This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2026.3737728

Canvas Beyond Words: Teaching Machines to Sketch Empowers Their Spatial Intelligence

Guanlin Wu, Boyan Su, Yang Zhao, Pu Wang, Yichen Lin and Hao Frank Yang†

Abstract—How to encode and evaluate spatial intelligence in foundation models remains an open challenge. Existing approaches often rely on textual proxies and VQA-style evaluation to assess visual-spatial intelligence (VSI), which can obscure geometric structure, encourage linguistic shortcuts, and hinder attribution to genuinely spatial reasoning abilities. To address this limitation, we introduce **spatial intelligence grid (SIG)**, a structured grid-based schema that explicitly represents object layouts, inter-object relations, and physically grounded priors. As a complementary modality to text, SIG offers a faithful and compositional representation of scene structure for foundation-model reasoning. Building on SIG, we further propose SIG-informed evaluation metrics for measuring a model's intrinsic VSI, thereby disentangling spatial competence from language priors. In few-shot in-context learning (ICL) experiments with state-of-the-art multimodal large language models (MLLMs), including GPT- and Gemini-family models, SIG consistently delivers larger, more stable, and more comprehensive improvements across all VSI metrics than VQA-only representations. These results highlight its potential as an effective annotation and training schema for learning VSI. We further demonstrate the generalizability of SIG across diverse scenarios and multiple MLLMs, and systematically examine how data sample selection influences SIG-based ICL through extensive experiments. In addition, we release `SIGBench`, a benchmark comprising 1.4K driving frames annotated with ground-truth SIG labels and human gaze traces, supporting both grid-based machine VSI tasks and attention-driven, human-like VSI tasks in autonomous-driving settings.

Index Terms—Multimodal large language models, Visual spatial intelligence, Autonomous driving

1 Introduction

Currently, Visual Question Answering (VQA) has emerged as the de facto format for evaluating Visual-Spatial Intelligence (VSI), as it necessitates the interpretation of imagery, the understanding of object relationships, and the inference of context to generate accurate textual responses [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. Within the computer vision community, the consensus holds that VSI entails the accurate perception, manipulation, and reasoning of visual-spatial attributes, such as size, location, and geometric correlations, mediated through textual descriptions [12]. For instance, state-of-the-art (SOTA) studies often utilize specific inquiries to probe this understanding, asking, *"What is the distance between the keyboard and the TV? How many cabinets are in the room? What is the height of the stool?"* [8]. In this paradigm, the model's ability to generate correct answers is interpreted as evidence of effective VSI acquisition.

Nevertheless, the question remains: **are VQA tasks the optimal metric for spatial intelligence?** From a cognitive science perspective, Gardner and Lohman define VSI more broadly as *"the ability to generate, retain, retrieve, and transform well-structured visual images"* [13]. In contrast to other well-formulated vision tasks, the core challenge of VSI lies in extracting 3D geometric insights from 2D observations and articulating them via natural language [8], [14]. While VQA inherently intertwines linguistic proficiency with spatial reasoning, human cognitive processes suggest these abilities can function independently. For example, human can construct mental spatial maps using non-visual cues, such as auditory reverberation, without relying on text [15]. Consequently, a growing number of research argues that VQA's reliance on textual output may fail to fully capture the underlying spatial intelligence [16], [17], [18]. These limitations advocate for the integration of additional physical priors or alternative representations that more accurately mirror the spatial reasoning mechanisms inherent to human cognition.

Historical artistic practices offer a compelling antecedent for the utilization of spatial priors in visual perception. During the Renaissance, prominent masters such as Albrecht Dürer, Leonardo da Vinci, and Vincent van Gogh famously employed grid-based "drawing machines" to facilitate accurate spatial construction, a practice widely documented in art history [23], [24], [25], [26], [27]. By projecting visual observations onto structured grids (see Fig. 1), these artists were able to systematically decompose complex 3D scenes into discrete spatial regions, thereby isolating and recording geometric and graphical correlations. As illustrated in the middle of Fig. 1, this methodology suggests that artists progressed from organized graphical representations to high-fidelity depictions—a process that parallels the cognitive trajectory of VSI, which naturally evolves from a structured visual framework to rich semantic or linguistic interpretation. Consequently, once the mechanism for perceiving spatial relationships is established via these grids, the specific subject matter becomes secondary. Any scene, regardless of its semantic content, is ultimately reducible to a unique configuration of graph-theoretic elements (nodes, shapes, edges, and color values). Collectively, these visual graphs define the underlying appearance of diverse objects, ranging from plaster casts to floral arrangements or human portraits [28].

† *Corresponding Author*

- *Guanlin Wu, Boyan Su, Yang Zhao, Pu Wang and Yichen Lin are with the Department of Civil and Systems Engineering (System Track), Johns Hopkins University, Baltimore, MD, 21218, USA. E-mail: {gwu32, bsu11, yzhao229, pwang80, ylin190}@jhu.edu.*
- *Hao Frank Yang is with the Department of Civil and Systems Engineering, Data Science and AI Institute, Johns Hopkins University, Baltimore, MD, 21218, USA. E-mail: haofrankyang@jhu.edu.*

Authorized licensed use limited to: Johns Hopkins University. Downloaded on November 10,2026 at 15:10:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2026.3737728



Fig. 1: **Examples of human VSI in painting.** Abraham Bosse, a French artist and theorist illustrates a systematic, grid-based method for achieving visual-spatial correlations in rendering three-dimensional space on a two-dimensional canvas (left) [19], and he employed this grid-based visual scheme in portrait painting (middle) [20], [21]. Procedures of drawing a cast with graphical priors from scratch (right) [22].

Can such grid-based graphical priors serve as robust machine representations for VSI? Analogous to the disciplined perception of a classical painter, we posit that endowing machines with grid or graph-based abstractions provides a structured inductive bias that effectively captures spatial hierarchies and relationships. This approach simplifies the decomposition of complex scenes and facilitates stronger language grounding. We propose that an optimal VSI representation requires a structured conceptual model that explicitly encodes geometric and topological constraints, achievable through grid or graph-based priors. Within this framework, visual primitives (*e.g.* shapes and edges) are mapped onto discrete spatial partitions and interconnected to preserve both local details and global context. By systematically organizing visual data into these relational structures, our proposed representation supports robust spatial reasoning and efficient manipulation, establishing a coherent pathway from low-level visual analysis to high-level semantic description.

Building upon these insights, we propose a novel VSI representation format, called **spatial intelligence grid (SIG)**, and then conduct foundational testing on existing Multimodal Large Language Models (MLLMs) in one of the most critical VSI application tasks: autonomous driving (AD) [29], [30], [31]. The AD domain necessitates real-time perception, the precise modeling of multi-agent dynamics, and high-fidelity situational awareness across heterogeneous environments. Consequently, the capacity to efficiently localize, track, and correlate dynamic elements, such as vehicles, pedestrians, and traffic infrastructure, is paramount for operational safety. Given its inherent complexity, high-dimensional variability, and safety-critical nature, AD serves as an optimal testbed for evaluating and refining SIG-based VSI frameworks. To this end, we present a novel benchmark, `SIGBench`, and systematically investigate three fundamental research questions: *1) SIG-Based VSI Assessment and Metric Design*: Asses whether current MLLMs can resolve spatial queries when grounded in SIG representations, and we propose a suite of novel evaluation metrics, derived from graph similarity theory, to quantify the fidelity of SIG-based reasoning. *2) SIG-based In-Context Learning (ICL) for VQA*: Investigate the extent to which SIG-informed features can augment spatial reasoning capabilities in VQA tasks, and further examine the impact of sample selection strategies on SIG-based ICL. *3) Human-Machine VSI Attention Analysis via SIG:* Examine the divergence between human VSI attention differs from machine-based approaches within SIG, and explore methods to achieve more human-like SIG in order to improve human–machine interactions and *4) Generalizability of SIG:* Evaluate the generalizability of SIG as a structured data schema for enhancing model VSI across diverse application domains and multiple MLLMs. In this research, our contributions can be summarized as follows:

- We introduce SIG, a novel grid-based representation for VSI, which integrates both explicit physical priors and human-inspired attention cues. We further develop a suite of structured, graph-based metrics for rigorous evaluation of VSI.
- We demonstrate SIG's efficacy as a data schema for enhancing VSI through few-shot ICL on MLLMs. Compared to ICL using VQA-style representation, SIG-based ICL yields larger, more stable and more comprehensive improvements in model's VSI.
- We investigate the performance of SIG-based few-shot ICL under different sample selection strategies and extend the analysis beyond AD scenarios, demonstrating the scalability and generalizability of SIG across tasks and domains.
- We release `SIGBench`, comprising 1.4K driving frames annotated with ground-truth SIG maps and corresponding human gaze data. This dataset supports comprehensive evaluation of grid-based machine VSI and human-aligned attention within our proposed evaluation metrics.

2 Related Work

General VSI in MLLMs. MLLMs have achieved remarkable success across diverse visual tasks, including object detection, semantic segmentation, and image captioning, by leveraging their unified vision–language representations and robust reasoning capabilities [37], [38], [39], [40], [41]. Building upon these foundations, recent research has adapted MLLM architectures and fine-tuning protocols to emphasize spatial relationships, enabling them to address visual–spatial reasoning questions such as *e.g. which object is at the left-most position*

Authorized licensed use limited to: Johns Hopkins University. Downloaded on November 10,2026 at 15:10:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2026.3737728

TABLE 1: **Comparison of existing VSI and AD benchmarks.** The Comp. indicates whether a benchmark provides comprehensive coverage. MC and AM stand for multiple-choice and attention-map evaluations, respectively. Size refers to the total number of words contained in the dataset. Ann. denotes original annotation, while Rep. indicates that the dataset has been repurposed which means assembled from previously published datasets. S.R.R. means spatial relation representation. The Text and Graph indicate whether the evaluation relies on textual formats (*e.g.*, multiple-choice questions or sentences) or on graph-based comparisons. Lastly, H-L signifies a human-like evaluation approach, whereas Auto means that the benchmark provides predefined metrics allowing for fully automated assessment.

Type	Benchmark	Benchmark Overview					Evaluation Metrics			
		Size	Source	#Tasks	Answer	S.R.R.	Text	Graph	H-L	A/G Eval.
Comp.	SEED [1]	1.682M	Ann.	12	MC	Text	✓	✗	✗	Auto
	MMBench [2]	0.096M	Rep.	20	MC	Text	✓	✗	✗	GPT
	MM-Vet [3]	0.003M	Rep.	-	MC	Text	✓	✗	✗	GPT
	MMMU [4]	0.324M	Ann.	30	MC/Open	Text	✓	✗	✗	Auto
VSI Only	VSR [6]	0.094M	Rep.	1	T/F	Text	✓	✗	✗	Auto
	GQA [32]	24.42M	Rep.	1	Open	Text	✓	✗	✗	Auto
	SRBench [7]	0.068M	Rep.	4	MC	Text	✓	✗	✗	Auto
Comp. AD	NuScenes-QA [33]	6.592M	Rep.	5	Open	Text	✓	✗	✗	Auto
	NuScenes-MQA [34]	43.43M	Rep.	4	Open	Text	✓	✗	✗	Auto
	NuPlanQA [35]	16.89M	Rep.	9	Open	Text	✓	✗	✗	Auto
VSI AD	NuScenes-SpatialQA [36]	78.75M	Rep.	2	MC/Open	Text	✓	✗	✗	Auto
	SIGBench	71.35M	Ann.	5	MC/SIG/AM	SIG	✓	✓	✓	Auto

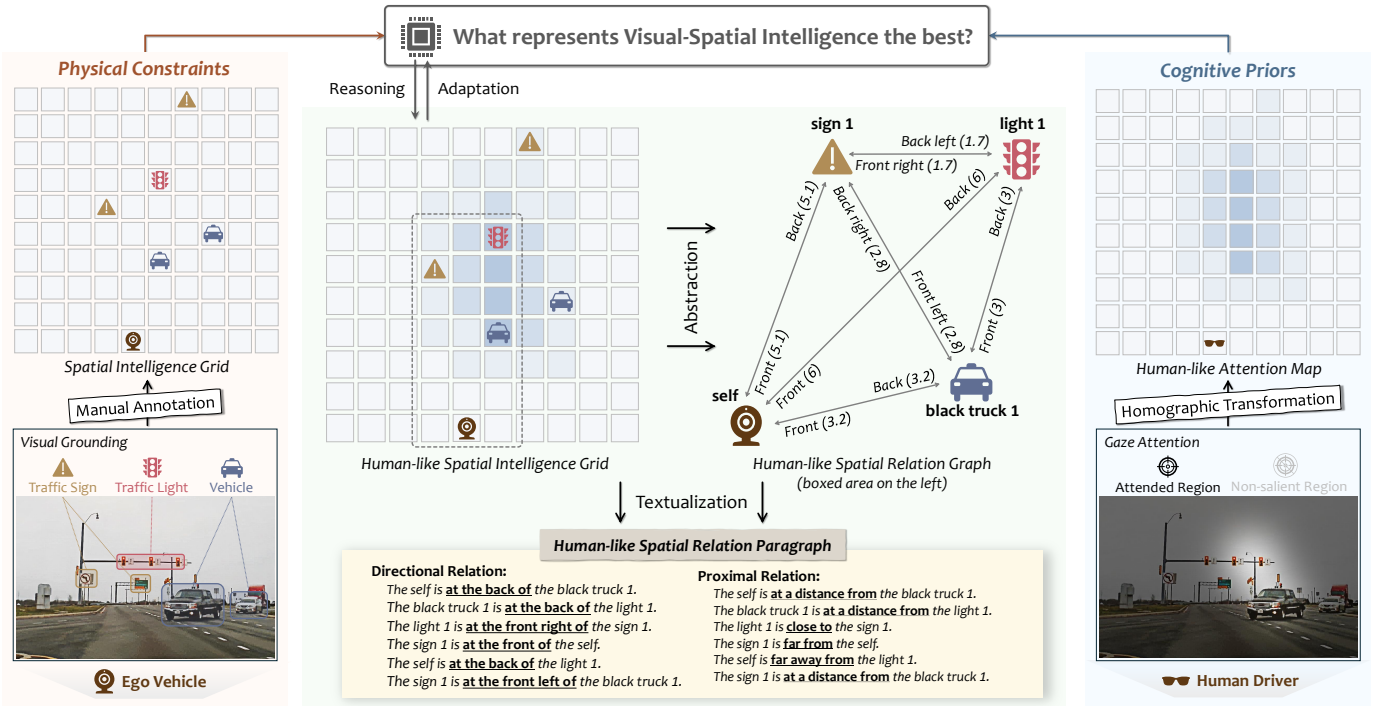


Fig. 2: **Overview of human-like SIG in AD scenario.** We use SIG to represent the spatial relation of traffic sign, traffic lights, vehicles and self (ego-vehicle) in the image shown in the left. In the right part, we apply homographic transformation to convert human gaze attention from image to SIG size. By integrating these two representations, we obtain the human-like SIG, from which the human-like SRG and SRP are derived and presented in the center. Objects of the same category are ordered based on their horizontal position in the image, for example, black truck 1 denotes the leftmost vehicle.

in this image? [42], [43], [44]. For example, SpatialVLM [43] introduces an automatic system for generating an Internet-scale 3D spatial VQA dataset and demonstrate that training VLMs on this data markedly improves both qualitative and quantitative spatial reasoning, while SpatialRGPT [44] enhances MLLM's visual–spatial reasoning by combining a 3D scene graph–driven regional representation pipeline with a flexible depth-integration plugin, enabling precise inference of relative directions and distances for user-specified regions. To rigorously assess general VSI, several large-scale benchmarks now include dedicated spatial-reasoning components within broader visual-task suites [1], [2], [3], [4], [5], [45], while others concentrate exclusively on spatial-reasoning in images [6], [7], [46] or videos [8], [9] using VQA-style formats. Although these benchmarks yield clear right/wrong judgments, they frequently conflate failures of basic visual grounding (e.g., missed object detections) with true shortcomings in spatial reasoning, especially in complex

Authorized licensed use limited to: Johns Hopkins University. Downloaded on November 10,2026 at 15:10:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2026.3737728

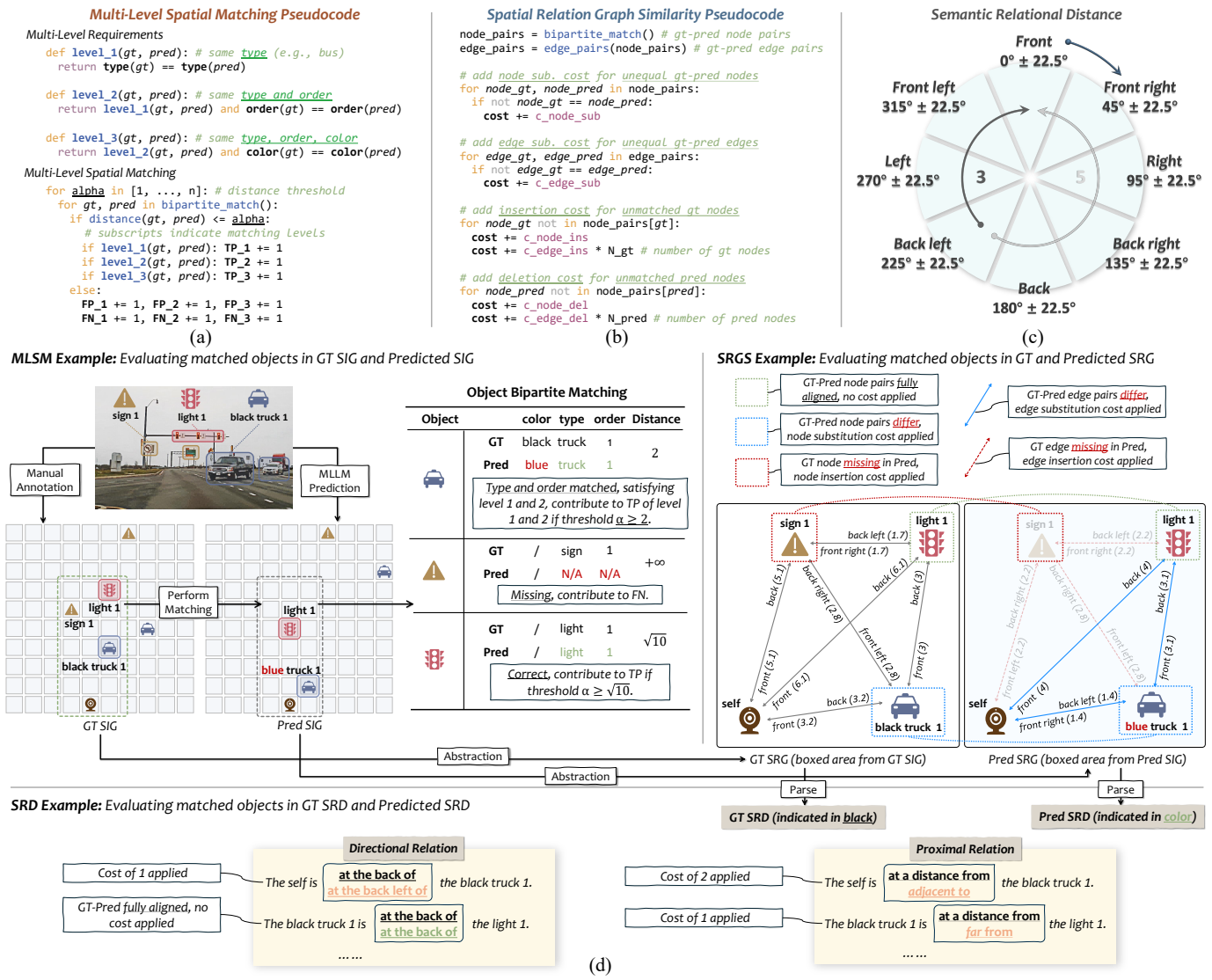


Fig. 3: **Detailed illustration of MLSM, SRGS and SRD.** (a) and (b) present the pseudocode for MLSM and SRGS, respectively. (c) illustrates how semantic relational distance is computed using the directional relation circle. Specifically, the semantic relational distance between two spatial prepositions is defined as the minimum number of steps along the circular ordering. For example, the distance from "at the back left of" to "at the front of" is 3 rather than 5. (d) provides an illustrative example of MLSM, SRGS, and SRD. All three metrics begin by aligning predicted SIG nodes with their ground-truth counterparts via bipartite matching. In the upper-left panel of (d), we show how true positives (TP), false positives (FP), and false negatives (FN) are computed for vehicles within the highlighted region for MLSM. In the upper-right panel, the computation of SRGS is illustrated by identifying the nodes and edges that must be inserted or substituted, together with the resulting total edit distance. The lower panel of (d) shows the example of how SRD is calculated.

scenes populated by numerous similar entities.

VSI with Scenario-based Physical Constraints. Real-world applications of VSI, whether in robotics [47], [48], human motion [49], [50] or AD [51], [52], must operate within strict domain-specific physical, environmental, and regulatory boundaries. In the context of AD, it aims to enable vehicles to perceive their surroundings, understand the spatial layout of the scene, and make safe driving decisions in real time [53]. A typical AD system relies on multiple sensors, such as cameras, LiDAR, radar, and maps, to support several core functions, including perception [54], planning [55], and control [56]. Among these, VSI is especially critical because the vehicle must not only recognize objects, but also reason about their 3D positions, motions, and relations to the road structure. To benchmark these capabilities, several VQA-style datasets [33], [34], [35], [36], [57] have been created based on large-scale, open-source AD collections such as nuScenes [58], Waymo Open Dataset [59], and nuPlan [60], etc. For instance, NuScenes-QA [33] compiles the first large-scale VQA benchmark for autonomous driving—comprising 34K street-view scenes and 460K question-answer pairs generated from 3D detection annotations and scene graphs, while NuScenes-SpatialQA [36] introduces the first large-scale, automated 3D scene-graph–driven QA benchmark built on the NuScenes dataset to systematically evaluate VLMs' spatial understanding and reasoning in driving scenarios. The detailed statistics of these datasets are summarized in Tab. 1. While these benchmarks provide valuable

Authorized licensed use limited to: Johns Hopkins University. Downloaded on November 10,2026 at 15:10:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2026.3737728

insight into text-image spatial reasoning under real-world constraints, their exclusive reliance on image-centric queries risks neglecting the richer, BEV representations that underpin modern AD perception and planning pipelines.

Human-Like VSI. Human scene understanding is guided by both overt gaze patterns (*where we look*) and covert cognitive maps (*how we represent spatial relations*). Gaze or saliency prediction models, which estimate pixel-level attention maps, have provided deep insights into visual prioritization in driving scenarios [61], [62], [63].Complementing gaze prediction, the notion of a cognitive map captures the internal spatial layout that humans use to reason about object relations [64] and has recently been integrated into MLLM frameworks for video understanding [8]. In our approach, we leverage an human-like SIG to emulate both gaze-driven saliency and structured visual-spatial representations, thereby enabling a human-like evaluation of visual-spatial reasoning for existing MLLMs, particularly within complex AD environments.

3 METHODS

3.1 Grid-based Visual-Spatial Intelligence

We introduce SIG, a grid-based representation format for VSI, as illustrated in Fig. 2, to encode road scenes in AD scenarios. Since AD scenes are inherently constrained by road geometry, our representation centers on the ego vehicle and nearby traffic-related entities, including other vehicles, traffic signs, traffic lights, and traffic lanes. Built upon SIG, we construct a directed spatial relation graph (SRG) to encode pairwise spatial relations using relative direction and grid distance, and a spatial relation paragraph (SRP) to express the same information in textual form. To assess VSI quantitatively, we propose three evaluation metrics: multi-level spatial matching (MLSM), spatial relation graph similarity (SRGS), and semantic relational distance (SRD). More specifically, MLSM measures absolute positional accuracy through direct comparison in the SIG domain. SRGS evaluates structural agreement between predicted and ground-truth (GT) SRGs by considering both node-level and edge-level correspondence. SRD measures the semantic discrepancy between the spatial expressions in predicted and GT SRPs, thereby evaluating how faithfully directional and proximity relations are conveyed in language. In order to emphasize fundamental spatial reasoning ability, the primary evaluation protocol considers only the ego vehicle, other vehicles, traffic lights, and traffic signs. Traffic lanes are excluded from the main metrics because lane evidence is frequently unreliable in real-world scenes, owing to worn, blurred, or absent markings. Nevertheless, we also provide supplementary results on a subset of scenes with clear lane annotations, where lane information is included in the evaluation. Prior to scoring, the predicted SIG is normalized with respect to the ground truth by matching the ego-vehicle location in the two representations. The same spatial offset is subsequently propagated to all remaining predicted objects.

Multi-level Spatial Matching. To evaluate SIG-based spatial representations independently of visual-grounding noise, we introduce a multi-threshold grid matching protocol inspired by tracking metrics such as MOTA [65] and HOTA [66]. MLSM first performs bipartite matching between predicted and GT objects (vehicles, traffic signs, traffic lights) using self-defined cost function c_v for vehicles and c_{sl} for traffic signs and traffic lights, which can be calculated by:

$$c_v = d * \omega_c * \omega_o * \omega_t, \quad c_{sl} = d * \omega_o \tag{1}$$

where $\omega_c, \omega_o, \omega_t$ are the weight for color, order and type matching, respectively. d is euclidean distance between objects' position on SIG, and each weight $\omega_c, \omega_o, \omega_t$ equals one when the attribute is unmatched and drops below one when it matches, thus granting larger spatial tolerances for objects that are matched in type, order, and/or color.

Once matching is performed, a pair is regarded as a true positive (TP) if its grid-cell centroids lie within a distance threshold α and all specified object attributes coincide. A GT instance without any matched prediction is counted as a false negative (FN), while any prediction lacking a corresponding ground-truth match is a false positive (FP). For vehicles, we apply three nested matching criteria: (1) identical type, (2) identical type + order, (3) identical type + order + color and one level for signs/lights (identical order). Over a set of n thresholds $\alpha \in [1, \ldots, n]$ based on euclidean distance, we can compute precision P_α, recall R_α, F1-score $\mathrm{F1}_\alpha$ and association accuracy AssA_α using:

$$\begin{aligned} \mathrm{P}_\alpha &= \frac{TP_\alpha}{TP_\alpha + FP_\alpha}, \quad \mathrm{R}_\alpha = \frac{TP_\alpha}{TP_\alpha + FN_\alpha}, \\ \mathrm{AssA}_\alpha &= \frac{TP_\alpha}{TP_\alpha + FP_{\alpha_k} + FN_{\alpha_k}}. \end{aligned} \tag{2}$$

Then, we normalize each of them through n thresholds to get overall $\mathrm{P}, \mathrm{R}, \mathrm{F1}$ and AssA, shown as:

$$\begin{aligned} \mathrm{P} &= \frac{1}{n}\sum_{\alpha=1}^{n} \mathrm{P}_\alpha, \quad \mathrm{R} = \frac{1}{n}\sum_{\alpha=1}^{n} \mathrm{R}_\alpha, \\ \mathrm{F1} &= \frac{2\mathrm{PR}}{\mathrm{P}+\mathrm{R}}, \quad \mathrm{AssA} = \frac{1}{n}\sum_{\alpha=1}^{n} \mathrm{AssA}_\alpha \end{aligned} \tag{3}$$

The pseudo-code and an example with detailed calculation of MLSM is illustrated in Fig. 3 (a) and (d) upper left.

Spatial Relation Graph Similarity. Unlike MLSM, SRGS directly evaluates the fidelity of individual relationships (*i.e.*, edges) between a predicted SRG and its GT counterpart. We operationalize this by computing the graph edit distance (GED) [67], which counts the minimal set of edit operations required to transform the predicted graph into the GT graph. Formally, let the directed SRG be denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_i\}_{i=1}^{n}$ represents the n object nodes (*e.g.* ego-vehicle, other vehicles, traffic signs, traffic lights) and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ comprises directed edges $e_{ij} = (v_i, v_j) \in \mathcal{E}$ encoding the spatial relation (direction + distance within the SIG) from node v_i to v_j. To compute SRGS, we first establish a set M that contains all matched pairs $(v_i, \hat{v}_{i'})$, where where $v_i \in \mathcal{V}$ is a GT node and $\hat{v}_{i'} \in \hat{\mathcal{V}}$ is its corresponding predicted node. We then decompose the edit distance into two components—node edit distance and edge edit distance. The node edit distance itself comprises three cost types:

- Substitution cost $\delta_{\mathrm{sub}}(v_i, \hat{v}_{i'})$: The cost of altering a predicted node $\hat{v}_{i'}$ so that its position and attributes match those of the GT node v_i.
- Deletion cost $\delta_{\mathrm{del}}(\hat{v}_{i'})$: The cost of removing a predicted node $\hat{v}_{i'}$ that has no corresponding GT match.

Authorized licensed use limited to: Johns Hopkins University. Downloaded on November 10,2026 at 15:10:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2026.3737728

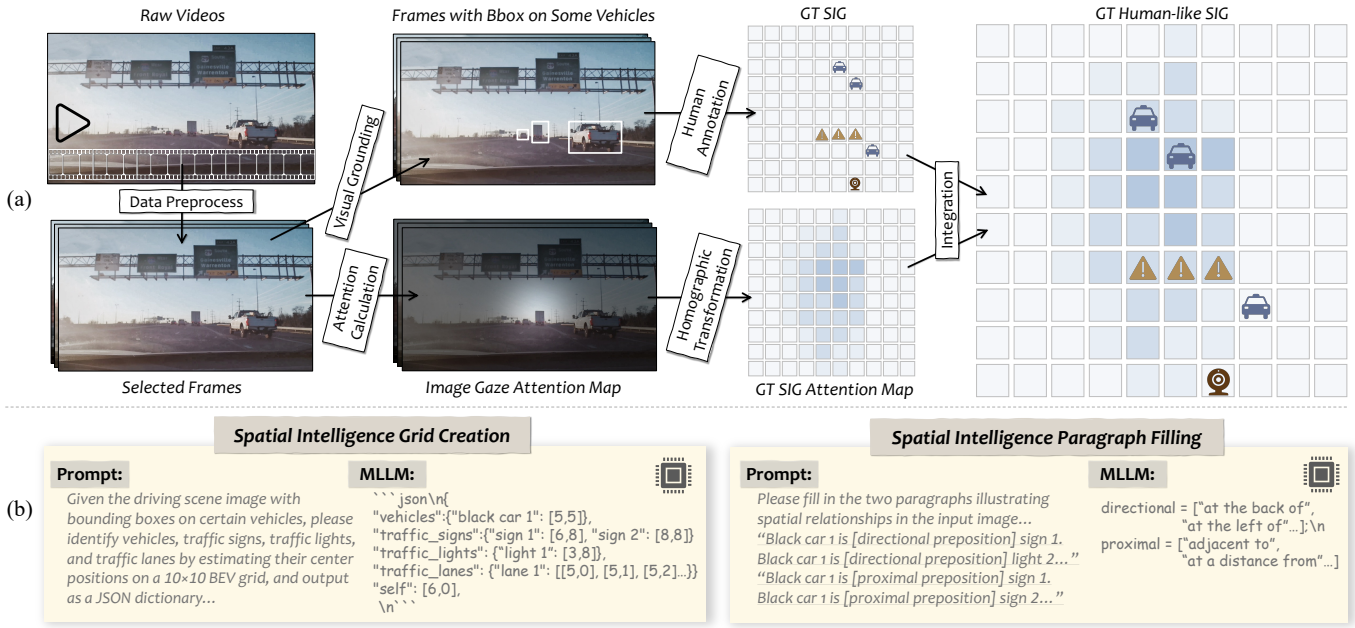


Fig. 4: The annotation pipeline of SIGBench is illustrated in (a). Starting from raw driving scenes, a subset of frames is selected for further process. For each scene, the GT SIG will be annotated by human and attention map in SIG size will be calculated from homographic transformation. Combining these, GT human-like SIG is produced. (b) illustrates the format of SIGC and SRPF tasks in SIGBench.

- Insertion cost $\delta_{\text{ins}}(v_j)$: The cost of adding an unmatched GT node v_j.

The detailed formulations of these cost functions $\delta_{\text{sub}}(v_i, \hat{v}_{i'}), \delta_{\text{del}}(\hat{v}_{i'})$ and $\delta_{\text{ins}}(v_j)$ are as follows:

$$\delta_{\text{sub}}(v_i, \hat{v}_{i'}) = \begin{cases} d^N_{v_i, v_{i'}} + \lambda_N \mathbb{I}\big[\text{attr}(v_i) \neq \text{attr}(\hat{v}_{i'})\big], & \\ & \text{if } (v_i, \hat{v}_{i'}) \in \mathcal{M}, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

$$\delta_{\text{del}}(\hat{v}_i) = \begin{cases} \eta^N_{\text{del}}, & \text{if } \hat{v}_i \notin \{\, v \mid (v, \hat{v}) \in \mathcal{M}\}, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

$$\delta_{\text{ins}}(v_j) = \begin{cases} \eta^N_{\text{ins}}, & \text{if } v_j \notin \{\, v \mid (v, \hat{v}) \in \mathcal{M}\}, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

where $\eta^N_{\text{del}}, \eta^N_{\text{ins}}$ are the cost for node insertion and deletion. The total node edit distance between the GT graph $\mathcal{G}$ (with node set $\mathcal{V}$) and the predicted graph $\hat{\mathcal{G}}$ (with node set $\hat{\mathcal{V}}$) is:

$$D_N(\mathcal{G}, \hat{\mathcal{G}}) = \sum_{(v_i, \hat{v}_{i'}) \in M} \delta_{\text{sub}}(v_i, \hat{v}_{i'}) + \sum_{v_i \in \mathcal{V}} \delta_{\text{del}}(v_i) + \sum_{\hat{v}_j \in \hat{\mathcal{V}}} \delta_{\text{ins}}(\hat{v}_j) \tag{7}$$

We then calculate the edge edit distance $D_E(\mathcal{G}, \hat{\mathcal{G}})$ similarly (details in Appendix A). Thus, combining D_N and D_E, we can calculate weighted total graph edit distance D_{total} and weighted graph similarity score S by:

$$D_{\text{total}} = \gamma\, D_N(\mathcal{G}, \hat{\mathcal{G}}) + \beta\, D_E(\mathcal{G}, \hat{\mathcal{G}}),$$
$$S = \max\left(0,\ 1 - \frac{D_{\text{total}}}{D_{\max}}\right) \in [0, 1] \tag{8}$$

where γ, β are weights and $D_{\max}$ denotes worst-case distance (all nodes and edges unmatched). The pseudo-code and an example with detailed calculation of SRGS is illustrated in Fig. 3 (b) and (d) upper right.

Semantic Relational Distance. To evaluate the precision of predicted prepositions in the SRP, we first map each preposition onto a discrete ordinal scale and then compute the minimal separation from the GT using either cyclic or linear distance measures. Directional prepositions are arranged on a circular topology, as shown in Fig. 3 (c), while proximal prepositions follow a linear ordering from nearest to farthest—"adjacent to," "close to," "at a distance from," "far from," and "far away from." For instance, the semantic distance between "adjacent to" and "far from" is 3 on this scale. An example is illustrated in Fig. 3 (d) lower. We assess performance using accuracy, mean absolute error (MAE), and mean squared error (MSE), which are calculated directly from the corresponding semantic distances as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\tilde{d}_i| \quad MSE = \frac{1}{n}\sum_{i=1}^{n} |\tilde{d}_i|^2. \tag{9}$$

Theoretical Runtime Analysis. In real-world application, some scenarios, such as AD, require real-time decisions. We provide the theoretical runtime analysis of our proposed evaluation metrics: MLSM, SRGS and SRD through time complexity and empirical runtime analysis in Sec. 5.2.

- **MLSM**: The time complexity for building cost matrix for vehicles, traffic signs and traffic lights is $O(n^2)$. According to [68], the time complexity for bipartite matching is $O(n^3)$. For multi-threshold matching, the time complexity is $O(n)$. Thus, the overall time for MLSM is $T_{\text{MLSM}}(n) = (n^2) + O(n^3) + O(n) \to O(n^3)$.
- **SRGS**: During the fully connected graph construction, we require $O(n)$ for all nodes and $O(n^2)$ for all edges. Similarly, the bipartite matching requires $O(n^3)$. For

Authorized licensed use limited to: Johns Hopkins University. Downloaded on November 10,2026 at 15:10:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2026.3737728

calculating GED, we are performing node-to-node and edge-to-edge comparison, which requires $O(n) + O(n^2)$. Thus, the time complexity for SRGS is $T_{\text{SRGS}} = 2(O(n) + O(n^2)) + O(n^3) \rightarrow O(n^3)$.

- **SRD**: We are computing pointwise distance. For SRD (Directional) and SRD (Proximal) there are 8 and 5 possible relations, respectively. Thus, the time complexity for SRD is $T_{\text{SRD}} = O(8n) + O(5n) \rightarrow O(n)$.

3.2 SIG-based In-context Learning for VSI

Whereas prior approaches enhance MLLMs' spatial reasoning by encoding depth maps, bounding boxes, or segmentation masks into textual QA formats [42], [43], [44], we explore the efficacy of using SIG directly within ICL paradigm. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denotes a dataset of N image-SIG paris. The ICL process for outputting answer y_q of query x_q can be formulated as:

$$y_q = \mathcal{F}_M(x_q; \mathcal{P}) \tag{10}$$

where $\mathcal{F}_M$ is the MLLM and example prompt $\mathcal{P} = \text{concat}(x_1, y_1, \ldots, x_k, y_k)$. In our experimental setup, we randomly select k image-SIG examples. Each x_i comprises the input image annotated with vehicle bounding boxes alongside a brief task description, while each y_i consists of the corresponding GT SIG (serialized as a JSON-like structure, *e.g.* `vehicles:{black truck 1:[5,3]}; traffic_signs:{sign 1:[3,5]}...`) and SRP derived from GT SIG. Our key insight is that, by exposing the model to paired visual inputs and explicit SIG outputs, the MLLM will internalize the mapping between image-based object locations and their structured grid representation. We validate the strong generalization capability of SIG in Sec. 5.2.

3.3 Human-Like VSI with Grid

In AD scenario, human decision-making relies on selective attention: drivers naturally concentrate on a limited set of scene elements, such as red traffic light or a pedestrian stepping off the curb, when making safety-critical decisions, rather than processing every object. To emulate this behavior, we incorporation the SIG representation with human gaze or saliency predictions [61], [69], which ideally represents the human attention range. To transform the human attention range form image size to SIG size, we use homographic transformation. Let each image pixel be represented by $p_i = (u_i, v_i, 1)^\top$ and its corresponding SIG grid coordinate by $w_i = (X_i, Y_i, 1)^\top$. We then determine the homographic matrix $H = [h_{ij}] \in \mathbb{R}^{3\times3}$ by solving the linear system of X_i and Y_i defined by the point correspondences as:

$$X_i = \frac{h_{11}u_i + h_{12}v_i + h_{13}}{h_{31}u_i + h_{32}v_i + h_{33}}, \quad Y_i = \frac{h_{21}u_i + h_{22}v_i + h_{23}}{h_{31}u_i + h_{32}v_i + h_{33}} \tag{11}$$

using singular value decomposition (SVD). We then project and normalize the raw attention map A_{Image} with image size into the SIG attention map A_{SIG} using

$$A_{\text{SIG}}(i,j) = \frac{A_{\text{Image}}(\lfloor x_{ij} \rfloor, \lfloor y_{ij} \rfloor) - \min_{u,v} A_{\text{Image}}(u,v)}{\max_{u,v} A_{\text{Image}}(u,v) - \min_{u,v} A_{\text{Image}}(u,v)}$$
$$\text{s.t. } [x_{ij}, y_{ij}, 1]^\top = H^{-1}[i,j,1]^\top \tag{12}$$

Thus, we can combine A_{SIG} and SIG to create SIG with attention weight (*a.k.a.* human-like SIG).

Human-Like Spatial Relation Graph Similarity. To incorporate human-like perceptual priorities into the SRGS metric, we modulate each node-edit cost by its attention weight derived from A_{SIG}. Likewise, each edge-edit cost is scaled by the mean attention weight of its two end-nodes. The resulting attention-aware SRGS therefore assigns heavier penalties to edit operations involving objects that receive greater human attention.

Human-Like Semantic Relational Distance. To apply human gaze attention as a weight for the calculation of human-like SRD, we weight each semantic distance by the average attention assigned to its two objects. Concretely, let the GT relation be *black car 1 is {at the back left of* white car 2*}* and the model's prediction be is *{at the front of}*. If black car 1 and white car 2 receive attention weights of 3 and 4, respectively, their semantic directional distance (Fig. 3 (c)) is 3. The resulting human-like SRD is therefore $3 \times \frac{3+4}{2} = 10.5$.

4 SIGBENCH DATASET

4.1 Overview

We introduce `SIGBench`, a benchmark designed to evaluate both grid-based and human-like VSI in MLLMs within AD scenario. `SIGBench` comprises 1,423 frames, each is annotated with (i) SIG and human-like SIG, (ii) SRP and human-like SRP and (iii) a corresponding gaze attention map in image size. The annotation pipeline is depicted as Fig. 4 (a). For each selected frame, we estimate a circular attention radius r based on:

$$r = min(r_w, r_h)$$
$$s.t.\ \ r_w = \frac{f * I_w}{s_w} * \tan(\frac{\text{fov}_w}{2}),\ r_h = \frac{f * I_h}{s_h} * \tan(\frac{\text{fov}_h}{2}) \tag{13}$$

where f denotes the camera focal length, I_w, I_h are image width and height in pixels, s_w, s_h are the corresponding sensor dimensions in millimeters, and fov_w and fov_h are the human horizontal and vertical field of view. The GT attention map is obtained by temporally aggregating attention maps over six consecutive frames. In the SIG annotation process, vehicles located within the bounding box are labeled using "color+type+order" (*e.g.* black car 1), where indices are assigned according to each vehicle's horizontal position from left to right in the image. Traffic signs and signal lights are labeled only when they are clearly visible, using the format "type+order" (*e.g.* sign 1, light 1) Further details of the annotation workflow are provided in Appendix B. `SIGBench` is organized into two principal components: (i) grid-based VSI (Sec. 4.2), containing tasks such as spatial intelligence grid creation (SIGC) and spatial relation paragraph filling (SRPF), and (ii) human-like VSI (Sec. 4.3), containing human-like SIGC and SRPF, and gaze prediction.

4.2 Grid-based Visual-Spatial Intelligence Tasks

Spatial Intelligence Grid Creation. In this task, the MLLM is tasked with generating a 10×10 SIG in which every vehicle with bbox, traffic sign, traffic light, and the self (ego-vehicle)

Authorized licensed use limited to: Johns Hopkins University. Downloaded on November 10,2026 at 15:10:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2026.3737728

TABLE 2: **Quantitative comparison of different MLLMs on SIGBench dataset for grid-based VSI tasks.** P and R are precision and recall, respectively. F1 is F1-score and AssA is association accuracy. S is similarity ($\gamma, \beta = 1$) and WS is weighted similarity ($\gamma \neq \beta$) in Eq. (8). Acc means accuracy. Dark blue and light blue mark the best and the second best result among all models.

Type	Models	MLSM				SRGS		SRD (Directional)			SRD (Proximal)		
		P↑	R↑	F1↑	AssA↑	S↑	WS↑	MAE↓	MSE↓	Acc↑	MAE↓	MSE↓	Acc↑
	Human-Level	0.918	0.932	0.938	0.869	0.897	0.901	0.568	2.372	0.753	0.720	2.508	0.760
Proprietary	Claude-3.5-Haiku	0.415	0.371	0.373	0.243	0.225	0.203	2.179	6.247	0.103	0.701	0.960	0.416
	Claude-3.7-Sonnet	0.511	0.427	0.450	0.306	0.299	0.299	2.285	6.712	0.092	0.691	0.928	0.420
	Gemini-1.5-Pro	0.471	0.470	0.454	0.309	0.282	0.272	2.073	5.899	0.126	1.241	2.720	0.298
	Gemini-2.0-Flash	0.472	0.472	0.456	0.311	0.283	0.270	2.344	6.923	0.083	1.092	1.835	0.246
	Gemini-2.5-Pro	0.473	0.586	0.507	0.355	0.232	0.226	1.316	2.937	0.254	0.790	1.225	0.398
	GPT-4o-mini	0.167	0.156	0.154	0.087	0.054	0.019	2.478	7.451	0.061	0.922	1.442	0.312
	GPT-4o	0.507	0.441	0.458	0.315	0.337	0.331	1.942	5.309	0.144	0.949	1.563	0.313
Open-source	InternVL3-9B	0.290	0.302	0.284	0.174	0.133	0.086	2.019	5.804	0.130	0.942	1.524	0.314
	InternVL3-14B	0.432	0.374	0.391	0.258	0.254	0.234	2.305	6.850	0.070	1.139	1.957	0.272
	InternVL2.5-26B	0.386	0.383	0.369	0.239	0.220	0.189	2.006	5.752	0.135	0.940	1.470	0.310
	Qwen-VL-2.5-7B	0.251	0.306	0.258	0.156	0.102	0.066	2.426	7.413	0.066	1.279	2.574	0.215
	Qwen-VL-2.5-32B	0.427	0.350	0.375	0.245	0.248	0.223	2.080	6.077	0.113	1.650	3.732	0.128

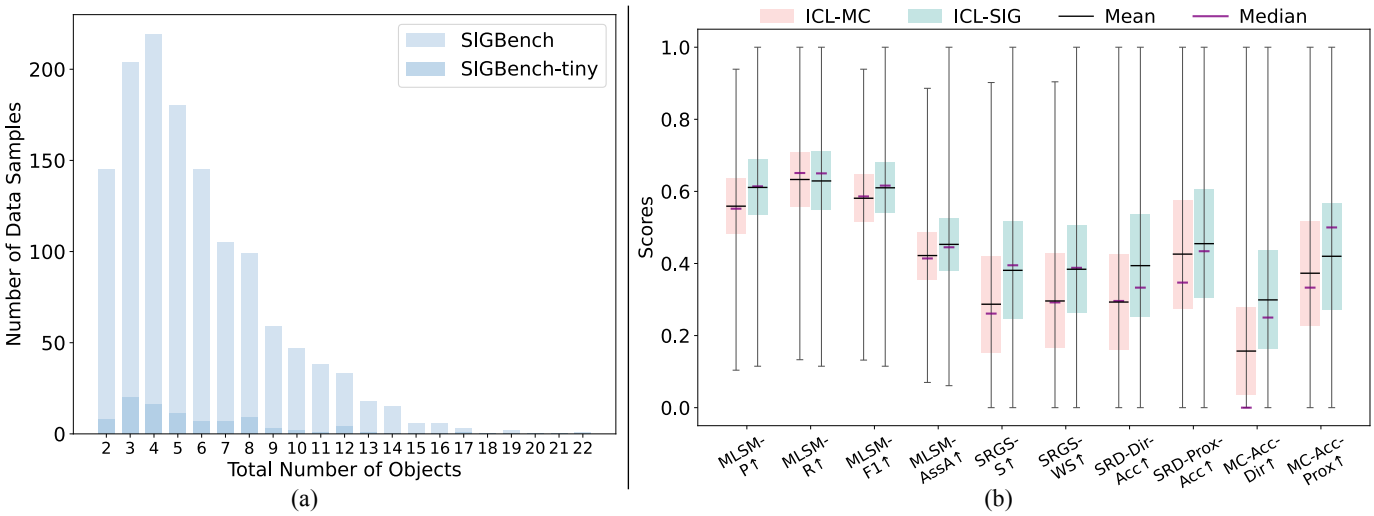


Fig. 5: (a) Data distribution of SIGBench and SIGBench-tiny with object numbers. (b) Error bar statistics of multi-run ICL on SIGBench-tiny using Gemini-2.5-Pro with random data selection. The bars show the range of $\mu \pm \frac{1}{2}\sigma$, where μ, σ means mean and standard deviation, respectively. The whiskers indicate min and max of each metrics. Across metrics, ICL-SIG generally yields higher means/medians than ICL-MC.

is placed according to its true world-coordinate position into a JSON file. The predicted SIG is then quantitatively compared against the GT representation using our MLSM and SRGS metrics. An example is depicted in Fig. 4 (b) left.

Spatial Relation Paragraph Filling. Differing from standard VQA setups, here each prompt consists of a fully-formed sentence in which only the prepositional phrase connecting two object mentions has been removed. The model's objective is to infer and insert the correct preposition based on contextual cues. We address two relation categories: directional and proximal. For every image, we construct two concise paragraphs: one featuring blank slots for directional prepositions and another for proximal prepositions, which has been illustrated in Fig. 4 (b) right.

4.3 Human-like Visual-Spatial Intelligence with Grid Tasks

Gaze Prediction. This task assesses an MLLM's capacity to forecast the driver's gaze attention map at frame i by conditioning on the sequence of attention maps from frames $i-5$ to $i-1$, thereby capturing the inherently spatiotemporal dynamics of human visual attention. We then evaluate how accurately the model can predict, from this historical gaze data, the specific regions of the scene to which a driver would most likely direct their focus.

Human-Like SIGC and SRPF. Both tasks mirror their grid-based VSI equivalents (Sec. 4.2) but integrate human gaze into the evaluation. Specifically, each object in frame i is assigned an attention weight from the SIG attention map A^i_{SIG} capturing its salience to a human observer. When computing errors, errors on highly weighted objects incur proportionally greater penalties, thereby emphasizing the model's capacity to accurately reason about the most behaviorally relevant elements. This attention-weighted scoring framework thus reveals whether an MLLM can prioritize spatial relations between objects in a manner similar as human visual focus.

5 Experiment

5.1 Model Selection and Evaluation Metrics

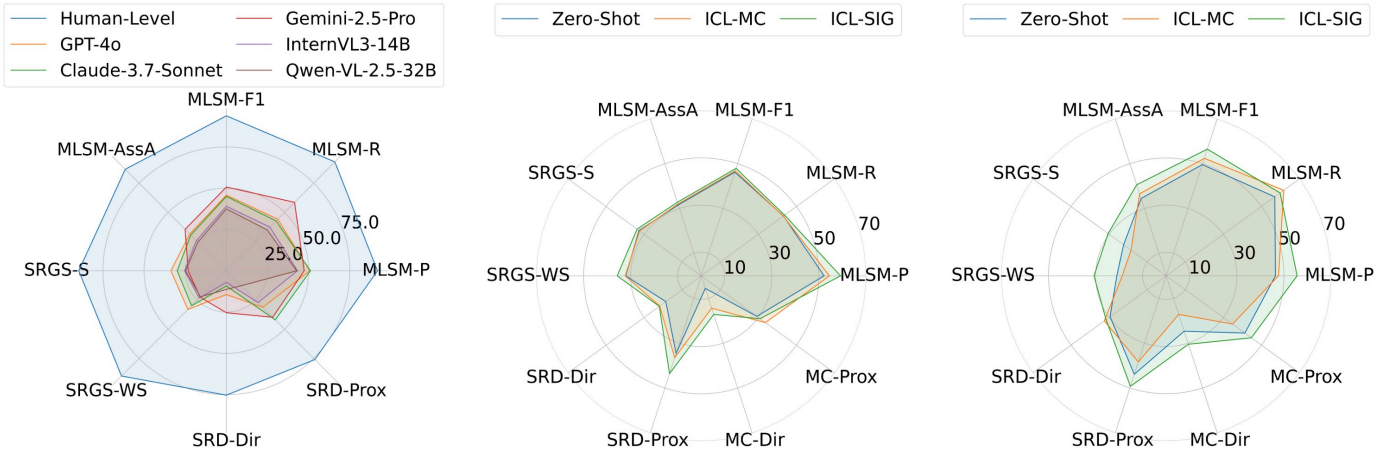
Models. We evaluate several top-tier MLLMs on SIGBench mainly from five modal families: 1) open-source models

Authorized licensed use limited to: Johns Hopkins University. Downloaded on November 10,2026 at 15:10:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2026.3737728

TABLE 3: **Quantitative comparison of 3-shot ICL for grid-based VSI tasks on `SIGBench`-tiny.** Z-S, ICL-MC and ICL-SIG are zero-shot inference, ICL with multiple-choice VQA and ICL with SIG, respectively. Light red denotes instances where performance after ICL falls below the zero-shot baseline. The number of objects in 3 data sample we used for ICL are 7, 7, 8.

Models	Type	MLSM				SRGS		SRD (Directional)			SRD (Proximal)			MC (Acc)	
		P↑	R↑	F1↑	AssA↑	S↑	WS↑	MAE↓	MSE↓	Acc↑	MAE↓	MSE↓	Acc↑	Dir.↑	Prox.↑
GPT-4o	Z-S	0.522	0.432	0.462	0.316	0.327	0.321	1.792	4.827	0.186	0.858	1.324	0.346	0.056	0.292
	ICL-MC	0.545	0.431	0.468	0.320	0.324	0.323	1.600	4.106	0.218	0.920	1.621	0.365	0.144	0.337
	ICL-SIG	0.592	0.438	0.479	0.328	0.337	0.357	1.593	4.094	0.220	0.775	1.271	0.436	0.172	0.309
Gemini-2.5-Pro	Z-S	0.464	0.570	0.496	0.345	0.224	0.210	1.151	2.426	0.295	0.721	1.100	0.439	0.247	0.413
	ICL-MC	0.477	0.617	0.524	0.366	0.185	0.187	1.174	2.667	0.325	0.845	1.387	0.384	0.172	0.348
	ICL-SIG	0.556	0.608	0.565	0.406	0.305	0.307	1.126	2.396	0.316	0.578	0.729	0.493	0.305	0.447



(a) Results of Grid-based VSI Tasks (b) ICL using GPT-4o (c) ICL using Gemini-2.5-Pro

Fig. 6: **Visualization of SIG-empowered VSI results.** SRD-Dir and SRD-Prox represent Acc in SRD (Directional) and SRF (Proximal), respectively. (a) illustrate the comparison between the performance of human and different models in grid-based tasks on `SIGBench`. Even for leading MLLMs, there is a substantial performance gap compared to human in VSI. (b) and (c) denotes the comparison of ICL results of GPT-4o and Gemini-2.5-pro on `SIGBench`-tiny. ICL-SIG outperforms the zero-shot baseline in all VSI metrics and achieves more improvements than ICL-MC in general.

such as InternVL [41] and Qwen-VL [38]; 2) Proprietary models including OpenAI GPT [37], Google Gemini [39] and Anthropic Claude [40]. The specific models in each model family with their complete names we used during experiment are listed in detail in Tab. 2.

Evaluation Metrics. For gaze prediction tasks, we follow the widely used metrics in gaze/saliency prediction task such as Person's Correlation Coefficient (PCC), KL-Divergence and Information Gain (IG) [61], [69]. For grid-based SIGC and SRPF, we use MLSM, SRGS and SRD mentioned in Sec. 3.1. For human-like SIGC and SRPF, we use human-like SRGS and SRD mentioned in Sec. 3.3.

5.2 Results and Analysis for Grid-based VSI

Zero-shot Inference on `SIGBench`, showing in Tab. 2. The table reports the zero-shot performance of multiple MLLMs on grid-based SIGC and SRPF tasks, evaluated across metrics for grid-based VSI tasks we mentioned in Sec. 5.1: MLSM, SRGS, and SRD with directional and proximal sub-tasks. Human-Level performance is included as an upper bound for reference, showing consistently high precision, recall, and relational accuracy. Overall, Gemini-2.5-Pro achieves the highest overall MLSM-F1 and the best directional SRD, highlighting its strength in localizing objects and estimating directional relations with precision. GPT-4o delivers the strongest results on SRGS, attaining the highest S and WS scores and ranks a close second on MLSM and directional SRD metrics, demonstrating its robust ability to infer inter-object spatial relationships. Meanwhile, Claude-3.7-Sonnet shows exceptional proficiency in modeling proximal relations, achieving the top MAE and MSE on SRD (Proximal), indicating strong performance in both semantic proximity and graph-based relation matching. Open-source models generally trail behind proprietary models across all tasks, with significantly lower MLSM F1 and SRD accuracy. However, larger variants (*e.g.*, InternVL3-14B, Qwen-VL-2.5-32B) do exhibit improvements over their smaller counterparts, potentially suggesting scaling benefits.

SIG-based ICL with Random Sample Selection, shown in Tab. 3. To assess SIG's advantage over traditional VQA-style encodings for VSI, we conduct 3-shot ICL on the two strongest zero-shot models: GPT-4o and Gemini-2.5-Pro. Our protocol is as follows: 1) Mini-benchmark construction: We randomly sample 90 frames from `SIGBench` to form `SIGBench`-tiny. For each image, we generate 4–8 multiple-choice questions probing directional and proximal relationships, closely following the style of existing VQA spatial reasoning datasets. Fig. 5 (a) compares the data distributions of `SIGBench` and `SIGBench`-tiny; 2) Prompt example creation: We selected three additional frames (not in `SIGBench`-tiny) and exhaustively annotated every object in their GT SIGs,

Authorized licensed use limited to: Johns Hopkins University. Downloaded on November 10,2026 at 15:10:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2026.3737728

TABLE 4: **Quantitative comparison of different MLLMs on `SIGBench` dataset for human-like VSI tasks.** H means human-like and KL-D means KL-Divergence. The results for H-SRGS, H-SRD are following the similar trend as the results in general VSI tasks.

Type	Models	Gaze Prediction			H-SRGS		H-SRD (Directional)			H-SRD (Proximal)		
		PCC↑	KL-D↓	IG↑	S↑	WS↑	MAE↓	MSE↓	Acc↑	MAE↓	MSE↓	Acc↑
Proprietary	Claude-3.5-Haiku	0.798	0.520	0.364	0.106	0.072	1.265	3.291	0.111	0.374	0.488	0.388
	Claude-3.7-Sonnet	0.772	0.344	0.623	0.157	0.127	1.346	3.607	0.096	0.350	0.428	0.398
	Gemini-1.5-Pro	0.914	0.125	0.946	0.161	0.123	1.209	3.081	0.119	0.713	1.430	0.259
	Gemini-2.0-Flash	0.861	0.231	0.803	0.160	0.121	1.439	3.956	0.084	0.631	0.969	0.232
	Gemini-2.5-Pro	0.868	0.314	0.637	0.148	0.118	0.784	1.621	0.222	0.450	0.659	0.349
	GPT-4o-mini	0.838	0.514	0.504	0.010	0.003	1.509	4.165	0.064	0.484	0.693	0.324
	GPT-4o	0.673	0.506	0.406	0.197	0.163	1.146	2.912	0.145	0.557	0.870	0.287
Open-source	InternVL3-9B	0.427	2.932	-2.652	0.059	0.030	1.303	3.450	0.105	0.525	0.669	0.250
	InternVL3-14B	0.859	0.860	-0.073	0.131	0.099	1.338	3.585	0.093	0.780	1.448	0.182
	InternVL2.5-26B	0.951	0.090	0.988	0.120	0.082	1.247	3.225	0.136	0.539	0.669	0.254
	Qwen-VL-2.5-7B	0.950	0.068	1.028	0.038	0.018	1.492	4.200	0.081	0.958	2.033	0.152
	Qwen-VL-2.5-32B	0.664	1.373	-0.908	0.123	0.087	1.302	3.433	0.102	1.231	2.875	0.081

TABLE 5: Average per-frame runtimes of three metrics on AMD Ryzen 5900X CPU with varying object numbers.

Metrics	Number of Objects									
	3	5	7	9	11	13	15	17	19	22
MLSM [s]	0.0001	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0003	0.0003	0.0004
SRGS [s]	0.0001	0.0002	0.0003	0.0004	0.0004	0.0005	0.0006	0.0006	0.0008	0.0009
SRD [s]	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

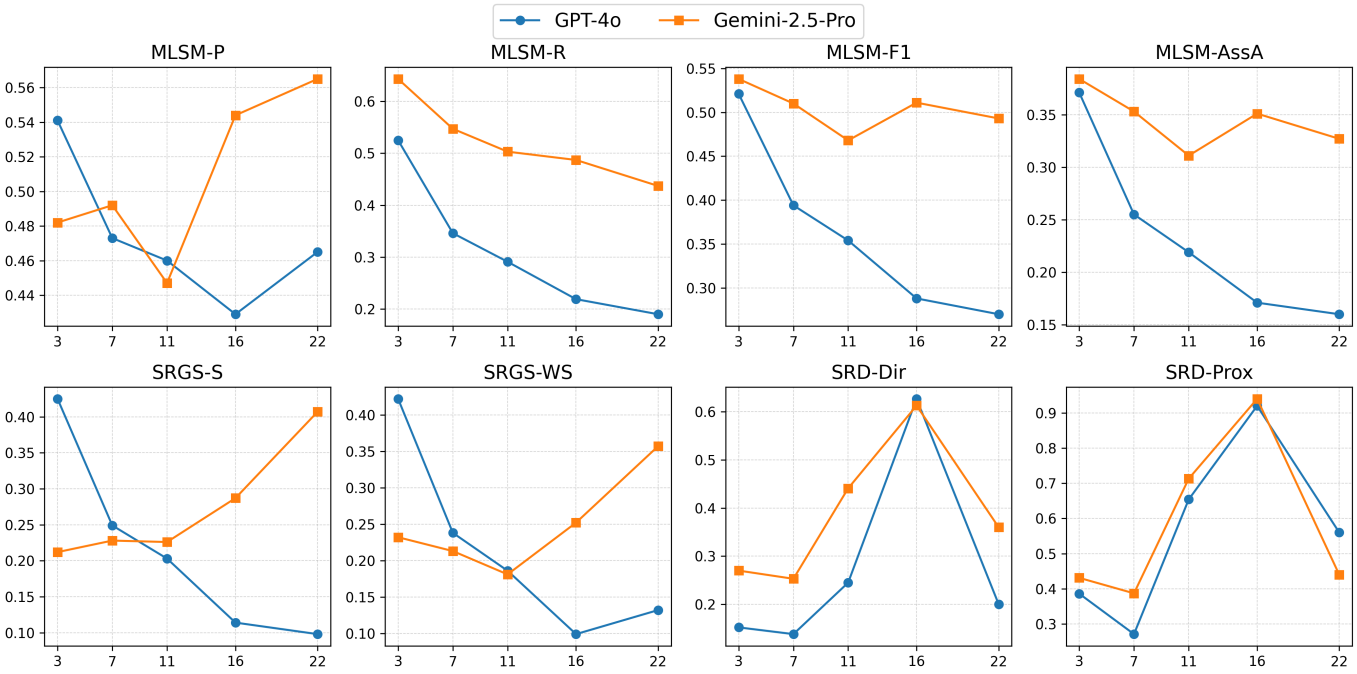


Fig. 7: **Quantitative Analysis of Failure Mode Cases using GPT-4o and Gemini-2.5-Pro.** We record the performance of GPT-4o and Gemini-2.5-Pro grouped different number of objects (3, 7, 11, 16, 22) in each data sample on `SIGBench`-tiny. As shown in the figure, GPT-4o's MLSM and SRGS scores steadily decline as the number of objects rises, while its SRD performance fluctuates, demonstrating that graph-based metrics are more sensitive to scene density and that raw text judgments alone can mask these trends. Gemini-2.5-Pro shows similar SRD variability but even improves on some graph metrics (MLSM-P, SRGS) with more objects, underscoring the complementarity of SIG's structured representation.

converting these into both SIG-based prompts and VQA-style QA pairs for ICL; 3) ICL conditions: ICL-SIG used prompts comprising full image and structured SIG representations, while ICL-MC used prompts with image and multiple-choice VQA pairs. Notably, SIG-based prompts offer a holistic scene-level encoding of spatial structure, in contrast to the more fragmented question-answer pairs in VQA-style prompts. Both settings were evaluated on `SIGBench`-tiny using our SIGC and SRPF metrics, alongside standard VQA accuracy.

The results in Tab. 3 reveal several important findings: 1) Consistent enhancement via ICL-SIG: both GPT-4o and Gemini-2.5-Pro show improvement on nearly every metric compared to zero-shot performance when using ICL-SIG. This is particularly clear for metrics such as MLSM-F1

Authorized licensed use limited to: Johns Hopkins University. Downloaded on November 10,2026 at 15:10:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2026.3737728

and AssA, where the structured SIG encoding yields better alignment with ground-truth spatial relations; 2) Robustness to sample selection: despite using only three randomly chosen frames as in-context examples, ICL-SIG reliably outperforms ICL-MC across models. Improvements are especially pronounced in MLSM and SRGS metrics, suggesting that richer, geometric-level representations provide more informative conditioning even under few-shot constraints; 3) Instability of ICL-MC compared to ICL-SIG: ICL-MC sometimes underperforms even the zero-shot baseline (as marked in light red), particularly on SRGS-S and MC (Directional) metrics for Gemini-2.5-Pro. This suggests that fragmented, question-focused prompts may fail to convey global spatial structure effectively in few-shot settings, confusing relational grounding. In contrast, SIG-based representations capture the full spatial configuration of objects in the scene, enabling the model to learn more human-like spatial reasoning strategies. To provide more insights about the stability of SIG-based ICL, we conduct five independent runs with randomly selected data samples using Gemini-2.5-Pro and provide the error bar statistics shown in Fig. 5 (b). As shown in Fig. 5 (b), SIG-based ICL shows a consistent improvement than ICL-MC across almost all metrics. These observations demonstrate that SIG provides a higher-fidelity and more transferable representation for VSI than conventional VQA formats, supporting its potential as a new paradigm for enhancing spatial reasoning in MLLMs.

Empirical Runtime Analysis. Besides the theoretical runtime analysis of our proposed evaluation in Sec. 3.1, we conduct experiment on AMD Ryzen 5900X CPU and record the empirical runtime as shown in Tab. 5. As we can see from the results, MLSM, SRGS, and SRD can be executed within sub-millisecond latency, which satisfies the demand of real-time applications such as AD and robotics.

5.3 Results and Analysis for Human-Like VSI with Grid

Tab. 4 summarizes model performance on human-like VSI tasks, including gaze prediction, human-like SIGC (H-SRGS), and human-like SRPF (H-SRD, Directional and Proximal).

Gaze prediction. This task evaluates how well models can anticipate human fixation locations in new frames, using three metrics: PCC, KL-Divergence, and IG. Surprisingly, InternVL2.5-26B and Qwen-VL-2.5-7B achieve the best scores in PCC and IG, suggesting a strong match to human gaze patterns. Upon inspecting their outputs, we observed that these models use a simple raw average of the previous five attention maps to predict the next fixation point, unlike other models that apply post-processing such as Gaussian blurring or edge detection. This simpler approach appears to avoid over-smoothing or edge-bias artifacts, yielding better alignment with natural human gaze behavior (see supplementary material for detailed visualizations and analysis).

Human-like SRGS and SRD. These tasks introduce attention-map-based weighting in their evaluation: predicted spatial graphs and relational descriptions are scored with higher weight on image regions where human drivers naturally focus. Despite this adjustment, the relative ranking of MLLMs on these human-like metrics remains broadly consistent with grid-based results in Tab. 2, suggesting that current MLLMs might be struggling at interpreting the prioritization of different objects in a scene. Overall, we conclude that current models still struggle to prioritize scene elements in a truly human-like manner within complex driving environments.

5.4 Analysis on Failure cases

As the results shown in Tab. 2 and Tab. 4, `SIGBench` is a challenging benchmark even for the world's leading MLLMs (*e.g.* GPT-4o, Gemini-2.5-Pro, etc). In this section, we want to further investigate where the challenges truly stem from both quantitative and qualitative perspective.

Quantitative Analysis. To investigate common failure modes, we analyze the zero-shot performance of GPT-4o and Gemini-2.5-Pro on `SIGBench` across scenes with different numbers of objects, as shown in Fig. 7. GPT-4o exhibits a clear degradation as scene complexity increases: its object-level MLSM scores generally decline, and its graph-based SRGS scores drop more noticeably, indicating reduced reliability in both object recognition and relational understanding in crowded scenes. Its SRD scores also fluctuate without a consistent trend, suggesting unstable spatial reasoning under higher visual complexity. In contrast, Gemini-2.5-Pro remains considerably more robust. Its MLSM scores stay relatively strong, while its SRGS scores improve as the number of objects increases, implying that richer scene structure provides useful relational cues rather than causing overload. Although some variation is also observed in SRD, Gemini-2.5-Pro maintains stronger overall spatial and relational consistency. These results suggest that GPT-4o is more sensitive to scene density, whereas Gemini-2.5-Pro scales more effectively with increasing visual complexity.

Qualitative Analysis. From qualitative perspective, we recognize two scenarios that current MLLMs are struggling to deal with: 1) The scenarios containing objects that are either very small (tiny objects) or positioned at periphery of the scene. In this case, these objects are frequently missed or mislocalized due to their low visual salience; 2) The scenarios containing objects with high overlapping area. High intersection over union (IoU) between objects often cause the models to conflate object identities or assign them to incorrect grid cells, resulting in swapped or merged scene-graph nodes. The visualization of both scenarios are shown in Fig. 9. We provide more detailed analysis in Appendix C.

6 Delving into SIG-based ICL

In previous sections, we demonstrate the superiority of SIG, compared to VQA, as a data schema to improve model's VSI capability using ICL with randomly selected data samples. Yet, as prior work has noted, the choice of in-context examples can dramatically affect performance [70], [71]. In this section, we systematically study how different data sample selection strategies influence SIG-based ICL from two perspectives: 1) Effect of different number of objects; 2) Effect of adaptive SIG samples.

6.1 Effect of Number of Objects

We first analyze the effect of object number on ICL performance. As shown in Fig. 5 (a), samples containing 3,

Authorized licensed use limited to: Johns Hopkins University. Downloaded on November 10,2026 at 15:10:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2026.3737728

TABLE 6: **Quantitative comparison of 3-shot ICL for grid-based VSI tasks on `SIGBench`-tiny with data selection based on number of objects.** Partial-empty means the number of objects in one or two of [vehicles, traffic signs, traffic lights] equals 0, while non-empty means the number of objects in vehicles, traffic signs and traffic lights in each data sample ≥ 1.

Type	#Objects	MLSM				SRGS		SRD (Directional)			SRD (Proximal)		
		P↑	R↑	F1↑	AssA↑	S↑	WS↑	MAE↓	MSE↓	Acc↑	MAE↓	MSE↓	Acc↑
Partial-empty	3	0.582	0.398	0.459	0.310	0.303	0.332	1.981	5.563	0.165	0.775	1.231	0.419
	4	0.562	0.405	0.458	0.314	0.326	0.341	1.777	4.667	0.180	0.845	1.314	0.367
	5	0.552	0.393	0.446	0.306	0.306	0.324	1.839	5.004	0.191	0.832	1.288	0.375
	8	0.571	0.465	0.499	0.349	0.352	0.365	1.668	4.340	0.203	0.797	1.244	0.393
	10	0.563	0.457	0.487	0.337	0.326	0.334	1.671	4.471	0.229	0.757	1.159	0.429
Non-empty	3	0.561	0.382	0.443	0.300	0.316	0.331	1.980	5.560	0.155	0.876	1.442	0.368
	4	0.584	0.419	0.476	0.331	0.340	0.357	1.881	5.164	0.167	0.760	1.125	0.398
	5	0.572	0.417	0.469	0.321	0.349	0.361	1.921	5.362	0.184	0.783	1.236	0.412
	8	0.543	0.484	0.499	0.350	0.362	0.365	1.667	4.359	0.204	0.848	1.313	0.357
	10	0.533	0.460	0.490	0.339	0.357	0.362	1.690	4.537	0.212	0.850	1.467	0.403
Combined	3+4+5	0.567	0.411	0.464	0.317	0.346	0.354	1.885	5.143	0.150	0.824	1.304	0.372
	3+4+8	0.600	0.442	0.501	0.350	0.365	0.380	1.849	5.000	0.171	0.780	1.241	0.421
	4+5+8	0.571	0.426	0.477	0.330	0.363	0.373	1.665	4.335	0.204	0.732	1.046	0.419
	4+8+10	0.598	0.455	0.505	0.353	0.360	0.374	1.566	4.007	0.229	0.779	1.253	0.430

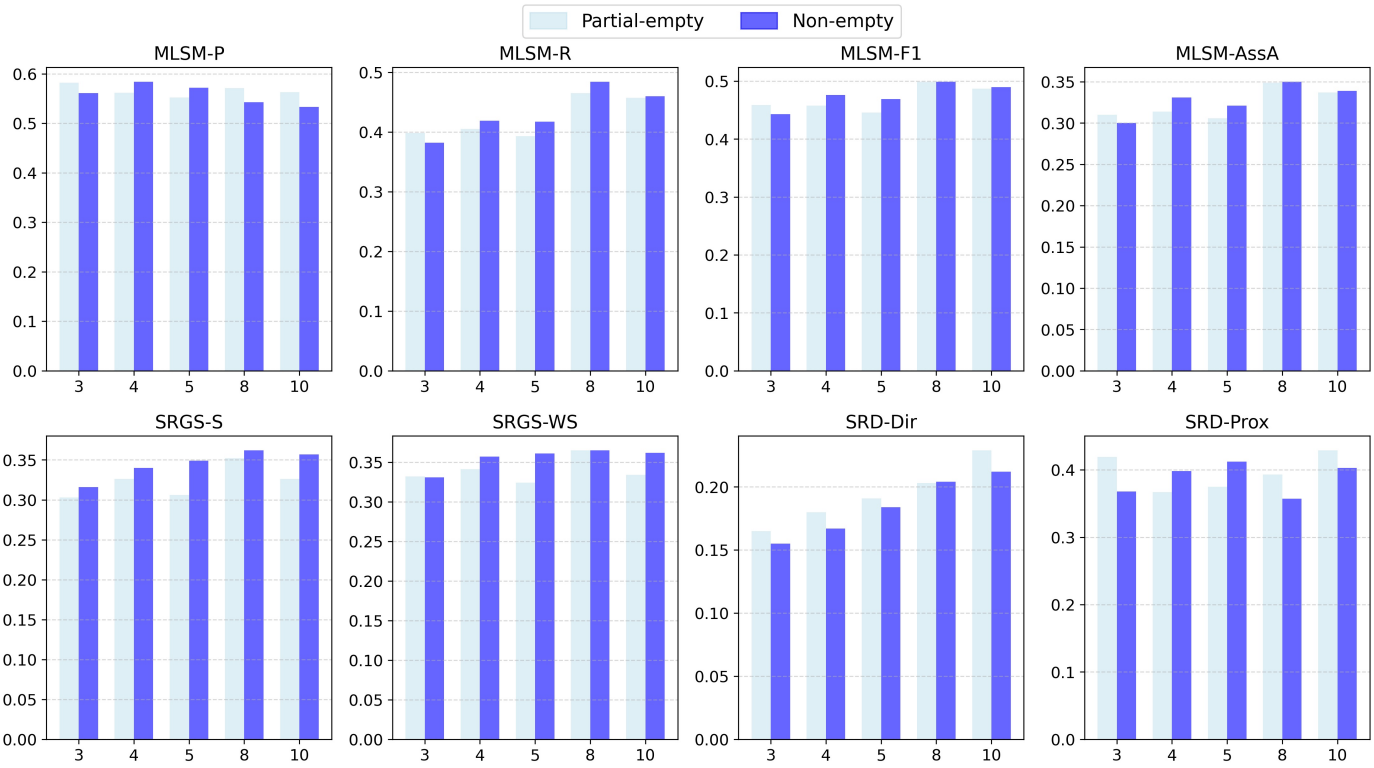


Fig. 8: **Comparison between partial-empty and non-empty data samples with different number of objects.** SRD-Dir and SRD-Prox shown here are the accuracy of SRD-Dir and SRD-Prox, respectively. In general, non-empty data samples achieves better performance than partial-empty data samples and the increasing number of objects results in an increase in performance.

4, and 5 objects dominate both `SIGBench` and `SIGBench`-Tiny, whereas samples with 8 and 10 objects are more prevalent in `SIGBench`-tiny. Based on this distribution, we evaluate SIG-based ICL on `SIGBench`-tiny using samples with 3, 4, 5, 8, and 10 objects. The results in Tab. 6 and Fig. 8 reveal two main observations: 1) Non-empty examples consistently outperform partial-empty ones across all object-count settings, indicating that examples containing all object categories provide more complete relational cues for learning spatial structure; 2) Performance generally improves as the number of objects increases, with 8-object samples achieving the strongest overall results among the evaluated settings.

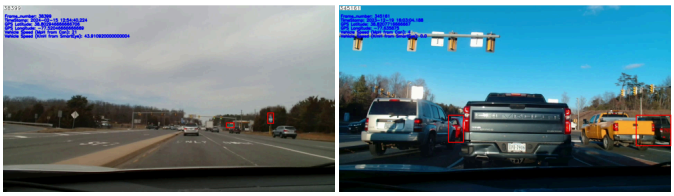
These findings highlight a fundamental limitation of current MLLMs in visual-spatial understanding: their generalization ability are substantially weaker than that of humans. Humans can readily transfer a learned spatial relation between one pair of objects to different object categories or to larger sets of similar objects. In contrast, the performance of existing MLLMs is noticeably affected by both the types and the number of objects present in ICL samples, suggesting that their VSI remains sensitive to instance composition rather

Authorized licensed use limited to: Johns Hopkins University. Downloaded on November 10,2026 at 15:10:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2026.3737728

TABLE 7: **Quantitative comparison of few shot ICL for grid-based VSI tasks on SIGBench-tiny with data selection based on adaptive SIG.** Different means we use different image and its corresponding GT SIG (*e.g.* 3-shot uses 3 different images), while adaptive means using the same image with different predicted SIG and GT SIG (*e.g.* 3-shot uses the same image with 3 different predicted SIG).

Shots	Type	MLSM				SRGS		SRD (Directional)			SRD (Proximal)		
		P↑	R↑	F1↑	AssA↑	S↑	WS↑	MAE↓	MSE↓	Acc↑	MAE↓	MSE↓	Acc↑
3	Different	0.561	0.415	0.460	0.318	0.329	0.343	1.673	4.379	0.211	0.722	1.044	0.422
	Adaptive	0.557	0.391	0.445	0.300	0.323	0.336	1.559	3.867	0.220	0.694	1.008	0.444
6	Different	0.570	0.417	0.469	0.320	0.331	0.346	1.563	3.933	0.223	0.699	0.995	0.429
	Adaptive	0.572	0.418	0.470	0.323	0.339	0.347	1.485	3.637	0.232	0.671	0.949	0.453
9	Different	0.588	0.447	0.499	0.346	0.364	0.379	1.607	4.040	0.213	0.657	0.954	0.470
	Adaptive	0.573	0.420	0.471	0.325	0.341	0.350	1.598	4.018	0.204	0.659	1.019	0.484



(a) Tiny Objects (b) Highly Overlapped Objects

Fig. 9: Selected failure cases with (a) tiny objects and (b) Objects with high overlap. We annotate objects using red bounding boxes. Please zoom in for details.

than abstracting relational concepts robustly. Motivated by this observation, we further examine mixed training sets composed of samples with different object numbers (*e.g.* 3+4+5). The top-performing combinations, 3+4+8, 4+5+8, and 4+8+10, all surpass any single-count setting, further demonstrating that heterogeneous object-number samples is beneficial for improving MLLMs' VSI.

6.2 Effect of Adaptive SIG Samples

We further study ICL with adaptive SIG samples. Unlike the heterogeneous sampling strategy used previously, where each shot contains a different image paired with its GT SIG, the adaptive strategy uses multiple shots from the same image by pairing several predicted SIGs with the corresponding GT SIG for direct comparison. Concretely, for each image, we rank candidate predicted SIGs by their MLSM precision scores and present them as few-shot examples in ranked order. The underlying intuition is that exposing the model to both weaker and stronger SIG predictions for the same scene can improve its ability to distinguish high-quality spatial descriptions and generate better SIG outputs. This design is consistent with human learning, where comparing imperfect solutions against a correct reference helps reveal the characteristics of an optimal answer.

As reported in Tab. 7, adaptive SIG samples produce clear gains over heterogeneous samples for MLSM and SRGS from 3- to 6-shot ICL, suggesting that they are particularly effective when the number of examples is limited. However, this advantage does not continue at larger shot numbers: under 9-shot ICL, heterogeneous samples achieve further improvement, whereas adaptive samples show little gain beyond 6 shots. This suggests that adaptive SIG samples are less scalable and that focusing repeatedly on a single scene

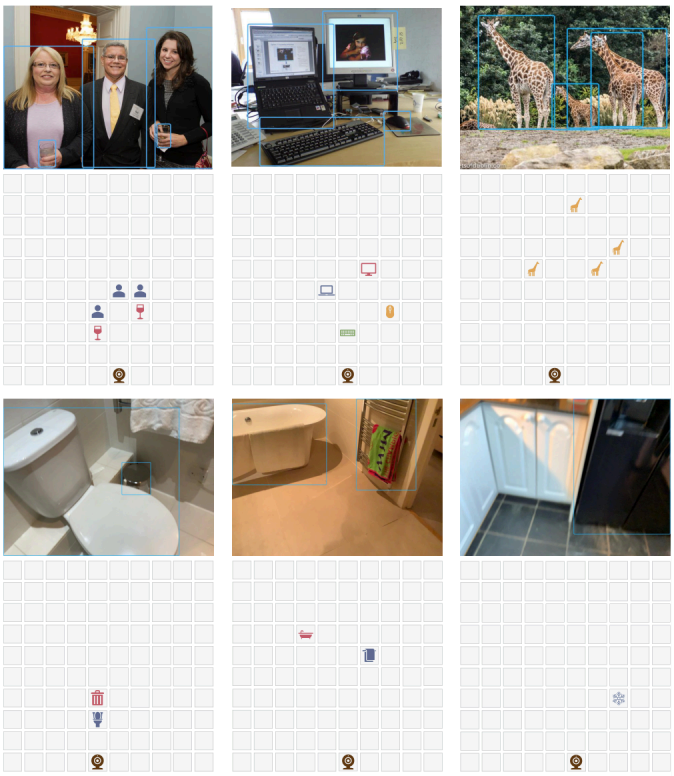


Fig. 10: **Visualization of SIG-COCO and SIG-ARKitScenes.** The top three samples are drawn from SIG-COCO, and the bottom three from SIG-ARKitScenes. Each sample comprises an image annotated with bounding boxes on prominent objects along with its corresponding SIG.

cannot fully replace the benefit of broader instance diversity. For SRD, adaptive SIG samples consistently outperform heterogeneous ones across all settings. Overall, these results suggest that the two strategies offer complementary benefits rather than serving as direct substitutes. To maximize the effectiveness of SIG-based ICL, adaptive and heterogeneous SIG samples should therefore be used jointly. Due to space limitation, we provide some additional visualizations of our results in Appendix D for more insights.

7 Generalizability of SIG as Data Schema

7.1 Generalizability across different domains

As a representation for VSI, SIG is not confined to autonomous driving and can be applied to many computer

Authorized licensed use limited to: Johns Hopkins University. Downloaded on November 10,2026 at 15:10:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

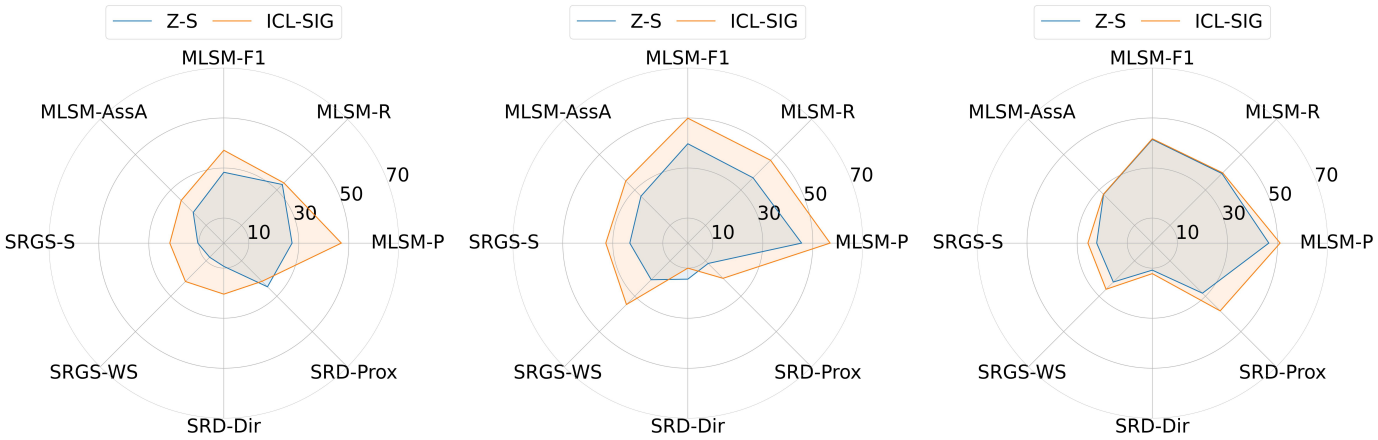
This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2026.3737728

TABLE 8: Results of zero-shot and 3-shot ICL on SIG-COCO and SIG-ARKitScenes for general VSI tasks using GPT-4o. Z-S means zero-shot and ICL-SIG means ICL using SIG.

Benchmark	Type	MLSM				SRGS		SRD (Directional)			SRD (Proximal)		
		P↑	R↑	F1↑	AssA↑	S↑	WS↑	MAE↓	MSE↓	Acc↑	MAE↓	MSE↓	Acc↑
SIG-COCO	Z-S	0.758	0.826	0.771	0.652	0.574	0.607	1.303	3.403	0.407	0.893	1.160	0.230
	ICL-SIG	0.860	0.840	0.849	0.749	0.711	0.746	0.713	1.827	0.640	0.357	0.360	0.643
SIG-ARKitScenes	Z-S	0.714	0.756	0.727	0.581	0.619	0.597	1.267	3.467	0.433	0.900	1.167	0.233
	ICL-SIG	0.809	0.850	0.823	0.719	0.729	0.737	0.467	0.533	0.567	0.300	0.500	0.800

TABLE 9: Quantitative comparison of 3-shot ICL on open-source models for general VSI tasks on `SIGBench-tiny`.

Models	Type	MLSM				SRGS		SRD (Directional)			SRD (Proximal)		
		P↑	R↑	F1↑	AssA↑	S↑	WS↑	MAE↓	MSE↓	Acc↑	MAE↓	MSE↓	Acc↑
Qwen-VL -2.5-7B	Z-S	0.273	0.331	0.283	0.173	0.104	0.080	2.129	6.084	0.093	1.230	2.461	0.247
	ICL-SIG	0.470	0.340	0.371	0.241	0.216	0.217	1.691	4.372	0.204	1.271	2.525	0.216
Qwen-VL -2.5-32B	Z-S	0.455	0.369	0.397	0.265	0.232	0.207	1.767	4.442	0.144	1.718	4.000	0.115
	ICL-SIG	0.569	0.468	0.499	0.351	0.328	0.347	2.161	6.120	0.100	1.447	3.183	0.199
InternVL3 -14B	Z-S	0.465	0.393	0.414	0.275	0.222	0.220	1.944	5.177	0.108	1.042	1.837	0.283
	ICL-SIG	0.510	0.397	0.417	0.276	0.257	0.261	1.856	4.795	0.122	0.763	1.082	0.383



(a) Qwen-VL-2.5-7B (b) Qwen-VL-2.5-32B (c) InternVL3-14B

Fig. 11: **Visualization of SIG-based ICL results on Open Source Models.** The experiments are conducted on `SIGBench-tiny`.

vision datasets with fixed object categories. To demonstrate this generality, we adapt our annotation pipeline to MS COCO [72] and ARKitScenes [73], and build two proof-of-concept benchmarks: SIG-COCO and SIG-ARKitScenes. For MS COCO, we select seven representative categories spanning people, objects, and animals: laptop, monitor, keyboard, mouse, person, wine glass, and giraffe. For ARKitScenes, a video-based indoor dataset, we focus on bathroom and kitchen scenes and use five categories: toilet, tub, clothes rack, fridge, and bin. Individual video frames are extracted from ARKitScenes to form the dataset. Together, these categories cover a diverse set of non-driving environments, including outdoor urban scenes and indoor household or office spaces, which are closely relevant to applications for both indoor and outdoor robotics (embodied AI) and augmented reality.

For each benchmark, we use three samples for ICL training and create a testing set. To keep the annotation process consistent with `SIGBench`, we employ Grounded-SAM-2 [74] to detect and localize target objects, retaining only prominent instances via confidence thresholding, and then apply the same annotation protocol to construct SIG. Unlike `SIGBench`, which boxes only vehicles requiring estimation, here we annotate all prominent target objects in the selected categories to assess SIG under more general scene conditions. Fig. 10 shows examples from SIG-COCO and SIG-ARKitScenes. The zero-shot and SIG-based ICL results of GPT-4o are summarized in Tab. 8. On SIG-COCO, SIG-based ICL improves MLSM by about 7% and SRGS by about 15%. It also substantially improves SRD, reducing MAE and MSE by nearly half and tripling the accuracy of SRD (Proximal). On SIG-ARKitScenes, despite the greater difficulty of indoor spatial relations, SIG-based ICL still yields strong gains: MLSM and SRGS increase by more than 10% on average, while SRD errors are reduced by more than half and SRD (Proximal) accuracy improves by over three times. These results indicate that the gains brought by SIG-based ICL transfer across domains, further validating the generalizability of SIG.

7.2 Generalizability across different MLLMs

Besides the promising performance shown by SIG-based ICL conducted on proprietary MLLMs such as Gemini-2.5-

Authorized licensed use limited to: Johns Hopkins University. Downloaded on November 10,2026 at 15:10:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2026.3737728

Pro and GPT-4o, we further examine whether SIG-based ICL generalizes to open-source models. To this end, we apply SIG-based ICL with random sample selection to two representative open-source MLLMs, Qwen-2.5-VL and InternVL-3. The results are reported in Tab. 9 and visualized in Fig. 11. Across all evaluated models, SIG-based ICL consistently improves performance on MLSM and SRGS. For Qwen-2.5-VL-7B, MLSM-P increases by nearly 20%, while SRGS improves by more than 10%. For Qwen-2.5-VL-32B, all MLSM and SRGS metrics show gains exceeding 10%, indicating stable benefits from SIG-based ICL. For SRD, the effect of SIG-based ICL is less consistent on the Qwen models. Qwen-2.5-VL-7B improves on directional SRD but declines on proximal SRD, whereas Qwen-2.5-VL-32B shows the opposite behavior. In contrast, InternVL3-14B achieves consistent gains across both directional and proximal SRD metrics. These results suggest that the transfer of SIG-based ICL to textual spatial reasoning can be achieved reliably in some open-source MLLMs.

8 Conclusion and Discussion

In this work, we propose a novel representation for visual-spatial intelligence (VSI), termed **Spatial Intelligence Grid (SIG)**, and a suite of graph-based evaluation metrics that leverage its structured topology to support more precise and general assessment of VSI. Extensive experiments across multiple Multimodal Large Language Models (MLLMs) show that few-shot in-context learning (ICL) with SIG consistently delivers larger, more stable, and more comprehensive VSI gains than conventional VQA-style prompts, underscoring its superior ability to encode complex spatial relationships. We further examine the role of sample selection in SIG-based ICL, and find that combining samples with diverse object counts and heterogeneous, quality-ranked adaptive SIGs yields the most robust improvements. We also demonstrate the generalizability of SIG across diverse scenarios and model families. Based on SIG, we construct `SIGBench`, a benchmark consists image with corresponding SIG and human gaze attention pairs in autonomous driving scenario, designed to evaluate both grid-based machine VSI and human-like attention-driven spatial reasoning under our proposed metrics. We posit that SIG offers an effective schema for VSI representation and that `SIGBench` will serve as a rigorous standard for genuine spatial-reasoning capability.

References

[1] B. Li, Y. Ge, Y. Ge, G. Wang, R. Wang, R. Zhang, and Y. Shan, "Seed-bench: Benchmarking multimodal large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 13 299–13 308.

[2] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, K. Chen, and D. Lin, "Mmbench: Is your multi-modal model an all-around player?" 2024. [Online]. Available: https://arxiv.org/abs/2307.06281

[3] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang, "Mm-vet: Evaluating large multimodal models for integrated capabilities," 2024. [Online]. Available: https://arxiv.org/abs/2308.02490

[4] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen, "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi," in *Proceedings of CVPR*, 2024.

[5] P. Xu, W. Shao, K. Zhang, P. Gao, S. Liu, M. Lei, F. Meng, S. Huang, Y. Qiao, and P. Luo, "Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 3, pp. 1877–1893, 2025.

[6] F. Liu, G. E. T. Emerson, and N. Collier, "Visual spatial reasoning," *Transactions of the Association for Computational Linguistics*, 2023.

[7] I. Stogiannidis, S. McDonagh, and S. A. Tsaftaris, "Mind the gap: Benchmarking spatial reasoning in vision-language models," 2025. [Online]. Available: https://arxiv.org/abs/2503.19707

[8] J. Yang, S. Yang, A. Gupta, R. Han, L. Fei-Fei, and S. Xie, "Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces," *arXiv preprint arXiv:2412.14171*, 2024.

[9] Y. Li, Y. Zhang, T. Lin, X. Liu, W. Cai, Z. Liu, and B. Zhao, "Sti-bench: Are mllms ready for precise spatial-temporal world understanding?" 2025. [Online]. Available: https://arxiv.org/abs/2503.23765

[10] S. K. Ramakrishnan, E. Wijmans, P. Kraehenbuehl, and V. Koltun, "Does spatial cognition emerge in frontier models?" in *International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=WK6K1FMEQ1

[11] S. Yang, R. Xu, Y. Xie, S. Yang, M. Li, J. Lin, C. Zhu, X. Chen, H. Duan, X. Yue, D. Lin, T. Wang, and J. Pang, "Mmsi-bench: A benchmark for multi-image spatial intelligence," *arXiv preprint arXiv:2505.23764*, 2025.

[12] J. M. Keller and X. Wang, "Learning spatial relationships in computer vision," in *Proceedings of IEEE 5th International Fuzzy Systems*, vol. 1. IEEE, 1996, pp. 118–124.

[13] I. Dennis and P. Tapsfield, *Human abilities: their nature and measurement*. Psychology Press, 2013.

[14] K. Janowicz, S. Gao, G. McKenzie, Y. Hu, and B. Bhaduri, "Geoai: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond," pp. 625–636, 2020.

[15] H. Gardner, *Frames of Mind: The Theory of Multiple Intelligences*. Basic Books, 2011. [Online]. Available: https://books.google.com/books?id=2IEfFSYouKUC

[16] J. Kuang, Y. Shen, J. Xie, H. Luo, Z. Xu, R. Li, Y. Li, X. Cheng, X. Lin, and Y. Han, "Natural language understanding and inference with mllm in visual question answering: A survey," *ACM Computing Surveys*, vol. 57, no. 8, pp. 1–36, 2025.

[17] B. S. Kim, J. Kim, D. Lee, and B. Jang, "Visual question answering: A survey of methods, datasets, evaluation, and challenges," *ACM Computing Surveys*, vol. 57, no. 10, pp. 1–35, 2025.

[18] J. Ji, R. Krishna, L. Fei-Fei, and J. C. Niebles, "Action genome: Actions as compositions of spatio-temporal scene graphs," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 236–10 247.

[19] A. Bosse, *Algemeene manier van de hr. Desargues, tot de practijck der perspectiven, gelijck tot die der meet-kunde, met de kleyne voet-maat, mitsgaders der plaatsen, en proportien van de stercke en flaauwe rakingen, of kleuren*. by Dancker Danckertsz.# woonende in de Kalver-straat, in de Danckbaarheyt#, vol. 1.

[20] K. Scott, "Screen wise, screen play: Jacques de lajoue and the ruses of rococo," *Art History*, vol. 36, no. 3, pp. 568–607, 2013.

[21] D. Rousar. Alberti's veil. [Online]. Available: https://www.sightsize.com/albertis-veil/

[22] J. C. Chase, *Drawing Made Easy*. EJ Clode, 1919.

[23] E. Panofsky, *The life and art of Albrecht Dürer*. Princeton University Press, 2023.

[24] J. C. Smith, *The" Invention" of Dürer as a Renaissance Artist*. Brill, 2010.

[25] A. Dürer and C. Dodgson, *Albrecht Dürer*. Medici Society, 1926.

[26] M. Kemp, *Art in History, 600 BC-2000 AD: Ideas in Profile*. Profile Books, 2015.

[27] Y. Yasuoka, "A new interpretation of the grid system reform in the late period," *The Journal of Egyptian Archaeology*, vol. 107, no. 1-2, pp. 265–280, 2021.

[28] R. Carter, "The art of romare bearden a resource for teachers," *National Gallery of Art, Washington*, 2003.

[29] Y. Duan, Q. Zhang, and R. Xu, "Prompting multi-modal tokens to enhance end-to-end autonomous driving imitation learning with llms," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 6798–6805.

[30] L. Liu, P. Wang, G. Wu, J. Jiang, and H. Yang, "Towards optimal mixture of experts system for 3d object detection: A game of accuracy, efficiency and adaptivity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–19, 2025.

Authorized licensed use limited to: Johns Hopkins University. Downloaded on November 10,2026 at 15:10:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2026.3737728

[31] P. Wang, G. Wu, Y. Zhao, Y. Lin, and H. F. Yang, "Llm agents can partially mimic human driving behaviors and decision values but better align with aggressive profiles," *Transportation Research Part C: Emerging Technologies*, vol. 186, p. 105606, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X2600094X

[32] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6693–6702.

[33] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang, "Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario," *arXiv preprint arXiv:2305.14836*, 2023.

[34] Y. Inoue, Y. Yada, K. Tanahashi, and Y. Yamaguchi, "Nuscenes-mqa: Integrated evaluation of captions and qa for autonomous driving datasets using markup annotations," *arXiv preprint arXiv:2312.06352*, 12 2023.

[35] S.-Y. Park, C. Cui, Y. Ma, A. Moradipari, R. Gupta, K. Han, and Z. Wang, "Nuplanqa: A large-scale dataset and benchmark for multi-view driving scene understanding in multi-modal large language models," 2025. [Online]. Available: https://arxiv.org/abs/2503.12772

[36] K. Tian, J. Mao, Y. Zhang, J. Jiang, Y. Zhou, and Z. Tu, "Nuscenes-spatialqa: A spatial understanding and reasoning benchmark for vision-language models in autonomous driving," 2025. [Online]. Available: https://arxiv.org/abs/2504.03164

[37] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[38] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A frontier large vision-language model with versatile abilities," *arXiv preprint arXiv:2308.12966*, 2023.

[39] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[40] Anthropic, "The claude 3 model family: Opus, sonnet, haiku," 2024, https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

[41] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma *et al.*, "How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites," *arXiv preprint arXiv:2404.16821*, 2024.

[42] W. Cai, Y. Ponomarenko, J. Yuan, X. Li, W. Yang, H. Dong, and B. Zhao, "Spatialbot: Precise spatial understanding with vision language models," *arXiv preprint arXiv:2406.13642*, 2024.

[43] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. Guibas, and F. Xia, "Spatialvlm: Endowing vision-language models with spatial reasoning capabilities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 14 455–14 465.

[44] A.-C. Cheng, H. Yin, Y. Fu, Q. Guo, R. Yang, J. Kautz, X. Wang, and S. Liu, "Spatialrgpt: Grounded spatial reasoning in vision-language models," in *NeurIPS*, 2024.

[45] B. Li, Y. Ge, Y. Chen, Y. Ge, R. Zhang, and Y. Shan, "Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension," *arXiv preprint arXiv:2404.16790*, 2024.

[46] J. Wang, Y. Ming, Z. Shi, V. Vineet, X. Wang, Y. Li, and N. Joshi, "Is a picture worth a thousand words? delving into spatial reasoning for vision language models," in *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 2024.

[47] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, "Rt-1: Robotics transformer for real-world control at scale," in *RSS*, 2023.

[48] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *CoRL*, 2023.

[49] Y. Zhang, J. Lin, A. Zeng, G. Wu, S. Lu, Y. Fu, Y. Cai, R. Zhang, H. Wang, and L. Zhang, "Motion-x++: A large-scale multi-modal 3d whole-body human motion dataset," *arXiv preprint arXiv:2501.05098*, 2025.

[50] Y. Zhang, G. Wu, L.-H. Chen, Z. Zhao, J. Lin, X. Jiang, J. Wu, Z. Li, H. F. Yang, H. Wang, and L. Zhang, "Humanmm: Global human motion recovery from multi-shot videos," 2025. [Online]. Available: https://arxiv.org/abs/2503.07597

[51] J. Ren, P. Sundaresan, D. Sadigh, S. Choudhury, and J. Bohg, "Motion tracks: A unified representation for human-robot transfer in few-shot imitation learning," 2025. [Online]. Available: https://arxiv.org/abs/2501.06994

[52] L. Liu, B. Su, J. Jiang, G. Wu, C. Guo, C. Xu, and H. F. Yang, "Towards accurate and efficient 3d object detection for autonomous driving: A mixture of experts computing system on edge," 2025. [Online]. Available: https://arxiv.org/abs/2507.04123

[53] Y. Zhao, P. Wang, Y. Zhao, H. Du, and H. F. Yang, "Safetraffic copilot: adapting large language models for trustworthy traffic safety assessments and decision interventions," *Nature Communications*, vol. 16, no. 1, p. 8846, Oct. 2025. [Online]. Available: https://doi.org/10.1038/s41467-025-64574-w

[54] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, T. Funkhouser *et al.*, "Openscene: 3d scene understanding with open vocabularies," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2023, pp. 815–824.

[55] M. Omama, P. Inani, P. Paul, S. C. Yellapragada, K. M. Jatavallabhula, S. Chinchali, and M. Krishna, "Alt-pilot: Autonomous navigation with language augmented topometric maps," *arXiv:2310.02324*, 2023.

[56] P. Wang, M. Zhu, H. Lu, H. Zhong, X. Chen, S. Shen, X. Wang, and Y. Wang, "Bevgpt: Generative pre-trained large model for autonomous driving prediction, decision-making, and planning," *arXiv preprint arXiv:2310.10357*, 10 2023.

[57] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, P. Luo, A. Geiger, and H. Li, "Drivelm: Driving with graph visual question answering," *arXiv preprint arXiv:2312.14150*, 2023.

[58] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proc. of IEEE/CVF conf. on comp. vision and pattern recognition*, 2020, pp. 11 621–11 631.

[59] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, "Scalability in perception for autonomous driving: Waymo open dataset," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2443–2451.

[60] H. Caesar, J. Kabzan, K. S. Tan, W. K. Fong, E. Wolff, A. Lang, L. Fletcher, O. Beijbom, and S. Omari, "Nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles," 2022. [Online]. Available: https://arxiv.org/abs/2106.11810

[61] A. Palazzi, D. Abati, F. Solera, and R. Cucchiara, "Predicting the driver's focus of attention: the dr (eye) ve project," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1720–1733, 2018.

[62] A. Biswas, P. Gupta, S. Khurana, D. Held, and H. Admoni, "Modeling drivers' situational awareness from eye gaze for driving assistance," in *Proceedings of The 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 3551–3567. [Online]. Available: https://proceedings.mlr.press/v270/biswas25a.html

[63] Y. Xia, D. Zhang, J. Kim, K. Nakayama, K. Zipser, and D. Whitney, "Predicting driver attention in critical situations," in *Computer Vision – ACCV 2018*, C. Jawahar, H. Li, G. Mori, and K. Schindler, Eds. Cham: Springer International Publishing, 2019, pp. 658–674.

[64] R. A. Epstein, E. Z. Patai, J. B. Julian, and H. J. Spiers, "The cognitive map in humans: spatial navigation and beyond," *Nature Neuroscience*, vol. 20, no. 11, pp. 1504–1513, 2017. [Online]. Available: https://www.nature.com/articles/nn.4656

[65] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *J. Image Video Process.*, vol. 2008, Jan. 2008. [Online]. Available: https://doi.org/10.1155/2008/246309

[66] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixe, and B. Leibe, "Hota: A higher order metric for evaluating multi-object tracking," *Int. J. Comput. Vision*, vol. 129, no. 2, p. 548–578, Feb. 2021. [Online]. Available: https://doi.org/10.1007/s11263-020-01375-2

[67] X. Gao, B. Xiao, D. Tao, and X. Li, "A survey of graph edit distance," *Pattern Anal. Appl.*, vol. 13, no. 1, p. 113–129, Feb. 2010.

[68] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms, Third Edition*, 3rd ed. The MIT Press, 2009.

Authorized licensed use limited to: Johns Hopkins University. Downloaded on November 10,2026 at 15:10:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2026.3737728

[69] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *arXiv preprint arXiv:1604.03605*, 2016.
[70] Y. Zhang, K. Zhou, and Z. Liu, "What makes good examples for visual in-context learning?" in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 17 773–17 794. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/398ae57ed4fda79d0781c65c926d667b-Paper-Conference.pdf
[71] Y. Sun, Q. Chen, J. Wang, J. Wang, and Z. Li, "Exploring effective factors for improving visual in-context learning," *IEEE Transactions on Image Processing*, vol. 34, pp. 2147–2160, 2025.
[72] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.
[73] G. Baruch, Z. Chen, A. Dehghan, T. Dimry, Y. Feigin, P. Fu, T. Gebauer, B. Joffe, D. Kurz, A. Schwartz, and E. Shulman, "ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. [Online]. Available: https://openreview.net/forum?id=tjZjv_qh_CE
[74] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang, "Grounded sam: Assembling open-world models for diverse visual tasks," 2024.



Guanlin Wu is currently a Ph.D. student at Johns Hopkins University, advised by Prof. Hao Frank Yang. He received the M.S. degree in Systems Engineering from Johns Hopkins University, Baltimore, MD, USA and the B.E. degree in Computer Science and Engineering from The Chinese University of Hong Kong, Shenzhen, China. His research interests include computer vision, multimodal large language models and multi-agent system. His works have been published in TPAMI, NeurIPS, CVPR and ICCV, etc.



Boyan Su received his Master's degree in Applied Mathematics and Statistics from Johns Hopkins University, Baltimore, MD, USA. His research interests include post-training and efficient fine-tuning strategies for multimodal large language models and developing efficient training and evaluation toolchains for MLLMs. His works have published his work in ICCV and NeurIPS, etc.



Yang Zhao is currently a Ph.D. student at Johns Hopkins University, advised by Prof. Hao Frank Yang. He received M.S. degree in Photogrammetry and Remote Sensing and B.S. degree in Geographical Information Science from Peking University, Beijing, China. His research focuses on trustworthy artificial intelligence agents for decision-making, with interests in large language models, human decision modeling, AI-assisted decision support systems, and AI agent behavior control. He is a recipient of the Amazon AI PhD Fellowship. His works have been published Nature Computational Science, Nature Communications, and NeurIPS, etc.



Pu Wang is currently a Ph.D. student at Johns Hopkins University, advised by Prof. Hao Frank Yang. He received the M.S. degree in Computer Science from New York University in 2025. His research interests include data-centric AI and applications of generative AI to systems, such as autonomous driving. He also investigates approaches for aligning diverse data modalities within multimodal LLMs toward more coherent and effective multimodal AI systems. His works have been published in Nature Communications, TPAMI, NeurIPS and Transportation Research Part C, etc.



Yichen Lin is currently a Ph.D. student at Johns Hopkins University, co-advised by Prof. Hao Frank Yang and Prof. Susu Xu. He received the M.S. degree in Data Science from Johns Hopkins University, Baltimore, MD, USA and the B.E. degree in Data Science and Big Data Technology from Shanghai University of Finance and Economics, Shanghai, China. His research interests include large language models (LLMs) and the applications of LLMs in decision making and human behavior. His works have been published in NeurIPS and Transportation Research Part C, etc.



Hao Frank Yang is currently an assistant professor with the Department of Civil and Systems Engineering, member of the Data Science and AI Institute, and Institute of Assured Autonomy at Johns Hopkins University. His research focuses on decision-focused computing methods and systems, particularly for real-world applications in transportation and public health. He received his Ph.D. degree in civil engineering (transportation) from the University of Washington, Seattle. He was a research scientist with the NSF AI Institute for Edge Computing and the Department of Electrical and Computer Engineering at Duke University. He received dual bachelor's degrees in electrical and computer (telecommunication) engineering from the Beijing University of Posts and Telecommunications and the University of London – Queen Mary. He has published more than 45 journal papers and conference proceedings in Nature Computational Science, Nature Communications, TPAMI, CVPR, ICLR, NeurIPS, ICCV, Information Fusion, Transportation Research Part C, and IEEE TITS, etc.

Authorized licensed use limited to: Johns Hopkins University. Downloaded on November 10,2026 at 15:10:32 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.