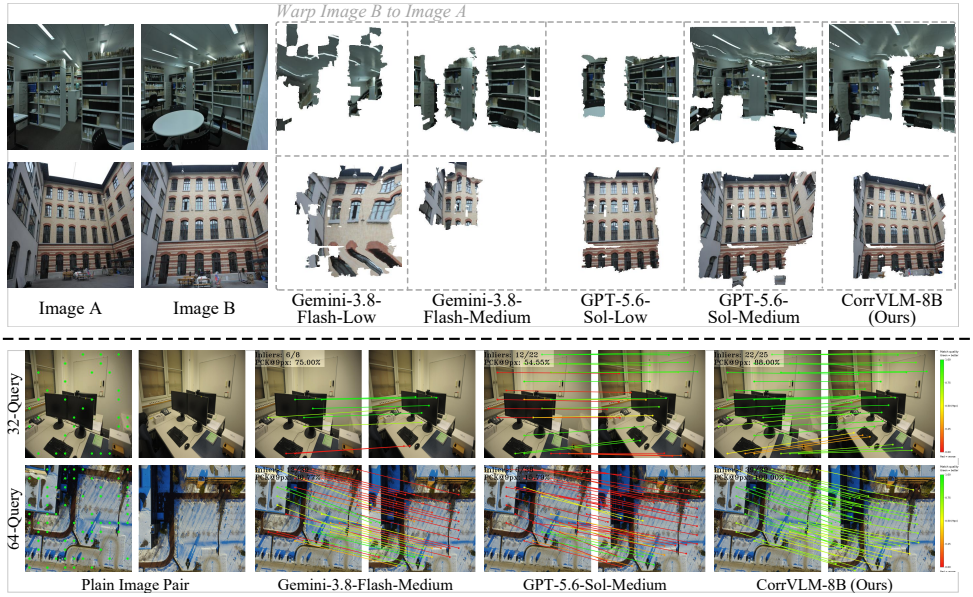


Figure 5: **Qualitative comparison of general-purpose VLMs and CorrVLM on AnyPoint-4K.** Top: Image B is warped to Image A using correspondences predicted under the 144-Query setting. Bottom: The two rows show 32-Query indoor and 64-Query aerial localization.

views, rather than uniformly stronger wide-baseline geometric reasoning. The low results on BLE and several simulated subsets further show that neither proprietary model provides reliable cross-domain correspondence.

The qualitative results in Fig. 5 support the same conclusion. Higher inference effort improves matching coverage, but general-purpose VLMs still produce fragmented warps and inconsistent correspondences. These failures become particularly clear in the aerial example. CorrVLM produces more complete and coherent mappings, showing that occasional correct predictions do not imply global cross-view understanding.

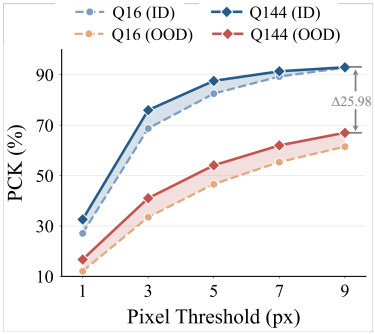


Figure 6: In-domain and out-of-domain performance of CorrVLM.

CorrVLM. To align visual-geometric matchers with AnyPoint-Bench, we use the 144 centers of a 12×12 patch grid as source queries and evaluate every method at these locations. This provides a common coordinate-level metric across the two paradigms. For CorrVLM, the task evaluates joint fine-grained cross-view understanding. For specialized matchers, it evaluates two-view geometric matching accuracy. The results in Tab. 3 show that CorrVLM-8B achieves the best mean PCK@9px of 66.98%, exceeding RoMa by 2.19 points and COTR Jiang et al. (2021) by 7.13 points. CorrVLM uses 700K task-specific training pairs, compared with 1.1M for others. The dataset breakdown reveals complementary strengths. CorrVLM leads on ETI, WEA, SCE, and ICL, with gains of 19.11 and 16.34 points on the two simulated indoor sets SCE and ICL. It also remains competitive on KIT, SEA, and NIG. This pattern suggests that the VLM-based representation transfers well to appearance changes and texture-poor indoor scenes. RoMa remains stronger on GL3, BLE, ETO, MUL, and GTA, where larger viewpoint and geometric changes favor matching-specific inductive biases.

The comparison in Fig. 6 reveals a clear domain gap. Under the 144-query setting, CorrVLM achieves 92.96% PCK@9px on the held-out DL3DV-10K split and 66.98% on AnyPoint-Full, a drop of 25.98 points. The near-saturated in-domain score shows that the 8B model has sufficient capacity to fit the task, while cross-domain generalization remains the main bottleneck. Across all thresholds, the 144-Query setting consistently outperforms the 16-Query setting in both domains. CorrVLM can leverage denser query context to improve multi-query localization accuracy.

Table 3: **Comparison with visual-geometric matchers.** CorrVLM-8B is evaluated against representative sparse and dense methods on AnyPoint-Full under the 144-Query setting.

Method	Mean	Real								Simulated			
		GL3	BLE	ETI	ETO	KIT	WEA	SEA	NIG	MUL	SCE	ICL	GTA
Visual-Geometric Models													
LoFTR	48.99	39.10	28.60	41.62	52.14	90.42	87.75	91.43	83.10	61.97	35.60	35.00	36.09
ELoFTR	49.52	39.12	29.20	41.20	54.38	90.34	87.69	91.52	83.58	65.26	34.72	34.67	37.56
JamMa	49.31	40.40	27.45	41.76	57.29	91.84	88.46	90.68	**86.03**	64.37	33.42	33.27	34.97
SLiM	49.58	39.79	29.54	44.38	51.50	88.71	87.97	90.09	83.57	61.84	36.92	36.87	36.54
COTR	59.85	51.44	41.98	57.03	72.58	**95.36**	91.61	**94.95**	84.21	68.80	41.24	47.89	50.87
DKM	45.16	48.39	46.42	43.25	72.27	43.48	44.64	54.92	20.38	58.59	28.06	24.47	61.31
RoMa	64.79	**62.38**	**58.72**	54.50	**82.50**	82.17	73.53	83.06	51.24	**80.56**	52.47	43.47	**75.49**
Vision-Language Models													
CorrVLM-8B	**66.98**	42.60	38.39	**66.72**	72.31	94.42	**93.34**	91.23	83.99	54.52	**71.58**	**64.23**	66.98

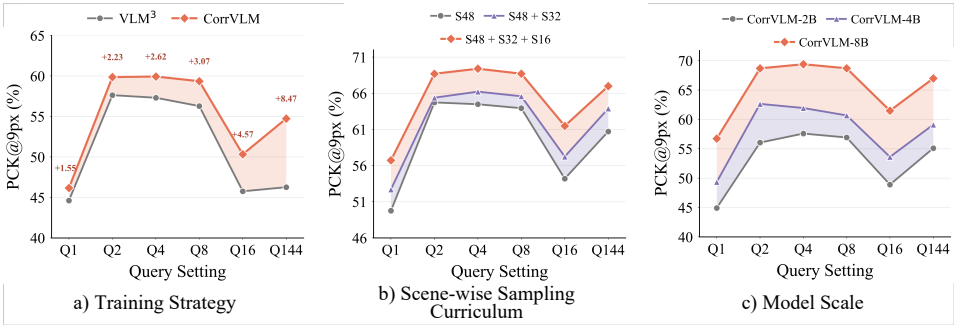


Figure 7: **Ablation studies of CorrVLM.** We report PCK@9px across six query settings. a) compares the direct coordinate training of VLM^3 Cai et al. (2026) with our CorrVLM training strategy. b) evaluates the progressive scene-wise sampling curriculum, where S_n denotes sampling n frames per scene. c) studies model scaling from 2B to 8B.

Ablation Study. As shown in Fig. 7, we study three key factors in CorrVLM, including the training strategy, scene-wise sampling curriculum, and model scale. CorrVLM consistently outperforms the direct coordinate training of VLM^3 Cai et al. (2026), with the margin increasing from 1.55 points at 1-Query to 8.47 points at 144-Query. This widening gap supports our motivation that direct coordinate prediction becomes increasingly difficult when more query-target bindings are generated jointly. CorrVLM's coarse-to-fine objective and compact spatial vocabulary are therefore particularly effective in the multi-query regime. The scene-wise curriculum provides complementary gains. Adding $S32$ and then $S16$ to $S48$ improves every query setting, showing that progressive exposure to lower-overlap pairs strengthens cross-view robustness. Performance also scales steadily with model size. Moving from 2B to 8B yields approximately 12 percentage points across all settings. The near-uniform gain from 1-Query to 144-Query suggests that scaling improves the underlying correspondence representation rather than only longer-output decoding. Notably, these improvements also extend to 2-Query, 4-Query, and 8-Query, which are unseen during training, further demonstrating generalization beyond the supervised query cardinalities.

5 Conclusion

We investigate whether VLMs possess fine-grained cross-view understanding, defined as the ability to localize arbitrary pixels across changes in viewpoint and appearance. We introduce AnyPoint-Bench, a unified benchmark that jointly evaluates pixel-level localization and global cross-view understanding through arbitrary multi-pixel queries. Our evaluation shows that this capability remains limited in general-purpose VLMs and reveals a pronounced gap between proprietary and open-source models. To examine whether this limitation is inherent to the VLM paradigm, we further develop CorrVLM with vocabulary repurposing and coarse-to-fine learning. CorrVLM substantially outperforms general-purpose VLMs, generalizes to unseen query cardinalities, and matches visual-geometric matchers with fewer training pairs. These results show that strong fine-grained cross-view understanding is attainable within the standard autoregressive VLM paradigm, although generalization across domains and under large viewpoint changes remains an important challenge.

AI USE STATEMENT

We used generative AI tools to assist with code generation for CorrVLM and to polish the language of this manuscript. The authors examined every AI-generated code suggestion and textual revision before incorporating it into the work. AI assistance improved the efficiency of implementation and writing but did not determine the research design, experimental results, or conclusions. The authors take full responsibility for the final code and manuscript.

REPRODUCIBILITY STATEMENT

We detail the construction and evaluation of AnyPoint-Bench and the training and inference of CorrVLM in Secs. 3 and 4. We will release AnyPoint-Full and AnyPoint-4K, CorrVLM weights, and code for data preparation, training, inference, and evaluation, together with the relevant prompts and configurations. For third-party images that cannot be redistributed, we will provide source identifiers and retrieval instructions. Reproducing proprietary VLM results depends on continued access to the corresponding services.

REFERENCES

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

Zhipeng Cai, Zhuang Liu, Yunyang Xiong, Zechun Liu, Vikas Chandra, and Yangyang Shi. VLM3: Vision language models are native 3D learners. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2026.

Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17765–17775, 2023.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19790–19800, 2024.

Zhiyuan Feng, Zhaolu Kang, Qijie Wang, Zhiying Du, Jiongrui Yan, Shubin Shi, Chengbo Yuan, Huizhi Liang, Yu Deng, Qixiu Li, et al. Seeing across views: Benchmarking spatial reasoning of vision-language models in robotic scenes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 155178–155227, 2026.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 148–166, 2024.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3354–3361, 2012.

Ankur Handa, Thomas Whelan, John McDonald, and Andrew J Davison. A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1524–1531, 2014.

Richard Hartley, Andrew Zisserman, et al. *Multiple view geometry in computer vision*, volume 2. Cambridge university press Cambridge, 2003.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023.

Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 6187–6197, 2021.

Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. Sti-bench: Are mllms ready for precise spatial-temporal world understanding? In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5622–5632, 2025.

Zongzhao Li, Zongyang Ma, Mingze Li, Songyou Li, Yu Rong, Tingyang Xu, Ziqi Zhang, Deli Zhao, and Wenbing Huang. Star-r1: Multi-view spatial transformation reasoning by reinforcing multimodal llms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12041–12051, 2026.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J Davison. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2697–2706, 2017.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 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 (CVPR)*, pp. 3260–3269, 2017.

Xuelun Shen, Wei Yin, Matthias Müller, Zijun Li, Kaixuan Wang, Xiaozhi Chen, Cheng Wang, et al. Gim: Learning generalizable image matcher from internet videos. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 13182–13197, 2024.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. *arXiv preprint arXiv:2104.00680*, 2021.

Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. Pdc-net+: Enhanced probabilistic dense correspondence network. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(8):10247–10266, 2023.

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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 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.

Haoning Wu, Xiao Huang, Yaohui Chen, Ya Zhang, Yanfeng Wang, and Weidi Xie. Spatialscore: Towards comprehensive evaluation for spatial intelligence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31029–31041, 2026.

Runsen Xu, Weiyao Wang, Hao Tang, Xingyu Chen, Xiaodong Wang, Fu-Jen Chu, Dahua Lin, Matt Feiszli, and Kevin J. Liang. Multi-spatialmllm: Multi-frame spatial understanding with multi-modal large language models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31078–31088, 2026.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643, 2025.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 157051–157088, 2026.

Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1787–1796, 2020.

Ching-Feng Yeh, Hu Xu, Zhuang Liu, Gregory P Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, Yangyang Shi, et al. Depthlm: Metric depth from vision language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 134491–134505, 2026a.

Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 14, pp. 12000–12008, 2026b.

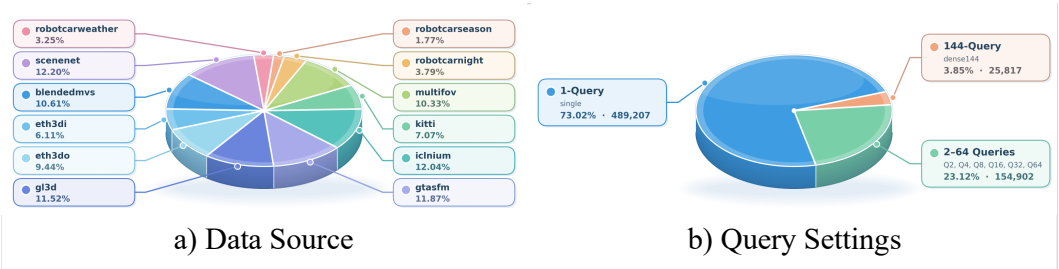
Xingyilang Yin, Chengzhengxu Li, Jiahao Chang, Chi-Man Pun, and Xiaodong Cun. Mllm-4d: Towards visual-based spatial-temporal intelligence. *arXiv preprint arXiv:2603.00515*, 2026.

Yuchen Zhang, Nikhil Keetha, Chenwei Lyu, Bhuvan Jhamb, Yutian Chen, Yuheng Qiu, Jay Karhade, Shreyas Jha, Yaoyu Hu, Deva Ramanan, et al. Ufm: A simple path towards unified dense correspondence with flow. *arXiv preprint arXiv:2506.09278*, 2025.

Qitao Zhao, Hao Tan, Qianqian Wang, Sai Bi, Kai Zhang, Kalyan Sunkavalli, Shubham Tulsiani, and Hanwen Jiang. E-rayzer: Self-supervised 3d reconstruction as spatial visual pre-training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7525–7535, 2026.

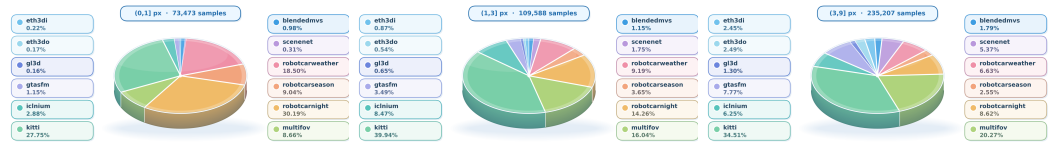
Hao Zhong, Muzhi Zhu, Shenyan Zeng, Anzhou Li, Cong Chen, Hua Geng, Duochao Shi, Wentao Ye, Tao Lin, Hao Chen, et al. Eliciting complex spatial reasoning in mllms through wide-baseline matching. *arXiv preprint arXiv:2606.03577*, 2026.

A Composition of AnyPoint-Bench-Full



a) Data Source b) Query Settings

Figure 8: **Composition of AnyPoint-Bench-Full**. (a) Relative contributions of the 12 data sources. (b) Distribution of 669,926 VQAs across eight query settings. The 1-Query setting accounts for 73.02% of the VQAs, while the six settings from 2 to 64 queries and the 144-Query setting account for 23.12% and 3.85%, respectively.

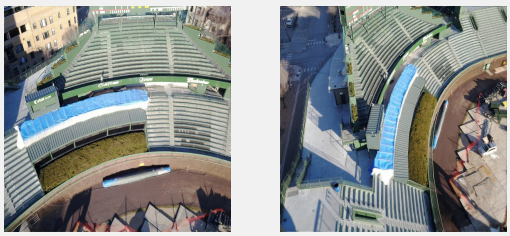


a) (0, 1]px, 6.30% b) (1, 3]px, 9.25% c) (3, 9]px, 15.57%

Figure 9: **Cross-view displacement statistics of AnyPoint-Bench-Full**. Panels (a)–(c) show the data-source composition of matched queries with displacements in (0,1], (1,3], and (3,9] pixels, respectively. These intervals account for 6.30%, 9.25%, and 15.57% of all benchmark queries, while the percentages around each pie describe the data-source composition within that interval. Small displacements are concentrated in driving data, particularly KITTI and RobotCar.

Data sources and query settings. AnyPoint-Bench-Full contains approximately 37K image pairs and 669,926 VQAs. As shown in Fig. 8(a), it shows their distribution across 12 data sources. Eight real-world sources contribute 53.56% of the VQAs, while four simulated sources contribute 46.44%. Together, they cover indoor and outdoor scenes, driving sequences under weather, seasonal, and nighttime changes, and simulated environments. Individual source shares range from 1.77% to 12.20%, so no single source dominates the benchmark. These sources are distinct from the DL3DV-10K data used to train CorrVLM, making AnyPoint-Bench-Full an out-of-domain evaluation set. As shown in Fig. 8(b), it shows that 489,207 VQAs use one query, 154,902 use two to 64 queries, and 25,817 use 144 queries. Although single-query VQAs are the most numerous, the benchmark includes 180,719 multi-query VQAs. In the 144-query setting, a model must return a target location for every point on the (12×12) query grid in one response. The eight settings therefore test precise localization across different query cardinalities, including five cardinalities unseen during CorrVLM training. We report results by cardinality because VQA counts differ substantially across settings. *For efficient evaluation, we construct AnyPoint-Bench-4K by randomly sampling 500 VQAs from each of the eight query settings in AnyPoint-Bench-Full across its 12 data sources, yielding 4,000 VQAs in total.*

Multi-Query Localization Prompt Template



Question: You are given two views of the same static 3D scene: frame 0 and frame 1. Frame 0 is divided into a 12x12 grid of patches. Every query coordinate below is a patch center in frame 0. Each valid target coordinate must be the precise corresponding pixel location in frame 1. It may lie anywhere in the image and is not restricted to a patch center. For both frames, the origin (0, 0) is at the top-left corner, x increases to the right, and y increases downward. Both query patch-center coordinates and target pixel coordinates are normalized integers in [0, 2000). For each query, find the pixel in frame 1 that observes the same visible physical surface point or object part. If it is occluded, outside the image, or has no reliable match, output no_match.
Query patch centers in frame 0:
q00: (750, 917)
q01: (1750, 1250)
q02: (417, 1583)
q03: (1917, 1583)
Output exactly 4 lines in the same order. Copy the frame 0 query coordinate on the left, on the right output either the precise frame 1 target pixel coordinate or no_match. Use exactly one of these formats for each line:
(source_x, source_y) -> (target_x, target_y)
(source_x, source_y) -> no_match
Output only the 4 correspondence lines.

Figure 10: **Multi-query localization prompt template.**

Cross-view displacement. As shown in Fig. 9, it examines the pixel displacement between each matched query and its target location after image resizing. The reported shares of the (0,1], (1,3], and (3,9]-pixel intervals are 6.30%, 9.25%, and 15.57% of all benchmark queries, respectively.

Each pie chart shows the distribution of data sources within its displacement interval. KITTI and the three RobotCar subsets jointly contribute 85.48% of the (0,1]-pixel interval, 67.04% of the (1,3]-pixel interval, and 52.31% of the (3,9]-pixel interval. Small displacements are thus concentrated in driving data. This pattern helps explain the strong *Same as Query* baseline on these subsets. Under PCK@9px, copying the source coordinate counts as correct whenever the true target lies within nine pixels. Weather, seasonal, and nighttime changes can still make these near-aligned pairs visually challenging. The full benchmark also contains larger displacements and no-match labels, exposing conditions in which coordinate copying is insufficient. The PCK experiments in Section 4 evaluate positive correspondences only, so the no-match cases in the full benchmark do not contribute to those scores.

Multi-Query Localization Template. The template specifies a common coordinate convention for both frames, with coordinates represented as normalized integers in [0, 2000). Repeating the source coordinate on each output line explicitly binds every prediction to its query, even when several points are localized in a single response. The 'no_match' option accounts for queries whose corresponding scene point is occluded, outside frame 1, or cannot be matched reliably.

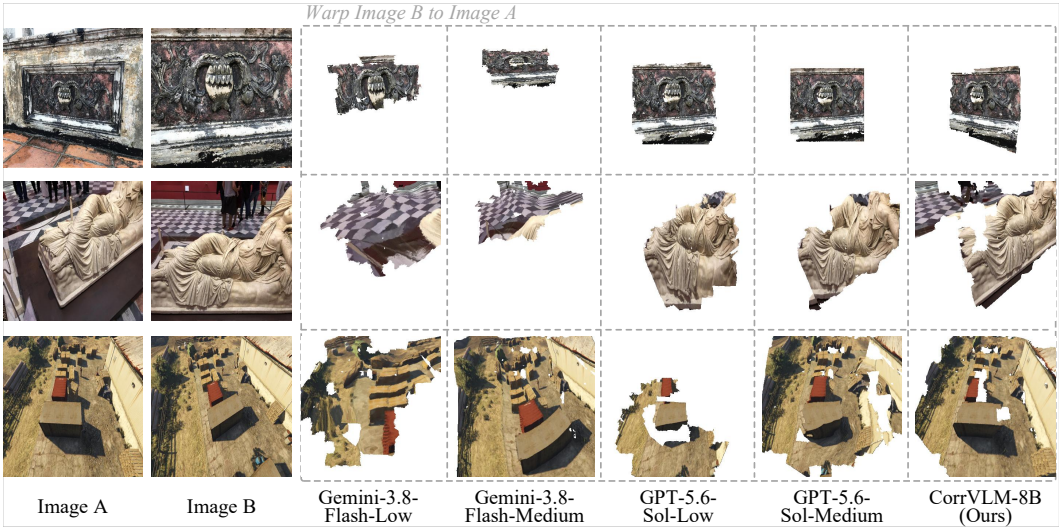


Figure 11: **Additional qualitative comparison of image warping on AnyPoint-4K.** Using each model's predicted correspondences, we warp Image B into the coordinate frame of Image A. The first two columns show the input pair, the remaining columns compare Gemini-3.8-Flash and GPT-5.6-Sol at low and medium inference effort with CorrVLM-8B. The rows show a carved wall relief, a sculpture, and an aerial scene. White regions are unfilled pixels in the warped outputs.

B Arbitrary-Query Inference

At inference time, we retain only the fine-level model $\mathcal{F}_{\theta_f^*}$. Given $\boldsymbol{\mathcal{I}_A}, \boldsymbol{\mathcal{I}_B}$, and an arbitrary query sequence $\boldsymbol{U}$ within the model's context-length limit, the model generates the same structured response used during fine-level optimization,

$$\hat{\boldsymbol{T}}^f = \left\{ \boldsymbol{u}_i \rightarrow \left(\langle c_{\hat{a}_i} \rangle, \langle c_{\hat{b}_i} \rangle \right) \right\}_{i=1}^{N} = \mathcal{F}_{\theta_f^*}(\boldsymbol{\mathcal{I}_A}, \boldsymbol{\mathcal{I}_B}, \boldsymbol{U}), \tag{9}$$

where $\hat{a}_i, \hat{b}_i \in \{0, \ldots, K-1\}$ are the predicted coordinate indices. Using the center of each predicted bin, we map the token pair back to the coordinate system of $\boldsymbol{\mathcal{I}_B}$ as,

$$\hat{\boldsymbol{v}}_i = \left(\frac{W_B}{K} \left(\hat{a}_i + \frac{1}{2} \right), \frac{H_B}{K} \left(\hat{b}_i + \frac{1}{2} \right) \right). \tag{10}$$

The query cardinality N is determined entirely by the input rather than being fixed by the model. Therefore, the same model naturally supports both single-query and variable-cardinality multi-query correspondence prediction through a unified coordinate-based interface.

C Additional Qualitative Comparisons

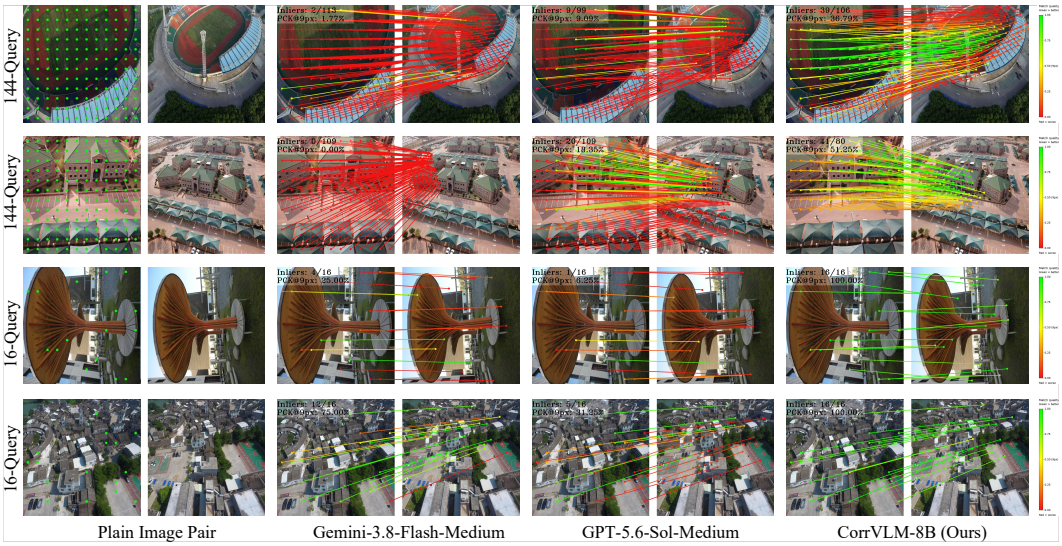


Figure 12: **Additional qualitative results for multi-query localization on AnyPoint-4K.** The first two rows show 144-Query aerial image pairs, and the last two show 16-Query pairs. The leftmost column shows the input pairs and query points in the reference view. The remaining columns visualize correspondences predicted by Gemini-3.8-Flash-Medium, GPT-5.6-Sol-Medium, and CorrVLM-8B. Line color indicates match quality from red (lower) to green (higher). Each result reports its inlier count and PCK@9px.

Warping. These examples extend the main-text warping comparison to different scene structures. In the wall-relief pair, GPT-5.6-Sol-Medium also produces a coherent alignment. For the sculpture and aerial pairs, CorrVLM-8B better preserves the subject's shape and the relative placement of scene elements, while several general-purpose outputs contain displaced or isolated fragments. Gaps remain in the CorrVLM-8B warps. Together, these results show how correspondence errors affect image-level alignment beyond the aggregate PCK@9px scores.

Multi-Query Localization. These examples extend the qualitative comparison in the main text by showing multi-query localization separately from image warping. In the two 144-Query aerial pairs, the general-purpose VLMs produce many low-quality correspondences across the source grid, while CorrVLM-8B localizes more points accurately throughout the scene. CorrVLM-8B also reaches 100% PCK@9px in both displayed 16-Query pairs. Its remaining errors in the 144-Query examples show that dense multi-query localization remains challenging even for a specialist model.

D Layer-wise Attention Analysis of CorrVLM

We extend the main-text attention analysis to LLM Transformer Layers 17–24 on three image pairs. Layers 17–21 respond to multiple structures in both views, including the corridor, entrance, and building facade. In Layers 22–24, attention becomes more selective around plausible matching regions. This progression is consistent with CorrVLM using cross-view scene context to identify a coarse location before focusing on local evidence for coordinate prediction. The maps provide qualitative support for this coarse-to-fine interpretation.

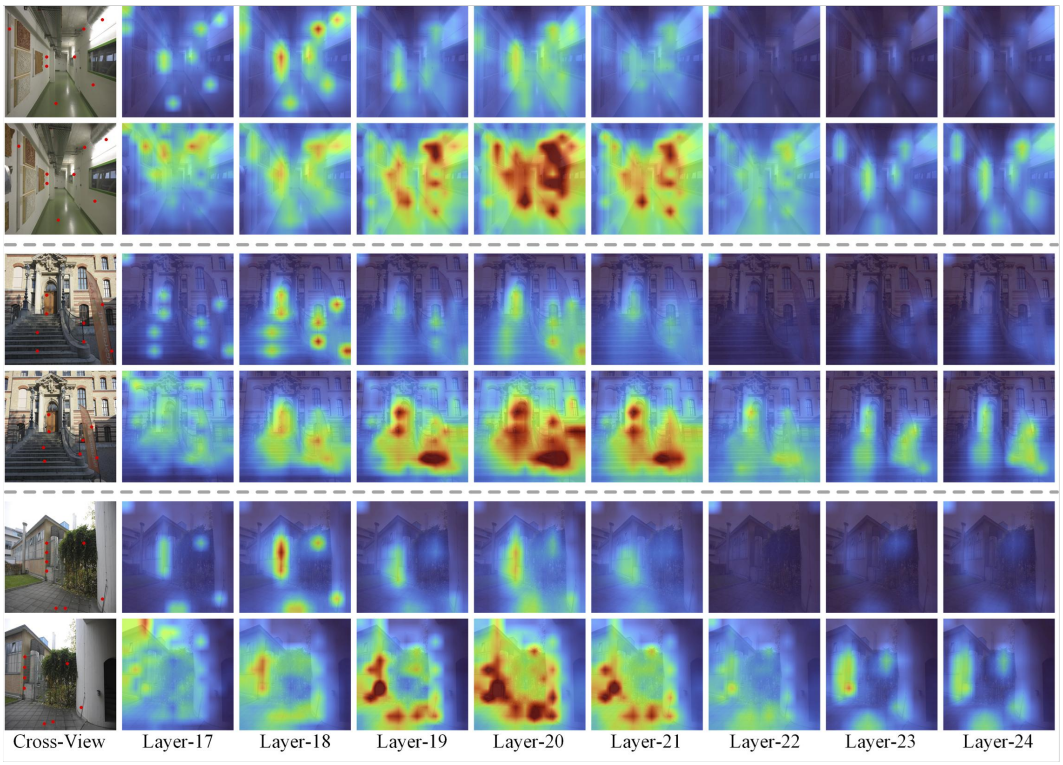


Figure 13: **Layer-wise attention of CorrVLM-8B during multi-query localization.** Three cross-view image pairs are shown, with the two views of each pair arranged in adjacent rows. Red dots mark selected points for visualization only, the model receives query coordinates as text. The remaining columns show attention maps from LLM Transformer Layers 17–24. Attention spans multiple scene regions in the earlier layers and becomes more concentrated around candidate matching regions in the later layers.