# ASKGEO: LEARNING TO QUERY GEOMETRY MODELS FOR SPATIAL REASONING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Spatial understanding requires vision-language models (VLMs) to identify the geometric evidence relevant to a given question, across both static and dynamic scenes. Despite their strong semantic capabilities, existing VLMs often struggle to recover 3D structure and reason about spatial relations directly from visual observations. Meanwhile, pretrained feed-forward geometry models encode rich geometric knowledge, but existing methods transfer it by extracting geometric features independently of the question and only afterwards selecting or fusing them, so the geometry model performs the same computation regardless of what is asked. We observe that query-based geometry models such as D4RT already provide a retrieval mechanism: their decoder retrieves geometric information for arbitrary spatiotemporal queries. Building on this, we introduce ASKGEO, a framework that lets a VLM learn to query a pretrained geometry model for the evidence a question requires. ASKGEO projects question-conditioned VLM features into the native query space of the geometry decoder, which retrieves geometric evidence tailored to the question and returns it to the VLM as a few geometry tokens. Because the queries are formed from the VLM's own representation of the question, the retrieved tokens are aligned with how the VLM interprets the question, and what to ask is learned end to end from question-answer supervision. Experiments on the static VSI-Bench and dynamic DSR-Bench show that ASKGEO achieves state-of-the-art performance on both benchmarks. Code will be released.

## 1 INTRODUCTION

Spatial understanding is a core capability for systems that must perceive and act in dynamic environments, such as autonomous vehicles and embodied agents. Beyond recognizing objects and their semantic attributes, it requires reasoning about scene geometry and how it evolves over time. Crucially, the geometric evidence needed for spatial reasoning is question-dependent. Identifying which object is closer requires comparing their positions, whereas determining whether an object is approaching the camera additionally requires disentangling object motion from camera motion. A general spatial reasoning system should therefore not only represent geometric structure, but also selectively retrieve the evidence relevant to each question.

Recent vision-language models (VLMs) have made substantial progress toward spatial intelligence. Existing approaches align language models with explicit 3D representations, including point clouds, reconstructed scenes, and 3D-aware visual tokens (Hong et al., 2023; Zhu et al., 2024; Zhang et al., 2026d), or improve spatial reasoning through geometrically grounded supervision, auxiliary geometric features or intermediate scene representations (Chen et al., 2024; Cheng et al., 2024; Yang et al., 2025; Qi et al., 2025; Zheng et al., 2026a; Fan et al., 2026). Nevertheless, current VLM representations remain primarily optimized for semantic recognition and language alignment, and often fail to reliably preserve metric geometry, camera and object motion (Zhou et al., 2025; Huang et al., 2026; Zhou & Lee, 2025; Chen et al., 2026b; Luo et al., 2026). Learning these capabilities entirely from visual-language supervision is also difficult, because precise geometric annotations and spatial question-answer data are expensive to obtain at scale.

In parallel, feed-forward geometry models have advanced rapidly. Recent models recover camera parameters, depth, and point maps from visual observations (Wang et al., 2025b;f; Lin et al., 2025; Keetha et al., 2026), while extensions to dynamic scenes additionally estimate temporal correspon-



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.

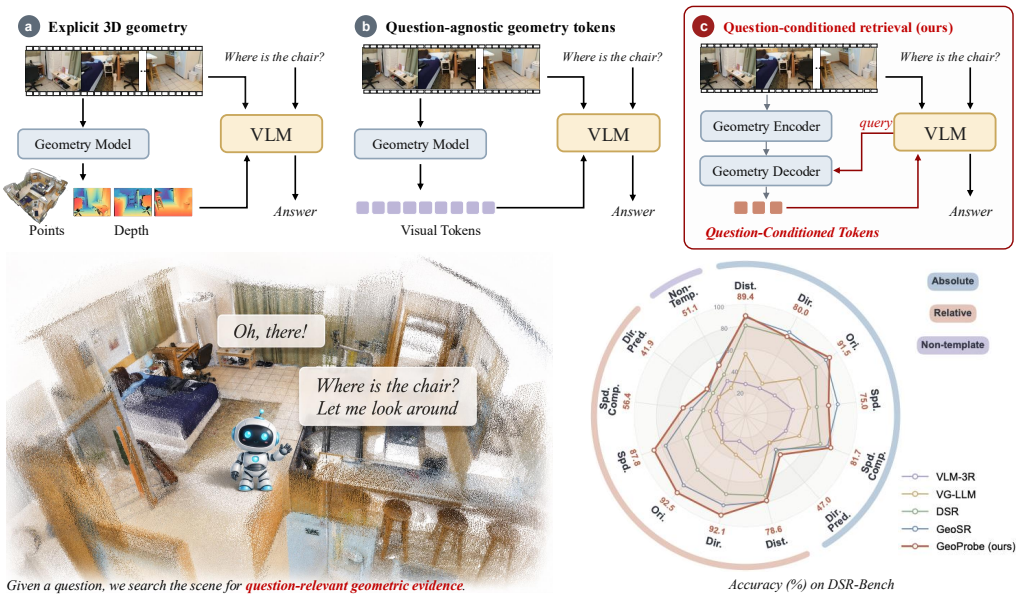


Figure 1: **Selecting geometry vs. asking for it.** Rather than first reconstructing explicit geometry (a) or providing a VLM with question-agnostic geometry features (b), AskGeo uses the language question to query the decoder of a pretrained geometry model (c) and retrieves compact, question-conditioned geometry tokens for spatial reasoning in both static and dynamic scenes.

dences and 3D point trajectories (Zhang et al., 2025a; Hu et al., 2026b; Wang et al., 2026; Chen et al., 2026a; Cong et al., 2026; Qian et al., 2026; Zhang et al., 2026b; Karhade et al., 2026). Although these models are typically trained for reconstruction rather than language understanding, their latent representations encode geometric cues required for spatial reasoning. This creates an opportunity to equip VLMs with strong geometric knowledge by reusing pretrained geometry models, rather than requiring them to relearn geometry solely through language-supervised training.

Effectively connecting geometry models and VLMs, however, remains nontrivial because the two model families are trained under different objectives and operate in different representation spaces. As illustrated in Figure 1(a), one strategy first predicts an explicit geometric representation, such as a depth map or point cloud, and then converts the reconstruction into an input that can be processed by the VLM (Chen et al., 2026c). While interpretable, this pipeline introduces additional reconstruction and serialization stages and may discard information retained in the latent representations of the geometry model. A second strategy, illustrated in Figure 1(b), directly transfers latent geometric features or geometry tokens into the VLM (Zheng et al., 2026a; Fan et al., 2026; Hu et al., 2026a; Zhang et al., 2026d;f). This avoids explicit conversion, but the geometric representations are generally extracted independently of the language question and are only subsequently selected, routed, or fused by an auxiliary module. Consequently, the geometry model performs the same geometric computation regardless of what information is actually requested.

We argue that *geometric information should be retrieved according to the question, rather than represented uniformly for all questions*. Our key observation is that query-based geometry models such as D4RT (Zhang et al., 2026b) already provide the mechanism for such retrieval: they encode a video once and decode geometric information for arbitrary spatiotemporal queries. A language question can be regarded as another form of query. As illustrated in Figure 1(c), we therefore map the question into the native query space of the geometry decoder, so that the VLM asks the geometry model for the evidence it needs instead of passively receiving generic geometry.

We introduce AskGeo, which enables a VLM to query a pretrained feed-forward geometry model for question-relevant geometric information. Given a language question, AskGeo extracts question-conditioned representations from the VLM and projects them into the native query space of the geometry decoder, where they replace coordinate-based queries to retrieve geometric evidence tailored to the question. The resulting features are fused across decoder layers and projected into

a small set of geometry tokens, which are processed by the VLM together with its visual and text tokens. Unlike Q-Former-based approaches (Zhou et al., 2026), which train an additional module to select from question-independent geometry features, AskGeo conditions the geometry decoding process itself. This design has two benefits. First, since the queries are formed from the VLM's own question representation, the retrieved geometry reflects how the VLM interprets the question and can be directly used by it. Second, although the geometry model remains frozen, gradients flow through its decoder to the query construction, allowing the model to learn *what to ask* from question-answer supervision, while the output projector aligns the retrieved geometry with the VLM.

We evaluate AskGeo on VSI-Bench (Yang et al., 2025), which evaluates spatial understanding in static scenes, and DSR-Bench (Zhou et al., 2026), which focuses on dynamic videos. AskGeo achieves state-of-the-art performance on both benchmarks, demonstrating the effectiveness of question-conditioned geometry retrieval across static scene structure and evolving spatiotemporal geometry. Ablation studies further show that our design outperforms Q-Former-based token selection with the same geometry model and VLM, and that the retrieved geometry depends on the question. Our contributions are summarized as follows.

- We introduce question-conditioned geometry retrieval, which incorporates the language question directly into geometry decoding, enabling the geometry model to retrieve geometric information relevant to the question.
- We propose AskGeo, which learns to query a pretrained geometry decoder with the VLM's own question representation using only question-answer supervision, so that the retrieved geometry tokens are both question-relevant and aligned with the VLM.
- We achieve state-of-the-art performance on both static and dynamic spatial reasoning benchmarks. Extensive experiments show that the retrieved geometry tokens contain useful question-relevant geometric information, and our retrieval is more effective than providing generic scene-level geometry or selecting geometry tokens with a Q-Former.

## 2 Related Work

### 2.1 Spatial Reasoning in Vision-Language Models

Existing approaches to spatial reasoning in VLMs can be broadly categorized into data-driven supervision, explicit spatial representation, and latent reasoning. Data-driven methods construct spatial question-answer pairs or geometry-aware instruction to teach metric, directional, and relational concepts through language supervision (Chen et al., 2024; Cheng et al., 2024; Fan et al., 2026). Representation-based methods augment visual tokens with depth, point clouds, spatial prompts, or 3D-aware representations (Hong et al., 2023; Zhu et al., 2024; Qi et al., 2025; Zheng et al., 2026a; Zhou & Lee, 2025). Latent reasoning methods instead encourage models to form and manipulate internal spatial states, such as cognitive maps, hierarchical spatial features, or continuous spatiotemporal representations (Yang et al., 2025; Chen et al., 2026b; Zhang et al., 2026d). Benchmarking efforts have followed a similar progression. Earlier benchmarks primarily evaluate metric relations, situated reasoning, and spatial understanding in static or persistent 3D environments (Azuma et al., 2022; Ma et al., 2023; Du et al., 2024; Yang et al., 2025), while recent benchmarks extend evaluation to dynamic scenes involving object motion, observer motion, temporal correspondence, and changing spatial relationships (Zhou et al., 2025; Zhang et al., 2025b; Huang et al., 2026).

### 2.2 Feed-Forward 3D and 4D Geometry Models

Feed-forward geometry models differ mainly in how they aggregate observations and represent scene state. Pairwise methods recover aligned pointmaps and correspondences from image pairs, forming the basis of optimization-free 3D reconstruction (Wang et al., 2024; Leroy et al., 2024). Global models process multiple observations to estimate cameras, depth, pointmaps, and tracks in a shared representation (Wang et al., 2025b;f; Lin et al., 2025; Keetha et al., 2026), while state-based models maintain recurrent scene states for online or streaming perception (Wang et al., 2025c). Dynamic models extend these formulations to time-varying scenes by jointly modeling camera motion, object motion, correspondence, and 3D trajectories (Zhang et al., 2025a; Karhade et al., 2026; Qian

et al., 2026; Liu et al., 2025). Query-based models formulate reconstruction and tracking as decoding arbitrary spatial and temporal queries from an encoded video representation (Zhang et al., 2026b; Jeon et al., 2026). Complementary depth and tracking models provide dense geometry or long-range trajectories, but typically focus on geometric outputs (Karaev et al., 2024; Zhang et al., 2026a; Yang et al., 2024). Collectively, these models organize visual information around geometry, motion, and correspondence rather than appearance alone.

### 2.3 Geometry-Enhanced Vision-Language Models

Existing geometry-enhanced VLMs differ mainly in how geometric knowledge is represented and transferred to the language model. Some methods convert geometric predictions into explicit representations, such as point clouds, reconstructed layouts, depth maps, structured scene descriptions, or persistent spatial memories, which are then encoded or queried by the VLM (Hong et al., 2023; Zhu et al., 2024; Qi et al., 2025; Sohn et al., 2026; He et al., 2026). Other methods transfer latent features from pretrained geometry models through visual-geometry fusion, hierarchical feature injection, geometry-aware reconstruction objectives, or representation distillation (Wu et al., 2026; Zheng et al., 2026a; Fan et al., 2026; Lee et al., 2025; Qu et al., 2026; Yu et al., 2026; Zhao et al., 2025; Zhang et al., 2026d;f;e; Wang & Huang, 2026; Hu et al., 2026a; Wang et al., 2025a; Zheng et al., 2026b; Chen et al., 2026c). These approaches preserve rich geometric information, but the transferred features are commonly determined by the visual input or fixed fusion architecture. Recent methods instead retrieve geometric evidence according to the current reasoning context. GeoThinker uses semantic visual representations to selectively query geometry features through cross-attention and gated integration (Li et al., 2026), while the GSM module of DSR Suite uses Q-Formers to summarize the language question and select relevant features from pretrained 4D geometry priors (Zhou et al., 2026; Li et al., 2023). A related idea appears in general-purpose VLMs, where InstructBLIP feeds the instruction into its Q-Former to extract instruction-aware visual features (Dai et al., 2023). Our method is most closely related to this active retrieval paradigm. However, instead of training an additional module to select from precomputed features, it directly projects question-conditioned language features into the native query space of a pretrained geometry decoder (Zhang et al., 2026b), allowing retrieval to be performed by the geometry model's original cross-attention mechanism.

## 3 Method

Spatial reasoning is inherently question-dependent: different questions may require different geometric evidence from the same visual observations. Rather than extracting a fixed geometric representation and relying on the VLM to identify the relevant information (Wu et al., 2026; Fan et al., 2026), AskGeo introduces the language question directly into the geometry decoding process, allowing the VLM to retrieve question-relevant geometric evidence from a pretrained geometry model. We first present the overall framework in Section 3.1, followed by question-conditioned geometry retrieval in Section 3.2 and geometry-token integration in Section 3.3.

### 3.1 Framework Overview

Given a video $V$ and a natural-language question $q$, our goal is to generate an answer $a$ that is grounded not only in visual appearance, but also in the geometric structure of the observed scene. As illustrated in Figure 2, our framework augments a pretrained VLM (Bai et al., 2025a) with geometric representations retrieved from a pretrained feed-forward geometry model. We consider query-based geometry models such as D4RT (Zhang et al., 2026b) and Point4D (Jeon et al., 2026), which share a common encoder–decoder formulation: a video encoder $\mathrm{Enc}$ first converts the input video into a spatiotemporal scene representation, $\mathbf{F} = \mathrm{Enc}(V)$, and a query-based decoder $\mathrm{Dec}$ subsequently retrieves geometric information from this representation.

The key idea is to introduce the language question directly into the geometry decoding process to retrieve question-relevant geometry. These decoders retrieve information from the encoded scene using explicit spatiotemporal queries; we instead construct queries from the question and map them into the pretrained decoder's query space, allowing the decoder to retrieve the geometric evidence required by the question. We keep the pretrained encoder and decoder unchanged, replace the original coordinate-based queries with question-conditioned queries, and use the resulting decoder features

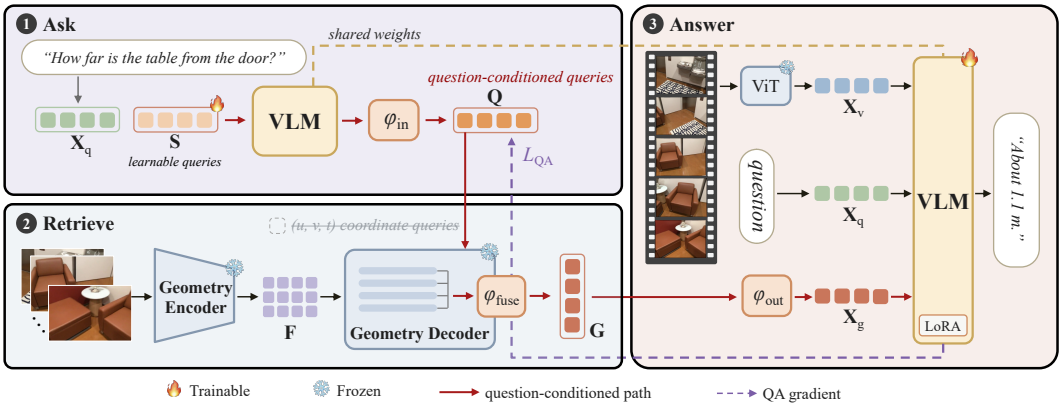


Figure 2: **Overview of AskGeo.** The language question is encoded together with a small set of learnable query slots and mapped into the native query space of a pretrained geometry decoder. The question-conditioned queries interact with the encoded spatiotemporal representation to retrieve relevant geometric information. Features from multiple decoder layers are fused and projected into the VLM embedding space as compact geometry tokens, which are combined with the original visual and language tokens for autoregressive answer generation.

as geometric tokens. These tokens are then integrated into the VLM to complement its visual representations with explicit 3D geometric information. We use D4RT as the main example, while the same formulation applies to other query-based geometry models.

### 3.2 Question-Conditioned Geometry Tokens

In models such as D4RT, the decoder takes space-time coordinates as queries and retrieves the corresponding geometric information from the encoded video. We repurpose this query mechanism by *replacing the coordinate-based queries with queries derived from the input question*. The resulting decoder features therefore encode geometric evidence conditioned on what is being asked.

**Geometry Query Construction.** A single question embedding forms a narrow bottleneck for retrieving all the geometric evidence required for spatial reasoning. We therefore represent each question with a small set of query slots. Specifically, we append $M$ learnable slot embeddings $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_M] \in \mathbb{R}^{M \times d}$ to the tokenized question $\mathbf{X}_\mathrm{q} \in \mathbb{R}^{L \times d}$ and process the resulting sequence with the language-model component of the VLM. The hidden states at the slot positions serve as question-conditioned query representations:

$$\mathbf{Z} = \mathrm{LM}\big([\mathbf{X}_\mathrm{q}; \mathbf{S}]\big)_{L+1:L+M} \in \mathbb{R}^{M \times d}, \tag{1}$$

where the subscript selects the last $M$ hidden states and LM denotes the language model. In practice, the slots are implemented as $M$ special `<geometry_query>` tokens appended to the question prompt. Because the language model uses causal attention and the slots appear after the question, each slot is conditioned on the complete question while being able to capture a different aspect of the required geometric information. The resulting representations describe what geometry should be retrieved in terms of the VLM's own understanding of the question, but they have not yet been grounded in the observed video.

**Question-Guided Geometry Decoding.** We map the question representations into the native query space of the pretrained geometry decoder using a lightweight projector,

$$\mathbf{Q} = \phi_{\mathrm{in}}(\mathbf{Z}) \in \mathbb{R}^{M \times d_g}. \tag{2}$$

These queries replace the original coordinate-based queries and interact with the encoded video representation $\mathbf{F}$ through the pretrained geometry decoder. In this way, the pretrained decoder acts as a geometry retrieval module whose output is directly conditioned on the question.

Rather than relying only on the final decoder layer, we aggregate query features from four evenly spaced layers to retain geometric information captured at different decoding depths. Let $\mathbf{H}^{(\ell)} =$

$\mathrm{Dec}^{(\ell)}(\mathbf{Q}, \mathbf{F})$ denote the query features at layer $\ell$. The final geometry tokens are obtained as

$$\mathbf{G} = \phi_{\mathrm{fuse}}\big(\mathbf{H}^{(\ell_1)}, \mathbf{H}^{(\ell_2)}, \mathbf{H}^{(\ell_3)}, \mathbf{H}^{(\ell_4)}\big) \in \mathbb{R}^{M \times d_g}, \tag{3}$$

where $\phi_{\mathrm{fuse}}$ is a lightweight MLP. The resulting $\mathbf{G}$ forms a compact set of question-conditioned geometry tokens for subsequent integration into the VLM.

**Relation to Geometry Token Selection.** Prior geometry-aware VLMs typically first extract a fixed set of geometric features and subsequently select or fuse them according to the question (Zhou et al., 2026; Zhang et al., 2026f). For example, GSM (Zhou et al., 2026) first obtains 3D tokens from a pretrained geometry encoder and then uses Q-Former modules to condense the question and select relevant information from these precomputed tokens. In contrast, ASKGEO injects question-derived queries directly into the native decoder of the pretrained geometry model. Question conditioning therefore occurs *during geometry decoding*, before the geometry tokens are formed, rather than as a subsequent selection step over a fixed geometry representation.

### 3.3 GEOMETRY TOKEN INTEGRATION

The retrieved representations $\mathbf{G}$ lie in the output space of the geometry decoder and therefore cannot be directly used by the VLM. We map them into the VLM hidden space using a lightweight projector $\phi_{\mathrm{out}}$, obtaining $M$ geometry tokens $\mathbf{X}_{\mathrm{g}}$. These tokens are then inserted between the original visual and language tokens:

$$\mathbf{X}_{\mathrm{g}} = \phi_{\mathrm{out}}(\mathbf{G}) \in \mathbb{R}^{M \times d}, \qquad \mathbf{X} = [\mathbf{X}_{\mathrm{v}}; \mathbf{X}_{\mathrm{g}}; \mathbf{X}_{\mathrm{q}}]. \tag{4}$$

Since the geometry tokens are retrieved by queries formed from the VLM's representation of the question and projected into its embedding space, they provide question-relevant geometric evidence in a form that the VLM can directly interpret. The VLM can therefore jointly reason over visual appearance, retrieved geometry, and the language question using its original attention and autoregressive generation architecture. This design requires no task-specific prediction head, leaves the VLM architecture unchanged apart from the added special tokens, and introduces only $M$ additional input tokens, independent of the video length.

### 3.4 TRAINING OBJECTIVE

We adopt a two-stage training strategy with the standard autoregressive language-modeling loss applied only to the answer tokens. Although the geometry model remains frozen, gradients are backpropagated through its decoder to $\phi_{\mathrm{in}}$ and the query slots, allowing the model to learn what geometric information to query, while $\phi_{\mathrm{fuse}}$ and $\phi_{\mathrm{out}}$ align the retrieved features with the VLM representation space. In *Stage 1*, both the pretrained VLM and geometry model are frozen, and only the geometry query slots and projection/fusion modules are trained. We remove the VLM's original visual tokens so that visual information is provided solely through the retrieved geometry tokens, encouraging the model to learn to retrieve and use geometric information for answering the question. In *Stage 2*, we restore the original visual tokens and adapt the language model of the VLM with LoRA (Hu et al., 2022), together with the geometry-related modules, while keeping the VLM visual encoder and geometry model frozen.

## 4 EXPERIMENTS

### 4.1 EXPERIMENTAL SETUP

**Dataset.** For **dynamic spatial reasoning**, we train on DSR-Train and evaluate on DSR-Bench (Zhou et al., 2026). DSR-Train contains approximately 50K question-answer pairs constructed from in-the-wild videos in Koala-36M (Wang et al., 2025d). DSR-Bench contains 1,484 question-answer pairs from 575 in-the-wild videos, where spatial relations evolve due to both object and camera motion. The benchmark covers absolute, relative, and non-template question types. Absolute subtasks include distance (*Dis.*), direction (*Dir.*), orientation (*Ori.*), and speed (*Spd.*), together with comparison and prediction variants such as speed comparison (*Spd. Comp.*) and direction prediction (*Dir. Pred.*). Relative subtasks evaluate analogous spatial relations with respect to

Table 1: Performance on DSR-Bench. Best results are in **bold**.

| Model | Abs. Dis. | Abs. Dir. | Abs. Ori. | Abs. Spd. | Abs. Spd. Comp. | Abs. Dir. Pred. | Rel. Dis. | Rel. Dir. | Rel. Ori. | Rel. Spd. | Rel. Spd. Comp. | Rel. Dir. Pred. | Non-Temp. Based | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Proprietary Models* | | | | | | | | | | | | | | |
| GPT-4o (Hurst et al., 2024) | 18.8 | 29.2 | 26.8 | 29.7 | 21.5 | 26.2 | 24.1 | 23.8 | 17.2 | 32.3 | 22.8 | 24.4 | 34.7 | 26.4 |
| GPT-5 (Singh et al., 2025) | 21.1 | 41.5 | 48.7 | 34.5 | 33.3 | 34.7 | 17.2 | 44.3 | 41.9 | 21.2 | 25.0 | 30.9 | 26.7 | 30.8 |
| Gemini-2.5-Flash (Comanici et al., 2025) | 18.8 | 27.6 | 19.5 | 25.0 | 23.6 | 22.0 | 11.2 | 28.4 | 30.8 | 23.2 | 17.1 | 22.6 | 38.8 | 24.9 |
| Gemini-2.5-Pro (Comanici et al., 2025) | 20.0 | 44.6 | 53.6 | 27.3 | 38.7 | 30.5 | 23.2 | 32.9 | 43.2 | 17.1 | 28.5 | 27.9 | 34.3 | 31.7 |
| *Video Understanding Models* | | | | | | | | | | | | | | |
| LLaVA-Video-7B (Zhang et al., 2024) | 22.3 | 16.9 | 25.6 | 33.3 | 45.1 | 24.5 | 24.1 | 15.9 | 17.2 | 19.1 | 24.2 | 21.4 | 33.9 | 25.9 |
| VideoRefer (Yuan et al., 2025) | 23.5 | 18.4 | 25.6 | 33.5 | 45.4 | 27.1 | 25.0 | 16.1 | 18.5 | 20.2 | 26.4 | 22.6 | 34.7 | 26.9 |
| LongVILA-R1 (Chen et al., 2025) | 20.0 | 21.5 | 23.1 | 21.4 | 37.6 | 22.8 | 24.1 | 26.1 | 28.3 | 22.2 | 17.8 | 20.8 | 33.9 | 25.3 |
| *General-Purpose Models* | | | | | | | | | | | | | | |
| Qwen2.5-VL-7B (Bai et al., 2025b) | 18.8 | 15.3 | 14.6 | 42.8 | 29.0 | 19.4 | 31.8 | 19.3 | 11.1 | 22.2 | 19.2 | 20.2 | 30.1 | 23.5 |
| Qwen2.5-VL-32B (Bai et al., 2025b) | 31.7 | 21.5 | 23.1 | 44.0 | 36.5 | 25.4 | 27.5 | 23.8 | 37.0 | 27.2 | 29.2 | 21.4 | 36.2 | 29.9 |
| Qwen3-VL-8B-Instruct (Bai et al., 2025a) | 23.5 | 24.6 | 42.6 | 29.7 | 27.9 | 33.8 | 18.1 | 28.4 | 34.5 | 24.2 | 22.1 | 27.9 | 33.5 | 28.7 |
| Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025a) | 25.8 | 27.6 | 46.3 | 30.9 | 31.1 | 34.7 | 20.6 | 29.5 | 37.0 | 28.2 | 24.2 | 31.5 | 35.4 | 31.1 |
| InternVL3.5-8B (Wang et al., 2025e) | 23.5 | 27.6 | 28.0 | 34.5 | 24.7 | 27.9 | 22.4 | 17.0 | 19.7 | 28.2 | 30.0 | 14.2 | 30.1 | 25.4 |
| InternVL3.5-38B (Wang et al., 2025e) | 25.8 | 27.8 | 29.2 | 34.2 | 24.7 | 28.5 | 26.7 | 16.3 | 23.4 | 29.2 | 32.1 | 15.4 | 31.3 | 26.7 |
| *Spatial Reasoning Models* | | | | | | | | | | | | | | |
| VLM-3R (Fan et al., 2026) | 28.2 | 27.6 | 31.7 | 42.8 | 38.7 | 33.0 | 34.4 | 23.8 | 30.8 | 22.2 | 26.4 | 29.1 | 35.0 | 31.4 |
| VG-LLM (Zheng et al., 2026a) | 55.2 | 32.3 | 58.5 | 57.1 | 51.6 | 32.2 | 56.0 | 36.3 | 32.0 | 30.3 | 32.1 | 29.1 | 27.9 | 38.4 |
| GSM (Zhou et al., 2026) | 87.0 | 73.8 | 84.1 | 73.8 | 72.0 | 35.5 | 75.8 | 76.1 | 77.7 | 60.6 | 37.1 | 35.1 | 46.4 | 58.9 |
| GeoSR (Zhang et al., 2026f) | 88.0 | **84.4** | 87.8 | **83.3** | 78.5 | 41.0 | **79.5** | 83.0 | 83.8 | 76.5 | 48.9 | 39.8 | **52.9** | 66.1 |
| AskGeo (Ours) | **89.4** | 80.0 | **91.5** | 75.0 | **81.7** | **47.0** | 78.6 | **92.1** | **92.5** | **87.8** | **56.4** | **41.9** | 51.1 | **68.8** |

another entity, time point, or reference frame. The non-template subset (*Non-Temp.*) further evaluates dynamic spatial reasoning with more flexible linguistic expressions beyond predefined question templates. For **static spatial reasoning**, we evaluate on VSI-Bench (Yang et al., 2025); the corresponding dataset details are provided in the Appendix.

**Training and Evaluation.** For **training**, we adopt a two-stage strategy for both static and dynamic spatial reasoning. In *Stage 1*, we freeze both the VLM and the geometry model and optimize only the query slots and the projectors $\phi_{\text{in}}$, $\phi_{\text{fuse}}$, and $\phi_{\text{out}}$ (40.6M trainable parameters). To isolate the contribution of the geometry representation, we *remove the original visual tokens of the VLM* and condition the model solely on the language question and the retrieved geometry tokens $\mathbf{X}_{\text{g}}$, so that the Stage 1 results in Table 2 directly reflect the quality of the retrieved geometry. In *Stage 2*, we restore the original visual tokens and use LoRA (Hu et al., 2022) to adapt the language model of the VLM, while keeping the geometry model frozen. The visual tokens are randomly dropped with probability $10\%$, encouraging the model to use the geometry in $\mathbf{X}_{\text{g}}$ rather than relying mainly on appearance cues. For **evaluation**, we follow the official protocols of both benchmarks, reporting mean relative accuracy for numerical questions and the average over all eight subtasks on VSI-Bench, and exact-match accuracy for each question type and overall on DSR-Bench.

**Implementation details.** We adopt Qwen3-VL-Instruct (Bai et al., 2025a) (8B) as the VLM and Open-D4RT,[1] an open-source implementation of D4RT (Zhang et al., 2026b), as the geometry model. Unless otherwise specified, we use $M = 32$ geometry queries and extract geometry features from $K = 4$ D4RT decoder layers. For Stage 2, LoRA (Hu et al., 2022) is applied to the VLM with rank 16, scaling factor 32, and dropout $0.05$. Both training stages use AdamW with a learning rate of $1 \times 10^{-4}$ and a global batch size of 32.

### 4.2 Main Results

Table 1 reports the results on DSR-Bench. Our method achieves the best overall accuracy of $68.8$, outperforming GSM and GeoSR by $9.9$ and $2.7$ points, respectively. The improvements are particularly substantial in relative direction, orientation, speed, and speed comparison, where reasoning requires understanding object motion and its relationship to camera motion. These results show that retrieving question-relevant geometry provides the VLM with useful geometric and motion cues for dynamic spatial reasoning. Additional results on VSI-Bench are provided in the Appendix.

### 4.3 Ablation Studies

We conduct ablation studies to examine the effectiveness of the retrieved geometry tokens. Specifically, we study whether the geometry tokens provide useful spatial evidence, how different ways of

[1] https://github.com/Lijiaxin0111/Open-d4rt

retrieving question-conditioned geometry affect performance, and whether the retrieved information is specific to the input question rather than a generic scene representation. Additional ablations on model configurations, including the number of geometry tokens, the decoder layers used for aggregation, and the hidden dimension of the projector $\phi_{\text{in}}$, are provided in the Appendix.

Table 2: Effectiveness of geometry tokens.

| Benchmark | Model | Acc. |
|---|---|---|
| DSR-Bench | Stage 1 | 65.7 |
| | Stage 2 | 68.8 |
| VSI-Bench | Stage 1 | 49.7 |
| | Stage 2 | 52.2 |

Table 3: Comparison of retrieval designs.

| Variant | Acc. |
|---|---|
| Q-Former selection | 60.9 |
| Dense-grid queries | 66.8 |
| Question token embedding | 65.3 |
| AskGeo | 68.8 |

Table 4: Question dependence of retrieval.

| Variant | Acc. |
|---|---|
| No question | 45.0 |
| Other question | 49.4 |
| Original question | 68.8 |

**Effectiveness of Geometry Tokens.** We first investigate whether the retrieved geometry tokens provide meaningful spatial evidence. As shown in Table 2, even without the original visual tokens (Stage 1), the model achieves strong performance on both benchmarks, demonstrating that the geometry tokens alone contain rich information for spatial reasoning. Since the language model is frozen in Stage 1, this also shows that the retrieved tokens can be directly interpreted by the original VLM, consistent with our design of forming the queries from its question representation.

**Comparison of Retrieval Designs.** We compare different ways of retrieving geometry conditioned on the input question on DSR-Bench, as shown in Table 3. For Q-Former-based token selection, we reproduce the GSM design (Zhou et al., 2026) using the same D4RT and Qwen3-VL backbones as our method. GSM first decodes question-independent geometric features and then selects relevant tokens with a question-conditioned Q-Former, underperforming our method by 7.9 points. This substantial gap highlights the benefit of our design, which introduces the question directly into geometry decoding rather than selecting from precomputed geometry afterwards. Replacing our question-conditioned decoder queries with dense-grid queries decreases accuracy by 2.0 points, further showing that targeted retrieval is more effective than extracting generic scene-level geometry. Finally, directly using question token embeddings instead of the VLM-derived question representation reduces accuracy by 3.5 points, showing that forming the queries from the VLM's own understanding of the question is important for retrieving geometry that the VLM can effectively use.

**Question Dependence of Retrieved Geometry.** We further analyze how strongly the retrieved geometry tokens depend on the input question. In Table 4, we keep the video input and LLM prompt fixed while modifying only the question used to construct geometry queries. Removing the question alone leads to a substantial drop of 23.8 points, indicating that language guidance plays a critical role in retrieving useful geometric evidence. Replacing the original question with another question type also reduces accuracy by 19.4, showing that different questions require different geometric information. Together, these results confirm that our method retrieves question-relevant geometric information, rather than simply encoding question-agnostic information about the entire scene.

**Ablation on Different Geometry Models and VLMs.** We ablate the choice of geometry model and VLM backbone in Table 5. Compared with VLM-only SFT, incorporating geometry consistently improves both VLM backbones. D4RT and Point4D achieve comparable results, suggesting robustness across geometry models. Point4D performs best with Qwen2.5-VL (69.2), while D4RT performs slightly better with Qwen3-VL (68.8). Overall, question-conditioned geometry tokens provide effective spatial cues across different VLM and geometry backbones.

Table 5: Performance with different geometry models and VLM.

| Geometry Model | VLM Backbone | |
|---|---|---|
| | Qwen3-VL | Qwen2.5-VL |
| None (SFT only) | 56.8 | 54.4 |
| D4RT (Zhang et al., 2026b) | 68.8 | 68.5 |
| Point4D (Jeon et al., 2026) | 68.4 | 69.2 |

### 4.4 Qualitative Results

Figure 3 compares AskGeo with GeoSR (Zhang et al., 2026f). Our method consistently handles cases requiring temporal motion understanding, including relative position changes, speed comparison, multi-stage motion variation, and pairwise distance reasoning. These examples show that our method better captures object correspondence and evolving spatial relations over time, enabling more reliable spatial reasoning. Additional qualitative results are provided in the Appendix.

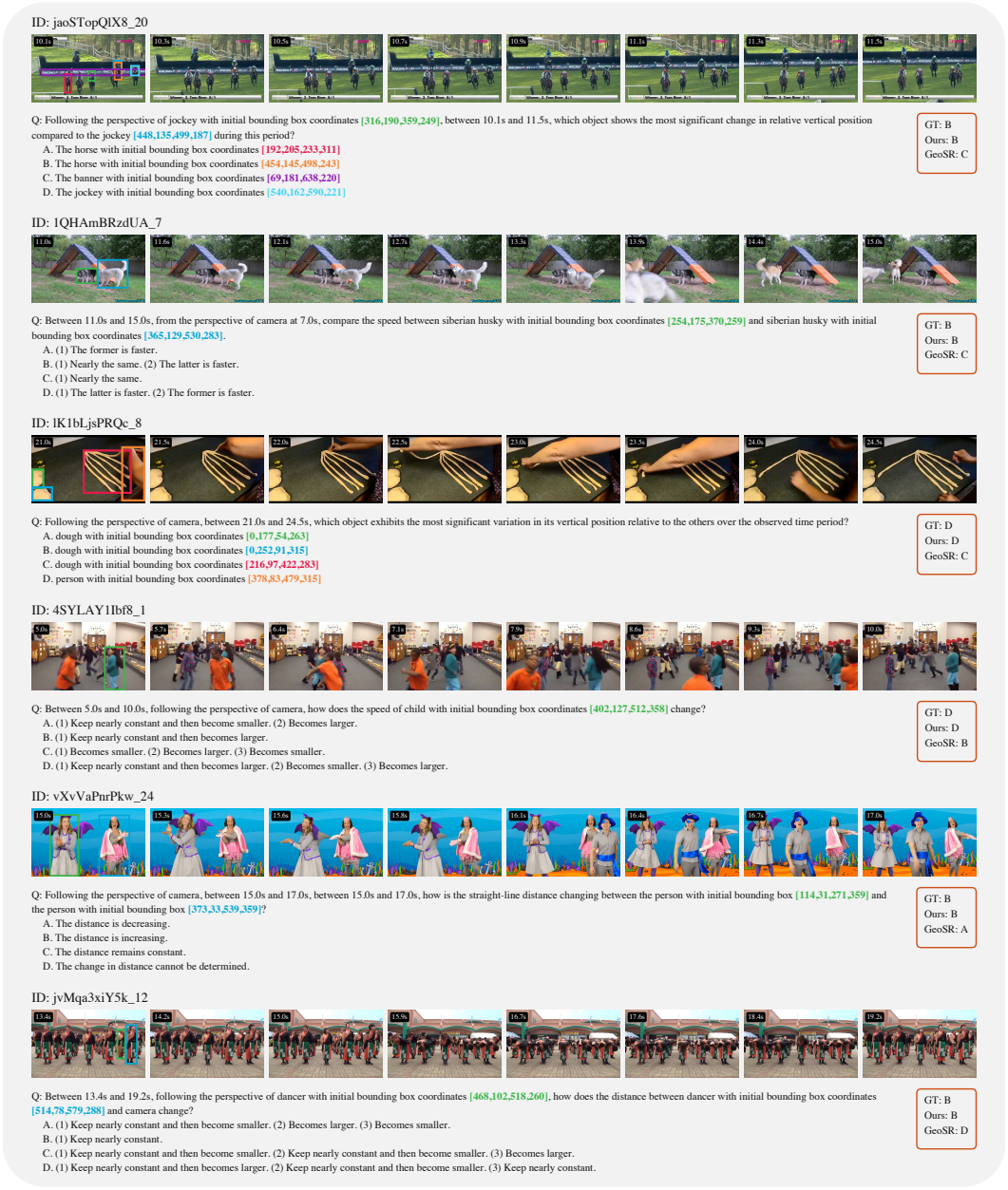


Figure 3: Qualitative comparisons with GeoSR (Zhang et al., 2026f) on dynamic spatial reasoning. Our method shows more reliable reasoning over temporal motion and spatial relations.

## 5 Conclusion

We present AskGeo, which enables a VLM to ask a pretrained geometry model for the geometric evidence required by each question. AskGeo introduces the question directly into the geometry decoding process and learns what to query from question-answer supervision, so that the retrieved geometry is both relevant to the question and aligned with the VLM. AskGeo achieves state-of-the-art performance on DSR-Bench and VSI-Bench, with particularly strong improvements on dynamic spatial reasoning involving relative direction, orientation, and speed. Our analysis further confirms that the retrieved geometry is conditioned on the question and provides effective spatial evidence for reasoning. These results demonstrate the effectiveness of question-conditioned geometry retrieval for leveraging pretrained geometric knowledge in VLM spatial reasoning.

## AI USE STATEMENT

We used Generative AI tools to polish the writing of author-drafted text and to assist in retrieving related literature. All AI-suggested edits were manually reviewed, and every cited reference was verified by the authors for existence. We take full responsibility for the content of this paper.

## REFERENCES

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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*, 2025a.

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

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024.

Xiaoxue Chen, Ziyi Xiong, Yuantao Chen, Gen Li, Nan Wang, Hongcheng Luo, Long Chen, Haiyang Sun, Bing Wang, Guang Chen, et al. Dggt: Feedforward 4d reconstruction of dynamic driving scenes using unposed images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1265–1276, 2026a.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. In *International Conference on Learning Representations*, volume 2025, pp. 18227–18246, 2025.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xiang An, Bo Li, Xin Xie, ZiDong Wang, Mingze Sun, Shuang Chen, Hongyu Li, et al. 4dthinker: Thinking with 4d imagery for dynamic spatial understanding. *arXiv preprint arXiv:2605.05997*, 2026b.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Xiang An, Yan Feng, Peng Pei, Xunliang Cai, et al. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2613–2624, 2026c.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Zhongxiao Cong, Qitao Zhao, Minsik Jeon, and Shubham Tulsiani. Flow3r: Factored flow prediction for scalable visual geometry learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 438–447, 2026.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

Mengfei Du, Binhao Wu, Zejun Li, Xuan-Jing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 346–355, 2024.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31054–31065, 2026.

Jixuan He, Xueting Li, Chieh Hubert Lin, and Ming-Hsuan Yang. Reasmory: 3d reconstruction as explicit memory for vlms spatial reasoning. *arXiv preprint arXiv:2606.00963*, 2026.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Wenbo Hu, Jingli Lin, Yilin Long, Yunlong Ran, Lihan Jiang, Yifan Wang, Chenming Zhu, Runsen Xu, Tai Wang, and Jiangmiao Pang. $G^2$vlm: Geometry grounded vision language model with unified 3d reconstruction and spatial reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9535–9546, 2026a.

Yu Hu, Chong Cheng, Sicheng Yu, Xiaoyang Guo, and Hao Wang. Vggt4d: Mining motion cues in visual geometry transformers for 4d scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 414–424, 2026b.

Yuzhi Huang, Kairun Wen, Rongxin Gao, Dongxuan Liu, Yibin Lou, Jie Wu, Jing Xu, Jian Zhang, Zheng Yang, Yunlong Lin, et al. Thinking in dynamics: How multimodal large language models perceive, track, and reason dynamics in physical 4d world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33446–33456, 2026.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Minsik Jeon, Jay Karhade, Deva Ramanan, and Shubham Tulsiani. Point4d: Long-range 4d motion reconstruction. *arXiv preprint arXiv:2609.09145*, 2026.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European conference on computer vision*, pp. 18–35. Springer, 2024.

Jay Karhade, Nikhil Keetha, Yuchen Zhang, Tanisha Gupta, Akash Sharma, Sebastian Scherer, and Deva Ramanan. Any4d: Unified feed-forward metric 4d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14578–14589, 2026.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *International Conference on 3D Vision (3DV)*. IEEE, 2026.

Kanghee Lee, Jungi Hong, Sion Lee, Injae Lee, Minseok Kwak, Kwonyoung Ryu, and Jaesik Park. Spatialmosaic: A multiview vlm dataset for partial visibility. *arXiv preprint arXiv:2512.23365*, 2025.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Haoyuan Li, Qihang Cao, Tao Tang, Kun Xiang, Zihan Guo, Jianhua Han, JiaWang Bian, Hang Xu, and Xiaodan Liang. Thinking with geometry: Active geometry integration for spatial reasoning. *arXiv preprint arXiv:2602.06037*, 2026.

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

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Xinhang Liu, Yuxi Xiao, Donny Y Chen, Jiashi Feng, Yu-Wing Tai, Chi-Keung Tang, and Bingyi Kang. Trace anything: Representing any video in 4d via trajectory fields. *arXiv preprint arXiv:2510.13802*, 2025.

Yueru Luo, Xu Yan, Changqing Zhou, Yiming Yang, Chao Zhan, Shuqi Mei, Chao Zheng, and Zhen Li. Reasoning to regulate: Chain-of-thought for traffic rule understanding. *arXiv preprint arXiv:2607.24199*, 2026.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=IDJx97BC38.

Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. *arXiv preprint arXiv:2501.01428*, 2025.

Shenhan Qian, Ganlin Zhang, Shangzhe Wu, and Daniel Cremers. Flow4r: Unifying 4d reconstruction and tracking with scene flow. *arXiv preprint arXiv:2602.14021*, 2026.

Kevin Qu, Haozhe Qi, Mihai Dusmanu, Mahdi Rad, Rui Wang, and Marc Pollefeys. Loc3r-vlm: Language-based localization and 3d reasoning with vision-language models. *arXiv preprint arXiv:2603.18002*, 2026.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, et al. Sat: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Tin Stribor Sohn, Maximilian Dillitzer, Jason J Corso, and Eric Sax. R4: Retrieval-augmented reasoning for vision-language models in 4d spatio-temporal space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38711–38721, 2026.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Haibo Wang and Lifu Huang. Learning geometric representations from videos for spatial intelligent multimodal large language models. *arXiv preprint arXiv:2606.05833*, 2026.

Haochen Wang, Yucheng Zhao, Tiancai Wang, Haoqiang Fan, Xiangyu Zhang, and Zhaoxiang Zhang. Ross3d: Reconstructive visual instruction tuning with 3d-awareness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9275–9286, 2025a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Jingkang Wang, Henry Che, Yun Chen, Ze Yang, Lily Goli, Sivabalan Manivasagam, and Raquel Urtasun. Flux4d: Flow-based unsupervised 4d reconstruction. *Advances in Neural Information Processing Systems*, 38:105673–105693, 2026.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10510–10522, 2025c.

Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8428–8437, 2025d.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20697–20709, 2024.

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

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. $\pi^3$: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025f.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *Advances in neural information processing systems*, 38: 13569–13597, 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 Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10371–10381, 2024.

Hanxun Yu, Xuan Qu, Lei Ke, Boqiang Zhang, Yuxin Wang, Jianke Zhu, and Dong Yu. Stream3d-vlm: Online 3d spatial understanding with incremental geometry priors. *arXiv preprint arXiv:2606.06891*, 2026.

Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, et al. Videorefer suite: Advancing spatial-temporal object understanding with video llm. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18970–18980, 2025.

Bowei Zhang, Lei Ke, Adam Harley, and Katerina Fragkiadaki. Tapip3d: Tracking any point in persistent 3d geometry. *Advances in Neural Information Processing Systems*, 38:135284–135303, 2026a.

Chuhan Zhang, Guillaume Le Moing, Skanda Koppula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle K Barral, Raia Hadsell, et al. Efficiently reconstructing dynamic scenes one d4rt at a time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7382–7392, 2026b.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Jilin Mei, Chunhui Chen, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *Advances in Neural Information Processing Systems*, 38, 2026c.

Jian Zhang, Shijie Zhou, Bangya Liu, Achuta Kadambi, and Zhiwen Fan. Spatialstack: Layered geometry-language fusion for 3d vlm spatial reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38678–38688, 2026d.

Jiaxin Zhang, Junjun Jiang, Haijie Li, Youyu Chen, Kui Jiang, and Dave Zhenyu Chen. Gap-mllm: Geometry-aligned pre-training for activating 3d spatial perception in multimodal large language models. *arXiv preprint arXiv:2603.16461*, 2026e.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=lJpqxFgWCM.

Shihua Zhang, Qiuhong Shen, Shizun Wang, Tianbo Pan, and Xinchao Wang. Make geometry matter for spatial reasoning. *arXiv preprint arXiv:2603.26639*, 2026f.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

Ziang Zhang, Zehan Wang, Guanghao Zhang, Weilong Dai, Yan Xia, Ziang Yan, Minjie Hong, and Zhou Zhao. Dsi-bench: A benchmark for dynamic spatial intelligence. *arXiv preprint arXiv:2510.18873*, 2025b.

Ruosen Zhao, Zhikang Zhang, Jialei Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. Spacemind: Camera-guided modality fusion for spatial reasoning in vision-language models. *arXiv preprint arXiv:2511.23075*, 2025.

Duo Zheng, Yanyang Li, Liwei Wang, et al. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *Advances in neural information processing systems*, 38:20560–20586, 2026a.

Hongpei Zheng, Lintao Xiang, Lin Qian, Zhenghao Li, Qijun Yang, and Hujun Yin. Reg3d: Reconstructive geometry instruction tuning for 3d scene understanding. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1301–1305. IEEE, 2026b.

Hanyu Zhou and Gim Hee Lee. Llava-4d: Embedding spatiotemporal prompt into lmms for 4d scene understanding. In *The Fourteenth International Conference on Learning Representations*, 2025.

Shengchao Zhou, Yuxin Chen, Yuying Ge, Wei Huang, Jiehong Lin, Ying Shan, and Xiaojuan Qi. Learning to reason in 4d: Dynamic spatial understanding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9637–9646, 2026.

Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. Vlm4d: Towards spatiotemporal awareness in vision language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8600–8612, 2025.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.