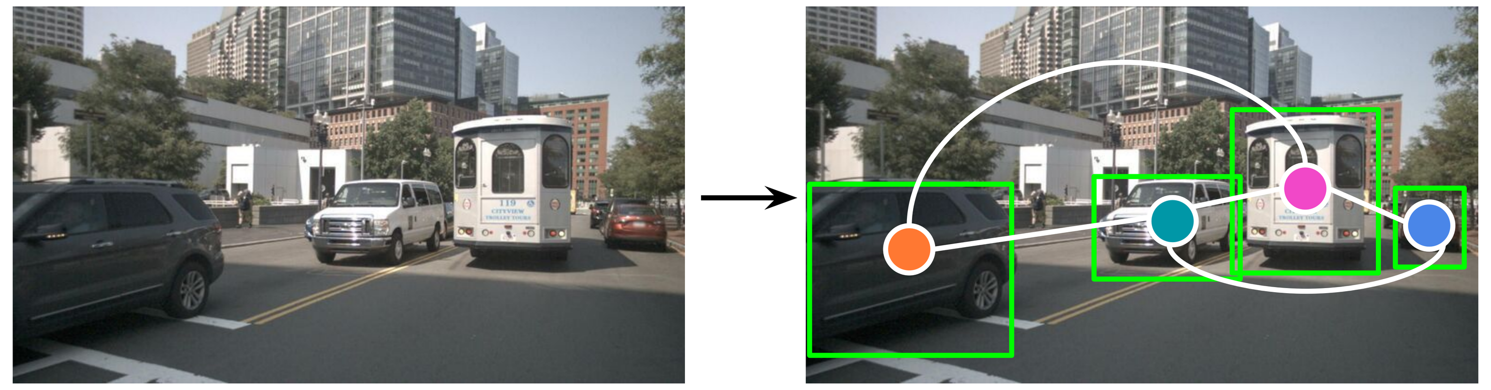
Enhanced Motion Forecasting with Visual Relation Reasoning

Sungjune Kim[1] Hadam Baek[1] Seunggwan Lee[1] Hyung-gun Chi[2] Hyerin Lim[3] Jinkyu Kim[1] Sangpil Kim[1]

[1]Korea University [2]Purdue University [3]Hyundai Motor Group

Paper Page

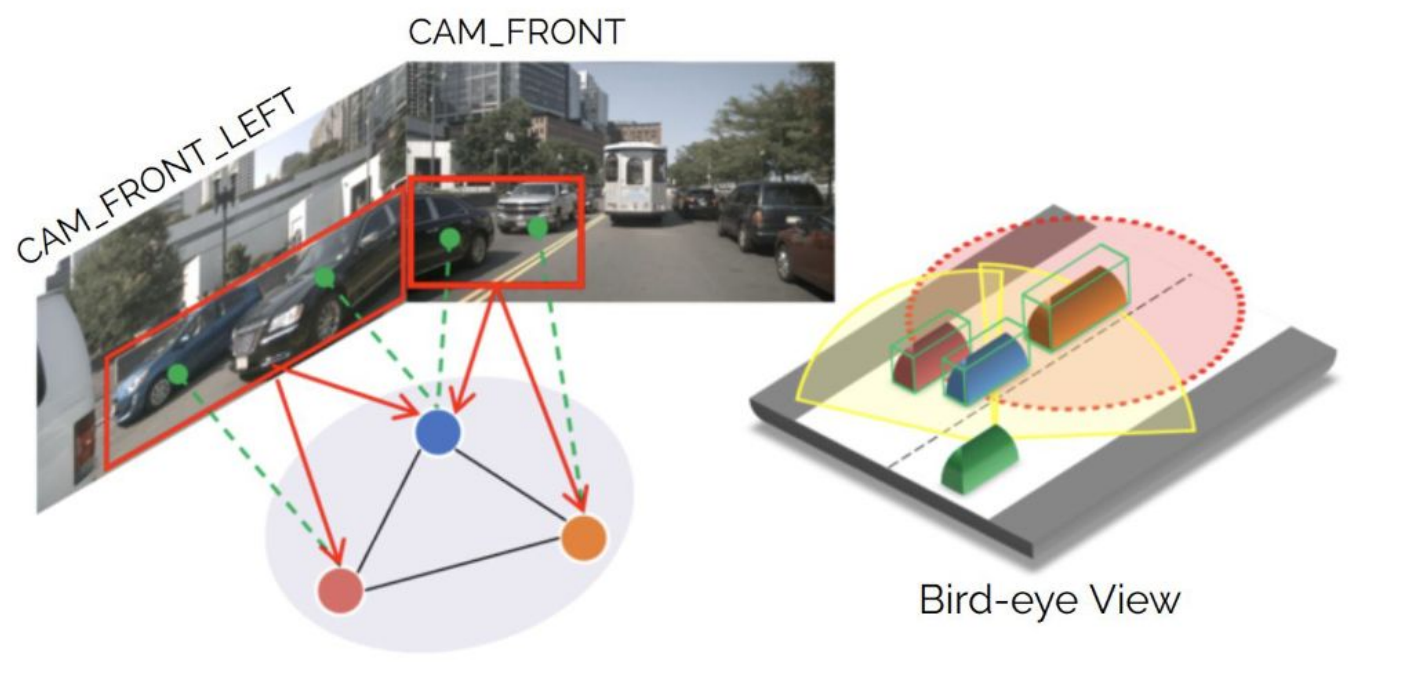
Motivation



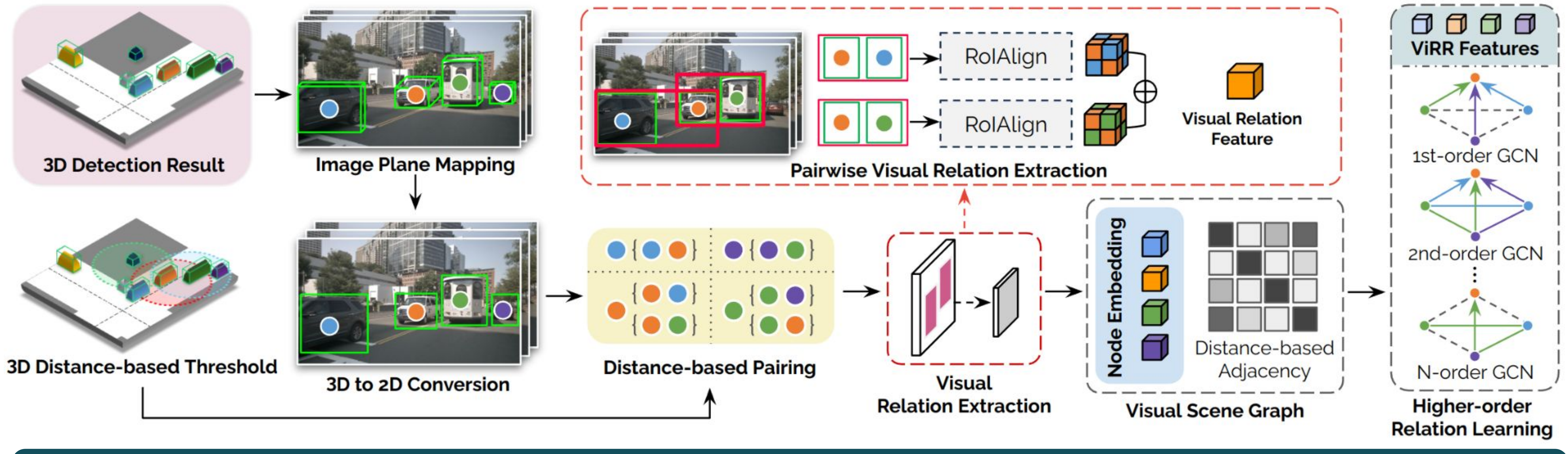
Vision-based autonomous driving is gaining more and more interests in the research field. However, explicit impact of visual information on motion/trajectory predictions are not explored in the literature. In this work, we specifically focus on how reasoning on the **visual relations** of road agents can improve the motion forecasting performance.

Contributions

- This work is the first to explore the benefits of reasoning **explicit visual relational semantics** for motion forecasting.
- We propose an innovative **visual scene graph architecture** that extracts **pairwise visual relations** of road agents and learns higher-order connectivity in the visual space.
- **ViRR enhances the motion forecasting performance** and provides a solid baseline for further research on the visual understanding for motion forecasting.

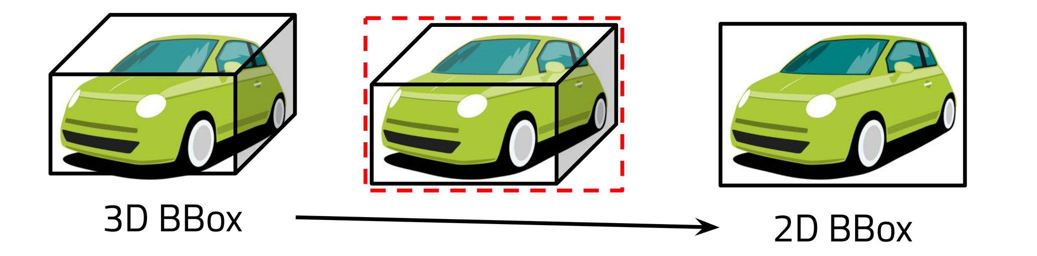


ViRR: Visual Relation Reasoning

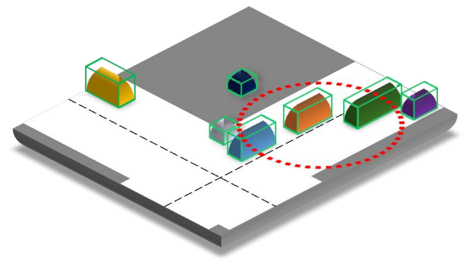


1. Visual Relation Extraction

A. 3D to 2D Conversion



B. 3D Distance-based Pairing



Identify pairs that have potential to influence the motion based on the 3D distance threshold

C. Visual Relation Feature Extraction

❖ **Pairwise Visual Relation Extraction**

$$\mathbf{v}_n = \frac{\sum_{m \in \mathcal{N}(n)} \mathcal{R}(F, \mathcal{B}^n, \mathcal{B}^m)}{|\mathcal{N}(n)|}$$

$\mathcal{R}$: RoIAlign Function
F : Image Feature Pyramids
$\mathcal{N}(n)$: Set of paired agents with n

- Utilize **RoIAlign technique** to extract the pairwise visual relation features
- Aggregate the pairwise features per agent, **transforming them into the agent node feature.**

❖ **Agent Node Feature**

$$\mathbf{X}^{(0)} = f\left(\begin{bmatrix} \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_N \end{bmatrix}\right) \in \mathbb{R}^{N \times d}$$

- The extracted pairwise visual features pass linear mapping function and **act as the graph node features.**
- These features **encapsulate rich local relations** between the neighboring agents in a visual space.

2. Higher-order Visual Relation Learning

A. 3D Distance-based Adjacency

- Local visual features **propagate globally** throughout the surrounding scenes.
- The information of the agents obtained in a single camera view can be **shared across agents in different viewpoints**.

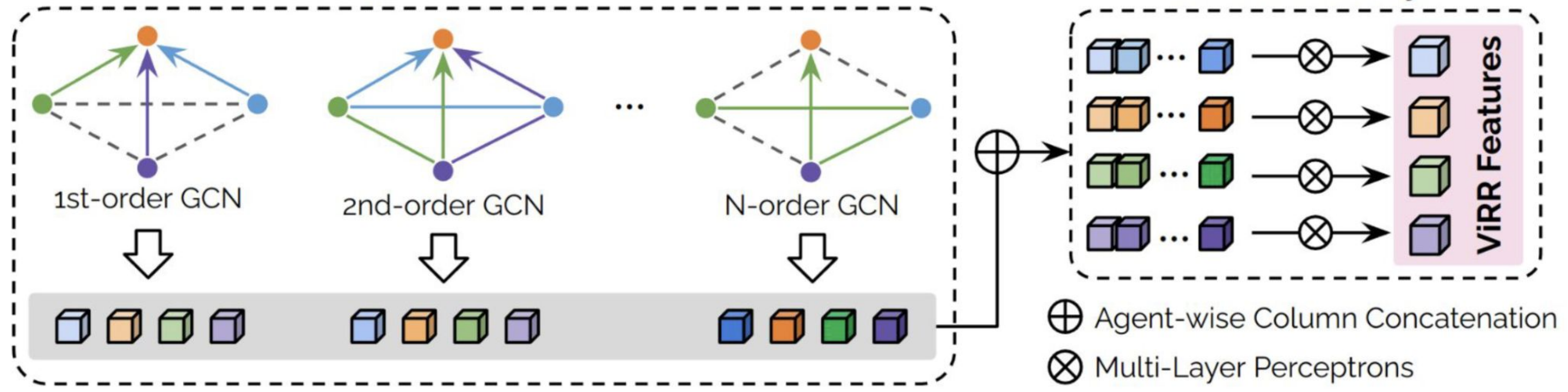
B. Higher-order Graph Convolution

❖ **Single-order Convolution**

$$\mathbf{H}_p^{(l)} = \sigma(\hat{\mathbf{A}}^p \mathbf{X}^{(l)} \mathbf{W}_p^{(l)})$$

❖ **Multi-order Convolution Combination**

$$\mathbf{X}^{(l+1)} = \sigma(\mathrm{MLP}(\lambda_1 \mathbf{H}_1^{(l)} \parallel \lambda_2 \mathbf{H}_2^{(l)} \parallel \ldots \parallel \lambda_P \mathbf{H}_P^{(l)}))$$



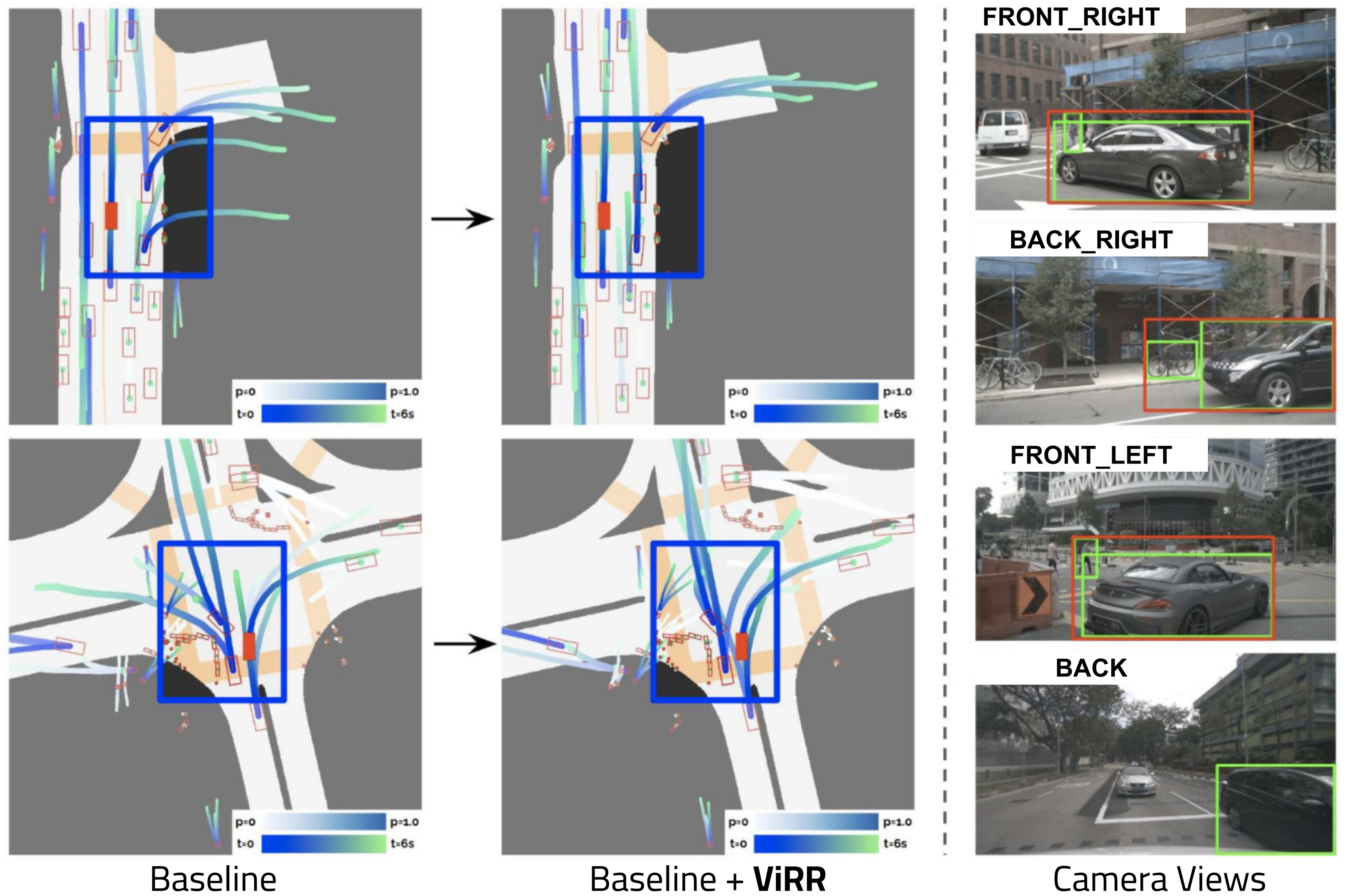
Limitation and Discussion

→ ViRR requires the detection results from perception stage, **limiting its generalizability on single-stage pipelines.**

→ How about incorporating **large vision-language models** as visual relation learner?

Experimental Results

※Quantitative Results



※Qualitative Results

	w/o ViRR			w/ ViRR		
	minADE ↓	minFDE ↓	MR ↓	minADE ↓	minFDE ↓	MR ↓
ViP3D [12]	2.051	2.862	0.244	1.690 (**+17.60%**)	2.075 (**+27.50%**)	0.191 (**+21.72%**)
UniAD (T+Mo) [19]	0.749	1.101	0.161	0.684 (**+8.68%**)	0.901 (**+18.17%**)	0.051 (**+68.32%**)
UniAD (T+M+Mo) [19]	0.732	1.063	0.158	0.628 (**+14.20%**)	0.891 (**+16.21%**)	0.040 (**+74.55%**)
MOTR + CVAE Motion	0.976	1.281	0.188	0.951 (**+2.47%**)	1.381 (**+16.16%**)	0.166 (**+11.68%**)

Tab 1. The motion forecasting performance enhancements with our proposed ViRR

Task	Metric	w/o ViRR	w/ ViRR
Tracking	AMOTA ↑	0.360	0.369 (+2.50%)
	AMOTP ↓	1.350	1.342 (+0.59%)
	IDS ↓	919	907 (+1.31%)
Motion Forecasting	minADE ↓	0.751	0.701 (+6.64%)
	minFDE ↓	1.109	0.954 (+13.98%)
	MR ↓	0.162	0.080 (+50.62%)

Tab 2. Benefits on Perception Stage

	minADE ↓	minFDE ↓	MR ↓
Baseline	0.732	1.063	0.158
GCN [25]	0.683	0.987	0.073
MixHop [1]	0.664	1.002	0.055
Cross-Attention [41]	0.655	0.929	0.049
ViRR ($P = 2$)	0.657	0.969	0.051
ViRR ($P = 4$)	**0.628**	**0.891**	**0.040**
ViRR ($P = 8$)	0.634	0.933	0.049

Tab 3. Relation reasoning algorithms

※ More results and analysis are reported in the paper. Check out the QR code above!

※ First author E-mail: ksjsungjune@korea.ac.kr