LOOK TO UNDERSTAND: SMALL-LARGE MODEL COLLABORATION FRAMEWORK FOR TRAFFIC SCENE UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) have demonstrated remarkable capabilities in traffic scene understanding and visual question answering. However, large VLMs still face several challenges in complex traffic scenes, including redundant visual information, insufficient utilization of spatial information, and unclear correspondence between question semantics and relevant visual regions. To address these challenges, we propose Look to Understand, a Small-Large Model Collaborative Framework that leverages small models for visual perception and large VLMs for scene semantic understanding and visual question answering. First, we design a Small Model-Guided Perception Chain for structured visual perception. Given a traffic scene image, a lightweight vision model extracts object categories and spatial bounding boxes. The detection results are encoded as structured perception tokens, providing the large VLM with explicit object category and spatial location information. Furthermore, we propose Small Model-Guided Question-Region Contrastive Learning, which leverages the object regions provided by the lightweight vision model to model their relevance to the question semantics and aggregate features. The question and visual region representations are then projected into a shared latent space, where contrastive learning is employed to constrain their semantic correspondence. Experiments on RoadSceneVQA and DRIVINGVQA demonstrate the effectiveness of the proposed framework.

1 INTRODUCTION

In recent years, Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) have demonstrated strong capabilities in natural language understanding, visual question answering, cross-modal reasoning, and complex decision-making, benefiting from large-scale parameters and massive pre-trained data. However, their substantial computational and storage requirements, particularly the high cost of processing high-resolution visual inputs, pose challenges for resource-constrained and real-time applications. In contrast, small models, such as object detection models and lightweight vision-language models, offer lower computational costs and faster inference, enabling efficient extraction of task-relevant information, but are generally limited in semantic understanding and complex reasoning due to their smaller model capacity. Therefore, large and small models exhibit complementary strengths, motivating Small-Large Model Collaboration to move beyond traditional one-way knowledge distillation toward collaborative paradigms that jointly exploit the strong reasoning capabilities of large models and the efficient perception capabilities of small models Chen et al. (2025); Li et al. (2025); Wang et al. (2026).

Existing research on Small-Large Model Collaboration has explored knowledge transfer, collaborative training, model routing, and inference-time cooperation. Early methods mainly transfer knowledge from large models to small models through knowledge distillation, while recent studies have investigated bidirectional collaboration, where small models can assist large-model training through predictions, confidence scores, and high-value training signals Rawat et al. (2024); Liu et al. (2024); Lu et al. (2024). Meanwhile, model routing and hierarchical reasoning dynamically allocate tasks among heterogeneous models according to task difficulty or model confidence, aiming to balance computational efficiency and task performance Zhou et al. (2026c); Wang et al. (2025). When extended to multimodal understanding, however, an important question is what information small

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

models should provide to large models. Existing approaches mainly exchange answers, soft labels, textual reasoning signals, or model-level features, while making limited use of the fine-grained visual perception capabilities of specialized small models. In particular, object detection models can efficiently provide explicit object categories and spatial locations, offering multimodal large models structured visual priors beyond conventional visual features. This is especially important for visual question answering, where complex scenes contain numerous objects and background regions but a specific question may concern only a small subset of them. Although recent studies have explored text-guided visual Token selection and visual evidence retrieval to retain task-relevant visual information Cai et al. (2025); Zhang et al. (2026b), these Token-based methods mainly focus on selecting which visual Tokens to retain and do not necessarily establish explicit correspondence between question semantics and specific target objects. This motivates the exploration of small model-guided structured visual perception and question-region alignment for multimodal understanding.

Furthermore, the mismatch between visual information and language-driven reasoning can limit the utilization of fine-grained visual evidence in multimodal models, as irrelevant visual information may be processed together with task-critical evidence while the correspondence between question semantics and visual regions remains implicit. Recent studies have therefore introduced structured visual primitives, such as visual evidence, points, and bounding boxes, to provide more explicit spatial grounding for multimodal reasoning Zhang et al. (2026b); Lu et al. (2026). This challenge is particularly pronounced in traffic scene understanding, where complex scenes contain numerous heterogeneous objects and spatial interactions, while specific questions often concern only a subset of objects and their relationships. RoadSceneVQA further demonstrates that traffic visual question answering requires comprehensive understanding across object perception, scene context, traffic behavior, and traffic rules Guan et al. (2026). These observations motivate us to reconsider the role of small models in multimodal collaboration: rather than serving only as independent predictors or knowledge-distillation sources, lightweight vision models can act as visual perception assistants for multimodal large models by identifying task-relevant object categories and their spatial locations. Accordingly, we introduce a small model-guided perception process in which lightweight models determine what visual information should be perceived, while object detection models determine where the relevant evidence is located, and the resulting structured perception information is provided to the multimodal large model for subsequent visual understanding and reasoning. This design is related to recent efforts on fine-grained multimodal collaboration, where small models provide task-specific visual heuristics to assist large models Sun et al. (2025).

Building upon these observations, we propose Look to Understand: Small-Large Model Collaboration for Traffic Scene Understanding, a fine-grained framework that integrates lightweight models with multimodal large models at semantic and feature levels. Specifically, we introduce a Small Model-Guided Perception Chain to provide structured object categories and spatial information, and a Small Model-Guided Question-Region Contrastive Learning mechanism to enhance the semantic correspondence between task-relevant visual regions and question representations. This collaboration enables lightweight models to provide explicit visual perception assistance for multimodal traffic scene understanding.

The main contributions are summarized as follows:

1) We propose a **Small Model-Guided Perception Chain** that enables collaboration among a lightweight vision model, an object detection model, and a multimodal large model for structured visual perception.

2) We develop **Small Model-Guided Question-Region Contrastive Learning**, which leverages object regions provided by the lightweight vision model to model their relevance to question semantics and enhance their semantic correspondence through contrastive learning.

3) We establish a **fine-grained Small-Large Model Collaborative framework** that integrates visual perception, spatial information, and visual features. Experiments on **RoadSceneVQA** and **DRIVINGVQA** demonstrate the effectiveness of the proposed framework.

2 RELATED WORK

2.1 LARGE-SMALL MODEL COLLABORATIVE LEARNING AND INFERENCE

Recent research on large-small model collaboration has gradually evolved from unidirectional knowledge distillation toward collaborative training, dynamic inference, and multimodal cooperation. Liu et al. (2024) and Lu et al. (2024) explore knowledge exchange and collaborative training between models of different scales, demonstrating the potential of mutual assistance beyond conventional teacher-student paradigms. For efficient multimodal inference, Cai et al. (2025), Zhang et al. (2026b), and Yang et al. (2026) investigate query-aware visual information selection and edge-cloud collaborative inference, aiming to reduce redundant visual computation and communication costs. In visual question answering, Sun et al. (2025) and Chen et al. (2026) further explore fine-grained collaboration between large and small vision-language models. These studies indicate a growing trend toward collaborative and query-guided multimodal reasoning, where models of different scales provide complementary capabilities for efficient visual understanding.

2.2 TRAFFIC SCENE UNDERSTANDING

In intelligent transportation, visual understanding has evolved from conventional perception tasks toward open-ended scene understanding and multimodal reasoning. Guan et al. (2026) introduce RoadSceneVQA for roadside visual question answering, while Vishal et al. (2025) and Zhou et al. (2026b) explore reasoning-rich traffic video understanding and structured scene modeling. Beyond traffic-specific studies, Lu et al. (2026) introduce points and bounding boxes as explicit visual primitives, Zhu et al. (2026) decompose visual reasoning into multiple stages, and Zhang et al. (2026b) enable active acquisition of task-relevant visual evidence. Zhou et al. (2026a) incorporate multimodal context into visual adaptation, while Zhang et al. (2026a) explore efficient and interpretable visual cognition for automatic train operation. Overall, these studies advance traffic scene understanding toward visual question answering, structured visual reasoning, and explicit spatial-semantic interaction.

2.3 RESEARCH GAP

In summary, existing research has advanced multimodal intelligent systems from two complementary perspectives: large-small model collaboration and traffic scene understanding. Large-small model collaboration has progressed from knowledge distillation to collaborative training, dynamic routing, query-guided visual token selection, and edge-cloud inference (Liu et al., 2024; Lu et al., 2024; Cai et al., 2025; Zhang et al., 2026b; Yang et al., 2026), while traffic scene understanding has expanded toward visual question answering, behavior understanding, visual evidence acquisition, and spatially grounded reasoning (Guan et al., 2026; Vishal et al., 2025; Zhou et al., 2026b; Lu et al., 2026; Zhu et al., 2026). However, fine-grained collaboration between lightweight visual models and multimodal large language models remains insufficiently explored. Existing methods mainly focus on visual token selection or evidence acquisition, with limited use of structured object information such as categories and spatial locations. Moreover, question-guided selection of task-relevant visual targets and their integration into MLLM reasoning remain underexplored, motivating the development of question-guided lightweight visual perception and multimodal reasoning collaboration.

3 METHODOLOGY

3.1 OVERALL FRAMEWORK

To address the issues of visual information redundancy, insufficient utilization of object-level spatial information, and unclear correspondence between question semantics and task-relevant visual regions in multimodal large language models (MLLMs) for complex traffic scenarios, we propose a large-small model collaborative framework, termed Small Model-Guided Collaboration (SGC) framework, for traffic scene understanding. The framework employs lightweight models as auxiliary modules for visual perception and an MLLM as the primary model for scene semantic understanding and visual question answering. By introducing structured perceptual information and

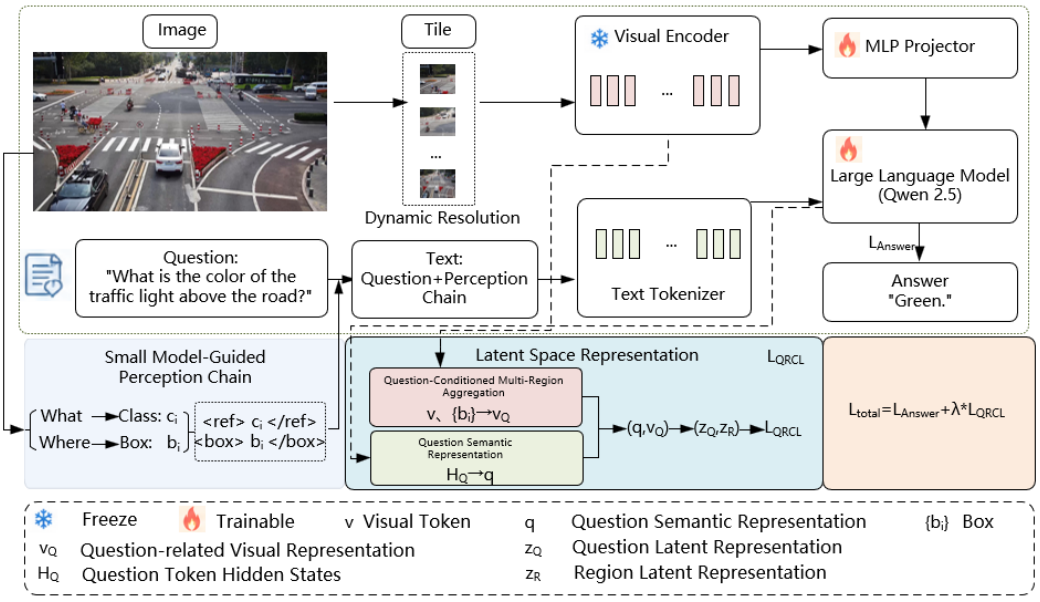


Figure 1: Overview of the proposed **Small Model-Guided Collaboration (SGC)** framework.

question-guided visual region alignment, the framework enables effective collaboration between lightweight models and MLLMs for fine-grained visual understanding.

3.2 Small Model-Guided Perception Chain

Complex traffic scenes contain numerous visual objects, including vehicles, pedestrians, and road infrastructure, while only a subset of these objects may be relevant to the subsequent visual understanding task. To provide explicit object-level visual information for the MLLM, we develop a structured visual perception method based on the **Perception Chain**. The method employs a lightweight vision model to extract object categories and spatial locations directly from the input image. The resulting detection information is then converted into structured perceptual Tokens and jointly provided to the MLLM together with the original image. In this way, the lightweight vision model serves as an auxiliary perception module that provides explicit object-level spatial priors for subsequent multimodal understanding.

Given an input image I, a lightweight vision model f_{vis} is first employed to extract the objects present in the scene:

$$D = f_{\text{vis}}(I) = \{(c_i, b_i)\}_{i=1}^{N}, \tag{1}$$

where c_i denotes the category of the i-th detected object, $b_i = [x_i^1, y_i^1, x_i^2, y_i^2]$ denotes its corresponding spatial bounding box, and N represents the total number of detected objects. Unlike conventional visual feature extraction, the proposed perception process explicitly preserves both the semantic category and spatial location of each detected object.

The detection results are further converted into structured perceptual Tokens. For the i-th detected object, its structured representation is defined as:

$$T_i = [\text{ref}], c_i, [\text{box}], b_i. \tag{2}$$

The complete perceptual sequence is then represented as:

$$T = \{T_1, T_2, \ldots, T_N\}. \tag{3}$$

Each structured perceptual Token contains both the object category and its spatial bounding box, thereby transforming the two-dimensional spatial information extracted by the lightweight vision model into explicit semantic representations that can be directly processed by the MLLM. Finally,

the original image I and the structured perceptual sequence T are jointly provided to the MLLM. The MLLM can therefore exploit both the original visual features and the explicit object-level spatial priors provided by the lightweight vision model for subsequent visual understanding and reasoning.

3.3 Small Model-Guided Question-Region Contrastive Learning

After obtaining the object categories and spatial information provided by the Perception Chain, different detected objects may exhibit different levels of semantic relevance to the current question. Therefore, object categories and bounding boxes alone are insufficient to explicitly characterize the fine-grained correspondence between the question and specific visual regions. To address this issue, we further propose a **Question-Region Contrastive Learning** method for fine-grained vision-language alignment. The method employs question semantics to guide the aggregation of candidate visual region features and uses bidirectional contrastive learning to explicitly constrain the cross-modal correspondence between question representations and task-relevant visual region representations.

Given a question

$$Q = \{w_1, w_2, \ldots, w_L\}, \tag{4}$$

a language encoder is first used to obtain the hidden representations of question Tokens:

$$H_Q = [h_1, h_2, \ldots, h_L], \tag{5}$$

where h_i denotes the semantic representation of the i-th question Token. A question attention mechanism is then employed to model the importance of different Tokens:

$$\alpha_i = \frac{\exp(g(h_i))}{\sum_{j=1}^{L} \exp(g(h_j))}, \tag{6}$$

where $g(\cdot)$ denotes a learnable attention mapping function, and α_i represents the importance of the i-th Token to the overall question semantics. The question-level semantic representation is subsequently obtained as:

$$q = \sum_{i=1}^{L} \alpha_i h_i. \tag{7}$$

Meanwhile, given the object bounding box set obtained from the Perception Chain,

$$B = \{b_1, b_2, \ldots, b_N\}, \tag{8}$$

the corresponding region features are extracted from the visual feature map:

$$R = \{r_1, r_2, \ldots, r_N\}, \tag{9}$$

where r_i denotes the visual representation of the region corresponding to the i-th detected object. The semantic relevance between the question and each candidate visual region is then computed as:

$$s_i = q^T W_r r_i, \tag{10}$$

where W_r denotes a learnable parameter matrix, and s_i represents the relevance between the question semantics and the i-th visual region. A Softmax function is further applied to obtain the region attention weights:

$$\beta_i = \frac{\exp(s_i)}{\sum_{j=1}^{N} \exp(s_j)}, \tag{11}$$

where β_i denotes the attention weight assigned to the i-th visual region. The candidate regions are then aggregated according to these weights to obtain the question-relevant visual representation:

$$v_Q = \sum_{i=1}^{N} \beta_i r_i. \tag{12}$$

To establish cross-modal correspondence between question semantics and task-relevant visual regions, Question and Region projection heads, denoted by $f_Q(\cdot)$ and $f_R(\cdot)$, respectively, are used to map the two modalities into a shared latent space:

$$z_Q = f_Q(q), \qquad z_R = f_R(v_Q). \tag{13}$$

The projected representations are then L_2-normalized:

$$\hat{z}_Q = \frac{z_Q}{\|z_Q\|_2}, \qquad \hat{z}_R = \frac{z_R}{\|z_R\|_2}. \tag{14}$$

For a training batch containing B samples, the Question–Region similarity matrix is constructed as:

$$S_{ij} = \frac{\hat{z}_{Q_i}^T \hat{z}_{R_j}}{\tau}, \tag{15}$$

where τ denotes the temperature coefficient. The question and visual region corresponding to the same sample are treated as a positive pair, while question-region pairs from different samples are treated as negative pairs. Based on this formulation, bidirectional contrastive learning objectives are constructed in the Region-to-Question and Question-to-Region directions:

$$\mathcal{L}_{R\to Q} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(S_{ii})}{\sum_{j=1}^{B}\exp(S_{ij})}, \tag{16}$$

$$\mathcal{L}_{Q\to R} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(S_{ii})}{\sum_{j=1}^{B}\exp(S_{ji})}. \tag{17}$$

The final bidirectional Question–Region alignment loss is defined as:

$$\mathcal{L}_{\mathrm{QRCL}} = \frac{1}{2}\left(\mathcal{L}_{R\to Q} + \mathcal{L}_{Q\to R}\right). \tag{18}$$

3.4 Training Objective

Within the proposed large-small model collaborative framework, the model jointly learns traffic scene visual question answering and the fine-grained correspondence between question semantics and task-relevant visual regions. Given a traffic scene image I, a question Q, and the ground-truth answer Y, the MLLM learns to generate the target answer from the visual information and question semantics. The visual question answering loss is defined as:

$$\mathcal{L}_{\mathrm{Answer}} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, I, Q, T), \tag{19}$$

where y_t denotes the t-th answer Token, $y_{<t}$ denotes the generated answer sequence before the current Token, and T denotes the structured perceptual sequence generated by the Perception Chain.

The Question–Region alignment loss is further incorporated into the visual question answering training process. By jointly optimizing the visual question answering objective and the cross-modal alignment objective, the model can learn answer generation while establishing explicit correspondence between question semantics and task-relevant visual regions. The overall training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{Answer}} + \lambda \mathcal{L}_{\mathrm{QRCL}}. \tag{20}$$

where $\mathcal{L}_{VQA}$ denotes the visual question answering generation loss, $\mathcal{L}_{QR}$ denotes the Question–Region alignment loss, and λ is the weighting coefficient of the alignment loss. Through joint optimization, the task-relevant visual information provided by the lightweight model can effectively complement the visual understanding and language reasoning capabilities of the MLLM, thereby enhancing its understanding of key objects and their spatial relationships in complex traffic scenes and improving its fine-grained visual question answering capability with respect to task-relevant visual regions.

4 EXPERIMENTS

Datasets. We conduct experiments on two traffic scene visual question answering datasets, RoadSceneVQA and DRIVINGVQA. The RoadSceneVQA dataset contains 34,736 question-answering

Table 1: Comparison of Different Models on the RoadSceneVQA Dataset.

Model	EM ↑	ROUGE-L ↑	BLEU-4 ↑	CIDEr ↑	METEOR ↑	SPICE ↑
Qwen2.5-VL (NSFT)	0.015	0.184	0.026	0.418	0.151	0.118
Qwen2.5-VL	0.152	0.366	0.183	1.750	**0.268**	**0.213**
MiniCPM-o 2.6	0.178	0.377	0.199	1.761	0.261	0.205
DeepSeek-VL-7B	0.121	0.334	0.163	1.566	0.239	0.197
SGC	**0.201**	**0.396**	**0.213**	**1.829**	0.267	0.209

samples, including 30,059 samples for training and 4,677 samples for testing. The DRIVINGVQA dataset contains 3,931 samples, with 2,752 samples used for training and 1,179 samples used for testing. Both datasets cover various traffic scene question answering tasks, including object recognition, object state understanding, spatial relationship reasoning, and traffic rule-related questions, enabling the evaluation of visual understanding and semantic reasoning capabilities in complex traffic scenarios from multiple perspectives.

Comparison Models and Experimental Settings. To evaluate the effectiveness of the proposed method, we select representative multimodal large language models as the base models, including Qwen2.5-VL-7B, MiniCPM-o 2.6-8B, DeepSeek-VL-7B, and InternVL3-8B. Each base model is fine-tuned according to its corresponding model architecture using the same data splits. Based on these base models, different experimental configurations are established for the two proposed core modules, including the base model, the base model with Perception Chain, the base model with Question-Region Contrastive Learning (QRCL), and the complete model equipped with both modules. These configurations are used to analyze the respective contributions of structured visual perception and question-region alignment to traffic scene visual question answering.

Training Settings. Following the training parameter settings of RoadSceneVQA Guan et al. (2026), we fine-tune the pretrained multimodal large language models for 1 epoch with an initial learning rate of 1×10^{-5}. The AdamW optimizer is employed for parameter optimization, together with a cosine learning rate scheduler and a warm-up strategy to improve training stability. For InternVL3-8B, training is conducted on a single GPU with a per-device batch size of 2 and 2 gradient accumulation steps, resulting in an effective batch size of 4. During training, bfloat16 precision, Flash Attention, and Gradient Checkpointing are employed to reduce memory consumption and improve training efficiency. The vision encoder is kept frozen, while the language model and the trainable modules introduced in this work are optimized. The maximum sequence length is set to 4096, with a warm-up ratio of 0.03 and a weight decay of 0.01.

For the proposed Perception Chain, the parameters of the lightweight vision model remain fixed during training, and only the object categories and spatial locations produced by the model are used for subsequent multimodal visual understanding. For Question-Region Contrastive Learning, the question projection head and region projection head map the question representation and candidate visual region representations into a shared latent space, respectively. The projection dimension is set to 256, and the temperature coefficient is set to $\tau = 0.07$. The weighting coefficient of the Question-Region alignment loss is set to $\lambda = 0.7$.

Evaluation Metrics. Following the training parameter settings of RoadSceneVQA Guan et al. (2026), we comprehensively evaluate the accuracy, textual consistency, and semantic similarity of model responses in traffic scene visual question answering using six evaluation metrics: Exact Match (EM), BLEU-4, ROUGE-L, METEOR, CIDEr, and SPICE.

Overall Experimental Results. We first evaluate different models on the RoadSceneVQA test set, as shown in Table 1. The results show that the proposed **Small Model-Guided Collaboration (SGC)** framework achieves the highest performance among the compared models on most evaluation metrics. Specifically, SGC achieves an EM score of 0.201, a ROUGE-L score of 0.396, a BLEU-4 score of 0.213, and a CIDEr score of 1.829, outperforming the corresponding baseline models. For METEOR and SPICE, SGC achieves scores of 0.267 and 0.209, respectively, which are comparable to the baseline results. These results demonstrate the effectiveness of incorporating structured visual perception and question-region alignment into multimodal large language models for traffic scene visual question answering. Compared with the pretrained Qwen2.5-VL without

Table 2: Comparison of Different Models on the DRIVINGVQA Dataset.

Model	EM ↑
Qwen2.5-VL	0.456
MiniCPM-o 2.6	0.279
DeepSeek-VL-7B	0.406
SGC	**0.465**

task-specific fine-tuning (NSFT), task-specific fine-tuning substantially improves the performance of Qwen2.5-VL across all evaluation metrics, indicating the importance of adapting the pretrained multimodal model to the traffic scene visual question answering task. Among the fine-tuned baseline models, Qwen2.5-VL, MiniCPM-o 2.6, and DeepSeek-VL-7B exhibit different performance characteristics across the evaluation metrics. In comparison, SGC further improves the overall performance, particularly in EM, ROUGE-L, BLEU-4, and CIDEr, indicating that the proposed large-small model collaborative framework can enhance the model's ability to utilize task-relevant visual information for traffic scene understanding.

We further evaluate the models on the DRIVINGVQA test set, with the results presented in Table 2. DRIVINGVQA is formulated as a multiple-choice visual question answering dataset, where each question is associated with a predefined set of candidate answers. Therefore, Exact Match (EM) is adopted as the sole evaluation metric, measuring whether the predicted choice exactly matches the ground-truth answer. SGC achieves an EM score of 0.465, compared with 0.456 for Qwen2.5-VL, 0.279 for MiniCPM-o 2.6, and 0.406 for DeepSeek-VL-7B. The results show that SGC obtains the highest EM score among the compared models on the DRIVINGVQA dataset, further demonstrating its effectiveness in multiple-choice traffic scene visual question answering.

Overall, the results on both RoadSceneVQA and DRIVINGVQA indicate that the proposed SGC framework can improve the performance of multimodal large language models across different traffic scene visual question answering datasets. By combining lightweight model-guided structured visual perception with question-region semantic alignment, SGC provides complementary visual information and enhances the model's ability to identify and utilize task-relevant visual content.

The performance differences among these configurations further provide insights into the complementarity between the structured spatial perception information provided by the lightweight vision model and the semantic understanding capability of the multimodal large language model. Meanwhile, Question-Region Contrastive Learning uses question semantics to guide the aggregation of candidate visual regions and employs contrastive learning to constrain the cross-modal correspondence between question representations and relevant visual regions, thereby enhancing the model's utilization of task-relevant visual content.

Ablation Studies. To investigate the contributions of the proposed components, we conduct ablation experiments on the RoadSceneVQA dataset, with the results presented in Table 3. The **Baseline** denotes the base multimodal large language model without the proposed structured perceptual information or Question-Region Contrastive Learning. We first introduce the Perception Chain to the baseline model, denoted as **+ Perception Chain**, and then incorporate both Perception Chain and Question-Region Contrastive Learning to obtain the complete **SGC** framework. Compared with the Baseline, introducing Perception Chain improves the performance across all six evaluation metrics. Specifically, EM increases from 0.188 to 0.191, ROUGE-L from 0.385 to 0.390, BLEU-4 from 0.204 to 0.206, CIDEr from 1.764 to 1.785, METEOR from 0.262 to 0.267, and SPICE from 0.205 to 0.210. These improvements indicate that providing structured object categories and spatial locations through the lightweight vision model can enhance the multimodal model's utilization of explicit object-level visual information.

When both Perception Chain and Question-Region Contrastive Learning are incorporated, the complete SGC framework further improves the overall performance. SGC achieves the best results on EM, ROUGE-L, BLEU-4, and CIDEr, with scores of 0.201, 0.396, 0.213, and 1.829, respectively. For METEOR, SGC achieves the same score of 0.267 as the Perception Chain variant, while its SPICE score of 0.209 is slightly lower than 0.210. Nevertheless, SGC achieves consistently strong performance across the evaluation metrics, indicating that Question-Region Contrastive Learning

Table 3: Ablation study of different components on the RoadSceneVQA dataset.

Model	EM ↑	ROUGE-L ↑	BLEU-4 ↑	CIDEr ↑	METEOR ↑	SPICE ↑
Baseline	0.188	0.385	0.204	1.764	0.262	0.205
+ Perception Chain	0.191	0.390	0.206	1.785	0.267	**0.210**
SGC	**0.201**	**0.396**	**0.213**	**1.829**	**0.267**	0.209

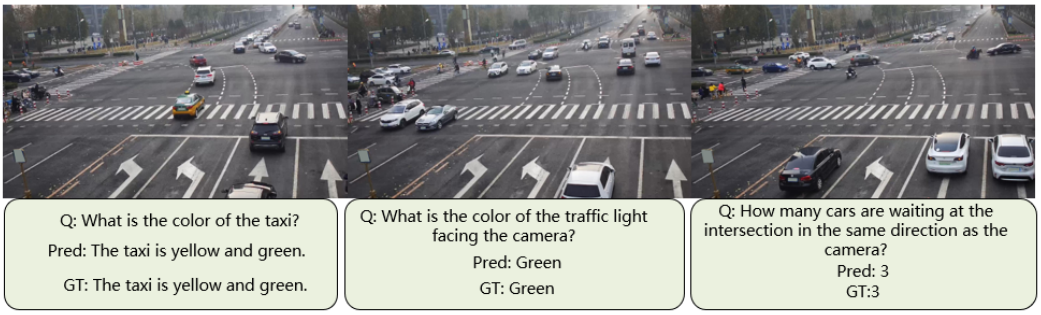


Figure 2: Visualization of the proposed SGC framework on representative traffic scenes.

provides additional benefits by encouraging the model to establish finer-grained semantic correspondence between the input question and task-relevant visual regions. Overall, the ablation results demonstrate that the structured perceptual information provided by Perception Chain contributes to improved traffic scene understanding, while the further incorporation of Question-Region Contrastive Learning enhances the model's ability to associate question semantics with relevant visual regions. The complementary effects of these two components lead to the improved overall performance of the proposed SGC framework.

Visualization. To provide an intuitive analysis of the impact of the proposed method on the visual understanding process, we further select representative traffic scenes for visualization, as shown in Fig. 2.

Overall, through quantitative experiments, ablation studies, and visualization analysis, we evaluate the proposed Look to Understand framework from multiple perspectives, including model performance, component contributions, and the visual understanding process, and further analyze the collaborative effects between the lightweight vision model and the multimodal large language model.

5 Conclusion

This paper presents Look to Understand: Small-Large Model Collaboration for Traffic Scene Understanding, a collaborative framework that integrates lightweight vision models with multimodal large language models (MLLMs) for fine-grained traffic scene understanding. The proposed framework introduces Small Model-Guided Perception Chain and Small Model-Guided Question-Region Contrastive Learning (QRCL) to provide structured visual information and enhance question-region correspondence. Experiments on RoadSceneVQA and DRIVINGVQA demonstrate the effectiveness of the proposed Small Model-Guided Collaboration (SGC) framework.

References

Kaitong Cai, Jusheng Zhang, Jing Yang, Yijia Fan, Pengtao Xie, Jian Wang, and Keze Wang. Flashvlm: Text-guided visual token selection for large multimodal models. *arXiv preprint arXiv:2512.20561*, 2025.

Qishen Chen, Wenxuan He, Xingyuan Chen, Chen Cheng, Minjie Bian, and Huahu Xu. Large-small model collaboration for medical visual question answering with task aware mixture of experts and

relation knowledge distillation. *Image and Vision Computing*, 165:105820, 2026. doi: 10.1016/j.imavis.2025.105820.

Y. Chen, J. Zhao, and H. Han. A survey on collaborative mechanisms between large and small language models. *arXiv preprint arXiv:2505.07460*, 2025.

Runwei Guan, Rongsheng Hu, Shangshu Chen, Ningyuan Xiao, Xue Xia, Jiayang Liu, Beibei Chen, Ziren Tang, Ningwei Ouyang, Shaofeng Liang, Yuxuan Fan, Wanjie Sun, and Yutao Yue. Roadscenevqa: Benchmarking visual question answering in roadside perception systems for intelligent transportation system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4366–4375, 2026.

S. Li, H. Wang, W. Xu, et al. Collaborative inference and learning between edge slms and cloud llms: A survey of algorithms, execution, and open challenges. *arXiv preprint arXiv:2507.16731*, 2025.

Zimo Liu, Kangjun Liu, Mingyue Guo, Shiliang Zhang, and Yaowei Wang. Cotuning: A large-small model collaborating distillation framework for better model generalization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10487–10496, 2024.

Ruijie Lu, Yiyang Ma, Xiaokang Chen, Lingxiao Luo, Zhiyu Wu, Zizheng Pan, Xingchao Liu, Yutong Lin, Hao Li, Wen Liu, Zhewen Hao, Xi Gao, Shaoheng Nie, Yixuan Wei, Zhenda Xie, Ting Chen, and Gang Zeng. Thinking with visual primitives. *arXiv preprint*, 2026.

Shichen Lu, Longteng Guo, Wenxuan Wang, Zijia Zhao, Tongtian Yue, Jing Liu, and Si Liu. Collaborative training of tiny-large vision language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024.

A. S. Rawat, V. Sadhanala, A. Rostamizadeh, et al. A little help goes a long way: Efficient llm training by leveraging small lms. *arXiv preprint arXiv:2410.18779*, 2024.

Zhongfan Sun, Kan Guo, Yongli Hu, Daxin Tian, Qingqing Gao, Jiapu Wang, Junbin Gao, Yanfeng Sun, and Baocai Yin. Large-small model synergy with multimodal fine-grained heuristics for knowledge-based visual question answering. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 935–944, 2025.

Joseph Raj Vishal, D. Basina, R. Patil, et al. Interact-video: Reasoning-rich video qa for urban traffic. *arXiv preprint arXiv:2507.14743*, 2025.

L. Wang, Z. He, J. Li, et al. Clash: Collaborative large-small hierarchical framework for continuous vision-and-language navigation. *arXiv preprint arXiv:2512.10360*, 2025.

S. Wang, Y. Liu, X. Tang, et al. Collaborative enhancement of large and small models for question answering via dual knowledge transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33630–33638, 2026.

Feng Yang, Xinrui Ju, Keyang Zhang, Xiandong Meng, Rongqun Lin, Howard Leung, Shiqi Wang, Haoliang Li, and Chris Xing Tian. Last: The last query token guides visual token pruning for edge-cloud collaborative mllm inference. *arXiv preprint arXiv:2607.27952*, 2026.

S. Zhang, R. Li, S. Deng, et al. Railvqa: A benchmark and framework for efficient interpretable visual cognition in automatic train operation. *arXiv preprint arXiv:2603.27112*, 2026a.

Yang Zhang, Xiaoshuai Sun, Rui Zhao, Wujin Sun, Yidong Chen, Jiayi Ji, Qian Chen, and Rongrong Ji. Look on demand: A cognitive scheduling framework for visual evidence acquisition in multimodal reasoning. *arXiv preprint arXiv:2605.28160*, 2026b.

Nan Zhou, Huiqun Wang, Yaoyan Zheng, and Di Huang. Covft: Context-aware visual fine-tuning for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a.

Xingcheng Zhou, Mingyu Liu, Walter Zimmer, Jiajie Zhang, and Alois Knoll. Sgta: Scenegraph based multi-modal traffic agent for video understanding. *arXiv preprint arXiv:2604.03697*, 2026b.

Y. Zhou, H. Lin, Y. Wu, et al. Before thinking, learn to decide: Proactive routing for efficient visual reasoning. *arXiv preprint arXiv:2606.30217*, 2026c.

Mengdan Zhu, Senhao Cheng, and Liang Zhao. Decompose, look, and reason: Reinforced latent reasoning for vlms. *arXiv preprint arXiv:2604.07518*, 2026.