2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

GSUC-VLM: Geometrically-Guided Spatial Understanding Chain of Vision Language Model for Autonomous Driving

Yifan Zhao[1], Ziyang Zheng[1], Congjia Chen[2], Shizhuo Zhang[1], Huixin Zhang[3], Wenrui Dai[1], Fan He[3], and Hongkai Xiong[1]

***Abstract*— Robust spatial understanding is crucial for Visual Question Answering (VQA) in autonomous driving that aims to enhance decision-making, reduce positional risks, and ensure road safety by providing answers based on the perception, prediction, and planning of driving scenarios. Despite remarkable success in semantic understanding of images and videos, existing Vision-Language Models (VLMs), as the prevailing paradigms for VQA, are limited in spatial understanding for multi-view scenes due to the lack of latent unified 3D reconstruction capability. They usually resort to additional spatial modalities such as point clouds or prior detection frameworks to enhance spatial understanding ability, but are still challenged by modality misalignment and degraded scalability. To overcome these limitations, in this paper, we propose a Geometrically-Guided Spatial Understanding Chain Framework (GSUC-VLM) for autonomous driving that leverages pretrained VLMs to jointly exploit semantic and spatial information in multi-view images. Specifically, we first design a dual-encoder architecture to fuse the semantic and spatial features separately extracted from multi-view images with a lightweight connector rather than introducing external spatial modalities. Subsequently, we align semantic and spatial features via distillation loss to generate semantic tokens enriched with the spatial information at the latent layer. Furthermore, we develop a projective feature conditioning method that incorporates camera intrinsic and extrinsic parameters to embed projection matrix encoding into the input vectors and introduce 3D position embeddings into the fusion layer for capturing complex spatial relationship across multiple views in autonomous driving. Experimental results show that the proposed GSUC-VLM achieves state-of-the-art performance in VQA tasks while providing Chain-of-Thought (CoT) understanding. Remarkably, GSUC-VLM demonstrates strong generalization on zero-shot VQA tasks.**

I. Introduction

Spatial understanding is essential for autonomous driving systems to accurately model 3D geometric information of real-world scenarios, and consequently, enable reliable driving decisions, risk mitigation, and overall road safety [1], [2], [3]. It encompasses a comprehensive grasp of the 3D environment, ranging from object recognition, object motion states, to relative positioning between objects and spatial relationship between the ego-vehicle and surrounding entities. Visual Question Answering (VQA) [4], [5], [6], [7] is emerging as a key task to evaluate spatial understanding in autonomous driving. It formulates driving-related questions grounded in real-world scenarios and provides standardized answers, covering spatial understanding, motion prediction, decision-making, and planning. The VQA benchmarks enable a holistic evaluation of spatial and semantic comprehension.

Fig. 1. Zero-shot spatial understanding by the proposed GSUC-VLM (green box) and existing VLMs like Qwen2.5-VL-3B (blue box) on the DriveBench dataset. Existing VLMs fail to accurately comprehend the spatial positions of objects and often focus on environmental descriptions not related to driving. They also struggle to achieve safety-critical reasoning and decision-making necessarily required for autonomous driving. In contrast, the proposed GSUC-VLM demonstrates superior zero-shot performance by effectively exploiting spatial relationships and leveraging Chain-of-Thought (CoT) reasoning to understand object saliency and make context-aware decisions for autonomous driving.

Inspired by the recent success in semantic understanding over images and videos, Vision-Language Models (VLMs) [8], [9] are deemed as promising alternatives to VQA and are widely applied to facilitate end-to-end autonomous driving systems [10], [11], [12], [13], [14]. However, existing VLMs primarily focus on single-view image or video understanding, and are restricted in spatial comprehension across multi-view inputs, a core requirement for autonomous driving. It remains an open and challenging problem to build a unified vision-language framework to simultaneously accomplish accurate semantic and spatial understanding from multi-view data.

*This work was supported in part by the National Natural Science Foundation of China under Grant 62401357, Grant 62371288, and Grant 62320106003, and National Key R&D Program of China under Grant 2025YFF0515604. (Corresponding Author: Ziyang Zheng)

[1]Yifan Zhao, Ziyang Zheng, Shizhuo Zhang, Wenrui Dai, and Hongkai Xiong are with Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: {zyf123456; zhengziyang; zhangshizhuo; daiwenrui; xionghongkai}@sjtu.edu.cn).

[2]Congjia Chen is with Beihang University, No.37 Xueyuan Road, Haidian District, Beijing 100191, China (e-mail: sy2317314@buaa.edu.cn).

[3]Huixin Zhang and Fan He are with Yinwang Intelligent Technology Co., Ltd., Shanghai, China (e-mail: {zhanghuixin; hf-inspire}@sjtu.edu.cn).

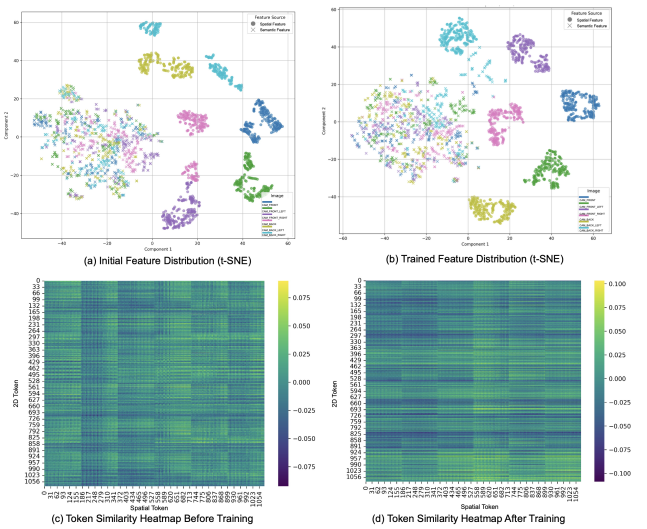

Fig. 2. t-SNE [19] feature distribution and token similarity heatmaps before and after training. Our model learns more compact and structured feature distributions, with clear six-view separation and improved alignment between semantic and spatial tokens. Heatmaps after training reveal soft global alignment, enabling semantically aware token interactions across views.

Recent VLMs have attempted to enhance spatial understanding by *integrating multi-modal datasets* or *incorporating detection priors*. One common approach is to *inject additional spatial* features using 3D data such as point clouds. However, aligning point clouds with multi-view image inputs and textual modalities is a significant challenge. Some existing methods align encoded point cloud features with image features in the bird's-eye view (BEV) space [15], while others directly align point cloud features with language embeddings [16]. These methods often discard or substantially alter the original visual encoder of VLMs, and compromise the ability of VLMs to leverage pre-trained semantic understanding from image data. It is infeasible to recover the ability via re-training VLMs due to limited availability of paired point clouds and images. These methods struggle to jointly achieve multi-view spatial understanding and high-level semantic reasoning.

Several approaches [17], [18], [4] attempt to improve spatial understanding through the use of detection priors. They rely on pre-trained detection networks to annotate bounding boxes or object coordinates in images or text prior to training. This helps the model focus on target regions but heavily depends on dataset-specific preprocessing and requires explicitly defined object references in either the visual or linguistic input. Moreover, they do not fundamentally improve the spatial understanding capabilities within the VLM framework itself, resulting in limited generalization performance.

To address these challenges, we propose a geometry-guided vision-language framework that integrates multi-view semantic and spatial information to construct a unified representation of these features and enhance feature fusion and alignment to better exploit the chain-of-thought and accommodate to VLMs, as depicted in Figure 2. The proposed method enables generalized and robust spatial understanding for autonomous driving without relying on point clouds or detection priors, as illustrated in Figure 1. Our contributions can be summarized as below.

- We propose GSUC-VLM, a novel geometry-guided spatial understanding chain of VLM framework for autonomous driving. It seamlessly integrates multi-view semantic and spatial information without the need for external spatial modalities such as point clouds.
- We design a dual-visual encoder architecture for extracting semantic and spatial features within the image space, followed by feature fusion and alignment via distillation loss, to preserve the zero-shot ability in both semantic and spatial understanding.
- We develop a projective feature conditioning mechanism that enables precise alignment between semantic and spatial features. We leverage virtual 3D position embeddings and camera projection matrix encodings to accurately capture complex spatial relationships across views and enhance ego-to-object positioning.
- Our method achieves state-of-the-art performance on various VQA benchmarks. Remarkably, it exhibits strong zero-shot generalization and provides interpretable reasoning chain that reflects robust spatial understanding.

II. Related Work

A. MLLMs for Spatial Understanding

Recently, multimodal large language models (MLLMs) have begun focusing on enhancing spatial understanding [20], [21], [22], [23], [24], enabling applications that are closer to real-world scenarios. Some approaches improve spatial understanding by leveraging multi-view images [25], [24] to reconstruct 3D scenes. While effective, these methods often lack rich semantic information from 2D images. They are typically successful in indoor environments with high view overlap, but perform poorly in autonomous driving scenarios where view overlap is limited, making direct application challenging. Other works enhance spatial understanding by introducing point cloud data [16], [26], [23], which provides explicit 3D spatial features. For example, Tod3cap [15] aligns point cloud and multi-view features in BEV space and uses 3D detection boxes as priors to improve captioning tasks. PointLLM [26] directly aligns point cloud data with language features to facilitate spatial understanding. However, these methods rely heavily on additional point cloud or depth data and are not seamlessly integrated with general VLM frameworks, making it difficult to fully leverage the visual-semantic capabilities of VLMs.

B. VQA in Autonomous Driving

In autonomous driving, VQA has become a critical component for vehicle spatial perception, as well as for decision-making and planning in complex scenarios. Existing VQA tasks span a wide range of topics, including multi-step reasoning across perception and decision-making [12], [27],

motion planning [28], and handling extreme conditions [17]. Recently, increasing attention has been given to improving spatial perception [13], [14], [29] in autonomous driving, as strong spatial understanding provides valuable information for human-vehicle interaction, trajectory planning, and scene interpretation. NuScenes-QA [13] emphasizes fine-grained spatial and temporal understanding using multi-modal inputs, such as LiDAR and camera data. Its questions are derived from 3D scene graphs, requiring an understanding of object locations and relationships. DriveLMM-o1 [14] further introduces step-by-step reasoning chains covering perception, prediction, and planning. Many of its questions explicitly test spatial comprehension, such as object positions, motion paths, and interactions in dynamic scenes. However, due to the limited spatial understanding capabilities of current VLMs, their performance on these VQA tasks remains suboptimal.

III. Proposed Method

Let $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^{M}$ denote a set of M RGB views captured at a specific moment in an autonomous driving scenario, where each view $\mathbf{I}_i \in \mathbb{R}^{H\times W\times 3}$ has a resolution of $H \times W$. Let $\{(Q_j, A_j)\}_{j=1}^{N}$ represent the corresponding N question-answer pairs. As illustrated in Figure 3, different from existing VLMs resorting to additional spatial data or priors, we unify the semantic and spatial representations in the image space for the input $\mathcal{I}$ through a dual-encoder and geometric conditioning module, and further align them with the reasoning chain of (Q, A) pairs in VLM to enable accurate spatial understanding. Specifically, we introduce the unified semantic and spatial representation through a dual-encoder and connector design in Section III-A, elaborate the projective feature conditioning module to encode complex geometric relationships in Section III-B, and finally present the geometry-guided feature alignment strategy in Section III-C.

A. Unified Representation of Semantic and Spatial Features

We achieve unified semantic and spatial representation by first extracting semantic and spatial features from $\mathcal{I}$ using dual visual encoders and then fusing the features with a lightweight connector.

1) Dual Visual Encoders: The dual visual encoders consist of a semantic encoder η_{sem} to capture semantic feature f_{sem} and a geometric encoder η_{spa} for spatial feature f_{spa}. The semantic encoder η_{sem} is initialized from the pre-trained vision backbone of Qwen2.5-VL [30] to obtain semantic features aligning vision and language modalities.

$$f_{\text{sem}} = \eta_{\text{sem}}(\mathcal{I}), \quad f_{\text{sem}} \in \mathbb{R}^{\lfloor \frac{H}{p} \rfloor \times \lfloor \frac{W}{p} \rfloor \times d} \tag{1}$$

where d and p denote the semantic feature dimension and the patch size. The geometric encoder η_{spa} aggregates geometric structures from multiple views and supports 3D-aware representation learning without requiring point cloud inputs.

$$f_{\text{spa}} = \eta_{\text{spa}}(\mathcal{I}), \quad f_{\text{spa}} \in \mathbb{R}^{\lfloor \frac{H}{p} \rfloor \times \lfloor \frac{W}{p} \rfloor \times d'} \tag{2}$$

where d' denote the spatial feature dimension. Here, we adopt VGGT [31] as the geometric encoder and extract its hidden features before task-specific heads (*e.g.*, DPT) as f_{spa}. These features incorporate fused spatial features from multiple views and could be more suitable for alignment with semantic features.

2) Lightweight Connector for Feature Fusion: Spatial and semantic features are fused using a lightweight connector consisting of a couple of two-layer MLPs to produce modality-specific representations f_{fused} capturing both semantic contents and spatial structures.

$$f_{\text{fused}} = \text{MLP}_1(f_{\text{sem}}) + \text{MLP}_2(f_{\text{spa}}). \tag{3}$$

The fused representation f_{fused} serves as the unified visual token fed into the VLM.

B. Projective Feature Conditioning

VLMs often struggle to accurately comprehend the spatial relationship between multi-view images captured from different perspectives. We then enhance the ability of spatial understanding of VLMs with the unified representation, and design a lightweight and plug-and-play method that injects multi-view geometric information into existing VLMs. The proposed method resorts to geometric conditioning at the input stage and 3D position embedding for multi-view data to maximize the retention of pre-trained weights without altering the core attention mechanism.

1) 3D Position Embedding for Multi-View Data: We treat a set of unordered multi-view images as an ordered video sequence to directly reuse the well-established, efficient spatiotemporal RoPE designed for videos. Each image patch is assigned a view index to distinguish its view source.

2) Geometric Conditioning Module at Input Stage: To further exploit the complex spatial relationship between views, we design a geometric conditioning module that encodes the camera projection matrix of each view into a low-dimensional vector. This vector contains all geometric information such as the position, orientation, and intrinsic parameters of a camera, and is appended to the patch embeddings of corresponding view. Precise geometric knowledge is then injected into the visual features at the input layer of the ViT to learn the association between geometry and visual content without altering subsequent Transformer layers.

Consider the camera for any i-th view with $i = 1, \cdots, M$, camera parameters include intrinsic parameters $\mathbf{L}_i$ and extrinsic parameters $\mathbf{S}_i = (\mathbf{R}_i, \mathbf{T}_i) \in SE(3)$ consisting of rotation $\mathbf{R}_i$ and translation $\mathbf{T}_i$. The intrinsic and extrinsic parameters determine the projection matrix $\mathbf{P}_i$ from the world coordinates to image coordinates, *i.e.*, $\mathbf{P}_i = [\mathbf{L}_i, \mathbf{0}]\,\mathbf{S}_i$. The inverse of $\mathbf{P}$ can be lifted and used to project the image back into the world coordinate system.

$$[\lambda \boldsymbol{d}_i, \mathbf{1}]^T = \hat{\mathbf{P}}_i^{-1}[\mathbf{I}_i, \mathbf{1}]^T, \tag{4}$$

where $\lambda \in \mathbb{R}$ is a scalar magnitude and $\boldsymbol{d}_i \in \mathbb{S}$ is a unit-norm ray direction. Inspired by [32], [33], we use Plücker raymaps as the geometric position encoding for multi-view images.

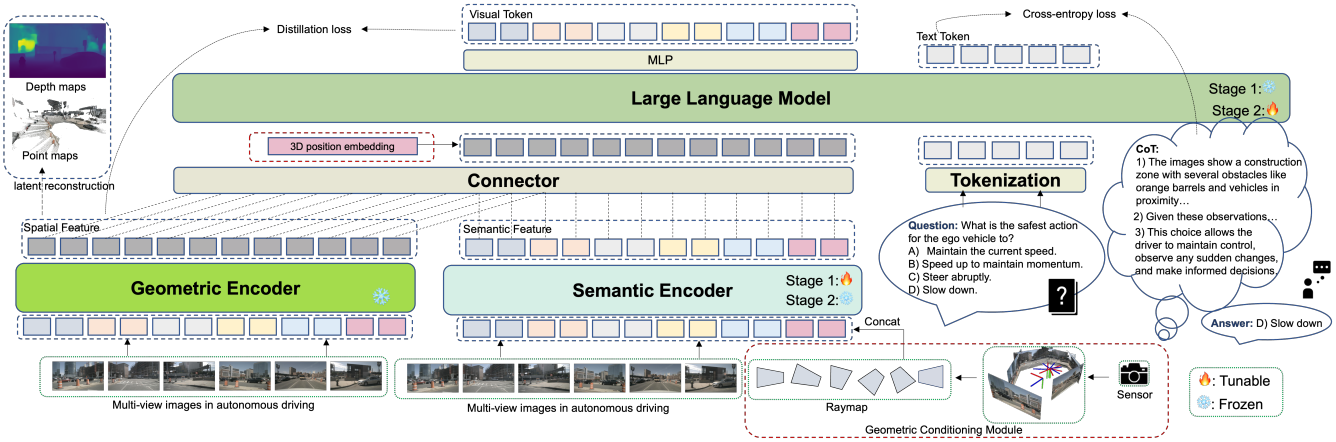

Fig. 3. Architecture of the proposed GSUC-VLM framework. It consists of a geometric encoder and a semantic encoder for extracting spatial and semantic features from multi-view images, respectively. A geometric conditioning module is introduced at the input of the semantic encoder to inject view-specific projection information, while a 3D position embedding is incorporated after feature concatenation to enhance spatial position understanding across views. The fused features are then processed by a Large Language Model to perform spatial understanding and question answering.

The position encoding of the i-th image can be represented as:

$$\mathbf{D}_i = [-(\mathbf{R})^T\mathbf{T}_i \times \boldsymbol{d}_i, \boldsymbol{d}_i]^T \in \mathbb{R}^6. \tag{5}$$

To integrate the raymaps tensor $\mathbf{D}_i$ with the vision transformer, we first resample both the RGB image and $\mathbf{D}_i$ to the required resolution using bilinear interpolation. The resampled image and raymap tensors are then partitioned into patches, and each patch is flattened and concatenated along the feature dimension. Thus, RGB data and ray direction information are combined into a 9-channel input (3 for RGB, 6 for ray directions) for each patch.

We redesign the patch embedding layer of ViT by setting the input channels to 9 to accommodate to additional ray direction data. For weight initialization, the weights for the RGB channels are copied from the pre-trained model, while the weights for the newly added ray direction channels are initialized to zero. This ensures the model behaves as the pre-trained version at the start of training, gradually incorporating ray direction information during fine-tuning without catastrophic forgetting.

C. Geometrically-Guided Feature Alignment

To sufficiently exploit the strong ability of the pre-trained geometric encoder in spatial modeling on multi-view images, we distill its knowledge into the visual encoder of the VLM. We introduce a distillation loss to align semantic features with spatial features within the visual representation. Specifically, we employ a cosine similarity-based distillation loss to align the semantic features $\mathrm{MLP}_1(f_{\mathrm{sem}})$ and spatial features $\mathrm{MLP}_2(f_{\mathrm{spa}})$. as formulated below.

$$\mathcal{L}_{\mathrm{distill}} = -\sum_{i=1}^{M} \cos\left(\mathrm{MLP}_1(f_{\mathrm{sem},i_p}), \mathrm{MLP}_2(f_{\mathrm{spa},i_p})\right), \tag{6}$$

where $\cos(\cdot,\cdot)$ denotes the cosine similarity function, f_{spa,i_p} and f_{sem,i_p} represent the spatial and semantic features of the p-th patch from the i-th view, respectively. This loss encourages the semantic features to be aligned with the spatial features in the embedding space.

We further incorporate the distillation loss $\mathcal{L}_{\mathrm{distill}}$ with a cross-entropy loss $\mathcal{L}_{\mathrm{cross}}$ from the language modeling objective to align the visual and language features for better VQA performance. The training loss is formulated as

$$\mathcal{L} = \mathcal{L}_{\mathrm{cross}} + \mathcal{L}_{\mathrm{distill}}. \tag{7}$$

A two-stage training strategy is adopted. In the first stage, we freeze the LLM and geometric encoder to train the semantic encoder to integrate spatial features with semantic features. In the second stage, we freeze the geometric and semantic encoders, and fine-tune only the LLM to better adapt to the VQA task.

IV. Experiments

A. Experimental Setting

1) Implementation Details: Our framework is built upon the backbone architectures of Qwen2.5-VL-3B [30] and VGGT [31], and has approximately 4B parameters. In each training step, a batch is randomly sampled from a single source within the mixed dataset. We refer to this as the first setting. The second setting fine-tunes only on DriveLMM-o1 and evaluates on NuScenes-QA. These two settings enable comprehensive comparisons with existing state-of-the-art VLMs under both fine-tuned and zero-shot scenarios. We adopt the Adam optimizer with a weight decay of 0.01 and a batch size of 64. The learning rate is linearly increased to 10^{-5} during a warm-up phase and subsequently decayed linearly. All experiments are conducted on 8 NVIDIA H200 GPUs (140GB), and each fine-tuning session lasts for approximately 12 hours.

2) Datasets: We train our network based on DriveLMM-o1 [14] and NuScenes-QA [13]. We evaluate our method on three representative autonomous driving VQA datasets: NuScenes-QA, DriveLMM-o1, and DriveBench [29]. These

TABLE I

PERFORMANCE COMPARISON OF DIFFERENT MODELS ON THE NUSCENES-QA [13] TEST SET. SINCE ALL THE ANSWERS TO THE QUESTIONS ARE WITHIN A SPECIFIED ANSWER SET, WE USE THE ACCURACY (%) OF GENERATED ANSWERS COMPARED TO THE STANDARD ANSWERS TO EVALUATE ALL RESULTS. H0 AND H1 DENOTE ZERO-HOP AND ONE-HOP REASONING QUESTIONS, RESPECTIVELY. ZS DENOTES ZERO-SHOT INFERENCE WITHOUT TASK-SPECIFIC FINE-TUNING.

Models	Exist			Count			Object			Status			Comparison			Acc
	H0	H1	All	H0	H1	All	H0	H1	All	H0	H1	All	H0	H1	All	
LLaMA-Adapter	34.2	6.3	19.3	5.0	0.1	2.7	23.7	4.6	7.6	9.8	11.3	10.8	2.6	1.5	1.6	9.6
LLaVA1.5-7B	38.9	51.9	45.8	7.7	7.6	7.7	10.5	7.4	7.8	7.0	9.9	9.0	64.5	50.8	52.1	26.2
Mulberry-7B	64.4	52.2	57.8	12.3	8.7	10.5	27.5	18.0	19.4	15.2	17.7	16.8	58.6	49.7	50.5	33.4
Qwen2.5-VL-3B	63.2	50.5	56.4	13.6	10.1	11.9	29.9	17.5	19.5	9.5	16.0	14.0	50.9	52.4	52.3	32.1
GSUC-VLM (ZS)	66.7	56.6	61.2	9.2	8.3	8.8	40.1	24.6	26.9	27.2	32.8	30.9	57.9	53.9	54.3	38.2
BEVDet	87.2	81.7	84.2	21.8	19.2	20.4	73.0	47.4	51.2	64.1	49.9	54.7	75.1	66.7	67.4	57.9
CenterPoint	87.7	82.3	84.8	22.5	19.1	20.8	71.3	49.0	52.3	66.6	56.3	59.8	**82.4**	**68.8**	**70.0**	59.5
MSMDFusion	89.0	82.3	85.4	23.4	21.1	22.2	**75.3**	50.6	54.3	**69.0**	56.2	60.6	78.8	**68.8**	69.7	60.4
LiDAR-LLM	79.1	70.6	74.5	15.3	14.7	15.0	59.6	34.1	37.8	53.4	42.0	45.9	67.0	57.0	57.8	48.6
PointLLM	80.2	77.1	78.5	16.1	16.5	16.3	57.4	33.6	37.1	47.7	41.7	43.8	77.3	61.8	63.2	50.2
GSUC-VLM	**89.7**	**84.1**	**86.7**	**25.5**	**24.3**	**24.9**	**75.3**	**55.0**	**57.9**	65.5	**58.1**	**60.7**	81.0	68.0	69.2	**62.0**

Fig. 4. Zero-shot Visualization Results on NuScenes-QA. We compared our method with Qwen2.5-VL-3B and LLaVA1.5-7B on the "Exist" and "Status" types. Since the comparison methods only output a single word after fine-tuning and lack reasoning and language capabilities, we visualize the zero-shot results of our method and the comparison methods. Our method accurately understands the relative position of targets in multi-view images, providing precise results along with a clear and complete chain of thought.

datasets collectively cover a wide range of spatial understanding tasks, including fine-grained localization, spatial understanding, and ego-motion analysis, based on multi-view visual inputs.

3) Evaluation Metrics: For closed-form answers such as Yes/No or multiple-choice questions where the answer lies within a predefined dictionary, we use accuracy as the evaluation metric. For open-ended answers, we adopt the system prompt provided in the original literature for scoring. All evaluations are conducted based on DeepSeek-R1-0528 [34].

B. Comparisons on DriveLMM-o1 and NuScenes-QA

We first compare our method with several state-of-the-art MLLMs on the NuScenes-QA dataset in Table I. The baselines include LLaMA-Adapter [35], LLaVA1.5 [36], Mulberry-7B [37], Qwen2.5-VL-3B [30], BEVDet [38], CenterPoint [39], MSMDFusion [40], LiDAR-LLM [16], and PointLLM [26]. Among them, BEVDet, CenterPoint, and MSMDFusion follow the fine-tuning protocol outlined in NuScenes-QA [13], using the MCAN [41] module for multi-modal fusion. LLaMA-Adapter, LLaVA1.5, Mulberry-7B and Qwen2.5-VL-3B are evaluated in a zero-shot setting, and the others are fine-tuned on NuScenes-QA.

Our method achieves superior zero-shot performance, especially in tasks involving object recognition and motion status understanding. After fine-tuning, our model further surpasses all baselines across most tasks, notably excelling in object-level recognition. This highlights the model's ability to retain rich semantic features while improving spatial understanding. On comparison-based questions, our model does not outperform MSMDFusion and CenterPoint. This may be attributed to their use of BEV point cloud representations to better preserve geometric precision and spatial relationships. To further demonstrate the spatial understanding capabilities of our model, we visualize zero-shot predictions on NuScenes-QA in Figure 4. Compared with Qwen2.5-VL-3B and LLaVA1.5-7B, our model provides more accurate responses grounded in multi-view visual context. It identifies object existence and states and produces coherent reasoning chains that reflect fine-grained spatial understanding.

In Table II, we further compared the performance of several state-of-the-art MLLMs on DriveLMM-o1,

TABLE II

PERFORMANCE COMPARISON WITH OPEN-SOURCE MODELS ON DRIVELMM-O1 [14]. ALL ANSWERS ARE SCORED ON DEEPSEEK-R1-0528 [34], USING THE SAME SCORING CRITERIA AS DRIVELMM-O1. ACCURACY IS COMPUTED FOR CLOSED-ANSWER QUESTIONS. OUR MODEL OUTPERFORMS OTHER OPEN-SOURCE MODELS AND IMPROVES ANSWER ACCURACY AND REASONING CAPABILITIES.

Model	Risk Assessment	Rule Adherence	Object Understanding	Relevance	Details	Overall Reasoning	Accuracy
Qwen2.5-VL-3B	51.6	65.2	48.8	54.1	47.8	56.2	37.51
Qwen2.5-VL-7B	51.2	62.1	50.5	54.7	46.0	56.5	37.81
Mulberry-7B	47.5	57.8	46.0	50.6	39.9	52.6	52.86
LLaVA-CoT-11B	49.1	60.7	48.1	51.9	43.2	54.2	49.27
LlamaV-o1	48.4	59.2	47.6	52.7	41.8	53.8	50.02
InternVL2.5-8B	50.5	61.6	48.4	53.5	47.1	55.3	54.87
DriveLMM-o1	58.8	69.9	60.2	63.8	52.1	63.0	62.36
GSUC-VLM-4B	**60.9**	**72.3**	**62.8**	**67.3**	**54.0**	**65.9**	**65.10**

TABLE III

PERFORMANCE COMPARISON IN PERCEPTION TASKS ON DRIVEBENCH [29]. AA: ACTION ALIGNMENT; MP: MOTION PRECISION; DCA: DRIVING CONTEXT APPROPRIATNESS; SA: SITUATIONAL AWARENESS; CC: CONCISENESS AND CLARITY; GR: GRAMMAR.

Models	AA	MP	DCA	SA	CC	GR	Total Score	MCQ Score
Qwen2.5-VL-3B	0.99	1.05	3.45	3.07	12.13	9.70	30.69	23.11
Qwen2.5-VL-7B	0.53	0.64	3.72	2.57	15.09	9.91	33.20	20.00
Mulberry-7B	1.21	0.97	4.75	3.51	12.12	9.62	32.54	36.70
LLaVA-CoT-11B	**1.34**	**0.98**	4.54	4.15	13.37	9.90	34.76	36.21
LlamaV-o1	1.24	0.41	4.71	2.90	15.35	**10.00**	35.41	23.31
InternVL2.5-8B	1.29	0.85	3.43	2.06	11.25	8.53	27.91	25.00
GSUC-VLM-4B	1.05	0.61	**4.96**	**4.92**	**15.50**	9.94	**37.60**	**39.42**

TABLE IV

PERFORMANCE COMPARISON WITH OPEN-SOURCE MODELS ON DRIVEBENCH [29]. OI: OBJECT IDENTIFICATION AND PRIORITY ORDER; SO: STATE OF THE OBJECT; RA: RECOMMENDED ACTION FOR EGO VEHICLE; ACC: MCQ SCORE OF PREDICTION TASK; AP: ACTION PREDICTION ACCURACY; RJ: REASONING AND JUSTIFICATION; CA: CONTEXTUAL AWARENESS AND SAFETY CONSIDERATIONS; CC: CONCISENESS AND CLARITY. AC: ANSWER CORRECTNESS; BU: BEHAVIORAL UNDERSTANDING AND DETAIL; RJ: REASONING AND JUSTIFICATION; CR: CONTEXTUAL RELEVANCE.

Models	Prediction					Planning					Behavior				
	OI	SO	RA	All	Acc	AP	RJ	CA	CC	All	AC	BU	RJ	CR	All
Qwen2.5-VL-3B	1.05	2.19	1.58	17.93	14.55	5.58	5.98	5.14	6.23	23.95	17.15	1.84	0.12	0.10	20.03
Qwen2.5-VL-7B	0.83	1.82	1.00	15.62	9.77	4.24	5.80	4.65	7.98	24.77	21.17	2.62	0.18	0.09	24.92
Mulberry-7B	1.23	1.10	0.98	15.93	8.76	8.57	**9.22**	6.25	7.28	32.66	15.05	5.61	5.52	4.80	41.00
LLaVA-CoT-11B	0.98	1.19	0.68	15.73	14.93	6.84	8.08	**6.80**	7.70	31.66	16.94	**6.60**	**6.66**	**9.99**	46.25
LlamaV-o1	1.23	1.23	1.07	17.97	8.74	6.84	5.30	3.87	7.33	24.41	12.35	1.46	5.57	0.06	20.74
InternVL2.5-8B	2.00	2.37	1.05	21.44	0.00	6.56	4.60	2.93	8.15	23.09	16.28	1.84	0.17	0.12	20.09
GSUC-VLM-4B	**2.63**	**2.42**	**2.69**	**24.80**	**17.20**	**10.91**	8.59	6.78	**9.00**	**37.66**	**21.20**	6.21	6.19	5.23	**49.02**

including Qwen2.5-VL-3B [30], Qwen2.5-VL-7B [30], Mulberry-7B [37], LLaVA-CoT-11B [42], LlamaV-o1 [43], InternVL2.5-8B [44], and DriveLMM-o1 [14]. The evaluation included accuracy on Multiple-Choice Questions (MCQ) questions as well as scores on all questions, with the scoring criteria encompassing Risk Assessment, Rule Adherence, Object Understanding, Relevance, and Missing Details, among other factors. All evaluations were performed using Deepseek-R1-0528. Our method significantly outperforms other approaches in terms of accuracy on multiple-choice questions and achieves notably higher semantic scores in the open-ended evaluation.

C. Zero-shot Results on DriveBench

We further evaluate the generalization capability of our method on the DriveBench dataset. Notably, neither our method nor the compared baselines are trained on DriveBench, and all baseline models remain consistent with those used in previous datasets. In Table III, we assess performance on perception tasks using both MCQ and Open-ended Questions, scored by the DeepSeek evaluator. Since MCQ questions often involve driving decisions, accuracy alone is insufficient to reflect actual performance. To address this, we adopt a composite scoring strategy, where accuracy serves as the primary metric, supplemented by the quality of the reasoning process. For open-ended questions, we follow the original DeepSeek evaluation metrics. Table III reports scores on key dimensions, the total score for open-ended questions (Total Score), and the final MCQ score (MCQ Score). Our method outperforms or closely matches other approaches on most metrics, demonstrating particularly strong results in object recognition and decision-making accuracy.

In Table IV, we further assess generalization across three

core task categories: *Prediction* (including both open-ended and MCQ questions), *Planning* (open-ended only), and *Behavior* (MCQ only). The table summarizes scores on representative indicators for each task. Our method consistently surpasses baselines in most categories, showing significant improvements in prediction and planning, while enhancing both semantic and spatial understanding. This lays a solid foundation for downstream autonomous driving tasks.

D. Ablation Study and Analysis

1) Effectiveness of the Proposed Architecture.: We first evaluated the impact of the dual semantic and geometric encoders, combined with distillation loss alignment. Table V presents a comparison of Qwen2.5-VL-3B's performance on NuScenes-QA, both before and after fine-tuning. Specifically, we compare the untrained Qwen2.5-VL-3B, the fine-tuned Qwen2.5-VL-3B-SFT, and the model incorporating dual encoders and distillation loss alignment, GSUC-VLM-4B w/o PFC. The results clearly demonstrate that fine-tuning significantly enhances the performance of Qwen2.5-VL-3B, particularly in tasks with stronger semantic content, such as target recognition. Moreover, after incorporating spatial features and alignment, the performance in the Status task shows a marked improvement, reflecting enhanced spatial understanding capabilities.

2) Effectiveness of Projective Feature Conditioning.: Next, we evaluated the contribution of Projective Feature Conditioning (PFC). As shown in the results, the introduction of PFC leads to performance improvements across most tasks, especially in the status task. Notably, the addition of PFC significantly improves the model's ability to understand the relative spatial relationships between objects in the scene, such as recognizing the position of a vehicle relative to the ego-vehicle, e.g., detecting that a vehicle is in front.

Furthermore, we validate the effectiveness of the 3D Position Embedding and the Geometric Conditioning Module in PFC. Specifically, GSUC-VLM-4B w/o PFC(D) denotes the model without the 3D Position Embedding, while GSUC-VLM-4B w/o PFC(G) denotes the model without the Geometric Conditioning Module. It can be observed that both components contribute to the final performance, with the Geometric Conditioning Module playing a particularly significant role.

3) Effectiveness of Feature Alignment.: We further validate the effectiveness of our Geometrically-Guided Feature Alignment. Specifically, GSUC-VLM-4B w/o FA denotes the proposed method without Geometrically-Guided Feature Alignment, i.e., training is conducted only in the second stage without the distillation-based alignment loss in the first stage. The results show that applying feature alignment consistently improves the metrics across different sub-tasks, with particularly notable gains in spatial position understanding related to status.

TABLE V

ABLATION STUDY OF GSUC-VLM-4B COMPONENTS ON NUSCENES-QA. QWEN2.5-VL-3B REPRESENTS THE UNTRAINED MODEL, QWEN2.5-VL-3B-SFT THE FINE-TUNED MODEL, GSUC-VLM-4B W/O FA THE PROPOSED METHOD WITHOUT GEOMETRICALLY-GUIDED FEATURE ALIGNMENT, GSUC-VLM-4B W/O PFC THE PROPOSED METHOD WITHOUT PROJECTIVE FEATURE CONDITIONING, AND GSUC-VLM-4B (FULL) THE COMPLETE MODEL PROPOSED IN THIS WORK.

Models	Exist	Object	Status	All
Qwen2.5-VL-3B	56.4	19.5	14.0	32.1
Qwen2.5-VL-3B-SFT	82.4	53.2	50.1	57.2
GSUC-VLM-4B w/o FA	86.1	56.9	57.6	61.5
GSUC-VLM-4B w/o PFC	85.4	54.5	58.0	60.9
GSUC-VLM-4B w/o PFC(D)	86.3	57.1	59.8	61.6
GSUC-VLM-4B w/o PFC(G)	85.9	55.6	58.3	61.3
GSUC-VLM-4B (full)	**86.7**	**57.9**	**60.7**	**62.0**

V. CONCLUSION

In this work, we present GSUC-VLM, a geometry-guided vision-language framework tailored for spatial understanding in autonomous driving. Our method effectively integrates multi-view semantic and spatial features without relying on point clouds or detection priors, addressing key limitations in prior VLM-based approaches. By introducing a dual-encoder architecture and a projective feature conditioning mechanism, GSUC-VLM enables precise spatial understanding and robust semantic alignment across views. Extensive experiments on multiple autonomous driving VQA benchmarks demonstrate that our approach achieves state-of-the-art performance and strong zero-shot generalization. These results highlight the potential of geometry-aware vision-language modeling for advancing safe and reliable autonomous driving systems.

REFERENCES

[1] C. Yang, Y. Chen, H. Tian, C. Tao, X. Zhu, Z. Zhang, G. Huang, H. Li, Y. Qiao, L. Lu, J. Zhou, and J. Dai, "BEVFormer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 17 830–17 839.

[2] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "BEVFormer: Learning bird's-eye-view representation from LiDAR-camera via spatiotemporal transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 3, pp. 2020–2036, 2025.

[3] S. Wang, Z. Yu, X. Jiang, S. Lan, M. Shi, N. Chang, J. Kautz, Y. Li, and J. M. Alvarez, "OmniDrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 22 442–22 452.

[4] L. Chen, O. Sinavski, J. Hünermann, A. Karnsund, A. J. Willmott, D. Birch, D. Maund, and J. Shotton, "Driving with LLMs: Fusing object-level vector modality for explainable autonomous driving," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 14 093–14 100.

[5] Y. Zhou, L. Huang, Q. Bu, J. Zeng, T. Li, H. Qiu, H. Zhu, M. Guo, Y. Qiao, and H. Li, "Embodied understanding of driving scenarios," in *Proceedings of the 18th European Conference on Computer Vision*, 2024, pp. 129–148.

[6] J. Xiang, T. Tao, Y. Gu, T. Shu, Z. Wang, Z. Yang, and Z. Hu, "Language models meet world models: Embodied experiences enhance language models," in *Advances in Neural Information Processing Systems 36*, 2023, pp. 75 392–75 412.

[7] A.-M. Marcu *et al.*, "LingoQA: Visual question answering for autonomous driving," in *Proceedings of the 18th European Conference on Computer Vision*, 2024, pp. 252–269.

[8] J.-B. Alayrac *et al.*, "Flamingo: a visual language model for few-shot learning," in *Advances in Neural Information Processing Systems 35*, 2022, pp. 23 716–23 736.

[9] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems 36*, 2023, pp. 34 892–34 916.

[10] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K.-Y. K. Wong, Z. Li, and H. Zhao, "DriveGPT4: Interpretable end-to-end autonomous driving via large language model," *IEEE Robotics and Automation Letters*, vol. 9, no. 10, pp. 8186–8193, 2024.

[11] W. Wang, J. Xie, C. Hu, H. Zou, J. Fan, W. Tong, Y. Wen, S. Wu, H. Deng, Z. Li, H. Tian, L. Lu, X. Zhu, X. Wang, Y. Qiao, and J. Dai, "DriveMLM: Aligning multi-modal large language models with behavioral planning states for autonomous driving," *arXiv preprint arXiv:2312.09245*, 2023.

[12] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, J. Beißwenger, P. Luo, A. Geiger, and H. Li, "DriveLM: Driving with graph visual question answering," in *Proceedings of the 18th European Conference on Computer Vision*, 2024, pp. 256–274.

[13] 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," in *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 2024, pp. 4542–4550.

[14] A. Ishaq, J. Lahoud, K. More, O. Thawakar, R. Thawkar, D. Dissanayake, N. Ahsan, Y. Li, F. S. Khan, H. Cholakkal, I. Laptev, R. M. Anwer, and S. Khan, "DriveLMM-o1: A step-by-step reasoning dataset and large multimodal model for driving scenario understanding," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 20 501–20 508.

[15] B. Jin, Y. Zheng, P. Li, W. Li, Y. Zheng, S. Hu, X. Liu, J. Zhu, Z. Yan, H. Sun, K. Zhan, P. Jia, X. Long, Y. Chen, and H. Zhao, "TOD3Cap: Towards 3D dense captioning in outdoor scenes," in *Proceedings of the 18th European Conference on Computer Vision*, 2024, pp. 367–384.

[16] S. Yang, J. Liu, R. Zhang, M. Pan, Z. Guo, X. Li, Z. Chen, P. Gao, H. Li, Y. Guo, and S. Zhang, "LiDAR-LLM: Exploring the potential of large language models for 3D LiDAR understanding," in *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, 2025, pp. 9247–9255.

[17] R. Tian, B. Li, X. Weng, Y. Chen, E. Schmerling, Y. Wang, B. Ivanovic, and M. Pavone, "Tokenize the world into object-level knowledge to address long-tail events in autonomous driving," in *Proceedings of the 8th Conference on Robot Learning*, 2025, pp. 3656–3673.

[18] Z. Zhang, X. Li, Z. Xu, W. Peng, Z. Zhou, M. Shi, and S. Huang, "MPDrive: Improving spatial understanding with marker-based prompt learning for autonomous driving," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 12 089–12 099.

[19] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[20] H. Huang, Y. Chen, Z. Wang, R. Huang, R. Xu, T. Wang, L. Liu, X. Cheng, Y. Zhao, J. Pang, and Z. Zhao, "Chat-Scene: Bridging 3D scene and large language models with object identifiers," in *Advances in Neural Information Processing Systems 37*, 2024, pp. 113 991–114 017.

[21] R. Fu, J. Liu, X. Chen, Y. Nie, and W. Xiong, "Scene-LLM: Extending language model for 3D visual reasoning," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025, pp. 2195–2206.

[22] Z. Qi, Z. Zhang, Y. Fang, J. Wang, and H. Zhao, "GPT4Scene: Understand 3D scenes from videos with vision-language models," in *The 14th International Conference on Learning Representations*, 2026.

[23] S. Chen, X. Chen, C. Zhang, M. Li, G. Yu, H. Fei, H. Zhu, J. Fan, and T. Chen, "LL3DA: Visual interactive instruction tuning for omni-3D understanding reasoning and planning," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 26 428–26 438.

[24] C. Zhu, T. Wang, W. Zhang, J. Pang, and X. Liu, "LLaVA-3D: A simple yet effective pathway to empowering lmms with 3d capabilities," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 4295–4305.

[25] Y. Hong, H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, and C. Gan, "3D-LLM: Injecting the 3D world into large language models," in *Advances in Neural Information Processing Systems 36*, 2023, pp. 20 482–20 494.

[26] R. Xu, X. Wang, T. Wang, Y. Chen, J. Pang, and D. Lin, "PointLLM: Empowering large language models to understand point clouds," in *Proceedings of the 18th European Conference on Computer Vision*, 2024, pp. 131–147.

[27] T. Wang, E. Xie, R. Chu, Z. Li, and P. Luo, "DriveCoT: Integrating chain-of-thought reasoning with end-to-end driving," *arXiv preprint arXiv:2403.16996*, 2024.

[28] Z. Xu, Y. Bai, Y. Zhang, Z. Li, F. Xia, K.-Y. K. Wong, J. Wang, and H. Zhao, "DriveGPT4-V2: Harnessing large language model capabilities for enhanced closed-loop autonomous driving," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 17 261–17 270.

[29] S. Xie, L. Kong, Y. Dong, C. Sima, W. Zhang, Q. A. Chen, Z. Liu, and L. Pan, "Are VLMs ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives," in *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[30] Qwen Team, Alibaba Group, "Qwen2.5-VL technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[31] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "VGGT: Visual geometry grounded transformer," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 5294–5306.

[32] R. Gao, A. Hoł yński, P. Henzler, A. Brussee, R. Martin-Brualla, P. Srinivasan, J. T. Barron, and B. Poole, "CAT3D: Create anything in 3D with multi-view diffusion models," in *Advances in Neural Information Processing Systems 37*, 2024, pp. 75 468–75 494.

[33] H. Jin, H. Jiang, H. Tan, K. Zhang, S. Bi, T. Zhang, F. Luan, N. Snavely, and Z. Xu, "LVSM: A large view synthesis model with minimal 3D inductive bias," in *The 14th International Conference on Learning Representations*, 2026.

[34] DeepSeek-AI, "DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[35] R. Zhang, J. Han, C. Liu, P. Gao, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, and Y. Qiao, "Llama-adapter: Efficient fine-tuning of language models with zero-init attention," *arXiv preprint arXiv:2303.16199*, 2023.

[36] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 26 296–26 306.

[37] H. Yao *et al.*, "Mulberry: Empowering MLLM with o1-like reasoning and reflection via collective monte carlo tree search," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[38] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "BEVDet: High-performance multi-camera 3d object detection in bird-eye-view," *arXiv preprint arXiv:2112.11790*, 2021.

[39] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 784–11 793.

[40] Y. Jiao, Z. Jie, S. Chen, J. Chen, L. Ma, and Y.-G. Jiang, "MSMDFusion: Fusing LiDAR and camera at multiple scales with multi-depth seeds for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 643–21 652.

[41] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6281–6290.

[42] G. Xu, P. Jin, Z. Wu, H. Li, Y. Song, L. Sun, and L. Yuan, "LLaVA-CoT: Let vision language models reason step-by-step," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 2087–2098.

[43] O. Thawakar, D. Dissanayake, K. P. More, R. Thawkar, A. Heakl, N. Ahsan, Y. Li, I. Z. M. Zumri, J. Lahoud, R. M. Anwer, H. Cholakkal, I. Laptev, M. Shah, F. S. Khan, and S. Khan, "LlamaV-o1: Rethinking step-by-step visual reasoning in LLMs," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 24 290–24 315.

[44] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, B. Li, P. Luo, T. Lu, Y. Qiao, and J. Dai, "InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 24 185–24 198.