MAVER: MEMORY-AUGMENTED VIEW-DECOUPLED EVIDENCE ROUTING FOR VIDEO SPATIAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video spatial reasoning requires integrating spatial cues that are distributed across time and viewpoints. Recent geometry-grounded video-language models provide increasingly rich reconstruction and camera-aware features, yet these dense representations are commonly compressed before language reasoning, which can obscure question-relevant spatial evidence. We introduce MAVER, a structured evidence interface that operates on high-resolution geometry-grounded video features before spatial pooling. MAVER first conditions current spatial features on previously observed context, then organizes them into cross-view shared and observation-specific evidence, and finally performs question-conditioned sparse routing. The selected evidence follows a dual-path injection scheme that both guides the dense visual stream before pooling and preserves selected evidence through a direct projected bypass to the language model. On VSI-Bench and VSTI-Bench, MAVER achieves average scores of 64.4 and 61.6, respectively, including a 7.1-point improvement over the controlled geometry-grounded baseline on VSI-Bench. These results suggest that structured evidence transformation before language reasoning provides an effective way to improve geometry-grounded video spatial reasoning.

1 INTRODUCTION

Video spatial reasoning often depends on observations separated across time and camera poses. As a camera moves, objects can leave the field of view, apparent relations can change with perspective, and evidence needed by a later question may have appeared only in an earlier observation. Recent work on spatial-temporal video reasoning and persistent 3D-aware memory therefore emphasizes preserving information beyond the current frame (Fan et al., 2026; Liu et al., 2026).

At the same time, geometry-grounded video-language models have substantially improved the spatial information available to language reasoning. Video-3D LLM introduces position-aware video representations (Zheng et al., 2025b); VG-LLM injects priors from a visual geometry encoder (Zheng et al., 2025a); VLM-3R couples video-language learning with instruction-aligned 3D reconstruction (Fan et al., 2026); and G^2VLM jointly models reconstruction and spatial reasoning (Hu et al., 2026). These methods strengthen geometric perception, but dense geometry-grounded features still need to be compressed and presented to the language model in a form that supports the question being asked.

This leaves an interface problem between geometric perception and language reasoning. Features collected at different times carry different temporal context, while a moving camera mixes scene properties that persist across views with observations tied to a particular viewpoint. Before reasoning, these heterogeneous cues also compete for a limited visual token budget. Existing work illuminates different aspects of this problem: persistent memory improves long-horizon spatial reasoning (Liu et al., 2026); GeoMind benefits from separating stable allocentric structure from view-conditioned egocentric evidence (Wei et al., 2026); and GeoThinker shows that geometric cues can be retrieved according to the model's reasoning demand instead of being fused indiscriminately (Li et al., 2026). However, these ideas are realized as separate memory, explicit-map, or geometry-integration mechanisms. How to transform dense geometry-grounded video features immediately before spatial pooling remains underexplored.

Motivated by this gap, we introduce **MAVER**, a **M**emory-**A**ugmented **V**iew-Decoupled **E**vidence **R**outing framework that operates on high-resolution geometry-grounded video features before spatial

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.

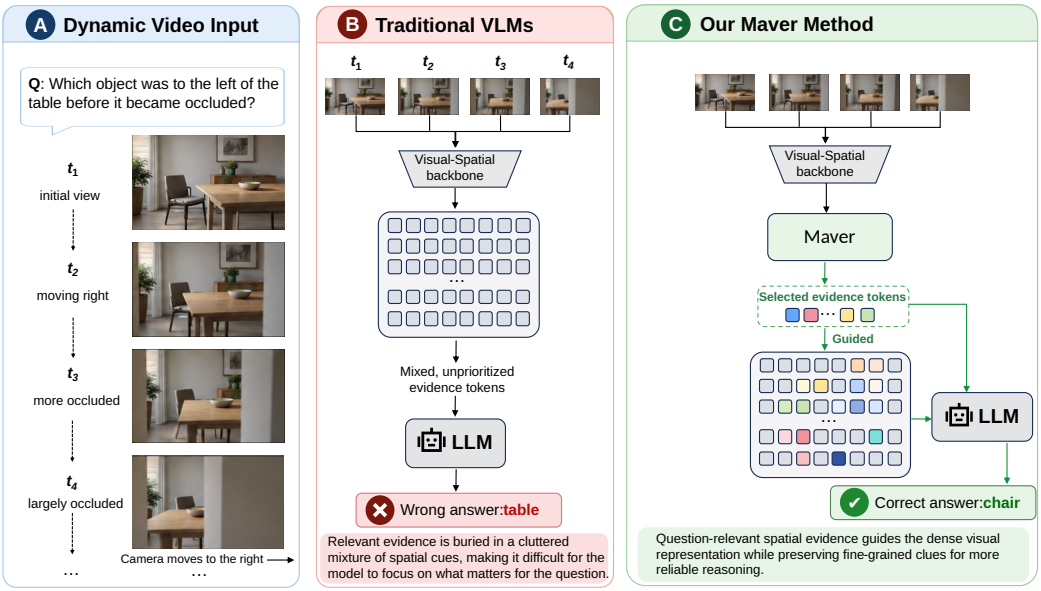

Figure 1: **Motivation and overview of MAVER.** (A) A spatial question may depend on evidence observed before viewpoint change or occlusion. (B) Directly compressing dense mixed evidence can bury the relevant clue. (C) MAVER organizes and selects question-relevant spatial evidence, using it both to guide the dense representation and to preserve selected evidence for language reasoning.

pooling. As illustrated in Figure 1, MAVER first uses **Causal Temporal-Spatial Memory (CTSM)** to condition current spatial tokens on recurrent history and local scene change. **Dual Evidence Decoupling (DED)** then organizes the temporally contextualized representation into view-invariant and view-specific evidence banks. Finally, inspired by active geometry integration (Li et al., 2026), **Sparse Evidence Routing (SER)** scores the structured evidence with the language question and retains a fixed Top-K subset. The selected evidence is used both to refine the dense visual stream before pooling and to provide a direct projected evidence stream to the language model.

MAVER achieves average scores of 64.4 on VSI-Bench and 61.6 on VSTI-Bench. Controlled ablations further show that the three stages contribute differently across temporal, relational, and routing-sensitive categories.

Our contributions are threefold:

- We introduce MAVER and a causal temporal-spatial memory that injects recurrent history and explicit local scene change into high-resolution geometry-grounded features before spatial pooling.
- We propose dual evidence decoupling, which constructs complementary view-invariant and view-specific evidence banks to organize persistent scene structure and observation-dependent cues in a compact latent space.
- We propose question-guided sparse evidence routing with dual-path injection, which selects structured evidence according to the query, uses it to refine the dense stream before pooling, and projects the selected evidence directly into the LLM input space.

2 Related Work

2.1 Geometry-Grounded Video Spatial Reasoning

General multimodal architectures connect visual representations to language models through learned cross-modal interfaces, including Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023), and LLaVA (Liu et al., 2023); MVBench systematizes evaluation across diverse video-understanding

tasks (Li et al., 2024). Spatial reasoning work extends this paradigm by introducing explicit geometric structure: 3D-LLM grounds language models in 3D scene representations (Hong et al., 2023), SpatialRGPT augments grounded reasoning with depth-aware spatial cues (Cheng et al., 2024), SpatialVLM scales spatial supervision for metric and qualitative reasoning (Chen et al., 2024a), and LLaVA-3D injects 3D-aware representations into multimodal language models (Zhu et al., 2025a).

Progress in visual geometry foundation models has further strengthened the geometric priors available from RGB observations. DUSt3R performs feed-forward multi-view reconstruction (Wang et al., 2024), MASt3R extends this line toward geometry-grounded image matching (Leroy et al., 2024), VGGT predicts cameras and 3D structure with a feed-forward transformer (Wang et al., 2025a), and CUT3R introduces a persistent recurrent state for continuous 3D perception (Wang et al., 2025b). Building on such priors, Video-3D LLM learns position-aware video representations (Zheng et al., 2025b), VG-LLM fuses visual tokens with geometry-encoder features (Zheng et al., 2025a), Spatial-MLLM combines semantic and spatial encoders with spatially informative frame sampling (Wu et al., 2025), VLM-3R injects reconstruction-aligned spatial and camera tokens into video-language reasoning (Fan et al., 2026), and G^2VLM unifies 3D reconstruction and spatial reasoning within one model (Hu et al., 2026). MAVER is complementary to these approaches: it does not introduce a stronger geometry encoder, but reorganizes already geometry-grounded high-resolution features before pooling and language reasoning.

2.2 Temporal and View-Structured Spatial Evidence

Long-video models have explored retaining information beyond a local clip. VideoStreaming propagates a compact memory through sequential clips and adaptively selects question-related historical memories for downstream reasoning (Qian et al., 2024). For spatial reasoning specifically, VLM^2 maintains working and episodic memories over view-consistent 3D-aware representations, demonstrating the value of persistent memory for long-horizon spatial questions (Liu et al., 2026). These methods motivate temporal retention. Our approach instead uses causal temporal conditioning to shape the *formation* of each current high-resolution evidence representation before subsequent pooling.

A complementary line structures evidence according to spatial reference or intermediate world representations. GeoMind explicitly separates allocentric topology from egocentric, view-conditioned evidence in a dual-reference map (Wei et al., 2026); Map2Thought performs spatial reasoning over an explicit metric cognitive map (Gao et al., 2026a); and 3DThinker introduces geometry-grounded imagination for reasoning from limited views (Chen et al., 2026). These works show that spatial evidence benefits from representation structure rather than an undifferentiated feature stream (Wei et al., 2026; Gao et al., 2026a; Chen et al., 2026). Our view-decoupled evidence construction follows this broader motivation but remains a latent, differentiable pre-pooling interface: shared queries and cross-frame aggregation bias one bank toward persistent evidence, while frame-conditioned queries preserve observation-specific evidence in the other.

2.3 Active and Question-Conditioned Geometry Selection

A growing line of work studies not only how geometry is represented, but also *which* geometric evidence should participate in reasoning. GeoThinker contrasts passive global fusion with active perception and selectively retrieves geometry according to semantic reasoning demand (Li et al., 2026). MSSR similarly formulates spatial reasoning around a minimal sufficient subset of 3D information rather than indiscriminate geometric context (Guo et al., 2026). Q-GeoMem makes memory writing question dependent by combining question relevance with geometric evidence management over time (Gao et al., 2026b). These approaches support the general principle that spatial information should be exposed selectively and in a task-dependent manner (Li et al., 2026; Guo et al., 2026; Gao et al., 2026b).

Our sparse routing design adopts this principle at a different interface. Instead of selecting raw geometry during feature fusion or deciding which frames enter an external memory, it routes a compact set of temporally contextualized and view-structured evidence directly from the language question before spatial pooling. The selected evidence then serves both as guidance for the dense stream and as a separate projected evidence stream to the LLM, tying question-conditioned selection to the representation that language reasoning actually receives.

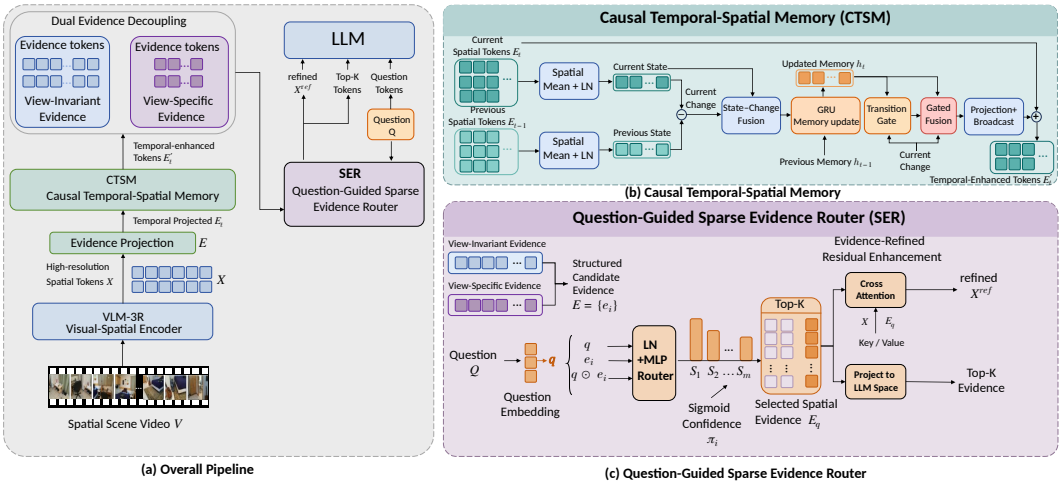

Figure 2: **MAVER architecture.** (a) High-resolution geometry-grounded tokens are temporally conditioned, decoupled into complementary evidence banks, and routed with the question before language reasoning. Selected evidence follows two paths: residual cross-attention refines the dense stream before pooling, while a separate projection maps the Top-K evidence directly into the LLM embedding space. (b) CTSM combines a current-frame state, first-order transition, and recurrent memory under causal updates. (c) SER scores structured candidates with question–evidence interactions and retains a fixed evidence budget.

3 METHOD

Given a monocular video and a natural-language question, MAVER inserts a structured evidence interface between a geometry-grounded video encoder and the language model. Figure 2 summarizes the pipeline. High-resolution spatial tokens are first temporally conditioned by CTSM, decoupled into view-invariant and view-specific evidence by DED, and sparsely selected by SER. As shown in Figure 2(c), the selected evidence then follows two paths: residual cross-attention uses it to refine the dense backbone stream before pooling, while a separate projection maps the same selected evidence into the LLM embedding space and appends it to the pooled visual sequence.

3.1 PROBLEM FORMULATION

Let $\mathcal{V} = \{I_t\}_{t=1}^{T}$ be a monocular RGB video and q a question. A geometry-grounded backbone produces high-resolution tokens $X = \{X_t\}_{t=1}^{T}$, $X_t \in \mathbb{R}^{N \times D}$, before its spatial pooling stage. MAVER operates at this point so that fine-grained evidence can be transformed before compression. With the default $T = 32$ and $N = 27 \times 27 = 729$, this stream contains 23,328 tokens. We use standard attention operators (Vaswani et al., 2017).

3.2 CAUSAL TEMPORAL-SPATIAL MEMORY

We first map each token into an evidence space, $\mathbf{e}_{t,n} = W_{\text{down}} X_{t,n}$. CTSM summarizes the current frame, explicitly models local change, and updates a GRU state (Cho et al., 2014) only from past and current observations:

$$\bar{\mathbf{e}}_t = \mathrm{LN}\left(\frac{1}{N}\sum_{n}\mathbf{e}_{t,n}\right), \quad \Delta\mathbf{e}_t = \bar{\mathbf{e}}_t - \bar{\mathbf{e}}_{t-1}, \quad \mathbf{h}_t = \mathrm{GRU}(W_s\bar{\mathbf{e}}_t + W_\Delta\Delta\mathbf{e}_t, \mathbf{h}_{t-1}), \tag{1}$$

where $\Delta\mathbf{e}_1 = \bar{\mathbf{e}}_1$. We fuse recurrent history and instantaneous change with a learned gate and broadcast the result back to every spatial token:

$$\mathbf{g}_t = \sigma(W_g[\mathbf{h}_t; \Delta\mathbf{e}_t]), \quad \boldsymbol{\tau}_t = \mathbf{g}_t \odot \mathbf{h}_t + (1 - \mathbf{g}_t) \odot \Delta\mathbf{e}_t, \quad \mathbf{e}'_{t,n} = \mathbf{e}_{t,n} + W_\tau\boldsymbol{\tau}_t. \tag{2}$$

Because $\mathbf{h}_t$ depends only on frames $1{:}t$, CTSM preserves causality. W_τ is zero-initialized, so the module initially reduces to the pretrained pathway.

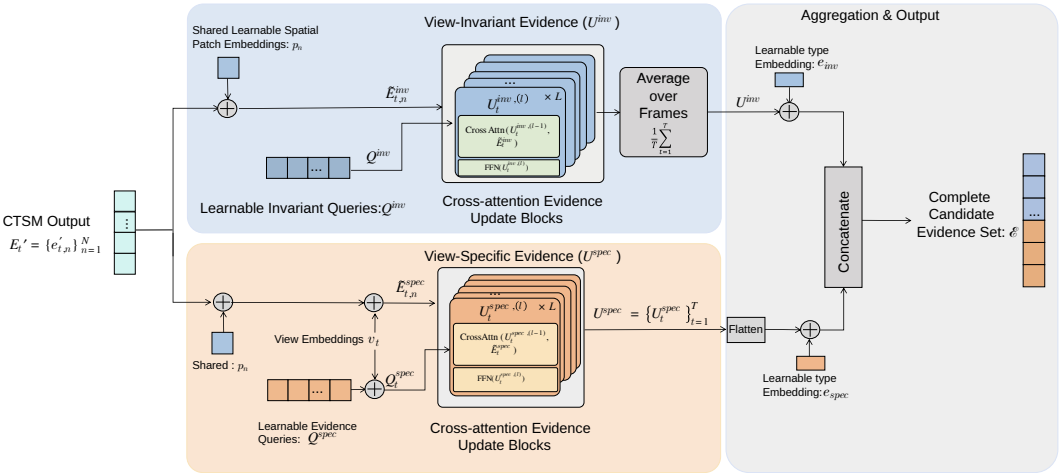

Figure 3: **Dual Evidence Decoupling.** The view-invariant bank uses shared queries and patch embeddings without view identity, then averages corresponding slots across frames to favor persistent evidence. The view-specific bank injects frame embeddings into both queries and token context and retains per-frame outputs. Type embeddings distinguish the two sources before concatenation into the candidate set routed by SER.

3.3 Dual Evidence Decoupling

Persistent scene structure and observation-dependent cues have different invariance requirements. DED therefore extracts them into two complementary banks using learned queries, following the general idea of attention bottlenecks and latent slots (Jaegle et al., 2021; Locatello et al., 2020). Let $\mathbf{P}$ denote shared spatial patch embeddings and XAttn a cross-attention extraction block.

The **view-invariant branch** uses shared queries Q^{inv} with no view identity and averages corresponding evidence slots across frames:

$$\mathbf{U}^{\mathrm{inv}} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{XAttn}\big(Q^{\mathrm{inv}}, \mathbf{E}'_t + \mathbf{P}\big)\,. \tag{3}$$

The **view-specific branch** adds a frame embedding $\mathbf{v}_t$ to both its queries and spatial context, and keeps each frame's output separate:

$$\mathbf{U}^{\mathrm{spec}}_t = \mathrm{XAttn}(Q^{\mathrm{spec}} + \mathbf{v}_t, \mathbf{E}'_t + \mathbf{P} + \mathbf{v}_t)\,. \tag{4}$$

Type embeddings identify the two sources before concatenation:

$$\mathcal{E} = \big[\mathbf{U}^{\mathrm{inv}} + \mathbf{e}_{\mathrm{inv}}; \mathrm{Flatten}(\mathbf{U}^{\mathrm{spec}}) + \mathbf{e}_{\mathrm{spec}}\big] = \{\mathbf{e}_i\}_{i=1}^{M}, \quad M = S_I + TS_S. \tag{5}$$

With $S_I = 16$, $S_S = 4$, and $T = 32$, DED produces 144 candidates from the original 23,328 high-resolution tokens.

3.4 Question-Guided Sparse Evidence Routing

Different questions require different subsets of the candidate bank. We form $\mathbf{q}$ by mean-pooling only prompt tokens visible at inference time, excluding image placeholders and ground-truth answer tokens. Each candidate is scored from the projected question, evidence, and their element-wise interaction:

$$s_i = \frac{f_{\mathrm{route}}([W_q\mathbf{q}; \mathbf{e}_i; (W_q\mathbf{q}) \odot \mathbf{e}_i])}{\tau_r}, \qquad \pi_i = \sigma(s_i). \tag{6}$$

We retain a fixed budget,

$$\mathcal{I}_q = \mathrm{TopK}(\{s_i\}_{i=1}^{M}, K), \qquad \mathbf{E}_q = \{\pi_i \mathbf{e}_i \mid i \in \mathcal{I}_q\}. \tag{7}$$

This question-conditioned selection follows the broader principle that spatial evidence should be exposed according to task demand rather than uniformly (Li et al., 2026; Guo et al., 2026; Gao et al.,

2026b). During training, Gumbel noise (Jang et al., 2017) is added only to the ranking scores before hard Top-K. We do not use a straight-through estimator: task gradients reach selected candidates through the multiplicative sigmoid gates π_i, while $\mathcal{L}_{\text{budget}}$ supervises all routing probabilities. Inference uses deterministic Top-K. We use $K = 16$ by default.

3.5 Dual-Path Evidence Injection

Top-K selection would be weakened if the retained information were immediately destroyed by pooling. We therefore use two complementary paths:

$$\mathbf{H}^{\text{vis}} = [\text{Pool}(\mathbf{X} + W_{\text{res}}\,\text{XAttn}(\mathbf{E}', \mathbf{E}_q))\,;\; W_{\text{up}}\mathbf{E}_q]\,. \tag{8}$$

The first path propagates question-relevant evidence back into the dense visual representation before pooling, producing a refined dense stream. The second maps the selected evidence itself through W_{up} into the LLM embedding space, where these projected evidence tokens are concatenated with the pooled visual tokens. W_{res} is zero-initialized, preserving the original visual path at initialization.

3.6 Training Objectives

Training uses the autoregressive answer loss together with compact objectives that make the two evidence banks complementary, information-preserving, and useful for sparse routing:

$$\mathcal{L} = \mathcal{L}_{\text{ans}} + \mathcal{L}_{\text{dec}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{budget}}\mathcal{L}_{\text{budget}} + \lambda_{\text{suff}}\mathcal{L}_{\text{suff}}, \tag{9}$$

where

$$\mathcal{L}_{\text{dec}} = \lambda_{\text{inv}}\mathcal{L}_{\text{inv}} + \lambda_{\text{bind}}\mathcal{L}_{\text{bind}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{view}}\mathcal{L}_{\text{view}} + \lambda_{\text{dis}}\mathcal{L}_{\text{dis}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}}. \tag{10}$$

For the invariant branch, same-index slots are encouraged to agree across views,

$$\mathcal{L}_{\text{inv}} = 1 - \mathbb{E}_{t,s}\left[\cos\left(\mathbf{U}^{\text{inv}}_{t,s}, \bar{\mathbf{U}}^{\text{inv}}_{s}\right)\right], \tag{11}$$

while $\mathcal{L}_{\text{bind}}$ uses Sinkhorn-based soft matching between adjacent-view slots and $\mathcal{L}_{\text{con}}$ applies scene-level InfoNCE between even/odd view aggregates. For the specific branch, $\mathcal{L}_{\text{view}}$ predicts normalized relative frame displacement. $\mathcal{L}_{\text{dis}}$ decorrelates normalized invariant/specific scene summaries, and $\mathcal{L}_{\text{div}}$ penalizes within-bank slot collapse.

To prevent compact evidence formation from discarding fine spatial content, we mask projected high-resolution visual features and reconstruct them from the complete candidate bank:

$$\mathcal{L}_{\text{rec}} = \text{SmoothL1}\Big(\hat{\mathbf{E}}_{\mathcal{M}}, \text{sg}(\mathbf{E}_{\mathcal{M}})\Big)\,. \tag{12}$$

SER is trained to be both sparse and answer-sufficient:

$$\mathcal{L}_{\text{budget}} = \frac{(\sum_i \pi_i - K)^2}{M^2}, \qquad \mathcal{L}_{\text{suff}} = \text{CE}(a \mid q, \mathbf{E}_q), \tag{13}$$

where the sufficiency branch predicts the ground-truth answer using only the selected evidence tokens. We use $(\lambda_{\text{inv}}, \lambda_{\text{bind}}, \lambda_{\text{con}}, \lambda_{\text{view}}, \lambda_{\text{dis}}, \lambda_{\text{div}}) = (0.10, 0.05, 0.05, 0.05, 0.01, 0.01)$, $\lambda_{\text{rec}} = 0.10$, and $\lambda_{\text{budget}} = 0.01$; λ_{suff} is 1.0 in the evidence-only stage and 0.5 thereafter. The implemented necessity objective is disabled in all reported runs. At inference, all auxiliary objectives and stochastic routing perturbations are removed; CTSM runs causally, DED constructs the two banks, and SER deterministically selects Top-K evidence before dual-path injection.

4 Experiments

4.1 Experimental Setup

Benchmarks and protocol. We evaluate MAVER on VSI-Bench (Yang et al., 2025) and VSTI-Bench (Fan et al., 2026). VSI-Bench evaluates eight video spatial reasoning categories: object counting, absolute distance, object size, room size, relative distance, relative direction, route planning, and appearance order. VSTI-Bench focuses on spatial-temporal reasoning under camera motion, covering camera–object absolute distance, camera displacement, camera movement direction, object–object relative position, and camera–object relative distance. We follow the official task-wise evaluation protocols. For Table 1, to avoid mixing source-specific aggregation conventions, we recompute Avg. for every method as the unweighted arithmetic mean of the eight displayed task-wise scores; the task-wise baseline results are taken from the cited sources.

Baseline methods. Following the benchmark comparison protocol, we group baselines into *Proprietary Models (API)*, *Open-sourced VLMs*, and *Specialized Spatial Reasoning Models*. The first two groups measure the spatial capability of general-purpose multimodal models, while the specialized group contains models explicitly trained or architected for spatial reasoning with geometric priors, spatial supervision, or reconstruction-aware representations (Zheng et al., 2025a; Wu et al., 2025; Yang et al., 2026a; Fan et al., 2026). Published baseline numbers are included as external reference under their reported settings; controlled attribution to MAVER uses the matched geometry-grounded baseline in Sec. 4.3.

Implementation and controlled baseline. We use the VLM-3R geometry-grounded stack (Fan et al., 2026), comprising SigLIP, CUT3R, spatial–visual–view fusion, an MLP projector, and Qwen2-7B. Both MAVER and our controlled geometry-grounded baseline use the same 250K-example training mixture sampled from VLM-3R-Data and VSI-590K (Fan et al., 2026; Yang et al., 2026b). To isolate the contribution of the proposed evidence interface, the controlled baseline uses the same geometry-grounded backbone and training mixture but directly applies the standard spatial-pooling pathway without CTSM, DED, or SER. MAVER starts from this matched setting and introduces the staged evidence interface before spatial pooling.

The evidence dimension is 1024. DED uses 16 invariant slots and 4 specific slots per frame, two 8-head update layers, and at most 32 views; SER selects $K = 16$ candidates with router temperature 0.7. Masked reconstruction uses ratio 0.30 (at most 64 masked tokens); the InfoNCE and Sinkhorn temperatures are 0.1 and 0.07, with three Sinkhorn iterations. CTSM uses a one-layer 256-dimensional recurrent state with zero dropout. Training uses AdamW, cosine decay, weight decay 0.01, BF16/TF32, and DeepSpeed ZeRO-2. Stage 1 freezes the geometry-grounded backbone and optimizes DED/SER and their auxiliary heads for 900 steps at 1×10^{-4} (effective batch size 48; 50 warm-up steps). Stage 2 jointly optimizes DED/SER and Qwen2 LoRA ($r = 32$, $\alpha = 64$, dropout 0.05) for 400 steps, with learning rates 5×10^{-5} and 2×10^{-5}, respectively (batch size 48; 20 warm-up steps). Stage 3 attaches CTSM and jointly optimizes CTSM, DED/SER, and the Stage-2 LoRA for 500 steps at 3×10^{-5} and 1×10^{-5} (batch size 32; 25 warm-up steps). SigLIP, CUT3R, fusion, and projector remain frozen in all three MAVER stages; the maximum language context is 32,768 tokens. The full MAVER interface adds 84.3M registered parameters (1.31M from CTSM). Trainable parameters are 83.0M in Stage 1, approximately 163.7M in Stage 2, and 165.0M in Stage 3 including the 80.7M LoRA parameters.

4.2 Main Results

Overall Performance on VSI-Bench. Table 1 compares MAVER with proprietary APIs, general open-sourced VLMs, and specialized spatial reasoning models. MAVER achieves the highest average score among the compared methods at 64.4, exceeding the strongest listed proprietary model by 19.0 points, the strongest general open-sourced VLM by 15.9 points, and the strongest specialized baseline by 3.5 points. The advantage is concentrated in the multiple-choice spatial reasoning categories: MAVER reaches 84.7 on relative direction, 48.2 on route planning, and 72.9 on appearance order, all the highest values in the table. In contrast, the numerical estimation categories are more mixed: several existing models retain stronger results on object counting, absolute distance, object size, room size, or relative distance. This pattern indicates that MAVER's main benefit lies in organizing temporally distributed and viewpoint-dependent evidence rather than uniformly improving metric estimation.

Performance on VSTI-Bench. Table 2 evaluates spatial-temporal reasoning under camera motion. MAVER obtains an average score of 61.6, compared with 44.0 for the strongest listed general open-sourced VLM and 58.8 for the strongest specialized baseline. The gains are largest on camera movement direction (65.2 versus 60.6) and camera–object relative distance (72.9 versus 68.6), while the differences on camera–object absolute distance (40.1 versus 39.4) and camera displacement (40.5 versus 39.6) are comparatively modest. Together with the VSI-Bench results, this distribution suggests that the proposed interface is most effective when reasoning depends on temporal change, viewpoint transitions, and relational evidence, whereas fine-grained metric distance estimation remains more dependent on the underlying geometric representation.

Table 1: **Evaluation on VSI-Bench.** Performance (%) of proprietary models, open-sourced VLMs, specialized spatial reasoning models, and MAVER across eight spatial reasoning categories. Task-wise baseline scores are taken from the cited sources; for consistency, Avg. is recomputed as the unweighted arithmetic mean of the eight displayed task-wise scores for every method.

Method	Avg.	Obj.C	Abs.D	Obj.S	Room.S	Rel.D	Rel.Dir	Route	Appr.
Proprietary Models (API)									
GPT-4o (OpenAI, 2024)	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5 Flash (Gemini Team et al., 2024)	42.1	49.8	30.8	53.5	54.4	37.7	41.0	31.5	37.8
Gemini-1.5 Pro (Gemini Team et al., 2024)	45.4	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6
Open-sourced VLMs									
LLaVA-NeXT-Video-7B (Zhang et al., 2024)	35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6
LLaVA-NeXT-Video-72B (Zhang et al., 2024)	40.9	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6
Qwen2.5-VL-7B (Bai et al., 2025)	31.2	40.9	14.8	43.4	10.7	38.6	38.5	33.0	29.8
LLaVA-OneVision-7B (Li et al., 2025)	32.4	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4
LLaVA-OneVision-72B (Li et al., 2025)	40.2	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6
InternVL3-78B (Zhu et al., 2025b)	48.5	71.2	53.7	44.4	39.5	55.9	39.5	28.9	54.5
Specialized Spatial Reasoning Models									
Spatial-MLLM-4B (Wu et al., 2025)	47.0	65.3	34.8	63.1	45.1	41.3	46.2	33.5	46.3
VG-LLM-4B (Zheng et al., 2025a)	47.3	66.0	37.8	55.2	59.2	44.6	45.6	33.5	36.4
VG-LLM-8B (Zheng et al., 2025a)	50.7	67.9	37.7	58.6	62.0	46.6	40.7	32.4	59.2
VST-7B-SFT (Yang et al., 2026a)	60.6	72.0	44.4	74.3	68.3	59.7	55.8	44.9	65.2
VLM-3R-7B (Fan et al., 2026)	60.9	70.2	49.4	69.2	67.1	65.4	80.5	45.4	40.1
MAVER	**64.4**	71.3	47.7	69.9	58.5	62.1	**84.7**	**48.2**	**72.9**

Table 2: **Evaluation on VSTI-Bench.** Performance (%) on five spatial-temporal reasoning tasks. CamObj-A: camera–object absolute distance; Disp.: camera displacement; Mov.: camera movement direction; ObjRel: object–object relative position; CamRel: camera–object relative distance.

Method	Avg.	CamObj-A	Disp.	Mov.	ObjRel	CamRel
Proprietary Models (API)						
GPT-4o (OpenAI, 2024)	38.2	29.5	23.4	37.3	58.1	42.5
Gemini-1.5 Flash (Gemini Team et al., 2024)	32.1	28.5	20.9	24.4	52.6	33.9
Open-sourced VLMs						
LLaVA-NeXT-Video-7B (Zhang et al., 2024)	40.0	28.2	1.8	49.8	64.7	55.6
LLaVA-OneVision-7B (Li et al., 2025)	41.7	29.9	19.3	47.5	62.1	49.8
LongVA-7B (Zhang et al., 2025)	32.3	13.5	5.1	43.7	57.9	41.2
InternVL2-8B (Chen et al., 2024b)	43.5	32.9	13.5	48.0	68.0	55.0
LongVILA-8B (Chen et al., 2025)	30.5	20.0	11.6	35.4	52.3	33.4
VILA-1.5-8B (Lin et al., 2024)	37.3	30.1	27.3	42.2	50.4	36.7
VILA-1.5-40B (Lin et al., 2024)	38.2	28.2	15.7	28.8	65.4	53.0
LLaVA-NeXT-Video-72B (Zhang et al., 2024)	44.0	32.3	10.5	48.1	78.3	50.9
Specialized Spatial Reasoning Models						
VLM-3R-7B (Fan et al., 2026)	58.8	39.4	39.6	60.6	86.5	68.6
MAVER	**61.6**	**40.1**	**40.5**	**65.2**	**89.4**	**72.9**

4.3 Ablation Studies

Table 3 compares module configurations against the controlled geometry-grounded baseline. The variants use the objectives associated with their active modules and characterize how the components interact within the MAVER training recipe.

Temporal and View Structure Contribute Differently. CTSM raises the VSI average from 57.3 to 60.6, with its dominant change on appearance order (35.9→60.5, +24.6), while most static metric categories move by less than one point. DED instead produces its clearest gains on relative direction (78.4→82.6, +4.2) and route planning (42.2→45.4, +3.2). Combining CTSM and DED reaches 63.2, including 70.1 on appearance order and 83.3 on relative direction, indicating complementary benefits from temporal contextualization and view-structured evidence formation.

Routing Benefits from Structured Candidates. SER alone obtains 56.6, slightly below the 57.3 baseline, whereas adding SER after CTSM+DED raises the average from 63.2 to 64.4. Relative to CTSM+DED, the full model improves every reported category, including +1.4 on relative direction, +1.6 on route planning, and +2.8 on appearance order. Relative to the controlled baseline, Full MAVER gains 7.1 points overall and 37.0 points on appearance order. This progression supports performing question-conditioned selection after the evidence has first been temporally contextualized and organized across viewpoints.

Table 3: **Core component ablation on VSI-Bench (%).** All variants use the same geometry-grounded backbone and data; objectives follow the active modules.

Variant	Avg.	Obj.C	Abs.D	Obj.S	Room.S	Rel.D	Rel.Dir	Route	Appr.
Geo. baseline	57.3	68.1	46.4	68.8	57.4	60.9	78.4	42.2	35.9
+ CTSM	60.6	67.7	45.9	69.0	57.3	61.3	79.2	43.7	60.5
+ DED	58.6	70.2	46.6	69.3	57.0	61.1	82.6	45.4	36.7
+ SER	56.6	67.4	45.5	67.2	56.9	60.6	77.1	42.0	35.7
+ CTSM + SER	61.2	67.9	46.0	68.6	57.4	61.4	78.6	43.9	65.8
+ DED + SER	58.9	71.0	46.7	68.9	57.7	61.0	83.4	46.8	35.8
+ CTSM + DED	63.2	70.5	46.8	69.3	57.6	61.7	83.3	46.6	70.1
Full MAVER	**64.4**	**71.3**	**47.7**	**69.9**	**58.5**	**62.1**	**84.7**	**48.2**	**72.9**

Table 4: **Mechanism-level ablations.** Benchmark averages (%); deltas are relative to the full model.

Variant	VSI	Δ	VSTI	Δ
Full MAVER	**64.4**	–	**61.6**	–
Only view-invariant	62.8	-1.6	59.6	-2.0
Only view-specific	63.5	-0.9	60.9	-0.7
Uniform selection	63.1	-1.3	60.3	-1.3
$K = 8$ / 12 / 32	61.7 / 62.5 / 62.3	–	–	–
w/o residual refinement	63.6	-0.8	–	–
w/o evidence bypass	64.0	-0.4	–	–
w/o local transition Δ_t	62.7	-1.7	60.1	-1.5
Bidirectional temporal processing	61.5	-2.9	59.9	-1.7

4.4 Mechanism Analysis

Table 4 examines the main design choices within the full configuration.

Evidence Composition and Routing. Using only the view-invariant bank gives 62.8/59.6 on VSI/VSTI, while the view-specific bank alone gives 63.5/60.9; neither matches the full 64.4/61.6 model. Uniform selection is also 1.3 points lower on both benchmarks, showing that the question-conditioned ranking matters beyond the candidate construction itself. The budget sweep is non-monotonic: $K = 8$, 12, and 32 obtain 61.7, 62.5, and 62.3 on VSI, respectively, whereas $K = 16$ reaches 64.4. The tested results therefore favor retaining a compact but non-minimal subset rather than either overly aggressive or less selective routing.

Injection and Temporal Update. Removing residual refinement or the direct evidence bypass reduces VSI to 63.6 and 64.0, respectively, suggesting smaller but directionally consistent contributions from the two injection paths. Temporal design has a larger effect: removing the local transition Δ_t lowers VSI/VSTI to 62.7/60.1, while bidirectional temporal processing gives 61.5/59.9. Since the bidirectional variant also changes the state-update regime, this comparison does not isolate causality alone, but the tested configuration favors explicit local change together with causal recurrent updates.

5 Conclusion

We introduced MAVER, a structured evidence interface for geometry-grounded video spatial reasoning. MAVER carries forward prior context with causal temporal memory, organizes persistent and observation-dependent cues through dual evidence decoupling, and selectively routes question-relevant evidence through complementary refinement and bypass paths. Across VSI-Bench and VSTI-Bench, its largest gains occur on appearance order, relative direction, route planning, and camera-motion reasoning, highlighting the value of organizing temporal and viewpoint-dependent evidence before spatial compression. A remaining limitation is that improvements on metric distance estimation are comparatively modest, suggesting that evidence organization alone does not solve fine-grained metric calibration. Future work could combine MAVER with stronger metric supervision and uncertainty-aware geometry, and extend structured evidence routing to longer, online, and embodied interaction.

AI USE STATEMENT

Generative AI tools were used in this work to generate synthetic datasets; create or modify scientific figures or images; suggest experimental parameters; create or edit software code; draft portions of the manuscript; summarize or analyze existing literature; identify research topics or gaps; support brainstorming and information retrieval; edit the manuscript for readability; identify relevant literature; format references; propose paper titles or keywords; assist with translation; and interpret experimental results. We did not use generative AI tools for the other research tasks that require disclosure under the ICLR policy, including formulating or proving mathematical claims, proposing or refining hypotheses, designing the research methodology or experiments, cleaning or reformatting datasets, or qualitative or thematic data analysis. All AI-assisted material was reviewed by the authors; code changes, figures, factual claims, bibliographic records, synthetic data, translations, and AI-assisted interpretations were checked before inclusion. The authors take responsibility for the final content of this work, including text, claims, code, data, and artifacts produced with AI assistance.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian 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, 2024a.

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*, 2025.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Xiang An, Yan Feng, Peng Pei, Xunliang Cai, and Ruqi Huang. 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, 2026.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67 (12):220101, 2024b.

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. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, Zhicheng Yan, Hongyu Xu, Justin Theiss, Tianlong Chen, Jiachen Li, Zhengzhong Tu, Zhangyang Wang, and Rakesh Ranjan. 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.

Xiangjun Gao, Zhensong Zhang, Dave Zhenyu Chen, Songcen Xu, Long Quan, Eduardo Pérez-Pellitero, and Youngkyoon Jang. Map2Thought: Explicit 3D spatial reasoning via metric cognitive maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 7154–7164, 2026a.

Xianqiang Gao, Qizhi Chen, Delin Qu, Haoming Song, Zhigang Wang, Bin Zhao, Dong Wang, and Xuelong Li. Q-GeoMem: Question-guided geometric memory for video spatial reasoning. *arXiv preprint arXiv:2605.27318*, 2026b.

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

Yejie Guo, Yunzhong Hou, Wufei Ma, Meng Tang, and Ming-Hsuan Yang. Pursuing minimal sufficiency in spatial reasoning. In *International Conference on Learning Representations*, 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. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Wenbo Hu, Jingli Lin, Yilin Long, Yunlong Ran, Lihan Jiang, Yifan Wang, Chenming Zhu, Runsen Xu, Tai Wang, and Jiangmiao Pang. G^2VLM: 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, 2026.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 4651–4664, 2021.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*, 2017.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision*, pp. 71–91, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025.

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. In *International Conference on Machine Learning*, 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 *Proceedings of the 40th International Conference on Machine Learning*, pp. 19730–19742, 2023.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26689–26699, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zuntao Liu, Yi Du, Taimeng Fu, Shaoshu Su, Cherie Ho, and Chen Wang. Vision-language memory for spatial reasoning. In *European Conference on Computer Vision*, pp. 297–316, 2026.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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*, pp. 5294–5306, 2025a.

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

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.

Xing Wei, Aoxiang Tian, Shaofan Liu, Jiansheng Peng, Chong Zhao, Xiang Bi, Yang Lu, Benhong Zhang, and Fan Yang. GeoMind: Explicit spatial reasoning via dual-reference geometric modeling. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, pp. 1821–1829, 2026.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10632–10643, 2025.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, Yi Lin, and Hengshuang Zhao. Visual spatial tuning. In *European Conference on Computer Vision*, pp. 192–211, 2026a.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-S: Towards spatial supersensing in video. In *International Conference on Learning Representations*, 2026b.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *Transactions on Machine Learning Research*, 2025.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. LLaVA-NeXT: A strong zero-shot video understanding model. LLaVA Blog, 2024.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3D world: Enhancing MLLMs with 3D vision geometry priors. In *Advances in Neural Information Processing Systems*, volume 38, pp. 23356–23382, 2025a.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3D LLM: Learning position-aware video representation for 3D scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8995–9006, 2025b.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. LLaVA-3D: A simple yet effective pathway to empowering LMMs with 3D capabilities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4295–4305, 2025a.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025b.