COMPOSITIONAL SPATIAL REASONING FOR MULTI-MODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial reasoning is a fundamental capability of multimodal large language models (MLLMs), requiring models to infer complex semantic, geometric, and temporal relationships from visual observations. Recent approaches enhance spatial reasoning by enriching visual representations with additional spatial priors from 3D reconstruction or depth estimation models. However, we find that this uniform representation alone does not guarantee better reasoning, as different queries require different compositions of spatial evidence. In this work, we identify two key challenges for effective spatial reasoning: *comprehensive representation* and *selective evidence composition*. We introduce **CoSpace**, a compositional spatial reasoning framework that follows the principle of representing comprehensively and reasoning selectively. CoSpace organizes heterogeneous visual information into structured semantic, geometric, and temporal representations, and performs query-adaptive evidence composition through hierarchical selection, allocation, and retrieval to construct query-relevant reasoning states. Extensive experiments across diverse spatial reasoning benchmarks demonstrate that CoSpace improves spatial reasoning performance, effectively exploits heterogeneous visual evidence, and enables query-adaptive composition of interpretable spatial representations. The code will be made publicly available.

1 INTRODUCTION

Spatial reasoning is emerging as a fundamental capability of multimodal large language models (MLLMs) (Yang et al., 2025a; Fan et al., 2026; Yang et al., 2026b). Beyond recognizing objects and visual semantics, MLLMs for spatial reasoning must reason about metric structure, spatial relations, and temporal dynamics from visual observations (Ma et al., 2025; Wang et al., 2025c; Ray et al., 2025). Recent methods therefore augment MLLMs with stronger spatial priors, substantially improving their spatial perception of images and videos (Zheng et al., 2025; Fan et al., 2026; Cai et al., 2024; Wu et al., 2025). However, stronger spatial perception does not necessarily translate into stronger spatial reasoning (Zhang et al., 2026b). A central challenge is therefore how to effectively leverage rich spatial information to advance the reasoning capability of MLLMs.

We trace this gap to a prevailing *uniform representation paradigm*. Existing methods typically integrate spatial priors into shared visual representations and use them uniformly across reasoning tasks (Zheng et al., 2025; Fan et al., 2026), implicitly assuming that richer representations will consistently improve reasoning. Yet spatial reasoning is inherently query-dependent, with different queries relying on different combinations of semantic, geometric, and temporal evidence (Li et al., 2026; Lin et al., 2026; Zhi et al., 2025). As shown in Fig. 1(a), adding VGGT (Wang et al., 2025a) geometry to Qwen2.5-VL (Bai et al., 2025b) improves only three of eight task categories while degrading five, indicating that richer geometry

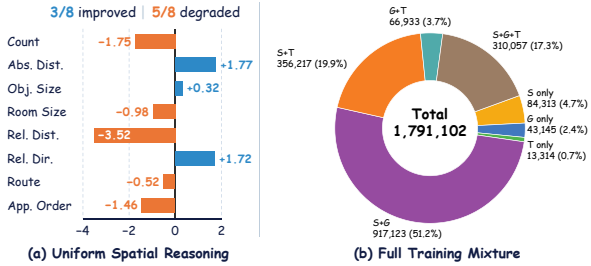


Figure 1: (a) Task-wise effects of geometry enhancement on VSI-Bench (VG-LLM minus Qwen2.5-VL-7B). (b) Capability composition of the full training mixture.

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.

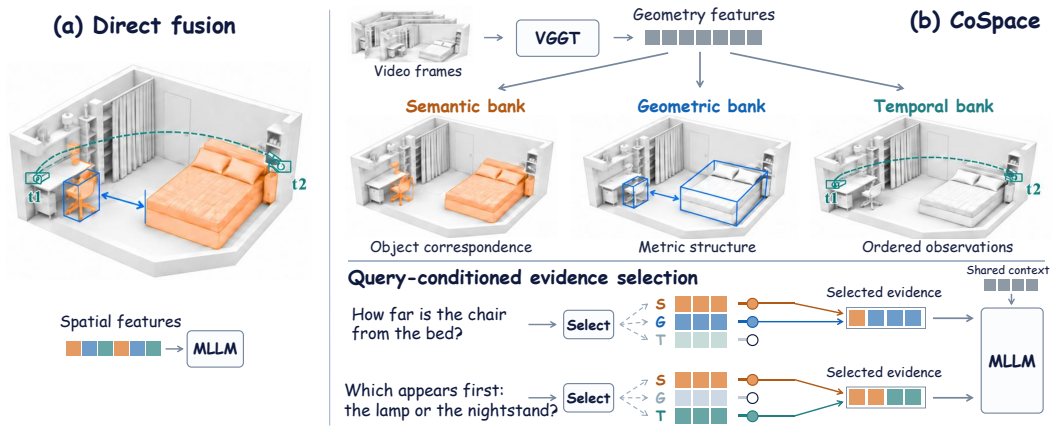


Figure 2: From uniform fusion to spatial reasoning composition. CoSpace organizes complementary spatial cues and composes them according to query-specific reasoning demands.

is not universally beneficial. Fig. 1(b) further shows that spatial reasoning queries involve diverse compositions of semantic, geometric, and temporal evidence. Together, these observations suggest that the key challenge is not simply to represent more spatial information, but to compose the *right evidence for each query*. This raises a fundamental question: *how can an MLLM represent spatial information comprehensively while reasoning over it selectively?*

We identify two complementary challenges underlying effective compositional spatial reasoning. The first is *comprehensive spatial representation*: preserving the heterogeneous cues required by different reasoning tasks. Semantic, geometric, and temporal cues capture complementary scene information across different spatial and temporal scales. These cues should therefore be explicitly organized and made separately accessible, providing a basis for subsequent query-dependent composition. The second is *selective evidence composition*: constructing a query-relevant reasoning state from these representations. Since different queries depend on different subsets and combinations of spatial evidence, uniformly exposing all available information can introduce redundancy and interference. The model must therefore determine which representations are relevant, how much each should contribute, and what evidence should be retrieved for the current query.

To this end, we introduce **CoSpace**, a compositional spatial reasoning framework built on the principle of *represent comprehensively, reason selectively*. As shown in Fig. 2, CoSpace decouples comprehensive scene representation from query-specific evidence composition. It organizes heterogeneous visual information into structured semantic, geometric, and temporal representation banks, preserving complementary cues about entity identity, spatial structure, and state evolution. Given a query, CoSpace then constructs a query-relevant reasoning state through hierarchical *selection*, *allocation*, and *retrieval*: it identifies relevant representation banks, determines their relative contributions, and retrieves the evidence most useful for the current reasoning process.

Extensive experiments across diverse spatial reasoning benchmarks demonstrate consistent improvements over strong MLLM baselines, while controlled analyses further show that query-adaptive evidence composition is more effective than uniformly exposing all spatial representations. These results suggest that advancing spatial intelligence requires not only richer representations of the visual world, but also the ability to dynamically compose the right evidence for each reasoning problem. The contributions of this work are summarized as follows:

- We identify the ineffective translation of rich spatial representations into reasoning gains as a key bottleneck in current spatial reasoning models and formulate two fundamental requirements: comprehensive representation and selective evidence composition.
- We propose **CoSpace**, a query-adaptive spatial reasoning framework that transforms heterogeneous visual information into structured semantic, geometric, and temporal representations and dynamically composes query-relevant evidence for reasoning.

- Extensive experiments demonstrate that CoSpace consistently advances spatial reasoning across diverse benchmarks and reveal query-adaptive evidence composition as a key factor in effective spatial reasoning.

2 Related Work

2.1 Spatial Reasoning for Multimodal Large Language Models

Spatial reasoning remains challenging for MLLMs despite advances in visual instruction tuning (Liu et al., 2023b; Li et al., 2024). Existing benchmarks evaluate spatial relations in images (Liu et al., 2023a), situated reasoning in 3D scenes (Azuma et al., 2022; Ma et al., 2023), and spatial understanding across videos and multiple views (Yang et al., 2025a; 2026b; Wang et al., 2025b). These tasks require reasoning about metric structure, viewpoints, cross-view correspondence, and temporal relations beyond object semantics (Li et al., 2025).

Existing methods improve spatial reasoning through stronger supervision and richer spatial representations. SpatialVLM (Chen et al., 2024), SPAR (Zhang et al., 2025), and SpatialRGPT (Cheng et al., 2024) introduce spatial instruction tuning and region-level grounding, while VG-LLM (Zheng et al., 2025), VLM-3R (Fan et al., 2026), and SpatialStack (Zhang et al., 2026a) incorporate geometric priors into MLLMs. Cambrian-S (Yang et al., 2025b) and Cambrian-P (Yang et al., 2026a) further enhance spatial reasoning with long-range visual context and camera-pose cues. More recent methods adapt how such information is used: GeoThinker (Li et al., 2026), GeoSR (Zhang et al., 2026b), and GeoWeaver (Miao et al., 2026) selectively integrate geometric features, while ViPS (Lin et al., 2026) and Q-GeoMem (Gao et al., 2026) adapt visual priors and geometric memory. In contrast, CoSpace jointly organizes semantic-, geometric-, and temporal-oriented cues and composes them according to query-specific reasoning demands.

2.2 3D Large Language Models

3D large language models extend language reasoning to 3D scenes through explicit or implicit 3D representations. PointLLM (Xu et al., 2024) connects point-cloud features with language models, while Scene-LLM (Fu et al., 2024) builds scene-level 3D representations for spatial understanding and reasoning. More recent approaches recover 3D-aware features directly from images or videos. Geometry foundation models such as DUSt3R (Wang et al., 2024), MASt3R (Leroy et al., 2024), and VGGT (Wang et al., 2025a) provide strong multi-view geometric priors, which are further incorporated into MLLMs by methods such as VG-LLM (Zheng et al., 2025) and VLM-3R (Fan et al., 2026). CoSpace builds on geometry-aware visual representations but targets a different problem: rather than learning a single representation for all queries, it organizes spatial cues with different functional emphases and selectively composes them for query-adaptive reasoning.

3 Method

3.1 Overview

CoSpace follows the principle of *represent comprehensively, reason selectively* by decoupling shared representation from query-adaptive spatial composition. Given visual observations $\mathcal{X} = \{I_t\}_{t=1}^{T}$ and a question query Q, CoSpace first organizes geometry-aware visual features into separately accessible representation banks with semantic, geometric, and temporal emphases. It then constructs a query-relevant reasoning state through three stages: *selection*, which determines which banks are relevant; *allocation*, which estimates their relative contributions; and *retrieval*, which extracts useful information from the selected banks. As shown in Fig. 3, this design preserves complementary spatial cues while allowing their contributions to adapt to the reasoning demand of each query.

Alongside these structured banks, CoSpace maintains a shared scene context that provides common visual information for all queries. The Qwen visual encoder (Bai et al., 2025b;a) produces visual tokens $\mathbf{H}_{\mathrm{2D}}$, while VGGT (Wang et al., 2025a) extracts geometry-aware features $\mathbf{G}$. A geometry merger projects $\mathbf{G}$ into the Qwen visual-token space to obtain $\mathbf{F}_{\mathrm{geo}}$, which is fused with the visual tokens as $\mathbf{H}_{\mathrm{base}} = \mathbf{H}_{\mathrm{2D}} + \mathbf{F}_{\mathrm{geo}}$. The resulting $\mathbf{H}_{\mathrm{base}}$ serves as a shared scene context and is

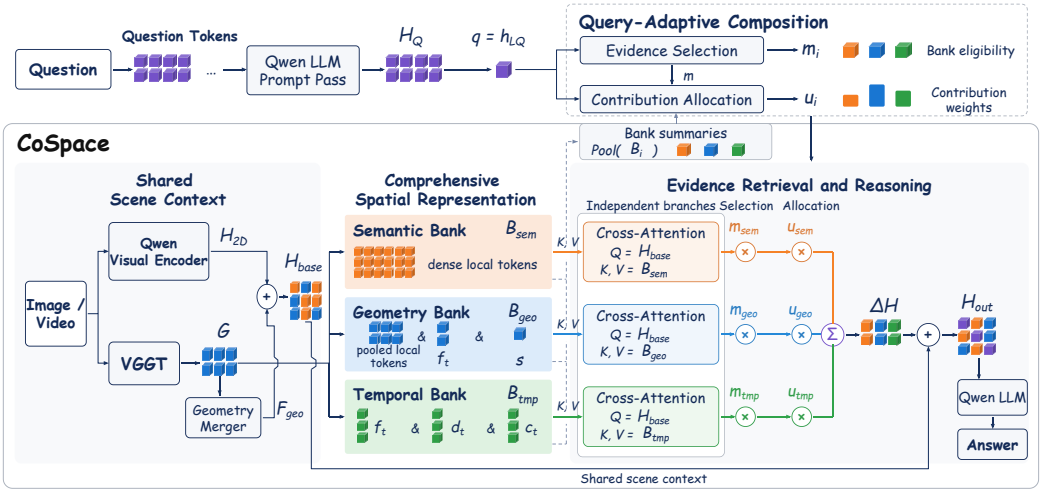


Figure 3: Overview of CoSpace. CoSpace follows the principle of *represent comprehensively, reason selectively*: shared geometry-aware features are organized into semantic-, geometric-, and temporal-oriented banks, while query-adaptive selection, allocation, and retrieval selectively compose these representations to refine the shared visual context for answer generation.

preserved through a residual pathway. Query-adaptive composition then selectively retrieves and weights complementary bank information to refine this context, yielding a reasoning representation tailored to the current question.

3.2 Comprehensive Spatial Representation

Different spatial queries rely on complementary cues at different spatial and temporal scales. To preserve these cues without collapsing them into a single homogeneous representation, CoSpace organizes shared VGGT features into three separately accessible banks with semantic, geometric, and temporal emphases. These banks differ in their granularity and functional organization, providing complementary representations for subsequent query-adaptive composition. Since they are derived from a shared geometry-aware backbone, their contents are complementary but not mutually exclusive. For each frame t, we project the VGGT features into the Qwen visual-token space as $\mathbf{Z}_t = \mathcal{P}_{\text{bank}}(\mathbf{G}_t)$. We denote the ordered sequence $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ by $\mathbf{X}_{1:n}$ and use Concat for token-wise concatenation.

Semantic bank. Fine-grained spatial reasoning requires preserving regional details and entity correspondences across observations. We therefore retain the dense projected tokens as $\mathbf{B}_{\text{sem}} = \text{Concat}(\mathbf{Z}_{1:T})$, which preserve local appearance, entity identity, and cross-frame correspondence cues for object-centric reasoning.

Geometry bank. Geometric reasoning requires structural information at multiple spatial scales, from local regions to the global scene. We therefore organize geometry-aware features at the local, frame, and sequence levels to capture complementary geometric context. Spatial average pooling of $\mathbf{Z}_t$ with a 2×2 kernel and stride 2 produces compact local tokens $\widetilde{\mathbf{Z}}_t$. We further compute frame summaries $\mathbf{f}_t = \text{Mean}(\mathbf{Z}_t)$ and a sequence summary $\mathbf{s} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{f}_t$. The geometry bank is then denoted as $\mathbf{B}_{\text{geo}} = \text{Concat}(\widetilde{\mathbf{Z}}_{1:T}, \mathbf{f}_{1:T}, \mathbf{s})$, providing multi-granularity geometric cues that capture both local structure and global scene context.

Temporal bank. Temporal reasoning requires modeling changes across observations while accounting for viewpoint variation. We compute adjacent-frame differences $\mathbf{d}_t = \mathbf{f}_{t+1} - \mathbf{f}_t$, for $t = 1, \ldots, T-1$, and project the VGGT camera representations (Wang et al., 2025a) into the visual-token space as $\mathbf{c}_t$. The temporal bank is $\mathbf{B}_{\text{tmp}} = \text{Concat}(\mathbf{f}_{1:T}, \mathbf{d}_{1:T-1}, \mathbf{c}_{1:T})$, which jointly captures frame states, inter-frame changes, and viewpoint information. For a single image, the difference-token sequence is empty, leaving only the frame and camera representations.

3.3 QUERY-ADAPTIVE SPATIAL COMPOSITION

The representation banks in Sec. 3.2 provide complementary spatial cues, but different queries require different subsets and combinations of them. CoSpace therefore composes these representations adaptively through two bank-level decisions: *selection*, which determines which banks should participate, and *allocation*, which determines their relative contributions. Together, these decisions control which spatial information is exposed to the subsequent reasoning process and how strongly each source influences the resulting representation.

Query encoding. To characterize the reasoning demand of each question, we obtain prompt hidden states $\mathbf{H}_Q = \{\mathbf{h}_1, \ldots, \mathbf{h}_{L_Q}\}$ through an auxiliary text-only forward pass of the same Qwen LLM used for answer generation. The input retains the system and user prompts together with the assistant generation header, while excluding visual tokens, visual boundary tokens, and answer tokens. We use the final prompt state $\mathbf{q} = \mathbf{h}_{L_Q}$ as the query representation shared by the selector and allocator. This auxiliary forward pass is performed without gradient computation.

Evidence selection. We first determine which banks are relevant to the current query. For each bank $i \in \mathcal{I}$, where $\mathcal{I} = \{\mathrm{sem}, \mathrm{geo}, \mathrm{tmp}\}$, a routing head predicts a relevance score

$$\rho_i = \sigma(\mathcal{R}_i(\mathbf{q})), \qquad m_i = \begin{cases} m_i^*, & \text{training}, \\ \mathbb{I}[\rho_i > \tau_r], & \text{inference}, \end{cases} \tag{1}$$

where σ denotes the sigmoid function and τ_r is the routing threshold. During training, a predefined binary mask m^* provides weak task-relevance supervision by specifying which banks are eligible for each question type. These masks supervise only coarse bank participation; they neither determine the relative contributions of selected banks nor prescribe the evidence retrieved from them. Using the predefined masks during training also prevents early routing errors from suppressing the learning of individual bank branches. At inference, eligibility is predicted solely from the question, without access to question-type labels.

Contribution allocation. Selection determines *which* banks may contribute, while the allocator determines *how much* each selected bank should contribute for the current input. For each bank, we combine the query representation with a pooled summary of its visual features to estimate an input-dependent utility:

$$a_i = \mathcal{U}_i(\mathbf{q}, \mathrm{Pool}(\mathbf{B}_i)), \qquad u_i = \frac{m_i \exp(a_i/\tau_u)}{\sum_{j \in \mathcal{I}} m_j \exp(a_j/\tau_u)}. \tag{2}$$

Here, $\mathcal{U}_i$ estimates the usefulness of each selected bank for the current query, while τ_u controls the weight normalization. The allocator assigns nonzero weights only to selected banks and normalizes them to sum to one. A single selected bank receives weight one; if no bank is selected, all u_i are set to zero, and no additional refinement is applied.

3.4 EVIDENCE RETRIEVAL

After selecting the relevant banks and determining their relative contributions, CoSpace retrieves fine-grained information from each participating bank. Since the semantic, geometric, and temporal banks organize spatial cues differently, we employ an independent cross-attention branch for each bank. Specifically, the shared visual representation $\mathbf{H}_{\mathrm{base}}$ serves as queries, while the corresponding bank tokens provide keys and values:

$$\mathbf{\Delta}_i = \mathcal{P}_i(\mathrm{CrossAttn}_i(\mathbf{H}_{\mathrm{base}}, \mathbf{B}_i, \mathbf{B}_i)), \tag{3}$$

where $\mathcal{P}_i$ projects the retrieved features to the channel dimension of $\mathbf{H}_{\mathrm{base}}$. Using the shared visual tokens as queries aligns the retrieved information with the original token positions, while the independent branches accommodate the different structures of the three banks. The retrieved features are then combined according to the query-adaptive contribution weights:

$$\mathbf{H}_{\mathrm{out}} = \mathbf{H}_{\mathrm{base}} + \sum_{i \in \mathcal{I}} u_i \mathbf{\Delta}_i. \tag{4}$$

Table 1: **VSI-Bench results comparison under the vanilla regime following Li et al. (2026).** Our method outperforms Qwen2.5-VL, VG-LLM, and GeoThinker under the same setting.

Model	LM	Avg.	Numerical Answer				Multiple-Choice Answer			
			Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Baselines										
Chance Level (Random)	–	–	–	–	–	–	25.0	36.1	28.3	25.0
Chance Level (Frequency)	–	34.0	62.1	32.0	29.9	33.1	25.1	47.9	28.4	25.2
Proprietary Models (API)										
GPT-4o (OpenAI, 2024)	Unk.	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5 Flash (Gemini Team, 2024)	Unk.	42.1	49.8	30.8	53.5	54.4	37.7	41.0	31.5	37.8
Gemini-1.5 Pro	Unk.	45.4	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6
Gemini-2.5 Pro (Gemini Team, 2025)	Unk.	51.5	43.8	34.9	64.3	42.8	61.1	47.8	45.9	71.3
Open-Sourced Models										
LLaVA-OneVision-7B (Li et al., 2024)	Qwen2-7B	32.4	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4
LLaVA-OneVision-72B	Qwen2-72B	40.2	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6
LLaVA-Video-7B (Zhang et al., 2024)	Qwen2-7B	35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6
LLaVA-Video-72B	Qwen2-72B	40.9	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6
InternVL2-8B (OpenGVLab Team, 2024)	InternLM2.5-7B	34.6	23.1	28.7	48.2	39.8	36.7	30.7	29.9	39.6
InternVL2-40B	Nous-Hermes-2-Yi-34B	36.0	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6
Qwen2.5-VL-3B (Bai et al., 2025b)	Qwen2.5-3B	28.6	32.7	19.5	17.3	25.1	37.3	44.9	30.4	21.8
Qwen2.5-VL-7B	Qwen2.5-7B	29.3	25.2	10.5	36.4	29.6	38.4	38.0	29.8	26.8
Vanilla Regime (Same Data and Protocol)										
Qwen2.5-VL-7B	Qwen2.5-7B	53.8	**69.6**	37.1	66.4	57.4	**54.2**	44.1	27.8	**73.5**
VG-LLM (Zheng et al., 2025)	Qwen2.5-7B	53.2	67.8	38.9	**66.8**	56.5	50.7	45.9	27.3	72.0
GeoThinker (Li et al., 2026)	Qwen2.5-7B	54.1	67.5	38.9	65.5	**61.0**	52.0	45.0	30.9	71.8
Ours	Qwen2.5-7B	**54.9**	68.2	**40.5**	64.7	59.0	53.2	**47.0**	**34.5**	72.2

This formulation preserves the shared scene context while selectively incorporating complementary information from the participating banks. Bank selection determines *which* representations are used, allocation determines *how much* each contributes, and cross-attention retrieves the corresponding fine-grained information. The resulting $\mathbf{H}_{\text{out}}$ forms a query-adapted visual representation and is passed, together with the question tokens, to the Qwen LLM for answer generation.

3.5 Training Objective

CoSpace supervises bank selection with task-relevance priors, while learning contribution allocation and evidence retrieval through answer prediction. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{ans}} + \frac{\lambda_r}{3} \sum_{i \in \mathcal{I}} \text{BCE}(\rho_i, m_i^*), \tag{5}$$

where $\mathcal{L}_{\text{ans}}$ is the autoregressive answer-generation loss, BCE denotes binary cross-entropy, and λ_r weights the routing loss. The routing objective aligns ρ_i with the eligibility mask m_i^*, while contribution allocation and evidence retrieval are optimized end-to-end through answer prediction.

4 Experiments

4.1 Experimental Setup

We evaluate CoSpace on VSI-Bench (Yang et al., 2025a), where the overall score averages Mean Relative Accuracy over four numerical tasks and accuracy over four multiple-choice tasks. We further report results on MMSI-Bench (Yang et al., 2026b), ViewSpatial (Li et al., 2025), SITE (Wang et al., 2025b), CV-Bench (Tong et al., 2024), and VSI-Debiased (Brown et al., 2025). For SITE, we report chance-adjusted accuracy (CAA). All aggregate scores are reported on a 0–100 scale with one decimal place. Score differences are computed before rounding.

CoSpace is built on Qwen2.5-VL-7B or Qwen3-VL-8B (Bai et al., 2025b;a). The language model and CoSpace modules are jointly optimized, while the Qwen visual encoder and VGGT (Wang et al., 2025a) remain frozen. Inference requires neither task labels nor reference answers. Following GeoThinker (Li et al., 2026), we replace the LLaVA-Hound subset of LLaVA-Video-178K (Zhang et al., 2024) with our spatial instruction data. We consider two regimes. The *vanilla regime* matches training data, backbone, frame sampling, optimization, and evaluation across Qwen2.5-VL, VG-LLM (Zheng et al., 2025), GeoThinker, and CoSpace, with geometry-enhanced methods sharing VGGT. All four models in this matched comparison use 32 frames for both training and inference.

Table 2: **VSI-Bench comparison under the scaled regime following Li et al. (2026) with varying backbones and input frames.**

Model	Backbone	Frames	Avg.	Numerical Answer				Multiple-Choice Answer			
				Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Reference											
Human Level†	–	–	79.2	94.3	47.0	60.4	45.9	94.7	95.8	95.8	100.0
Prior Spatial Reasoning Models											
Cambrian-S-7B (Yang et al., 2025b)	Qwen2.5-7B	128	67.5	73.2	50.5	74.9	72.2	71.1	76.2	41.8	80.1
GeoThinker-7B (Li et al., 2026)	Qwen2.5-VL-7B	32	68.5	–	–	–	–	–	–	–	–
SenseNova-SI-8B (Cai et al., 2026)	InternVL3-8B	64	68.8	72.0	53.5	76.8	72.8	69.6	80.8	48.5	76.4
SpaceMind (Zhao et al., 2026)	InternVL3-8B	32	69.6	73.3	61.4	77.3	74.2	67.2	88.4	44.3	70.6
OneCanvas (Baranowski et al., 2026)	Qwen3-VL-8B	32	70.1	68.4	59.5	75.4	76.5	71.0	84.8	**59.8**	65.7
OmniStream-7B (Yan et al., 2026)	Qwen2.5-7B	32	70.6	73.2	55.7	76.9	74.8	72.3	82.1	45.4	84.6
SpatioLM (Wu et al., 2026)	SenseNovaSI-8B	32	71.6	**81.9**	55.7	75.2	71.5	**78.2**	74.4	47.9	**87.8**
GeoThinker-8B	Qwen3-VL-8B	32	72.6	74.6	62.8	76.9	74.9	75.9	90.2	48.4	77.5
Cambrian-P (Yang et al., 2026a)	Qwen2.5-7B	32	72.0	–	–	–	–	–	–	–	–
Cambrian-P	Qwen2.5-7B	128	73.7	74.9	60.1	76.0	76.9	74.8	89.5	52.6	85.0
CoSpace											
Ours	Qwen2.5-VL-7B	32	71.1	71.9	59.2	74.7	76.7	69.9	88.0	47.9	80.4
Ours	Qwen3-VL-8B	32	73.4	72.2	64.5	76.5	**79.8**	73.1	89.1	47.9	83.7
Ours⋆	Qwen3-VL-8B	128	**74.0**	70.6	**67.2**	**77.9**	78.5	76.3	**92.7**	44.8	84.0

† Human performance is reported on VSI-Bench Tiny. ⋆ The same Qwen3-VL-8B checkpoint trained with 32 frames is evaluated with 128 frames, without retraining. Frames denote the reported number of input frames at evaluation. "–" denotes unavailable results.

The *scaled regime* uses approximately 1.79M training examples with Qwen2.5-VL-7B and Qwen3-VL-8B. CoSpace uses 32 frames for training and inference unless otherwise specified; competing methods retain their reported settings.

4.2 Main Results

Vanilla regime. As shown in Tab. 1, CoSpace achieves the best overall score of **54.9** under matched settings, surpassing Qwen2.5-VL, VG-LLM, and GeoThinker by 1.2, 1.7, and 0.9 points, respectively. It further leads in absolute distance, relative direction, and route planning, highlighting the effectiveness of query-adaptive spatial composition.

Scaled regime. Under the scaled regime, CoSpace achieves **71.1** with Qwen2.5-VL-7B and **73.4** with Qwen3-VL-8B using 32 frames shown in Tab. 2, with the latter outperforming all compared 32-frame models. Without retraining, increasing the inference context to 128 frames further improves CoSpace to **74.0**, surpassing the 128-frame Cambrian-P baseline and achieving the best overall performance among the compared models.

Cross-benchmark performance. CoSpace achieves the best results on MMSI-Bench and ViewSpatial (Table 3), outperforming the Qwen3-VL-8B-Instruct backbone by 3.8 and 8.7 points, respectively. Together with its strong VSI-Bench performance, these results demonstrate robust cross-benchmark generalization of query-adaptive spatial composition.

Table 3: **Cross-benchmark comparison across diverse spatial intelligence benchmarks.** * denotes the S1+S2 variant of VG-LLM; its reported results are taken from GeoThinker (Li et al., 2026).

Model	VSI-Bench	MMSI-Bench	ViewSpatial	SITE(CAA)	CV-Bench
Qwen3-VL-8B-Instruct	57.7	28.8	39.0	**37.8**	85.1
Cambrian-S-7B	67.5	25.8	40.9	36.4	76.9
VG-LLM-8B*	62.2	30.0	45.8	–	81.7
GeoThinker Qwen3-VL-8B	72.6	30.9	45.9	36.1	85.1
Cambrian-P-7B	73.7	28.0	40.0	33.3	78.7
CoSpace Qwen3-VL-8B	73.4	**32.6**	**47.7**	36.9	80.5

4.3 Ablation Studies and Further Analysis

The component ablations in Tab. 4 are retrained from the same Qwen2.5-VL-7B initialization under the vanilla regime, with each ablated component removed throughout both training and inference.

Bank selection and contribution allocation. Removing the router reduces the overall score from 54.9 to 54.3, while removing the allocator lowers it to 53.9. Eliminating both components further decreases performance to 53.7. These results show that selective bank access and adaptive contribution weighting are both important, with their combination providing the strongest performance.

Complementarity of representation banks. Removing the semantic, geometric, or temporal bank lowers the overall score by 1.6, 1.4, and 0.9, respectively. Semantic and geometric banks are par-

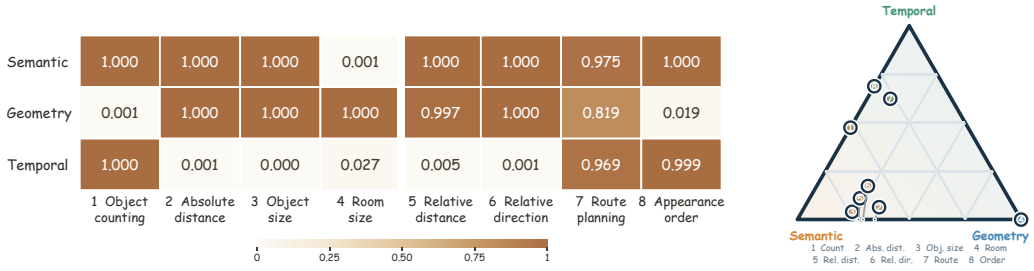


Figure 4: **Task-dependent bank utilization on VSI-Bench.** For CoSpace (Qwen3-VL-8B, scaled regime), the heatmap and ternary plot show mean routing probabilities and allocation weights, respectively, computed per task over the full benchmark.

Table 4: **Ablation study of CoSpace on VSI-Bench under the vanilla regime.** Bank-removal variants exclude the specified bank. All variants are retrained. Δ denotes the overall score change relative to the full model.

Variant	Router	Allocator	Numerical Answer				Multiple-Choice Answer				Avg.	Δ
			Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order		
w/o Router & Allocator	–	–	68.4	41.1	65.3	55.1	52.5	45.4	32.0	69.7	53.7	−1.2
w/o Router	–	✓	68.6	39.4	65.2	59.0	53.9	44.3	32.5	71.4	54.3	−0.6
w/o Allocator	✓	–	68.3	40.9	65.0	54.9	49.2	46.7	34.0	72.5	53.9	−1.0
w/o Semantic Bank	✓	✓	68.9	41.4	61.8	59.6	49.0	43.9	34.0	68.1	53.3	−1.6
w/o Geometry Bank	✓	✓	68.5	39.1	63.1	59.3	50.3	44.2	32.0	72.0	53.6	−1.4
w/o Temporal Bank	✓	✓	66.2	40.7	63.9	58.5	53.0	45.9	35.1	69.4	54.1	−0.9
Full Model	✓	✓	68.2	40.5	64.7	59.0	53.2	47.0	34.5	72.2	**54.9**	–

ticularly important for relative distance and direction, while the temporal bank contributes most to appearance order and object counting. These task-specific degradations highlight the complementary roles of the three representation banks.

Inference-frame scaling and debiased evaluation. Tab. 5 compares CoSpace with VG-LLM, GeoThinker, and Cambrian-S/P (Yang et al., 2025b; 2026a) across inference-frame counts on VSI-Bench and VSI-Debiased. CoSpace achieves the highest or joint-highest VSI-Bench score at every frame count, and leads VSI-Debiased at 8, 32, 64, and 128 frames. Using the same Qwen3-VL-8B checkpoint trained with 32 frames, its VSI-Bench score improves from 73.4 at 32 frames to **74.2** at 64 frames and remains stable at 128 frames (74.0). On VSI-Debiased, performance increases from 67.9 at 32 frames to **69.1** at 128 frames. CoSpace also remains strong with fewer frames, indicating robustness to varying visual context lengths. The consistent gains on VSI-Debiased further suggest that the improvements are not driven by benchmark-specific shortcuts. Overall, CoSpace effectively exploits additional visual observations beyond its training frame count without retraining.

Table 5: **Performance comparison and inference-frame scaling of different methods.** Checkpoints are fixed while varying input frames at inference. * denotes the S1+S2 VG-LLM variant trained with 8 frames, with results from GeoThinker (Li et al., 2026).

Model	Benchmark	# Inference Frames				
		8	16	32	64	128
VG-LLM-8B*	VSI-Bench	–	60.5	62.2	63.7	63.1
	VSI-Debiased	–	51.6	52.4	55.2	55.1
GeoThinker$_{32F}$	VSI-Bench	62.7	69.2	72.6	73.4	73.4
	VSI-Debiased	56.1	**64.3**	66.3	67.7	68.1
Cambrian-S$_{128F}$	VSI-Bench	53.4	58.6	63.6	66.4	67.5
	VSI-Debiased	43.8	49.7	55.6	59.1	59.9
Cambrian-P$_{128F}$	VSI-Bench	57.9	61.0	67.2	70.3	73.7
	VSI-Debiased	49.4	53.4	60.8	63.4	66.9
CoSpace$_{32F}$	**VSI-Bench**	**63.3**	**69.2**	**73.4**	**74.2**	**74.0**
	VSI-Debiased	**57.7**	63.4	**67.9**	**68.4**	**69.1**

Inference-time routing intervention. We further evaluate routing robustness using the same Qwen3-VL-8B checkpoint without retraining while retaining the learned allocator (Tab. 6). *All Banks* activates all three banks, *Random Valid* samples a nonempty routing mask uniformly, and *Swapped* samples a mask excluding the predefined relevance pattern for the current task. These interventions reduce the overall score from 73.4 to 71.5, 70.8, and 70.5, respectively. The consistent degradation, together with the retrained *w/o Router* ablation, confirms that query-adaptive bank

Table 6: **Inference-time routing intervention on VSI-Bench.** All settings use the same Qwen3-VL-8B checkpoint without retraining and retain the learned allocator. Δ denotes the score change relative to the full model.

Setting	Appr. Order	Abs. Dist.	Obj. Count	Rel. Dist.	Obj. Size	Room Size	Route Plan	Rel. Dir.	Overall	Δ
Full Model	83.7	64.5	72.2	73.1	76.5	79.8	47.9	89.1	**73.4**	—
All Banks	84.3	60.4	73.0	74.1	76.6	67.4	47.9	88.5	71.5	−1.8
Random Valid	84.0	61.6	72.1	72.3	75.8	68.2	44.3	88.4	70.8	−2.5
Swapped	83.7	60.2	72.3	73.0	75.5	66.3	44.9	88.0	70.5	−2.9

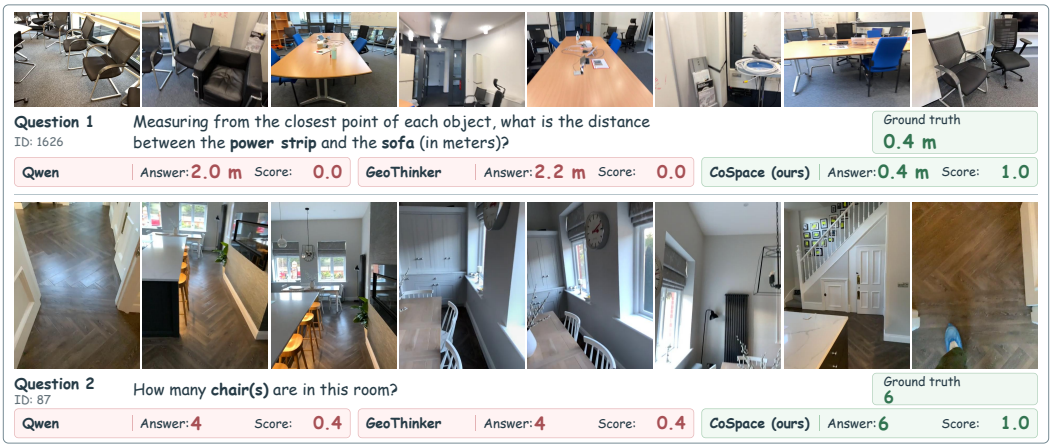


Figure 5: Qualitative comparison of Qwen, GeoThinker, and CoSpace. Predictions are shown alongside ground-truth answers and per-example MRA scores (0–1).

selection provides benefits beyond contribution weighting. We note that part of the inference-time drop may also arise from mismatch with the routing patterns observed during training.

Task-dependent bank utilization. Fig. 4 reveals clear task-specific routing and weighting patterns. Room-size estimation relies almost exclusively on geometry, while object counting and appearance order primarily combine semantic and temporal cues; route planning activates all three banks. Even tasks with similar routing masks exhibit different allocation weights. Appendix F.2 provides examples of router decisions and allocator weights for individual questions.

4.4 Qualitative Analysis

Fig. 5 compares CoSpace with Qwen (Bai et al., 2025b) and GeoThinker (Li et al., 2026) on representative spatial reasoning cases. In the first example, both baselines substantially overestimate the distance between the power strip and sofa, predicting 2.0 m and 2.2 m, whereas CoSpace correctly predicts 0.4 m. In the second example, both baselines undercount the chairs as four, while CoSpace correctly identifies all six. These examples illustrate CoSpace's stronger handling of both metric estimation and object-centric spatial reasoning.

5 Conclusion

We presented **CoSpace**, a compositional spatial reasoning framework for bridging the gap between rich spatial representations and effective reasoning in MLLMs. CoSpace follows the principle of *represent comprehensively, reason selectively*, organizing semantic, geometric, and temporal representations into structured banks and composing them through query-adaptive selection, allocation, and retrieval. Experiments across diverse benchmarks, together with controlled ablations and routing analyses, consistently support the effectiveness of this design. Our results suggest that advancing spatial reasoning requires not only richer representations, but also query-adaptive composition of the right information for each reasoning problem.

AI USE STATEMENT

We used AI tools to assist with data filtering after dataset construction. These tools also assisted with manuscript drafting and polishing, preparation of illustrations and architecture diagrams, code revisions, experimental parameter selection, and literature searches. The authors take full responsibility for the final content of this work, including the data, code, figures, and reported results.

ETHICS STATEMENT

This research adheres to the ethical principles of the ICLR community and focuses on advancing spatial reasoning in multimodal large language models. The study builds on existing datasets and pretrained models, with CoSpace-Instruct-60K derived from visual annotations through the data construction and quality control procedures documented in Appendix A. The original licenses and usage restrictions of these resources remain applicable to their use and redistribution. Improved spatial reasoning has the potential to benefit scene understanding and embodied intelligence; however, these benefits must be considered alongside the risks of inherited biases, erroneous spatial judgments, and misuse in applications involving personal privacy. We acknowledge these potential concerns and emphasize that the proposed framework is intended for academic research and beneficial applications. Deployment in settings where spatial predictions inform consequential decisions requires further assessment of reliability and the relevant ethical implications.

REPRODUCIBILITY STATEMENT

We are committed to supporting the reproducibility of the reported results through transparent documentation of the methodology and experimental procedures. The formulation of CoSpace and the experimental setup are presented in Sections 3 and 4.1, complemented by the data preparation and implementation details in Appendices A–C. Together, these materials describe the construction of CoSpace-Instruct-60K, the generation of routing supervision, and the training configurations for the vanilla and scaled regimes, including data composition, optimization settings, trainable components, and integration with the Qwen backbones. The protocols for inference-time bank interventions and the assumptions and proofs underlying the theoretical analysis are provided in Appendices D and E, respectively. The code will be made publicly available to facilitate independent reproduction of the results and further research.

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*, 2022. URL https://arxiv.org/abs/2112.10482.

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

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

Bartłomiej Baranowski, Dave Zhenyu Chen, and Matthias Nießner. OneCanvas: 3D scene understanding via panoramic reprojection. *arXiv preprint arXiv:2606.19253*, 2026. URL https://arxiv.org/abs/2606.19253.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitScenes: A diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3D: A large benchmark and model for 3D object detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Ellis Brown, Jihan Yang, Shusheng Yang, Rob Fergus, and Saining Xie. Benchmark designers should "train on the test set" to expose exploitable non-visual shortcuts. *arXiv preprint arXiv:2511.04655*, 2025. URL https://arxiv.org/abs/2511.04655.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. SpatialBot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024. URL https://arxiv.org/abs/2406.13642.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, Tongxi Zhou, Jiaqi Li, Hui En Pang, Oscar Qian, Yukun Wei, Zhiqian Lin, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Xiangyu Fan, Hanming Deng, Lewei Lu, Liang Pan, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Scaling spatial intelligence with multimodal foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2511.13719.

Meng Cao, Haoran Tang, Haoze Zhao, Hangyu Guo, Jiaheng Liu, Ge Zhang, Ruyang Liu, Qiang Sun, Ian Reid, and Xiaodan Liang. PhysGame: Uncovering physical commonsense violations in gameplay videos. *arXiv preprint arXiv:2412.01800*, 2024. URL https://arxiv.org/abs/2412.01800.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, 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. URL https://arxiv.org/abs/2401.12168.

Dongping Chen, Yue Huang, Siyuan Wu, Jingyu Tang, Liuyi Chen, Yilin Bai, Zhigang He, Chenlong Wang, Huichi Zhou, Yiqiang Li, Tianshuo Zhou, Yue Yu, Chujie Gao, Qihui Zhang, Yi Gui, Zhen Li, Yao Wan, Pan Zhou, Jianfeng Gao, and Lichao Sun. GUI-World: A video benchmark and dataset for multimodal GUI-oriented understanding. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/3926df1a00c9abf056df7bcf253d026a-Paper-Conference.pdf.

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. URL https://arxiv.org/abs/2406.01584.

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*, 2026. URL https://arxiv.org/abs/2505.20279.

Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-LLM: Extending language model for 3D visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024. URL https://arxiv.org/abs/2403.11401.

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*, 2026. URL https://arxiv.org/abs/2605.27318.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. URL https://arxiv.org/abs/2403.05530.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. URL https://arxiv.org/abs/2507.06261.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2017. doi: 10.1007/s11263-016-0981-7.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *Proceedings of the European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2406.09756.

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. *arXiv preprint arXiv:2408.03326*, 2024. URL https://arxiv.org/abs/2408.03326.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. ViewSpatial-Bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025. URL https://arxiv.org/abs/2505.21500.

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. URL https://arxiv.org/abs/2602.06037.

Xiao Lin, Xiaohu Huang, and Kai Han. Beyond single expert: Harmonizing diverse visual priors in MLLMs for spatial understanding. *arXiv preprint arXiv:2607.15054*, 2026. URL https://arxiv.org/abs/2607.15054.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023a. URL https://arxiv.org/abs/2205.00363.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023b. URL https://arxiv.org/abs/2304.08485.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3DSRBench: A comprehensive 3D spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6924–6934, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Ma_3DSRBench_A_Comprehensive_3D_Spatial_Reasoning_Benchmark_ICCV_2025_paper.html.

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://arxiv.org/abs/2210.07474.

Farzaneh Mahdisoltani, Guillaume Berger, Waseem Gharbieh, David Fleet, and Roland Memisevic. On the effectiveness of task granularity for transfer learning. *arXiv preprint arXiv:1804.09235*, 2018.

Deshui Miao, Xingsen Huang, Yameng Gu, Xin Li, Haijun Zhang, and Ming-Hsuan Yang. GeoWeaver: Grounding visual tokens with geometric evidence before scene reasoning. *arXiv preprint arXiv:2605.22558*, 2026. URL https://arxiv.org/abs/2605.22558.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. URL https://arxiv.org/abs/2410.21276.

OpenGVLab Team. InternVL2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy. InternVL project blog, July 2024. URL https://internvl.github.io/blog/2024-07-02-InternVL-2.0/.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. SAT: Dynamic spatial aptitude training for multimodal language models. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=DW8U8ZWa1U.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Tim J. Schoonbeek, Tim Houben, Hans Onvlee, Peter H. N. de With, and Fons van der Sommen. IndustReal: A dataset for procedure step recognition handling execution errors in egocentric videos in an industrial-like setting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4365–4374, 2024. URL https://arxiv.org/abs/2310.17323.

Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, pp. 279–287, 2019.

Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2406.16860.

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*, 2025a. URL https://arxiv.org/abs/2503.11651.

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. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wang_DUSt3R_Geometric_3D_Vision_Made_Easy_CVPR_2024_paper.html.

Wenqi Wang, Reuben Tan, Pengyue Zhu, Jianwei Yang, Zhengyuan Yang, Lijuan Wang, Andrey Kolobov, Jianfeng Gao, and Boqing Gong. SITE: towards spatial intelligence thorough evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025b. URL https://arxiv.org/abs/2505.05456.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. HoloAssist: An egocentric human interaction dataset for interactive AI assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20270–20281, 2023. URL https://holoassist.github.io/.

Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M. de Melo, Jieneng Chen, and Alan Yuille. Spatial457: A diagnostic benchmark for 6D spatial reasoning of large mutimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24669–24679, 2025c. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Spatial457_A_Diagnostic_Benchmark_for_6D_Spatial_Reasoning_of_Large_CVPR_2025_paper.html.

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. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/142b166c8b18b43774c4f0e0c1e9948e-Abstract-Conference.html.

Jing Wu, Jianhua Wu, Jiayi Guan, Jiahong Chen, Jinghui Lu, Hangjun Ye, Bingzhao Gao, and Long Chen. SpatioLM: Towards general physical spatial intelligence in vision-language models. In *Proceedings of the International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2608.01899.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. PointLLM: Empowering large language models to understand point clouds. In *Proceedings of the European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2308.16911.

Yibin Yan, Jilan Xu, Shangzhe Di, Haoning Wu, and Weidi Xie. OmniStream: Mastering perception, reconstruction and action in continuous streams. In *Proceedings of the European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2603.12265.

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*, 2025a. URL https://arxiv.org/abs/2412.14171.

Jihan Yang, Zifan Zhao, Xichen Pan, Shusheng Yang, Junyi Zhang, Bingyi Kang, Hu Xu, Shang-Wen Li, and Saining Xie. Cambrian-P: Pose-grounded video understanding. *arXiv preprint arXiv:2605.22819*, 2026a. URL https://arxiv.org/abs/2605.22819.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-S: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025b. URL https://arxiv.org/abs/2511.04670.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, 2026b. URL https://arxiv.org/abs/2505.23764.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1910.01442.

Amir R. Zamir, Alexander Sax, William B. Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. From flatland to space: Teaching vision-language models to perceive and reason in 3D. *arXiv preprint arXiv:2503.22976*, 2025. URL https://arxiv.org/abs/2503.22976.

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*, 2026a. URL https://arxiv.org/abs/2603.27437.

Shihua Zhang, Qiuhong Shen, Shizun Wang, Tianbo Pan, and Xinchao Wang. Make geometry matter for spatial reasoning. *arXiv preprint arXiv:2603.26639*, 2026b. URL https://arxiv.org/abs/2603.26639.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. URL https://arxiv.org/abs/2410.02713.

Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16811–16822, 2026. URL https://arxiv.org/abs/2511.23075.

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, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/1d8dc55c1f6cf124af840ce1d92d1896-Abstract-Conference.html.

Hongyan Zhi, Peihao Chen, Junyan Li, Shuailei Ma, Xinyu Sun, Tianhang Xiang, Yinjie Lei, Mingkui Tan, and Chuang Gan. LSceneLLM: Enhancing large 3D scene understanding using adaptive visual preferences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3761–3771, 2025. URL https://arxiv.org/abs/2412.01292.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics*, 37(4), 2018. doi: 10.1145/3197517.3201323.

APPENDIX

A CoSpace-Instruct-60K **16**

B Routing Supervision **18**

C Training Details **20**

D Inference-time Bank Intervention **21**

E Theoretical Analysis of Query-Conditioned Spatial Composition **23**

E.1 Problem Formulation 23

E.2 Risk Dominance of Conditional Composition 23

E.3 Relation to the CoSpace Function Class 24

E.4 Selective Composition as Nuisance Suppression 24

E.5 Scope of the Analysis 25

F Additional Analysis on VSI-Bench **26**

F.1 Task-wise Performance Gains 26

F.2 Query-dependent Bank Composition 26

A COSPACE-INSTRUCT-60K

We construct **CoSpace-Instruct-60K**, a spatial instruction dataset comprising approximately 60K image and video question–answer pairs organized around semantic, geometric, and temporal capabilities. The dataset provides supervision for the foundational visual capabilities that support spatial reasoning. Questions and answers are derived from structured visual annotations, including object attributes, bounding boxes, surface normals, camera trajectories, and action labels. The dataset contains 18K semantic, 22K geometric, and 20K temporal examples, with the task composition summarized in Tab. 7. These categories indicate the primary supervision target of each question, while allowing individual examples to involve multiple capabilities.

Semantic data. Semantic questions are constructed from the object, attribute, and region annotations in Visual Genome (Krishna et al., 2017). They connect appearance understanding with entity identification and region grounding. Color and visible-state questions ask the model to recognize annotated properties, while attribute-based identification requires identifying an entity that satisfies a combination of descriptive attributes. Grounding questions establish correspondence in both directions, predicting a region from an object name or identifying an object from a region, alongside verification of proposed object–region matches. Questions use direct-answer, binary, or multiple-choice formats where appropriate. Negative answers and distractors are constructed from incompatible attributes or alternative entities, with checks to reduce ambiguous references and answer leakage through object names.

Geometric data. Geometric supervision combines object-level 3D annotations accessed through Omni3D (Brazil et al., 2023), surface normals from Taskonomy (Zamir et al., 2018), and shape attributes from Visual Genome. The Omni3D sources include Hypersim (Roberts et al., 2021), nuScenes (Caesar et al., 2020), SUN RGB-D (Song et al., 2015), ARKitScenes (Baruch et al., 2021), and KITTI (Geiger et al., 2012). Object dimensions and depth information support comparisons of spatial scale, height, and distance, retaining candidates with sufficiently distinct measurements. Spatial-relation questions use bounding-box geometry to derive containment, support, contact, proximity, and occlusion labels. These relations are annotation-based geometric proxies,

Table 7: **Task composition of CoSpace-Instruct-60K.** Sizes denote question–answer pairs in thousands (K). Individual entries are rounded independently.

Task	Source	Size
Semantic		**18K**
Color recognition	Visual Genome	2.8K
Visible-state recognition	Visual Genome	3.9K
Attribute-based identification	Visual Genome	4.5K
Object-to-region grounding	Visual Genome	2.6K
Region-to-object grounding	Visual Genome	2.6K
Grounding verification	Visual Genome	1.5K
Geometry		**22K**
Relative spatial scale	Omni3D	3.5K
Relative height	Omni3D	2.0K
Relative depth	Omni3D	4.8K
Object spatial relations	Omni3D	4.8K
Surface relations	Taskonomy	2.4K
Surface orientation	Taskonomy	1.4K
Shape recognition	Visual Genome	3.1K
Temporal		**20K**
Temporal phase	VidSTG / VidOR	1.7K
Visibility continuity	VidSTG / VidOR	2.1K
Start/end presence	VidSTG / VidOR	2.3K
Camera motion	RealEstate10K	5.4K
Directional displacement	Something-Something V2	3.9K
Deformation	Something-Something V2	1.9K
Rotation	Something-Something V2	1.1K
Gravity-related actions	Something-Something V2	0.8K
Pretense and failed actions	Something-Something V2	0.9K
Total		**60K**

rather than exact physical measurements. Surface questions use local normals to determine orientation and parallel or orthogonal relationships, while shape questions associate entities with their annotated geometric forms. Together, these tasks cover both object-level structure and local surface geometry.

Temporal data. Temporal supervision draws on VidSTG annotations over VidOR videos (Zhang et al., 2020; Shang et al., 2019), RealEstate10K camera trajectories (Zhou et al., 2018), and Something-Something V2 action labels (Mahdisoltani et al., 2018). VidSTG relation intervals are aggregated for each target to construct questions about temporal phase, continuity, and presence near the beginning or end of a video. These intervals serve as proxies for temporal presence, rather than exhaustive frame-level visibility annotations. RealEstate10K trajectories provide translation and rotation labels for camera-motion questions, with clips segmented according to changes in motion. Something-Something V2 contributes questions about directional displacement, deformation, rotation, gravity-related actions, and pretense or failed actions. Their answers use instantiated action descriptions, linking the observed motion to the participating objects.

Quality control and supervision. Task-specific filtering removes invalid annotations, ambiguous entity references, and spatial or temporal relationships that do not support a sufficiently clear answer. Selected subsets receive additional review with Qwen2.5-VL-7B-Instruct to assess consistency between the visual content and the proposed annotation. These checks cover object attributes, region grounding, and video action descriptions. All retained examples follow a common visual-input, question, and final-answer format. Language supervision is restricted to the final answer, without intermediate reasoning targets. The structured source annotations are used for data construction rather than supplied as additional model inputs. Representative examples from CoSpace-Instruct-60K are shown in Figure 6.



Figure 6: **Representative examples from CoSpace-Instruct-60K.** Visual inputs, questions, and final answers illustrate the semantic, geometric, and temporal tasks.

Capability composition analysis. Beyond the primary task categories, individual examples may require multiple types of spatial evidence. Fig. 7 summarizes the capability composition of CoSpace-Instruct-60K.

For CoSpace-Instruct-60K, the three categories are explicitly balanced to provide complementary supervision for semantic, geometric, and temporal reasoning. For the full training mixture, we further analyze the distribution of activated capability combinations according to the routing masks described in Appendix B. The statistics show that multi-capability examples constitute a substantial portion of the training mixture, reflecting the compositional nature of spatial reasoning tasks.

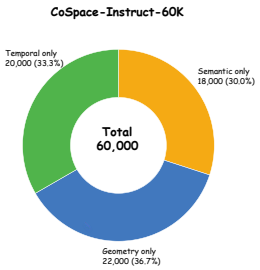


Figure 7: Composition of CoSpace-Instruct-60K across semantic, geometric, and temporal supervision.

B Routing Supervision

We supervise the router with task masks constructed through manually specified rules for each training source. These routing masks also enable the capability composition analysis of the full training mixture shown in Fig. 1(b), where different combinations of semantic, geometric, and temporal capabilities are summarized. Each mask is a multi-hot vector $\mathbf{m} = (m_{\text{sem}}, m_{\text{geo}}, m_{\text{tmp}}) \in \{0,1\}^3$, whose entries indicate the eligibility of semantic, geometric, and temporal banks, respectively. Semantic labels concern entity identity, attributes, and reference resolution; geometric labels concern spatial properties and relations; temporal labels concern cross-frame association, motion, and event progression. An example may activate one, two, or all three banks. The labels are assigned offline using structured task information where available, supplemented by source priors and textual rules for heterogeneous instruction data.

CoSpace-Instruct-60K and SPAR-derived tasks. Our task mixture combines examples sampled from CoSpace-Instruct-60K (Appendix A) with spatial tasks derived from SPAR (Zhang et al., 2025). The foundational semantic, geometric, and temporal tasks retain their respective single-bank labels. SPAR-derived tasks involving object references and spatial reasoning activate semantic and geometric banks, including distance and scale comparisons, spatial imagination, and cross-view inference. Object counting, appearance order, and frame localization activate semantic and temporal

Table 8: **Routing labels for data with structured task information.** Mask entries denote semantic, geometric, and temporal bank eligibility, in that order. Assignments follow the task definitions of each source.

Task group	Mask
CoSpace-Instruct-60K and SPAR-derived mixture	
Foundational semantics: attributes, identification, and grounding	100
Foundational geometry: depth, height, scale, shape, and surfaces	010
Foundational temporal tasks: event progression and motion	001
Object-centered spatial reasoning, spatial imagination, and cross-view inference	110
Object counting, appearance order, and frame localization	101
Room-scale estimation	010
Indoor navigation and route planning	111
VSI-590K	
Absolute/relative object counts and appearance order	101
Object distance, direction, and size	110
Room size	010
VLM-3R	
Object counting	101
Room size	010
Object size, distance, direction, and relative position	110
Camera displacement and movement direction	011
Camera–object distance across frames	111
Route planning	111

banks to associate entities across observations. Room-scale estimation receives a geometric label, while navigation activates all three banks. Assignments follow the structured task markers retained during construction, including those associated with records lacking an explicit question-type field. Tab. 8 summarizes these mappings alongside those used for VSI-590K and VLM-3R.

VSI-590K. For VSI-590K (Yang et al., 2025b), we use a fixed mapping over its 12 question types. Absolute and relative counting, together with appearance order, receive semantic–temporal masks. Object distance, direction, and size questions receive semantic–geometric masks, including questions expressed relative to the camera. Room-size estimation activates only the geometric bank. These assignments encode task-level priors and are applied consistently across image and video examples of each question type.

VLM-3R. For VLM-3R (Fan et al., 2026), object counting, room size, and object spatial properties follow the mappings above. Camera displacement and movement direction activate geometric and temporal banks, whereas camera–object distance tasks additionally activate the semantic bank to account for the referenced entity. Object–object position questions retain semantic–geometric labels, and route planning activates all three banks. For indoor questions without existing task labels, we infer their types through ordered question-template rules before applying the mapping. These rules identify expressions for counting, room area, object dimensions, distance, direction, and navigation.

PhysInstruct. For PhysInstruct (Cao et al., 2024), we construct routing labels from the question and its annotated final answer, with the semantic bank always eligible. Temporal labels are determined by cues for motion, event order, interactions, anomalies, and causal changes. Geometric labels use spatial and physical cues, including position, contact, collision, gravity, and trajectories. The rules combine question-specific cues with evidence from the question–answer pair: for example, motion accompanied by spatial descriptions activates the geometric label, while anomaly questions supported by motion, physical, or spatial evidence activate both geometric and temporal labels.

Rule precedence accounts for questions whose focus is primarily semantic. Product or advertising questions receive 100 directly. Appearance or style questions suppress geometric and temporal labels under the appearance-only conditions, although subsequent anomaly rules can restore them when supporting evidence is present. If neither label remains active but anomaly or causal cues are

Table 9: Training data composition. Counts indicate the number of examples used from each source.

Source	Vanilla	Scaled
CoSpace-Instruct-60K	40,000	40,000
SPAR-derived spatial data	251,000	251,000
VSI-590K	–	590,667
VLM-3R	–	340,226
PhysInstruct	–	139,209
Cambrian-S-3M subset	–	430,000
Total	291,000	1,791,102

present, the temporal label is enabled. The resulting masks therefore reflect an ordered set of rules rather than the union of all keyword matches.

Cambrian video instructions. For the 430K subset of Cambrian-S-3M (Yang et al., 2025b), we combine source priors with user questions and task metadata, keeping the semantic bank eligible throughout. CLEVRER (Yi et al., 2020), Something-Something V2 (Mahdisoltani et al., 2018), HoloAssist (Wang et al., 2023), and IndustReal (Schoonbeek et al., 2024) examples receive 111 directly. Other designated action- and event-oriented sources start from `101`, with geometric labels added when spatial or physical cues are present. Question and task descriptions further identify counting, event order, actions, causality, prediction, and spatial relations to refine the temporal and geometric labels.

GUI-World (Chen et al., 2025) receives a separate treatment: the geometric label remains inactive, while interface operations such as clicking, typing, scrolling, and navigation determine temporal eligibility. Text-recognition questions without geometric cues retain an inactive geometric label. For appearance-only questions without general temporal or geometric cues, temporal eligibility is retained only when supported by the source prior; this refinement can remove labels triggered solely by counting expressions. Multi-turn examples receive a single record-level mask based on all user turns and the associated metadata. Assistant answers are excluded from this process.

Interpretation and use. These masks provide weak supervision under the task conventions of each source. Related task names can therefore receive different labels: foundational camera-motion questions use `001`, SPAR-derived view-change inference uses `110`, and VLM-3R camera-motion tasks use `011`. Video input alone does not determine temporal eligibility. During training, the masks supervise the router and specify which bank outputs participate in fusion, while the allocator determines their relative contributions. Language supervision remains restricted to final answers. At inference, the router predicts bank eligibility from the question, without access to the annotated answers used in offline label construction.

C Training Details

Training data. Table 9 summarizes the data used in the vanilla and scaled regimes. The vanilla regime combines 40,000 examples from CoSpace-Instruct-60K with 251,000 SPAR-derived spatial examples, for a total of 291,000 examples. The scaled regime additionally incorporates VSI-590K, VLM-3R, PhysInstruct, and a subset of Cambrian-S-3M, yielding 1,791,102 training examples. The construction of CoSpace-Instruct-60K and the routing labels are described in Appendices A and B, respectively. All configurations use final-answer language supervision.

Training configuration. We use Qwen2.5-VL-7B for the vanilla regime and both Qwen2.5-VL-7B and Qwen3-VL-8B for the scaled regime. All configurations are trained for one epoch on eight GPUs with 32 input frames, a global batch size of 64, and a learning rate of 10^{-5}. We fully fine-tune the language model while keeping the Qwen visual encoder and VGGT frozen. The geometry merger, bank projections, router, allocator, and cross-attention refinement branches are optimized jointly with the language model. For Qwen3-VL, the additional geometric projections and scalar gates along the DeepStack pathway are also trainable.

Table 10: Layer correspondence in the Qwen3-VL DeepStack integration. All indices are zero-based.

VGGT layer	Qwen visual layer	Language layer
11	8	0
17	16	1
23	24	2

Geometry feature extraction. We use VGGT (Wang et al., 2025a) as a geometry feature encoder, combining per-image feature extraction with cross-frame aggregation to capture geometric relationships across input views. The resulting latent patch features are projected for shared geometry fusion and spatial bank construction, while camera tokens provide additional viewpoint information for the temporal bank.

Routing and allocation. During training, annotated masks determine bank participation, while the router is supervised by the mean binary cross-entropy across the three banks with weight $\lambda_r = 0.2$. At inference, bank eligibility is determined by the strict threshold $m_i = \mathbb{I}[\rho_i > 0.5]$. The allocator normalizes utility scores over eligible banks with temperature $\tau_u = 1.0$; a single eligible bank receives unit weight. Contribution weights are learned through the answer-generation objective without additional weight targets.

If no bank is eligible, all contribution weights are set to zero without applying normalization over the empty subset, yielding $\mathbf{H}_{\text{out}} = \mathbf{H}_{\text{base}}$. This rule applies during training, inference, and ablations. For bank-removal ablations, the corresponding bank is disabled throughout training and inference, and all training examples are retained. In particular, removing the geometry bank changes a room-size training mask from 010 to 000, leaving only the shared additive fusion of Qwen visual and VGGT features in the bank-refinement pathway.

Qwen3-VL integration. We augment the DeepStack pathway with geometric features from VGGT layers 11, 17, and 23, aligned with Qwen3-VL visual layers 8, 16, and 24, respectively. The resulting features are injected into language layers 0, 1, and 2. Table 10 lists the correspondence; all indices are zero-based. Each geometric branch uses an independent learnable scalar gate. These contributions are independent of the router masks and allocator weights and remain active when the bank mask is empty.

D Inference-time Bank Intervention

Complementing the retrained component ablations in Table 4 and the routing interventions in Table 6, we examine how the trained model responds to interventions on individual bank refinement branches. All experiments use the same Qwen3-VL-8B checkpoint on VSI-Bench, with all learned parameters fixed and no additional training.

Following the notation in the main text, let $\mathcal{I} = \{\text{sem}, \text{geo}, \text{tmp}\}$ denote the bank index set, m_i the original eligibility mask, a_i the utility score, u_i the normalized contribution weight, and Δ_i the refinement residual of bank i. The original composition is

$$H_{\text{out}} = H_{\text{base}} + \sum_{i \in \mathcal{I}} u_i \Delta_i. \tag{6}$$

We distinguish residual suppression after allocation from eligibility removal before weight normalization.

Zero Bank: suppressing the residual after allocation. For a target bank k, we retain the original routing decisions and contribution weights, but set its refinement residual to zero:

$$\Delta_i^{\text{zero}(k)} = \begin{cases} 0, & i = k, \\ \Delta_i, & i \neq k, \end{cases} \qquad H_{\text{out}}^{\text{zero}(k)} = H_{\text{base}} + \sum_{i \in \mathcal{I} \setminus \{k\}} u_i \Delta_i. \tag{7}$$

The remaining weights are not renormalized. Thus, bank k may retain a nonzero allocation weight while contributing no refinement residual.

Table 11: Inference-time bank intervention on VSI-Bench using the same Qwen3-VL-8B checkpoint without retraining. Zero preserves the original allocation weights, whereas Disabled renormalizes weights over the remaining eligible banks.

Setting	Appearance	Abs. Dist.	Counting	Rel. Dist.	Size	Room	Route	Rel. Dir.	Overall
Full Model	83.7	64.5	72.2	73.1	76.5	79.8	47.9	89.1	**73.4**
Zero Semantic	82.9	64.7	72.9	74.1	76.2	79.8	46.4	89.3	73.3
Zero Geometry	83.7	61.7	72.2	74.4	76.5	65.7	46.9	88.6	71.2
Zero Temporal	83.2	64.5	72.6	73.1	76.5	79.8	46.4	89.1	73.2
Disabled Semantic	83.7	63.1	71.7	73.0	75.0	79.8	44.9	87.2	72.3
Disabled Geometry	83.7	62.1	72.2	74.1	76.4	65.7	46.4	89.1	71.2
Disabled Temporal	83.0	64.5	73.0	73.1	76.5	79.8	44.9	89.1	73.0

Disabled Bank: removing eligibility before normalization. For a target bank k, we set its eligibility to zero while retaining the original eligibility decisions and utility scores for the remaining banks. Under the bank-specific scoring functions in the main text, removing bank k does not change the inputs to the other banks' utility estimators. The remaining eligible subset is

$$\mathcal{S}^{(-k)} = \{i \in \mathcal{I} \setminus \{k\} : m_i = 1\}. \tag{8}$$

The contribution weights and output become

$$\begin{aligned} u_i^{\text{disabled}(k)} &= \begin{cases} \dfrac{\exp(a_i/\tau_u)}{\sum_{j\in\mathcal{S}^{(-k)}} \exp(a_j/\tau_u)}, & i \in \mathcal{S}^{(-k)}, \\ 0, & \text{otherwise}, \end{cases} \\ H_{\text{out}}^{\text{disabled}(k)} &= H_{\text{base}} + \sum_{i\in\mathcal{I}} u_i^{\text{disabled}(k)} \Delta_i. \end{aligned} \tag{9}$$

If $\mathcal{S}^{(-k)}$ is empty, normalization is skipped and all weights are set to zero, yielding $H_{\text{out}} = H_{\text{base}}$. Otherwise, the remaining eligible weights sum to one and satisfy $u_i^{\text{disabled}(k)} = u_i/(1-u_k)$ for $i \in \mathcal{S}^{(-k)}$. This renormalization preserves the relative weight ratios among the remaining eligible banks; it does not introduce a new estimate of their relative importance.

Both interventions preserve the shared geometry-enhanced representation H_{base} and the independent geometric contributions in the Qwen3-VL DeepStack pathway. Disabling or zeroing the geometry bank therefore removes its refinement branch contribution, rather than all geometric information available to the model.

Results and interpretation. Table 11 reports the results. Geometry-bank interventions produce the largest observed overall decrease: both Zero Geometry and Disabled Geometry obtain 71.2, compared with 73.4 for the full model. In both cases, the room-size score falls from 79.8 to 65.7, indicating particular sensitivity to geometry-bank refinement for room-size estimation. Although their rounded overall scores coincide, their task-level outcomes differ.

Zero Semantic obtains 73.3, a decrease of 0.1 points, whereas Disabled Semantic obtains 72.3, a decrease of 1.1 points. This contrast shows that the intervention outcome depends on whether the remaining contributions are renormalized. The two settings therefore probe different perturbations of the same learned composition. Zero Temporal and Disabled Temporal obtain 73.2 and 73.0, corresponding to decreases of 0.2 and 0.4 points. We interpret these small aggregate changes cautiously in the absence of uncertainty estimates.

These interventions characterize the checkpoint's sensitivity to bank refinement removal and contribution normalization. Since training uses task-dependent bank combinations, the observed changes may also partly reflect a mismatch with training-time composition patterns. They complement the retrained ablations in Table 4 by examining the behavior of an already trained model under targeted interventions.

E THEORETICAL ANALYSIS OF QUERY-CONDITIONED SPATIAL COMPOSITION

We examine two properties: the flexibility of input-dependent composition relative to a fixed mixture, and the effect of adding task-irrelevant residuals under a simplified noise model. The analysis concerns composition with the representation and prediction modules held fixed; it does not establish a guarantee on the optimization outcome of the complete MLLM.

E.1 PROBLEM FORMULATION

Let $Z = (\mathcal{X}, Q)$ denote a visual input–question pair and Y its target answer. For the analysis, we hold the shared representation, bank-specific refinement functions, and downstream predictor fixed, and vary only the composition. The output representation is

$$\mathbf{H}_{\text{out}}(Z) = \mathbf{H}_{\text{base}}(Z) + \sum_{i \in \mathcal{I}} c_i(Z)\mathbf{\Delta}_i(Z), \qquad \mathcal{I} = \{\text{sem}, \text{geo}, \text{tmp}\}. \tag{10}$$

Here, $c_i(Z) = m_i(Z)u_i(Z)$ combines bank eligibility and contribution weight. Because the allocator already assigns zero weight to excluded banks, $c_i(Z) = u_i(Z)$ under the implementation in Sec. 3.3. A nonempty selection produces nonnegative coefficients summing to one; an empty selection gives $\mathbf{c} = \mathbf{0}$.

Let $\mathcal{C}$ be a nonempty set of feasible compositions. For input $Z = z$ and composition $\mathbf{c} \in \mathcal{C}$, define the conditional prediction risk

$$r_z(\mathbf{c}) = \mathbb{E}\left[\ell\left(Y,\, f\left(\mathbf{H}_{\text{base}}(z) + \sum_{i \in \mathcal{I}} c_i \mathbf{\Delta}_i(z), Q\right)\right) \mid Z = z\right], \tag{11}$$

where f denotes the fixed downstream predictor and ℓ the answer-generation loss. We assume that the risks and their pointwise infimum are measurable and that the expectations below are finite.

Fixed composition. A single composition is shared across all inputs. Its optimal population risk is

$$\mathcal{R}^*_{\text{fixed}} = \inf_{\mathbf{c} \in \mathcal{C}} \mathbb{E}_Z\left[r_Z(\mathbf{c})\right]. \tag{12}$$

Conditional oracle. Allowing the composition to vary freely with the input gives the pointwise oracle risk

$$\mathcal{R}^*_{\text{cond}} = \mathbb{E}_Z\left[\inf_{\mathbf{c} \in \mathcal{C}} r_Z(\mathbf{c})\right]. \tag{13}$$

This quantity is an idealized lower bound for feasible input-dependent composition rules, rather than CoSpace's training objective or achieved risk. Attaining it requires a measurable rule that selects pointwise optimal compositions, which a learned router–allocator need not represent or recover.

E.2 RISK DOMINANCE OF CONDITIONAL COMPOSITION

Proposition 1. Under the assumptions above,

$$\mathcal{R}^*_{\text{cond}} \le \mathcal{R}^*_{\text{fixed}}. \tag{14}$$

Proof. For every fixed $\mathbf{c} \in \mathcal{C}$ and input z,

$$\inf_{\mathbf{c}' \in \mathcal{C}} r_z(\mathbf{c}') \le r_z(\mathbf{c}). \tag{15}$$

Taking expectations gives

$$\mathbb{E}_Z\left[\inf_{\mathbf{c}' \in \mathcal{C}} r_Z(\mathbf{c}')\right] \le \mathbb{E}_Z\left[r_Z(\mathbf{c})\right]. \tag{16}$$

The inequality holds for every fixed composition. Taking the infimum over its right-hand side therefore yields

$$\mathbb{E}_Z\left[\inf_{\mathbf{c}' \in \mathcal{C}} r_Z(\mathbf{c}')\right] \le \inf_{\mathbf{c} \in \mathcal{C}} \mathbb{E}_Z\left[r_Z(\mathbf{c})\right], \tag{17}$$

which proves Eq. (14). □

Strict improvement. Suppose an optimal fixed composition $\mathbf{c}^*_{\text{fixed}}$ exists. Its input-dependent excess risk is

$$g(Z) = r_Z(\mathbf{c}^*_{\text{fixed}}) - \inf_{\mathbf{c}\in\mathcal{C}} r_Z(\mathbf{c}) \geq 0. \tag{18}$$

Consequently,

$$\mathcal{R}^*_{\text{fixed}} - \mathcal{R}^*_{\text{cond}} = \mathbb{E}_Z[g(Z)]. \tag{19}$$

If $\Pr(g(Z) > 0) > 0$, the expected excess risk is positive, and hence

$$\mathcal{R}^*_{\text{cond}} < \mathcal{R}^*_{\text{fixed}}. \tag{20}$$

Thus, strict improvement requires more than allowing the composition to vary: the best fixed composition must be suboptimal on a set of inputs with positive probability. For example, metric-distance and appearance-order questions may favor different mixtures. When no optimal fixed mixture accommodates both demands, the conditional oracle can achieve a lower population risk.

E.3 Relation to the CoSpace Function Class

The pointwise oracle permits unrestricted input-dependent choices. CoSpace instead uses a parameterized router and allocator, whose capacity depends on their inputs and functional forms.

Let $\mathcal{F}_{\text{fixed}}$ be the family of constant composition functions:

$$\mathcal{F}_{\text{fixed}} = \{c : \; c(Z) = \bar{\mathbf{c}}, \quad \bar{\mathbf{c}} \in \mathcal{C}\}. \tag{21}$$

Let $\mathcal{F}_{\text{cond}}$ be a family of measurable composition functions conditioned on the question and visual bank summaries. Define $\mathcal{R}(c) = \mathbb{E}_Z[r_Z(c(Z))]$. If

$$\mathcal{F}_{\text{fixed}} \subseteq \mathcal{F}_{\text{cond}}, \tag{22}$$

then minimizing the same risk over the larger family gives

$$\mathcal{R}^*_{\text{cond}} \leq \inf_{c\in\mathcal{F}_{\text{cond}}} \mathcal{R}(c) \leq \inf_{c\in\mathcal{F}_{\text{fixed}}} \mathcal{R}(c) = \mathcal{R}^*_{\text{fixed}}. \tag{23}$$

The first inequality follows from the pointwise oracle definition; the second follows from set inclusion.

The inclusion in Eq. (22) is compatible with the router–allocator design whenever the heads can produce constant routing scores and utility logits. For a fixed nonempty support $\mathcal{S}$ and positive weights $\bar{c}_i$ summing to one on that support, constant routing scores can select $\mathcal{S}$. Setting $a_i = \tau_u \log \bar{c}_i$ for $i \in \mathcal{S}$ then recovers the desired weights through masked softmax. Selecting no banks recovers the zero composition. These are capacity statements about inference-time composition, not claims that routing supervision will lead training to these solutions.

Finite data, imperfect supervision, and nonconvex optimization may prevent the learned model from reaching the best member of $\mathcal{F}_{\text{cond}}$. Moreover, question and bank summaries may not retain all information available to the pointwise oracle. The analysis therefore motivates conditional composition without guaranteeing an empirical advantage for an arbitrary learned router.

E.4 Selective Composition as Nuisance Suppression

We next consider a simplified additive model for the effect of task-irrelevant refinements. Fix a question type and let $\mathcal{A} \subseteq \mathcal{I}$ denote its relevant banks. All expectations in this subsection are over inputs and residual variability within this setting. For notational convenience, residual representations are treated as vectors.

Let $\mathbf{\Delta}^*$ be the ideal residual, and write the residual from relevant banks as

$$\mathbf{\Delta}_{\text{rel}} = \mathbf{\Delta}^* + \mathbf{e}, \tag{24}$$

where $\mathbf{e}$ is its approximation error. Holding this relevant residual fixed, suppose additional bank contributions act as nuisance terms:

$$\mathbf{\Delta}_{\text{all}} = \mathbf{\Delta}_{\text{rel}} + \sum_{j\notin\mathcal{A}} \alpha_j \boldsymbol{\epsilon}_j. \tag{25}$$

Here, the coefficients α_j are fixed constants. Assume finite second moments and

$$\begin{aligned} \mathbb{E}[\boldsymbol{\epsilon}_j] &= 0, \\ \mathbb{E}\left[\boldsymbol{\epsilon}_j^\top \boldsymbol{\epsilon}_k\right] &= 0, \qquad j \neq k, \\ \mathbb{E}\left[\mathbf{e}^\top \boldsymbol{\epsilon}_j\right] &= 0. \end{aligned} \tag{26}$$

These assumptions describe zero-mean nuisance components that are orthogonal in expectation to one another and to the relevant residual error.

Proposition 2. Under these assumptions,

$$\begin{aligned} &\mathbb{E}\left[\|\boldsymbol{\Delta}_{\mathrm{all}} - \boldsymbol{\Delta}^*\|_2^2\right] \\ &\quad = \mathbb{E}\left[\|\boldsymbol{\Delta}_{\mathrm{rel}} - \boldsymbol{\Delta}^*\|_2^2\right] + \sum_{j \notin \mathcal{A}} \alpha_j^2 \mathbb{E}\left[\|\boldsymbol{\epsilon}_j\|_2^2\right]. \end{aligned} \tag{27}$$

Therefore, adding these nuisance refinements cannot decrease expected squared representation error. The increase is strict whenever some $\alpha_j \neq 0$ and $\mathbb{E}[\|\boldsymbol{\epsilon}_j\|_2^2] > 0$.

Proof. Equations (24) and (25) give

$$\boldsymbol{\Delta}_{\mathrm{all}} - \boldsymbol{\Delta}^* = \mathbf{e} + \sum_{j \notin \mathcal{A}} \alpha_j \boldsymbol{\epsilon}_j. \tag{28}$$

Expanding the squared norm,

$$\begin{aligned} \left\| \mathbf{e} + \sum_{j \notin \mathcal{A}} \alpha_j \boldsymbol{\epsilon}_j \right\|_2^2 &= \|\mathbf{e}\|_2^2 + \sum_{j \notin \mathcal{A}} \alpha_j^2 \|\boldsymbol{\epsilon}_j\|_2^2 \\ &\quad + 2 \sum_{j \notin \mathcal{A}} \alpha_j \mathbf{e}^\top \boldsymbol{\epsilon}_j \\ &\quad + 2 \sum_{\substack{j<k \\ j,k \notin \mathcal{A}}} \alpha_j \alpha_k \boldsymbol{\epsilon}_j^\top \boldsymbol{\epsilon}_k. \end{aligned} \tag{29}$$

Taking expectations eliminates the cross terms. Substituting $\mathbf{e} = \boldsymbol{\Delta}_{\mathrm{rel}} - \boldsymbol{\Delta}^*$ establishes Eq. (27). □

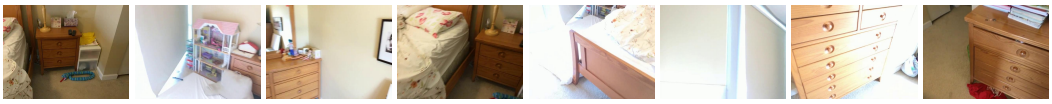
Interpretation and limitations. Relevance is question-dependent: a bank can provide useful information for one task while contributing nuisance variation for another. The proposition identifies a mechanism by which suppressing an irrelevant refinement can reduce error, while the shared representation remains available.

This comparison holds the relevant residual and its coefficients fixed. In CoSpace, changing bank eligibility can also renormalize the weights of the remaining banks. Furthermore, learned weights depend on the input and may correlate with residual errors. For input-dependent coefficients, the cross terms vanish only under suitable weighted or conditional orthogonality assumptions; the unweighted assumptions above are not sufficient. Proposition 2 is therefore an explanatory model, not an exact prediction of an all-bank intervention.

E.5 Scope of the Analysis

Proposition 1 compares a pointwise conditional oracle with the best fixed composition under a common representation and predictor. Its function-class counterpart additionally requires the inclusion in Eq. (22). Neither result guarantees that finite-sample training recovers an optimal composition, nor does it establish dominance over all existing spatial reasoning methods.

Proposition 2 concerns squared representation error under fixed-coefficient additive nuisance assumptions. It does not directly bound the autoregressive answer-generation loss or account for changes to other branches during retraining.



Question	Semantic	Geometric	Temporal	Prediction	Ground truth
Which is closest to the mirror: window, TV, nightstand, or lamp?	0.76	0.24	—	nightstand	nightstand
What is the size of this room (m^2)?	—	1.00	—	16.0	16.0
In what order do lamp, trash bin, TV, and nightstand first appear?	0.30	—	0.70	nightstand → lamp → TV → trash bin	nightstand → lamp → TV → trash bin
How many pillows are in this room?	0.52	—	0.48	3	3
Which is closest to the closet: bed, TV, lamp, or mirror?	0.80	0.20	—	mirror	bed

Figure 8: Qualitative examples of CoSpace on VSI-Bench under the scaled regime. Selected video frames are shown with five questions, bank allocation weights u_i, predictions, and ground-truth answers. Weights are normalized over the selected banks; "—" denotes an unselected bank. Green and red predictions indicate correct and incorrect answers, respectively.

Together, these results explain two potential benefits of query-adaptive composition: accommodating different evidence requirements across inputs and suppressing irrelevant residual variation. Whether these benefits are realized by CoSpace is assessed through the empirical comparisons and ablations.

F Additional Analysis on VSI-Bench

F.1 Task-wise Performance Gains

Fig. 9 shows the task-wise score differences between CoSpace and VG-LLM under the vanilla regime, based on Tab. 1. CoSpace improves on seven of the eight tasks, indicating that the benefit of compositional spatial reasoning extends across diverse reasoning categories rather than being concentrated on a single task. The largest gain appears in route planning (+7.2), followed by room-size estimation (+2.6) and relative-distance estimation (+2.5), all of which require integrating broader scene structure with task-specific spatial cues. CoSpace also improves absolute distance (+1.6), relative direction (+1.2), object counting (+0.4), and appearance order (+0.2), showing consistent benefits across both metric and relational reasoning. The only degradation occurs on object-size estimation (-2.1), suggesting that adaptive composition is not uniformly beneficial for every spatial attribute. Overall, the task-wise pattern supports our central observation that different reasoning problems benefit from different spatial information, and that selectively composing such information can improve performance across a broad range of tasks. Differences are computed before rounding and reported to one decimal place.

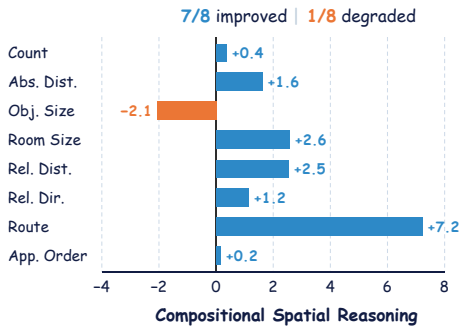


Figure 9: Task-wise gains of CoSpace over VG-LLM on VSI-Bench under the vanilla regime.

F.2 Query-dependent Bank Composition

Fig. 8 presents examples from the same scene under the scaled regime, together with bank allocation weights and predicted answers. Within this shared visual context, the room-size question selects only

the geometric bank, whereas appearance-order and object-counting questions combine semantic and temporal evidence. The two relative-distance questions select semantic and geometric banks with different allocation weights, illustrating how the composition varies across individual queries. The final question provides a failure case: the model predicts the mirror as the object closest to the closet, while the reference answer is the bed.