Li Auto | Base Model

Dual-Pathway Geometry-Aware MLLM for Spatial Intelligence

Yufei Zheng[1,2] **Xuhan Zhu**[2] **Zide Liu**[2] **Chunpeng Zhou**[2] **Chenfeng Wang**[1,2] **Yongchao Xu**[1]
Yunnan Wang[3] **Jiawei Liu**[1,†] **Pengfei Yu**[2,†] **Wei Zhai**[1,†] **Yang Cao**[1]
Zheng-Jun Zha[1]

[1]University of Science and Technology of China [2]Li Auto Inc. [3]Shanghai Jiao Tong University

Spatial understanding of the physical world from 2D visual inputs hinges on two complementary forms of geometric knowledge: holistic 3D structural perception and fine-grained metric scale estimation. Existing multimodal large language models (MLLMs) typically address only one facet, ingesting either depth maps or point clouds as additional model inputs, which incurs substantial computational overhead and inherits the generalization limitations of upstream prediction models. We propose GAMSI, a dual-pathway Geometry-Aware MLLM for Spatial Intelligence that takes only RGB images as input while internalizing both forms of geometric prior within a unified autoregressive backbone. Specifically, we introduce Metric-Structure Decoupled Queries (MSDQ) which employ two groups of learnable queries to respectively extract dense metric signals and sparse structural cues from the shared visual context, with a task-decoupled attention mask further preventing the two pathways from contaminating each other. Building on this, an Expert-Guided Visual Grounding (EVG) module projects the aggregated cues back to frame-level visual features and aligns them with vision foundation models, which serve purely as training-time supervision, rather than as model inputs. We further build a multi-task spatial instruction-tuning dataset (MTS) comprising 152,776 samples spanning 13 task types and three visual modalities, consolidated from six public datasets. Trained with a two-stage curriculum, GAMSI achieves state-of-the-art performance on seven spatial intelligence benchmarks.

Date: May 25, 2026
Correspondence: jwliu6@ustc.edu.cn, yupengfei1@lixiang.com, wzhai056@ustc.edu.cn
Project Page: https://github.com/zzzyf01/GAMSI

1 Introduction

Spatial understanding [1, 2, 3, 4] refers to the ability to perceive 3D structure, infer metric scale, and reason about geometric relationships from 2D observations. It underpins a wide range of downstream applications, including embodied manipulation [5, 6], autonomous navigation [7], and AR/VR interaction [8]. In these settings, a system must not only recognize objects but also quantify their distances, sizes, and spatial arrangements, often from limited visual input. Recently, Multimodal Large Language Models (MLLMs) [9, 10, 11, 12, 13, 14] have emerged as the default interface for such tasks, unifying perception, language, and reasoning within a single autoregressive backbone queried via natural language. Whether MLLMs genuinely possess such spatial cognition, rather than merely producing vague, ungrounded spatial descriptions, has become a central research question in the field.

†Corresponding author.

Preprint.

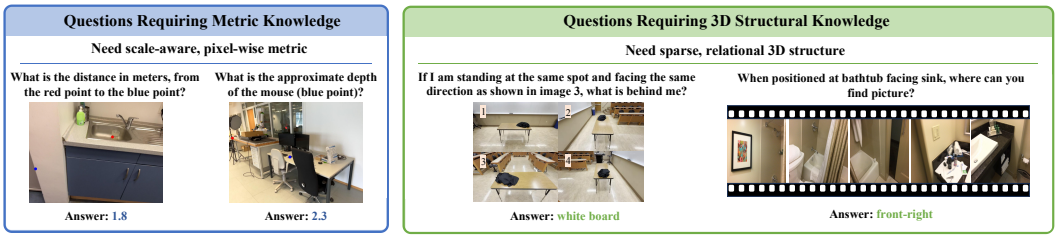

Figure 1: Two complementary categories of spatial questions. Metric questions require scale-aware, pixel-wise cues, whereas 3D structural questions require sparse, relational cues.

Despite the encouraging progress made by existing studies, endowing MLLMs with reliable spatial cognition remains fundamentally challenging. A closer examination of real-world spatial questions reveals that they essentially fall into two complementary categories [2, 15], each demanding a distinct form of geometric knowledge that purely 2D semantic features alone are insufficient to provide, as illustrated in Fig. 1. (i) A large class of spatial questions concerns absolute physical quantities, such as asking the distance in meters between two points or the depth of a specific object. Answering them relies on scale-aware, pixel-wise metric cues, as produced by monocular depth estimators [16, 17], where every pixel is anchored to a concrete physical unit. Without such a dense, unit-bearing signal, MLLMs can offer at best qualitative guesses, since RGB features alone suffer from inherent scale ambiguity. (ii) An equally important class involves relational, viewpoint-dependent geometric information, such as inferring what lies behind the camera or locating an object from a specified viewpoint. Answering them depends on sparse, relational 3D structural cues, as recovered by multi-view 3D reconstruction models such as VGGT [18], which encode inter-frame correspondences, camera motion, and global scene layout in a point-cloud-like feature space. Such structural information differs fundamentally from pixel-wise depth: it is relational rather than dense, and viewpoint-aware rather than unit-bound.

However, existing approaches typically address only one facet of this picture. The first family of methods [15, 19, 20] feeds predicted depth maps into the model as an additional input modality; while such methods improve the perception of distance and size, they offer little benefit for relational or viewpoint-dependent questions. The second family [2, 21] instead exploits 3D structural priors, such as explicit point clouds or features extracted from 3D reconstruction foundation models; these methods strengthen layout and viewpoint understanding, yet lack the absolute scale information required for metric tasks. To date, very few works have sought to jointly leverage these two complementary sources of geometric knowledge. Moreover, the vast majority of existing methods [2, 15] rely on auxiliary prior models to generate additional information such as depth maps or 3D point clouds, and feed them into the model as part of its input. This not only amplifies dependence on the generalization capability of external models, but also substantially increases the inference-time overhead.

To address the above limitations, we propose GAMSI, a dual-pathway Geometry-Aware MLLM framework for Spatial Intelligence. Unlike prior approaches, GAMSI operates on raw RGB images alone at inference, requiring no depth maps, point clouds, or camera parameters, and thus removes any inference-time dependence on auxiliary inputs or external models. Its core idea is to route two heterogeneous types of geometric knowledge through two decoupled pathways and internalize these priors into the MLLM's own semantic space during training, enabling the model to spontaneously invoke the corresponding geometric cognition from images alone at inference. To this end, we introduce two sets of learnable query embeddings dedicated to metric depth and 3D structure, respectively. Since a naive causal mask would allow the two query groups to leak into each other via shared self-attention, our Metric-Structure Decoupled Queries (MSDQ) adopt a task-decoupled attention mask under which each group attends only to the image tokens while remaining mutually invisible, cleanly disentangling the dense metric signal from the sparse structural cue at the representation level. Furthermore, since the queried information resides in the LLM's abstract semantic space and is inherently misaligned with the geometric feature space required for precise spatial perception, we further introduce an Expert-guided Visual Grounding (EVG) module, in which the visual queries project the aggregated depth and 3D cues into a frame-level feature space supervised by pretrained vision foundation models during training only, leaving inference entirely free of any external model.

To progressively inject and activate the dual-path geometric priors of GAMSI, we adopt a two-stage training paradigm that advances from low-level perceptual alignment to high-level task-oriented

intelligence. In the first stage, we leverage the SenseNova-SI-800K dataset [22] to align the MSDQ and EVG modules via metric-depth and 3D-structure supervision. However, SenseNova-SI is perception-centric by design: its samples are heavily concentrated on a handful of foundational tasks such as general visual understanding, metric measurement, and perspective-taking, while higher-order categories such as spatial imagination, object counting, and appearance-order recall are extremely scarce, leaving its instruction-format coverage too narrow on its own. To bridge this gap, in the second stage we construct MTS (Multi-Task Spatial), an instruction-tuning dataset consolidated from six complementary public sources [3, 20, 23, 24, 25, 26] to replenish the high-order task samples missing from SenseNova-SI. MTS contains 152,776 instances across 13 curated task types, organized into metric and structural spatial perception and spanning both simple and complex tasks to jointly supervise the two branches of GAMSI. It is further balanced across single-image, multi-image, and video modalities to prevent overfitting to any particular input form, and will be publicly released to support future research on general-purpose spatial intelligence.

Our contributions are summarized as follows: (1) We propose GAMSI, a dual-pathway geometry-aware MLLM that internalizes metric-depth and 3D structural priors into a single autoregressive backbone, enabling reliable spatial cognition from RGB images alone. (2) We introduce two sets of learnable queries with a task-decoupled attention mask that disentangles metric and structural signals, preventing cross-contamination under shared self-attention. (3) We design an EVG module in which visual queries project the depth and 3D cues aggregated within the LLM back to the visual space, supervised by vision foundation models. (4) We build MTS, a multi-task spatial intelligence dataset of 152,776 samples covering 13 tasks and three modalities. Extensive experiments show that GAMSI achieves state-of-the-art performance across diverse spatial intelligence benchmarks.

2 Related Work

Visual Spatial Intelligence. Enhancing the spatial understanding capability of MLLMs has attracted increasing attention. Extensive benchmark studies [1, 3, 27] reveal that, despite notable progress on general visual tasks, existing models still exhibit clear deficiencies in spatial understanding. Prior efforts can be broadly grouped into two categories. The first feeds predicted depth maps as an auxiliary input [15, 19, 20], which improves the perception of distance and size but offers little benefit for relational or viewpoint-dependent questions. The second exploits 3D structural priors from reconstruction foundation models, as in 3DSR [28], 3DThinker [29], and Spatial-MLLM [2], strengthening layout and viewpoint understanding but lacking the absolute scale needed for metric tasks. Very few works jointly leverage these two complementary cues, leaving the modeling of geometric layout and depth scale underexplored. Moreover, most existing methods [2, 15] rely on auxiliary prior models to produce depth maps or point clouds as inputs, which amplifies dependence on external models and increases inference-time overhead. In contrast, we propose a spatial intelligence architecture that takes only images and text as input, requiring no auxiliary 3D data, and adaptively extracts both 3D structural features and metric depth features from visual inputs to jointly support spatial perception. With this design, our model achieves a unified understanding of geometric layout and depth scale in a purely RGB-based, annotation-efficient paradigm.

Vision Foundation Model. Vision foundation models (VFMs) have become a cornerstone of modern visual understanding, producing transferable representations that adapt to a wide range of downstream tasks [20, 30]. On the semantic side, contrastive vision-language models such as CLIP [31] and ALIGN [32] align images with natural language at scale and endow visual encoders with open-vocabulary priors; self-supervised models such as DINO [33] and DINOv2 [34] learn dense features with strong semantic information and part-level structure; promptable segmentation models such as SAM [35] and SAM 2 [36] further provide class-agnostic mask priors that can be flexibly plugged into perception pipelines. In parallel, a growing body of geometry-oriented foundation models extends this paradigm to 3D: monocular depth predictors such as ZoeDepth [37], and Depth Anything [16, 17] achieve zero-shot depth estimation by training on massive heterogeneous data, while feed-forward reconstruction models such as MASt3R [38], and VGGT [18] unify pose estimation, dense matching, and point cloud regression within a single network, turning multi-view geometry into a learnable prior. Building on this line of research, our work integrates the capabilities of these VFMs in monocular depth estimation and 3D point cloud reconstruction into a unified large model, which directly extracts holistic spatial layout and metric depth information from raw images and uses them as structured priors to support spatial intelligence and spatial understanding tasks.

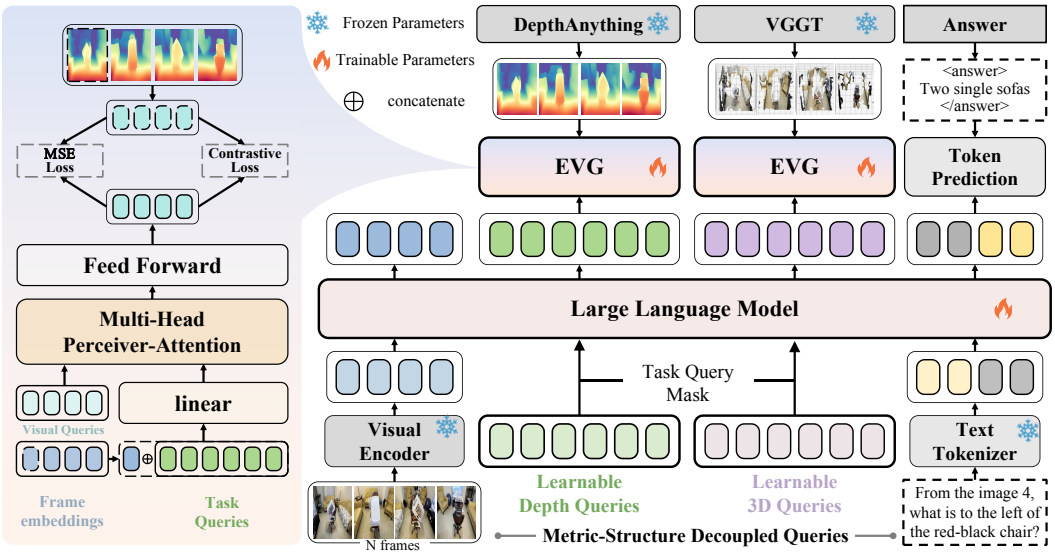

Figure 2: The overall architecture of the proposed GAMSI. It consists of two key modules: (1) Metric-Structure Decoupled Queries (MSDQ): routes two sets of learnable queries into independent pathways via a task-decoupled attention mask, separately encoding metric-depth and 3D structural cues. (2) Expert-Guided Visual Grounding (EVG): anchors each pathway to the feature space of a complementary visual expert (Depth Anything and VGGT), injecting task-specific priors for reliable spatial understanding.

3 Methodology

As illustrated in Figure 2, GAMSI jointly learns 3D structural and metric-depth representations from image–text inputs through two tightly coupled designs. The first is Metric-Structure Decoupled Queries (MSDQ), a dual-pathway query mechanism that employs task-decoupled attention to independently encode geometry-level structural cues and scale-aware metric-depth cues, while a shared self-attention backbone preserves cross-pathway interaction. The second is Expert-Guided Visual Grounding (EVG), which anchors each pathway's queries in the feature spaces of complementary visual experts, namely a 3D reconstruction model for structural grounding and a metric depth estimator for scale grounding, thereby injecting pathway-specific priors. Built upon a general MLLM backbone, GAMSI is trained in two stages: spatial perception alignment, followed by multi-task spatial understanding fine-tuning.

3.1 Metric-Structure Decoupled Queries

To extract two complementary forms of spatial representations directly from image–text inputs, we introduce two sets of task-specific learnable query embeddings into the LLM input sequence. Formally, given a scene $\mathcal{S} = \{I^i\}_{i=1}^N$ consisting of N frames with $I^i \in \mathbb{R}^{H\times W\times 3}$, a textual question $\mathcal{Q}$, and its corresponding answer $\mathcal{A}$, the vision encoder produces per-frame patch embeddings $F_v = \{f_v^i\}_{i=1}^N$, where $f_v^i \in \mathbb{R}^{P\times C}$. On the text side, $\mathcal{Q}$ is first tokenized into a sequence of token ids and then mapped through the token embedding layer to obtain $F_{t_q} \in \mathbb{R}^{L_q\times C}$, while $\mathcal{A}$ is tokenized into the id sequence $T_{\mathcal{A}} \in \mathbb{R}^{L_a}$, which serves as the supervision target.

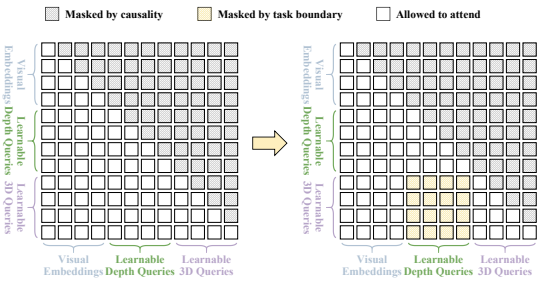

Figure 3: Comparison between the original causal mask (left) and our task-decoupled mask (right), which additionally blocks cross-task attention between Depth and 3D queries (yellow).

We then introduce two sets of task-specific learnable queries, metric depth queries $Q_m \in \mathbb{R}^{K\times C}$ and 3D structural queries $Q_s \in \mathbb{R}^{K\times C}$, each consisting of K vectors dedicated to capturing metric-depth

cues and 3D structural relations, respectively. The two query groups are placed after the visual tokens to form the full LLM input sequence $[F_v,\ Q_m,\ Q_s,\ F_{t_q}]$, so that both query groups can causally attend to the shared visual context. Whether F_{t_q} appears before or after the queries does not affect the design; we place it after the queries solely for notational simplicity. Since both query groups participate in the same causal self-attention at every transformer layer while pursuing heterogeneous learning objectives, unconstrained attention from Q_s to Q_m would contaminate the structural pathway with metric-depth signals. To prevent such leakage, we introduce a task-decoupled attention mask, as illustrated in Figure 3. Let $[S_m, E_m]$ and $[S_s, E_s]$ denote the index ranges of the metric and structural queries in the input sequence. We modify the standard causal mask $\mathbf{M}_{\text{causal}}$ by setting:

$$\mathbf{M}_{\text{causal}}[i, j] = -\infty, \quad \forall\, i \in [S_s, E_s],\ j \in [S_m, E_m]$$

Note that causal masking inherently prevents Q_m from attending to Q_s; the above constraint further blocks the reverse direction, so that the two query streams are fully decoupled from each other while both still fully attend to the shared visual embeddings, thereby routing metric-depth and 3D structural cues into two independent pathways. After forward propagation through the LLM, the last hidden states at the query positions are extracted as $\hat{Q}_m, \hat{Q}_s \in \mathbb{R}^{K\times C}$, and the last hidden states at the visual token positions of the i-th frame are extracted as the LLM-refined patch embeddings $\hat{f}_v^i \in \mathbb{R}^{P\times C}$. All of them are forwarded to the Expert-Guided Visual Grounding (EVG) module.

3.2 Expert-Guided Visual Grounding

Although $\hat{Q}_m$ and $\hat{Q}_s$ carry task-relevant context aggregated by the LLM, there is no guarantee that they encode the specific spatial information required for structural or metric perception, so explicit grounding in expert visual knowledge is necessary to activate these capabilities. Moreover, the LLM's abstract semantic space is fundamentally misaligned with the geometric feature spaces demanded by precise spatial perception. To bridge this gap, the Expert-Guided Visual Grounding (EVG) module anchors each query output in frame-level visual features and supervises it with representations from a corresponding pretrained vision foundation model. Since a single linear projection is insufficient for such cross-domain alignment, we adopt a Perceiver-style attention mechanism that first fuses the query output with patch features and then distills task-relevant spatial information through a set of learnable visual queries.

We describe the metric pathway as representative; the structural pathway is processed symmetrically. Given the LLM-refined patch embeddings of the i-th frame $\hat{f}_v^i \in \mathbb{R}^{P\times C}$ and the LLM-refined metric queries $\hat{Q}_m$, we first obtain a query-conditioned visual feature via a linear projection:

$$f_m^i = \text{Linear}\Big([\hat{f}_v^i; \hat{Q}_m]\Big), \tag{1}$$

where $[\cdot\,;\cdot]$ denotes concatenation along the token dimension. We then apply a perceiver attention block [39] followed by an output projection:

$$f_{vq}^i = \text{Linear}(\text{PerceiverAttn}\left(Q_v, [f_m^i; Q_v]\right)), \tag{2}$$

where $Q_v \in \mathbb{R}^{K_v\times C}$ is a set of learnable visual queries. The aligned representation f_{vq}^i is then supervised against the expert feature f_{gt}^i extracted by a pretrained vision foundation model. We combine an MSE regression loss for point-wise proximity and an InfoNCE contrastive loss [40] for discriminative alignment:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(\left\|f_{vq}^i - f_{\text{gt}}^i\right\|_2^2) \tag{3}$$

$$\mathcal{L}_{\text{CL}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}(f_{vq}^i, f_{\text{gt}}^i)/\tau\right)}{\sum_{f_{\text{gt}}\in\mathcal{B}}\exp\left(\text{sim}(f_{vq}^i, f_{\text{gt}})/\tau\right)}, \tag{4}$$

where $\mathcal{B}$ denotes the set of expert features from all frames within the current training batch, $\text{sim}(\cdot,\cdot)$ is cosine similarity on the flattened representations, and τ is a learnable temperature. The total alignment loss is:

$$\mathcal{L}_{\text{Align}} = \mathcal{L}_{\text{MSE}} + \lambda\mathcal{L}_{\text{CL}} \tag{5}$$

where λ balances the two objectives, and we set it to 0.01 based on magnitude considerations. The same procedure is applied symmetrically to $\hat{Q}_s$ under supervision from the structural expert.

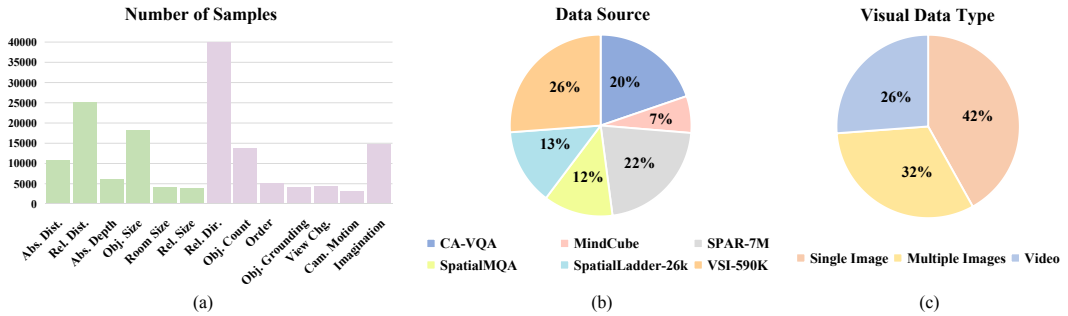

Figure 4: Composition of the MTS Dataset. (a) Distribution of Samples across Spatial Task Types; (b) Proportion of samples contributed by each source dataset; (c) Proportion of samples across three visual modalities (single image, multiple images, and video).

3.3 Two-Stage Training Strategy

We adopt a two-stage training strategy to progressively equip GAMSI with spatial understanding capabilities. Stage 1 instills foundational 3D structural and metric-depth perception, while Stage 2 generalizes these priors to diverse downstream spatial intelligence tasks. Both stages share an identical optimization objective, which we formalize at the end of this subsection for clarity.

Stage 1: Spatial Perception Alignment. In the first stage, we train GAMSI on the open-source SenseNova-SI-800K dataset [22] to instill foundational 3D structural and metric-depth perception capabilities. Through the visual alignment supervision imposed by the EVG module, this stage establishes the geometric priors that are essential for subsequent task-conditioned intelligence.

Stage 2: Multi-task Spatial Understanding Fine-tuning. Building upon the foundational 3D structural and metric-depth perception acquired in Stage 1, the second stage aims to generalize GAMSI's capabilities across a broad spectrum of downstream spatial intelligence tasks. To this end, we construct a large-scale multi-task instruction-tuning corpus, termed the MTS (Multi-Task Spatial) Dataset, by consolidating and re-organizing samples from six publicly available spatial intelligence datasets: CA-VQA [20], MindCube [23], SPAR-7M [24], SpatialMQA [25], SpatialLadder-26k [3], and VSI-590K [26]. As summarized in Figure 4, the MTS Dataset contains 152,776 instruction-following samples covering 13 spatial task types, and spans three visual modalities in comparable proportions: single image (42%), multiple images (32%), and video (26%). To further structure the task space, we organize the 13 task types into two complementary groups: (i) *Metric Perception* (absolute/relative distance, absolute depth, object- and room-level absolute size, relative size); (ii) *Structural Spatial Perception* (relative direction, object count, appearance order, object grounding, view change inference, camera motion inference, spatial imagination). Together, these two groups cover the spectrum from low-level geometric measurement to high-level structural and viewpoint-aware perception. The MTS Dataset will be publicly released to facilitate future research on unified spatial intelligence.

Unified Training Objective. Both training stages share an identical optimization objective, which couples the visual alignment supervision from the EVG module with the standard autoregressive language modeling loss over the answer tokens. The language modeling loss maximizes the likelihood of each answer token conditioned on all preceding context:

$$\mathcal{L}_{\text{LM}} = -\sum_{t=1}^{L_a} \log p\left(T_{\mathcal{A},t} \,\middle|\, T_{\mathcal{A},<t},\, F_v,\, Q_s,\, Q_m,\, F_{t_q}\right), \tag{6}$$

where L_a is the answer length, $T_{\mathcal{A},t}$ denotes the t-th answer token, and $T_{\mathcal{A},<t}$ represents all preceding answer tokens. The overall training objective used in both stages is:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \mathcal{L}_{\text{Align}}, \tag{7}$$

where $\mathcal{L}_{\text{Align}}$ is the visual alignment loss defined in the EVG. The two stages differ only in the training dataset employed: SenseNova-SI for Stage 1 and the MTS Dataset for Stage 2. The loss formulation and optimization pipeline remain consistent across stages, enabling a smooth transfer of geometric priors from spatial perception learning to multi-task fine-tuning.

Table 1: Comparison with state-of-the-art methods on seven spatial intelligence benchmarks. All benchmark scores and the macro-average are reported in percentage (%). "SI Dataset" denotes the size of the spatial-intelligence training set used by each method, "Avg." is the macro-average across all benchmarks, and MindCube* refers to the MindCube-Tiny. **Bold** and underline indicate the best and second-best result per column. Subscripts $S1$ and $S2$ denote Stage-1 and Stage-2, respectively.

Method	SI Dataset	Avg.	VSI-Bench	MindCube*	ViewSpatial	CV-Bench	SPBench-SI	SPBench-MV	SparBench
Open-source General Models									
Qwen2.5-VL-3B-Instruct [13]	-	40.3	27.5	42.5	36.2	63.0	40.9	41.8	30.0
Bagel-7B-MoT [41]	-	40.7	32.5	34.2	41.3	56.7	38.1	43.1	39.1
InternVL3-2B [14]	-	40.9	33.5	36.8	32.7	68.1	41.2	46.7	27.2
Qwen2.5-VL-7B-Instruct [13]	-	43.4	33.3	34.2	37.4	75.3	47.4	42.7	33.8
Qwen3-VL-2B-Instruct [13]	-	46.8	48.5	35.7	36.4	66.3	51.6	55.9	33.3
InternVL3-8B [14]	-	47.5	43.3	41.8	38.6	68.4	51.5	52.9	35.8
Qwen3-VL-8B-Instruct [13]	-	52.7	56.9	29.9	42.2	82.4	51.7	65.7	39.9
Open-source Spatial Intelligence Models									
SpatialLadder-3B [3]	260K	54.5	45.1	43.5	44.8	73.6	70.0	71.7	32.9
Spatial-MLLM-4B [2]	135K	43.8	46.2	33.5	34.6	54.7	45.3	56.7	35.3
SpaceR-7B [42]	151K	46.8	40.9	32.1	38.4	74.9	47.9	56.2	37.3
ViLaSR-7B [43]	81K	43.3	44.9	35.2	35.7	43.5	48.9	57.5	37.3
VST-3B-SFT [9]	4M	54.6	51.7	36.0	52.9	84.6	53.7	65.0	38.1
VST-7B-SFT [9]	4M	57.0	55.5	39.1	50.7	85.5	52.2	68.5	47.2
Cambrian-S-3B [26]	590K	51.0	55.9	39.0	40.9	75.2	57.7	55.4	32.9
Cambrian-S-7B [26]	590K	52.3	63.0	38.1	41.3	76.9	54.6	54.4	38.0
SenseNova-SI-Qwen3-VL-8B [22]	8M	63.4	62.9	73.8	51.2	83.4	60.4	71.1	41.1
SenseNova-SI-InternVL3-8B [22]	800K	64.0	60.6	66.7	55.1	84.0	62.5	70.3	48.9
Ours									
GAMSI$_{S1}$	800K	67.4	66.0	75.1	59.6	83.9	60.6	76.9	49.6
GAMSI$_{S1+S2}$	950K	**75.8**	**68.5**	**94.1**	**63.0**	**85.7**	**79.1**	**80.4**	**59.9**

4 Experiments

4.1 Experimental Settings

Implementation Details. We build all our experiments upon Qwen3-VL-8B-Instruct [13], and adopt a two-stage supervised fine-tuning pipeline. We set the number of metric and structural queries to $K = 40$ each. Stage 1 trains on SenseNova-SI-800K [22] to jointly aligns the multimodal architecture and injects general spatial-intelligence priors into the model; this stage is optimized for up to 3 epochs with a batch size of 64 and a learning rate of 8×10^{-6}. In the second stage, we curate the MTS Dataset by combining CA-VQA [20], MindCube [23], SPAR-7M [24], SpatialMQA [25], SpatialLadder-26K [3], and VSI-590K [26], and continue fine-tuning for 2 epochs with a batch size of 64 on 8 NVIDIA H200 GPUs and a learning rate of 4×10^{-6}. For both stages, we employ the AdamW optimizer.

Evaluation Benchmarks. We evaluate GAMSI on seven widely-used spatial-intelligence benchmarks using the EASI toolkit [44], namely VSI-Bench [1] (5,155 video questions), MindCube-Tiny [23] (1,040 multi-image questions on viewpoint transformation and spatial imagination), ViewSpatial-Bench [45] (5,712 questions for perspective-dependent reasoning), CV-Bench [27] (2,638 single-image questions on 2D/3D vision tasks), SPBench-SI/SPBench-MV [3] (1,009 single-image and 319 multi-image questions), and SPAR-Bench [24] (7,211 questions for multi-difficulty spatial reasoning).

4.2 Main Results

Table 1 compares GAMSI with open-source general-purpose MLLMs [13, 14, 41] and spatial-intelligence-specialized models [2, 3, 9, 22, 26, 42, 43] on seven benchmarks covering single-image, multi-image, and video inputs. Overall, GAMSI$_{S1+S2}$ achieves the best result on every benchmark and improves the macro-average from 64.0% (the strongest prior model, SenseNova-SI-InternVL3-8B) to 75.8%, an absolute improvement of 11.8%. The largest margins over the best prior result on each benchmark are observed on benchmarks dominated by relational and viewpoint-dependent reasoning ($+20.3\%$ on MindCube-Tiny over SenseNova-SI-Qwen3-VL-8B, $+11.0\%$ on SPAR-Bench over SenseNova-SI-InternVL3-8B, and $+9.1\%$ on SPBench-SI over SpatialLadder-3B), corresponding to the regime in which 2D semantic features alone are insufficient and in which the proposed 3D structural pathway is most effective. The advantage of GAMSI is further corroborated from two complementary, partially controlled angles. Under the same SenseNova-SI-800K training data, GAMSI$_{S1}$ outperforms SenseNova-SI-InternVL3-8B on most benchmarks and improves the macro-average from 64.0% to 67.4% ($+3.4\%$), demonstrating that, when the data scale is held fixed, the

Table 2: Ablation study of the dual-pathway design on a 10K balanced subset of SenseNova-SI-800K. All benchmark scores and the macro-average are reported in percentage (%). "Q_s" and "Q_m" denote the 3D structural and metric-depth query pathways. "Mask" indicates the task-decoupled attention mask in MSDQ. MindCube* refers to the MindCube-Tiny subset. **Bold** marks the best result.

Q_s	Q_m	Mask	**Avg.**	**VSI-Bench**	**MindCube***	**ViewSpatial**	**CV-Bench**	**SPBench-SI**	**SPBench-MV**	**SparBench**
-	-	-	54.1	55.0	28.6	42.3	86.3	50.8	68.9	46.9
✓	-	-	55.0	56.4	35.7	43.8	81.6	52.3	68.6	46.5
✓	✓	-	55.1	56.9	33.8	43.0	83.2	52.8	68.9	46.8
✓	✓	✓	**55.7**	56.4	35.7	42.1	86.3	52.3	69.9	46.9

proposed metric-structure decoupled pathways and EVG grounding yield gains that purely data-driven instruction tuning cannot achieve. Under the identical Qwen3-VL-8B backbone, GAMSI$_{SI}$ further surpasses SenseNova-SI-Qwen3-VL-8B by 4.0% on the macro-average (67.4% vs. 63.4%) while consuming roughly $10\times$ less spatial data (800K vs. 8M), with the largest gains again on structure- and viewpoint-heavy tasks ($+8.4\%$ on ViewSpatial, $+8.5\%$ on SPAR-Bench, $+5.8\%$ on SPBench-MV, and $+3.1\%$ on VSI-Bench), indicating that injecting an appropriate architectural prior is a substantially more sample-efficient route to spatial competence than scaling training data alone. Building on this foundation, Stage-2 multi-task fine-tuning on the MTS Dataset further improves the macro-average from 67.4% to 75.8% ($+8.4\%$), with the largest gains concentrated on benchmarks well aligned with our 13 task types ($+19.0\%$ on MindCube-Tiny, $+18.5\%$ on SPBench-SI, and $+10.3\%$ on SPAR-Bench), confirming that the foundational metric and structural priors instilled in Stage-1 are not benchmark-specific and transfer effectively across task formulations and input modalities once exposed to a diverse, well-organized multi-task instruction dataset.

4.3 Ablation Experiment

To dissect the contribution of each component in GAMSI, we conduct controlled ablations on a balanced 10K subset sampled from SenseNova-SI-800K. All variants share the Qwen3-VL-8B backbone and identical training hyper-parameters, and only the components introduced by MSDQ and EVG are toggled. Results on the seven spatial-intelligence benchmarks are reported in Table 2.

Effect of dual-pathway geometric priors. Adding the 3D structural pathway alone improves the macro-average from 54.1% to 55.0% ($+0.9\%$), with the largest gains concentrated on benchmarks that explicitly target viewpoint transformation and perspective-dependent reasoning ($+7.1\%$ on MindCube-Tiny and $+1.5\%$ on ViewSpatial), together with a moderate gain on VSI-Bench ($+1.4\%$). This pattern confirms that VGGT-grounded structural cues address the regime in which 2D semantic features alone are insufficient. However, this single-pathway variant regresses on CV-Bench (from 86.3% to 81.6%), a benchmark in which 3D depth and 3D distance subtasks account for a substantial portion of the questions, indicating that structural priors alone are inadequate, and may even be detrimental, for absolute scale perception. Activating the metric-depth pathway under task-decoupled mask isolation on top further improves the macro-average to 55.7% and restores CV-Bench to 86.3% ($+4.7\%$ over the structural-only variant), while largely preserving the structural gains, with only a marginal drop on ViewSpatial. This confirms that structural and metric priors address disjoint subsets of spatial questions, and their joint internalization yields balanced spatial cognition.

Effect of the task-decoupled attention mask. Comparing the two dual-pathway variants, removing the mask reduces the macro-average from 55.7% to 55.1% (-0.6%), and produces consistent regressions across most benchmarks (-3.1% on CV-Bench, -1.9% on MindCube-Tiny, and -1.0% on SPBench-MV). This indicates that, under unconstrained causal self-attention, signals from Q_m leak into Q_s and contaminate each pathway with the other's objective, eroding the gains from injecting two heterogeneous priors. The task-decoupled mask separates the two streams at the representation level, restoring the full benefit of dual-pathway routing without introducing additional parameters.

4.4 Qualitative Visualization

To verify that the two pathways indeed extract the geometric information they are designed for, we visualize the predictions of Q_m and Q_s on four indoor scenes in Figure 5, where the *Depth Target* and *VGGT Target* are produced by the EVG experts as supervision. The depth maps decoded from Q_m closely match the depth target: distant regions such as windows are assigned cool tones

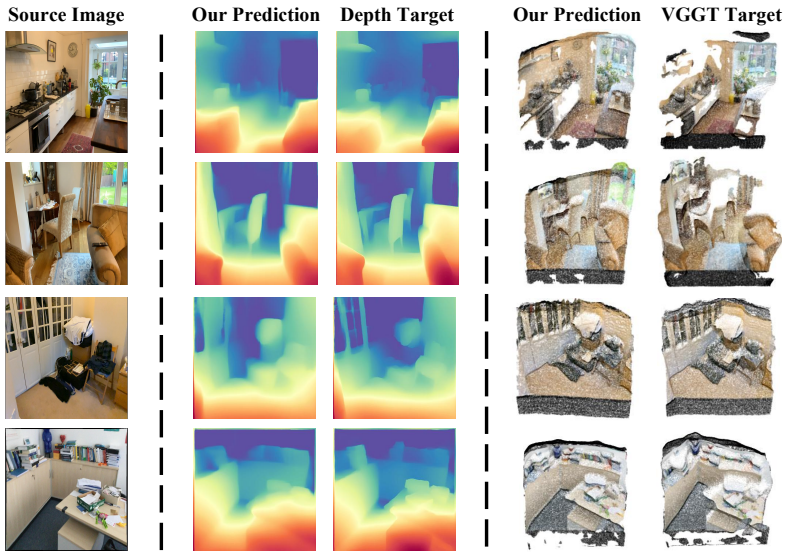

Figure 5: Qualitative comparison of the dual-pathway predictions (Q_m depth and Q_s point cloud) against their respective depth and VGGT targets on four scenes.

while nearby furniture is rendered in warm tones, with no noticeable blurring or scale drift and only minor, barely perceptible differences from the target. In parallel, the point clouds decoded from Q_s faithfully reproduce both the global scene layout and the camera viewpoint of the VGGT target, including furniture arrangement, wall and floor orientation, and finer details such as plants and books. Moreover, the two pathways behave as complementary rather than redundant channels: Q_m does not produce sparse structural artifacts, and Q_s does not encode dense pixel-wise depth, indicating that the task-decoupled attention mask successfully isolates the two streams. Together, these observations confirm that Q_m carries metric depth and Q_s carries 3D structural information, validating the dual-pathway design at the representation level.

5 Conclusion

We presented GAMSI, a dual-pathway geometry-aware MLLM that internalizes two complementary forms of geometric knowledge, namely fine-grained metric depth and holistic 3D structural perception, into a unified autoregressive backbone, while taking only RGB images as input at inference time. The Metric-Structure Decoupled Queries (MSDQ) employ two groups of learnable queries together with a task-decoupled attention mask to cleanly disentangle the metric and structural pathways at the representation level, and the Expert-Guided Visual Grounding (EVG) module aligns the queried cues with vision foundation models as training-time-only supervision, eliminating any inference-time dependence on auxiliary inputs. To foster multi-task tuning for higher-order spatial intelligence, we further construct MTS, a 152,776-sample dataset covering 13 task types and three visual modalities. Trained with a two-stage curriculum, GAMSI achieves state-of-the-art results on seven spatial intelligence benchmarks, with ablations and visualizations confirming that the two pathways capture their intended metric and structural information.

6 Limitations

Despite the promising results, the proposed GAMSI has several limitations: (1) While GAMSI performs strongly on a wide range of spatial intelligence tasks, its behavior in more complex scenarios (e.g., long-horizon spatial planning), which jointly demand fine-grained geometric perception and high-level reasoning, warrants deeper investigation. (2) Constrained by computational resources, our experiments are conducted on an 8B-scale backbone; the scalability of GAMSI to larger models and its adaptability to diverse multimodal architectures remain to be explored.

References

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

[2] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025.

[3] Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.08531*, 2025.

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

[5] Zirui Song, Guangxian Ouyang, Mingzhe Li, Yuheng Ji, Chenxi Wang, Zixiang Xu, Zeyu Zhang, Xiaoqing Zhang, Qian Jiang, Fengxian Ji, et al. Maniplvm-r1: Reinforcement learning for reasoning in embodied manipulation with large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 18558–18566, 2026.

[6] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[7] Saeid Nahavandi, Roohallah Alizadehsani, Darius Nahavandi, Shady Mohamed, Navid Mohajer, Mohammad Rokonuzzaman, and Ibrahim Hossain. A comprehensive review on autonomous navigation. *ACM Computing Surveys*, 57(9):1–67, 2025.

[8] Xincheng Huang, Michael Yin, Ziyi Xia, and Robert Xiao. Virtualnexus: Enhancing 360-degree video ar/vr collaboration with environment cutouts and virtual replicas. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–12, 2024.

[9] Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025.

[10] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[11] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[12] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

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

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

[15] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9490–9498. IEEE, 2025.

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

[17] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

[18] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

[19] Yang Liu, Ming Ma, Xiaomin Yu, Pengxiang Ding, Han Zhao, Mingyang Sun, Siteng Huang, and Donglin Wang. Ssr: Enhancing depth perception in vision-language models via rationale-guided spatial reasoning. *arXiv preprint arXiv:2505.12448*, 2025.

[20] Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, et al. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7395–7408, 2025.

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

[22] Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, et al. Scaling spatial intelligence with multimodal foundation models. *arXiv preprint arXiv:2511.13719*, 2025.

[23] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025.

[24] Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *arXiv preprint arXiv:2503.22976*, 2025.

[25] Jingping Liu, Ziyan Liu, Zhedong Cen, Yan Zhou, Yinan Zou, Weiyan Zhang, Haiyun Jiang, and Tong Ruan. Can multimodal large language models understand spatial relations? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 620–632, 2025.

[26] Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis L Brown II, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. In *The Fourteenth International Conference on Learning Representations*, 2025.

[27] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

[28] Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. Mllms need 3d-aware representation supervision for scene understanding. *arXiv e-prints*, pages arXiv–2506, 2025.

[29] Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Xiang An, Yan Feng, Peng Pei, Xunliang Cai, et al. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. *arXiv preprint arXiv:2510.18632*, 2025.

[30] Yunnan Wang, Fan Lu, Kecheng Zheng, Ziyuan Huang, Ziqiang Li, Wenjun Zeng, and Xin Jin. Vision-centric activation and coordination for multimodal large language models. *arXiv preprint arXiv:2510.14349*, 2025.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[32] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[33] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

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

[35] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[36] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[37] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

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

[39] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

[40] Evgenia Rusak, Patrik Reizinger, Attila Juhos, Oliver Bringmann, Roland S Zimmermann, and Wieland Brendel. Infonce: Identifying the gap between theory and practice. *arXiv preprint arXiv:2407.00143*, 2024.

[41] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

[42] Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

[43] Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*, 2025.

[44] Zhongang Cai, Yubo Wang, Qingping Sun, Ruisi Wang, Chenyang Gu, Wanqi Yin, Zhiqian Lin, Zhitao Yang, Chen Wei, Oscar Qian, et al. Holistic evaluation of multimodal llms on spatial intelligence. *arXiv preprint arXiv:2508.13142*, 2025.

[45] Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025.

Table 3: Statistics of the MTS Dataset by task type. The dataset contains 152,776 instruction-following samples covering 13 spatial task types consolidated from six public spatial intelligence datasets.

Type	Data Source	Number
absolute distance	CA-VQA, SPAR-7M, SpatialLadder-26k	10,629
relative distance	CA-VQA, SPAR-7M, SpatialLadder-26k, VSI-590K	25,109
absolute depth	SPAR-7M	6,000
absolute size (object)	CA-VQA, SpatialLadder-26k, VSI-590K	18,229
absolute size (room)	SpatialLadder-26k, VSI-590K	3,959
relative size	CA-VQA	3,847
relative direction	CA-VQA, SPAR-7M, SpatialMQA, SpatialLadder-26k, VSI-590K	40,102
object count	CA-VQA, SpatialLadder-26k, VSI-590K	13,664
appearance order	SpatialLadder-26k, VSI-590K	5,126
object grounding	CA-VQA, SPAR-7M	4,111
view change inference	MindCube, SPAR-7M	4,302
camera motion inference	SPAR-7M	3,000
spatial imagination	MindCube, SPAR-7M	14,698
13 Types	CA-VQA, MindCube, SPAR-7M, SpatialMQA, SpatialLadder-26k, VSI-590K	152,776

A Details of the MTS Dataset

To support comprehensive spatial reasoning capabilities in our model, we construct the Multi-Task Spatial (MTS) instruction tuning dataset by consolidating and re-organizing samples from six publicly available spatial intelligence datasets, namely CA-VQA [20], MindCube [23], SPAR-7M [24], SpatialMQA [25], SpatialLadder-26K [3], and VSI-590K [26]. The resulting corpus contains 152,776 instruction-following samples spanning 13 fine-grained spatial task types. Table 3 reports the detailed composition of MTS, including the source dataset(s) and the number of samples associated with each task type.

A.1 Construction Pipeline

Since each source dataset is already a well-curated spatial intelligence training dataset, we only perform the following lightweight pre-processing steps on each of them to obtain a unified instruction-following format:

Task mapping. We manually inspect the original category annotations of each source dataset. For datasets without explicit category labels, we perform keyword-based matching according to the characteristics of the questions to assign a category. We then select samples belonging to our 13 target task types from the categorized pool.

Format unification. All samples are converted into a consistent dictionary format consisting of the system-user dialogue (`messages`), the task type (`type`), and the associated `images`.

A.2 Data Source Summary

The six source datasets contribute complementary aspects to MTS:

CA-VQA provides a broad mixture of metric and qualitative spatial questions, contributing to 7 of the 13 task types.

SPAR-7M offers large-scale pixel-level and viewpoint-level annotations, covering 8 task types including depth, grounding, view change, camera motion, and spatial imagination.

SpatialLadder-26k supplies hierarchical spatial questions ranging from object counting to absolute distance estimation.

VSI-590K focuses on video-based spatial intelligence and significantly enlarges the pool of relative direction, object count, and room-level size samples.

SpatialMQA contributes relative direction questions.

MindCube provides mental rotation and viewpoint-transformation samples that are crucial for the spatial imagination and view change inference tasks.

B Detailed Experimental Settings

B.1 Details of Visual Expert model

In our implementation, we instantiate the metric and structural experts with Depth Anything V2 [17] and VGGT [18], respectively.

B.2 Details of Evaluation Benchmarks

VSI-Bench. [1] is a comprehensive evaluation benchmark designed to assess visual-spatial intelligence in Multimodal Large Language Models (MLLMs) through egocentric video understanding. The benchmark comprises 5,155 video question-answer pairs derived from 288 real-world videos. The videos span diverse environments across multiple geographic regions, providing broad coverage for evaluating spatial understanding in realistic scenarios. By leveraging continuous egocentric video streams rather than isolated images, VSI-Bench targets the spatial integration abilities required for embodied scene understanding.

MindCube-Tiny [23] is a multi-image evaluation benchmark comprising 1,040 question-answer pairs focused on viewpoint transformation and spatial imagination. It assesses models' abilities to mentally manipulate spatial configurations across different viewpoints, requiring integrated reasoning over multiple images to infer spatial relationships not directly observable from any single view. MindCube-Tiny is organized into three settings of increasing complexity: (1) Rotation, where a stationary camera rotates in place to capture 2 to 4 orthogonal views of a central foreground object; (2) Occlusion, which leverages occlusion phenomena to require object permanence under partial visibility, transformation of lateral (left-right) relationships into depth (front-back) relationships across views, and multi-view integration for coherent 3D understanding; and (3) Among, where the camera rotates around a central object positioned between it and several surrounding objects, capturing four orthogonal views that each place the central object in front of one surrounding object.

ViewSpatial-Bench [45] is a comprehensive evaluation framework comprising 5,712 question-answer pairs across more than 1,000 3D scenes. This benchmark evaluates VLMs' spatial localization capabilities from both egocentric and allocentric viewpoints, addressing the critical gap in perspective-dependent reasoning abilities that are essential for embodied interaction and multi-agent collaboration. By jointly probing first-person and third-person perspectives, ViewSpatial-Bench provides a systematic testbed for evaluating whether models can reconcile spatial references across different observer frames.

CV-Bench [27] addresses the limitations of existing vision-centric benchmarks through 2,638 manually-inspected single-image examples. The benchmark repurposes established vision datasets to evaluate MLLMs on fundamental 2D and 3D computer vision tasks. The evaluation encompasses 2D spatial comprehension through spatial relationships and object counting, while 3D understanding is assessed via depth ordering and relative distance estimation. By repurposing well-curated vision datasets, CV-Bench ensures high annotation reliability and offers a principled reference point for measuring foundational visual-spatial competence.

SPBench-SI & SPBench-MV [3] are evaluation benchmarks constructed using the SpatialLadder-26k pipeline applied to the ScanNet validation set. Both benchmarks undergo rigorous quality control through standard pipeline filtering strategies supplemented by manual curation to ensure data disambiguation and high-quality annotations. SPBench-SI serves as a single-image evaluation benchmark designed to assess models' spatial understanding and reasoning capabilities from individual viewpoints, encompassing four task categories—absolute distance, object size, relative distance, and relative direction—with a total of 1,009 samples. SPBench-MV constitutes a multi-image evaluation benchmark that requires models to perform joint spatial modeling across multiple viewpoints, with a total of 319 samples. In addition to the tasks included in SPBench-SI, SPBench-MV incorporates object counting tasks to evaluate models' capabilities in identifying and enumerating objects within multi-view scenarios.

SPAR-Bench [24] constitutes a comprehensive evaluation framework for systematically assessing spatial perception and reasoning capabilities in Vision-Language Models (VLMs) across multiple difficulty levels. The benchmark encompasses 20 diverse spatial understanding tasks spanning single-view, multi-view, and video-based modalities, incorporating 7,211 manually verified question-answer pairs to ensure annotation quality and reliability. By covering tasks of varying granularity and input

modality, SPAR-Bench enables fine-grained diagnosis of model strengths and weaknesses across the full spectrum of spatial reasoning challenges encountered in real-world visual understanding.

B.3 Ablation on the Number of Queries

To examine the effect of the query count K in MSDQ, we vary $K \in \{8, 16, 24, 32, 40, 48, 56\}$ for both metric and structural query groups while keeping all other hyper-parameters fixed. Following the protocol of the main ablation (Table 2), all variants are trained on the same 10K balanced subset of SenseNova-SI-800K. As reported in Table 4, the macro-average rises from 54.3% at $K \leq 16$ to a peak of 55.7% at $K = 40$, then plateaus at 55.2% for $K \geq 48$ with no further benefit, indicating that further enlarging K adds computational overhead without measurable gain. We therefore adopt $K = 40$ as the default in all main experiments.

Table 4: Ablation on the number of learnable queries K per group in MSDQ, evaluated on the same 10K balanced subset of SenseNova-SI-800K used in Table 2. All scores are reported in percentage (%). **Bold** marks the best result per column.

K	Avg.	VSI-Bench	MindCube*	ViewSpatial	CV-Bench	SPBench-SI	SPBench-MV	SparBench
8	54.3	56.0	31.6	42.2	82.5	52.1	69.4	46.3
16	54.3	56.6	32.2	43.1	82.4	52.9	65.9	46.7
24	54.1	56.5	32.9	43.1	81.4	51.5	67.0	46.6
32	55.0	56.0	35.2	42.6	81.8	52.3	70.2	46.8
40	**55.7**	56.4	35.7	42.1	86.3	52.3	69.9	46.9
48	55.2	56.6	35.0	43.1	83.2	52.7	68.9	47.1
56	55.2	56.4	38.0	43.4	82.9	51.9	67.4	46.7

C Broader Impacts

Positive Societal Impacts. Our work advances the spatial cognition of multimodal large language models, enabling reliable reasoning about 3D structure and metric scale directly from RGB images. By removing the inference-time dependence on auxiliary depth maps, point clouds, or camera parameters, GAMSI lowers the hardware and sensor requirements for deploying spatially intelligent systems, which could democratize access to such capabilities in resource-constrained settings and benefit applications such as assistive technologies for visually impaired users, safer robotic navigation, and AR/VR interaction.

Negative Societal Impacts. Despite these benefits, GAMSI produces quantitative spatial estimates (e.g., distances and sizes) in natural language, and such outputs may appear authoritative even when incorrect, particularly in out-of-distribution scenes; over-reliance in safety-critical pipelines such as autonomous driving or robotic manipulation could lead to harmful decisions. Moreover, stronger spatial cognition from raw RGB inputs lowers the effort required to infer 3D scene layouts from casually captured images or publicly shared videos, raising privacy concerns regarding the potential reconstruction of private environments without consent. To mitigate these risks, we recommend that deployments include calibrated uncertainty estimation and human oversight in high-stakes decision-making.