Holistic Order Prediction in Natural Scenes

NeurIPS 2025

Paper arXiv link

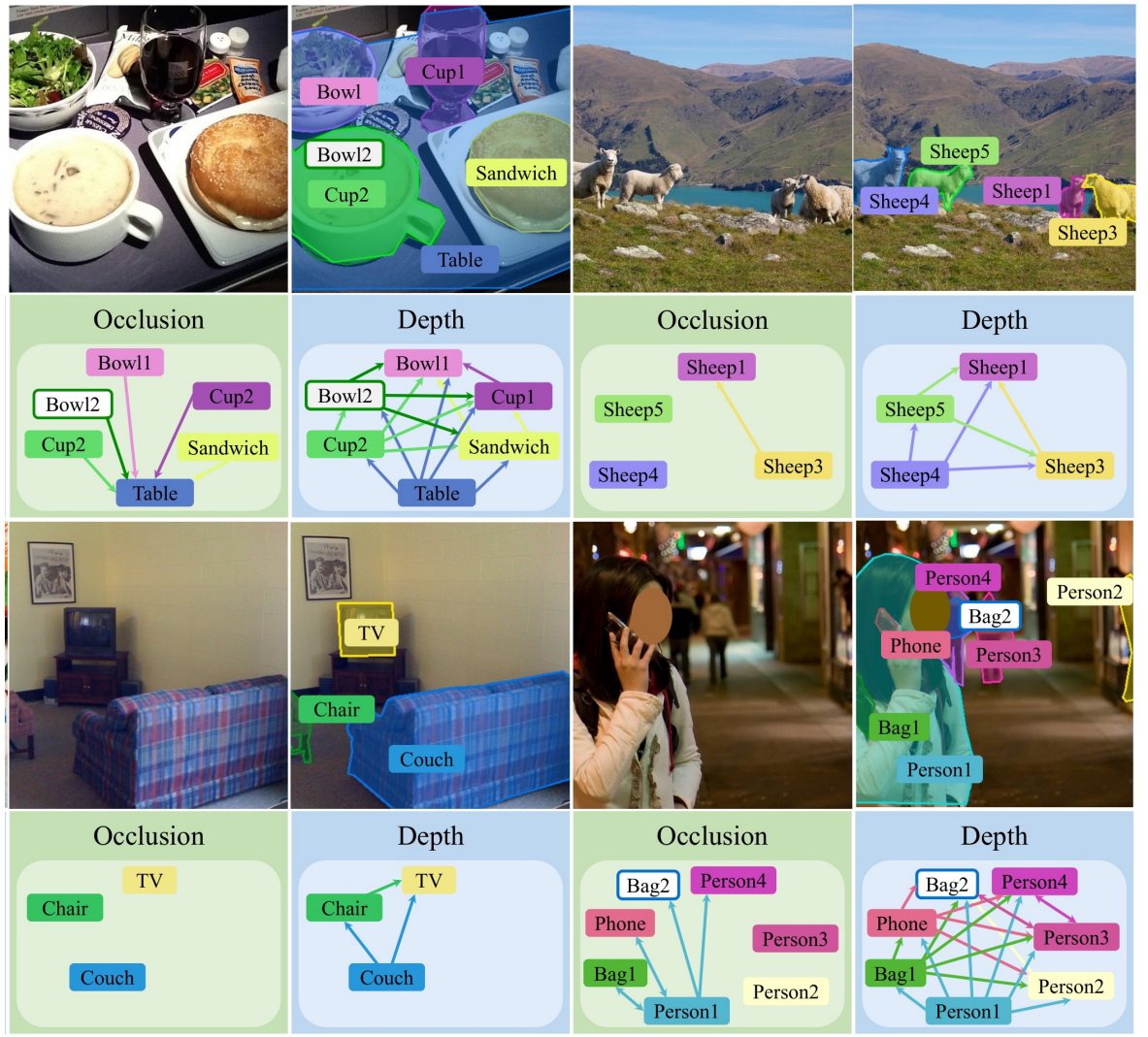
Pierre Musacchio[1], Hyunmin Lee[2], Jaesik Park[1]

[1]Seoul National University
[2]LG AI Research



Motivations

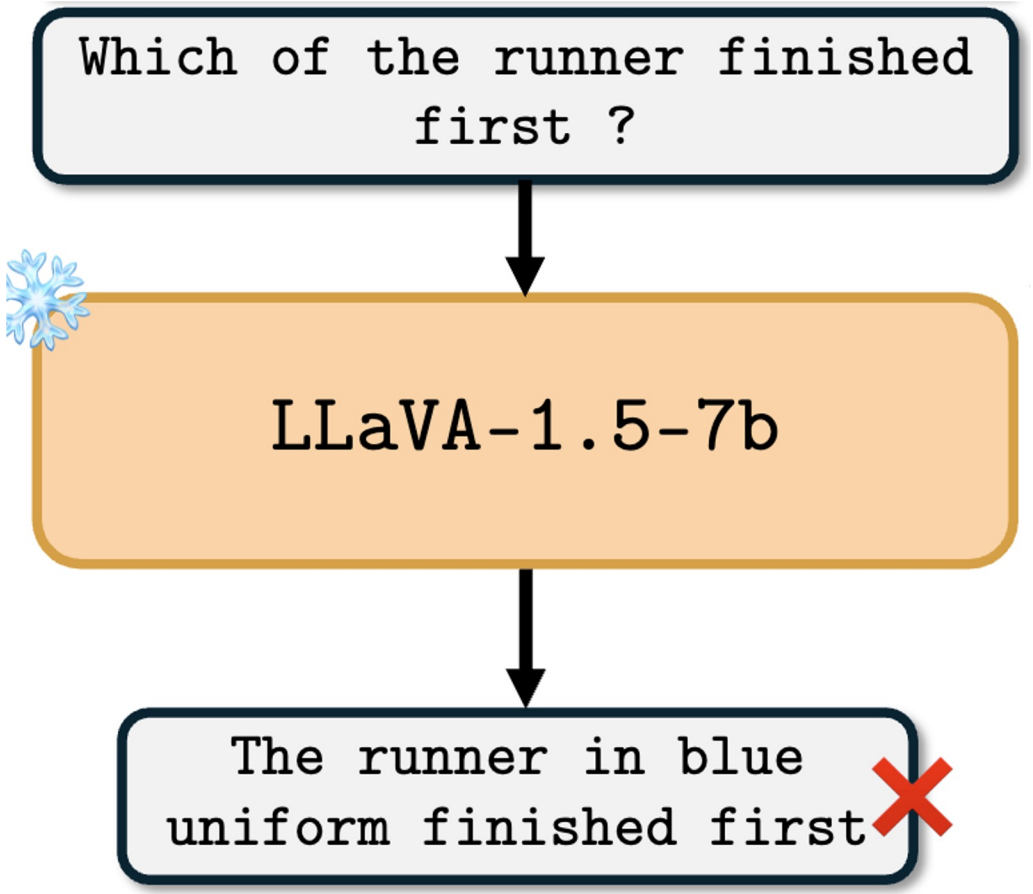
- Instance-wise geometry is everywhere
- Modern baseline does not exist
- Useful for VLM understanding



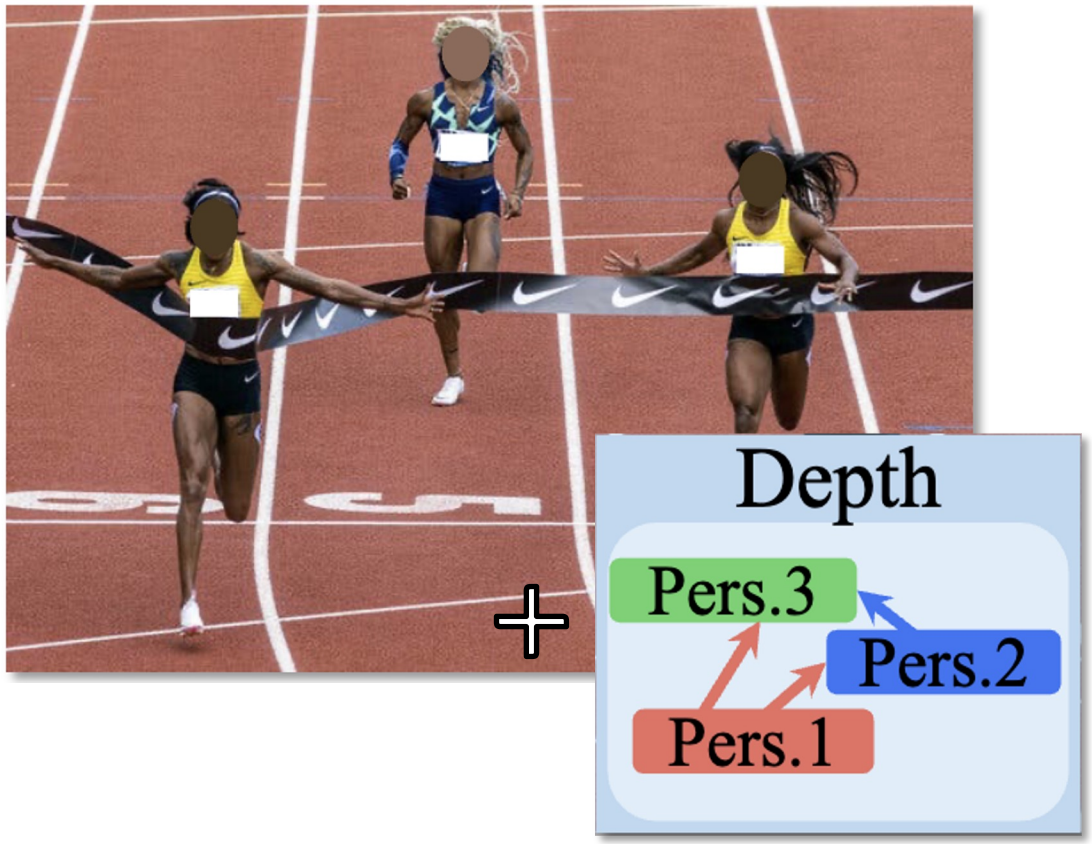
Visual Language Model Prompting

Which of the runner finished first ?

Visual Language Model Prompting

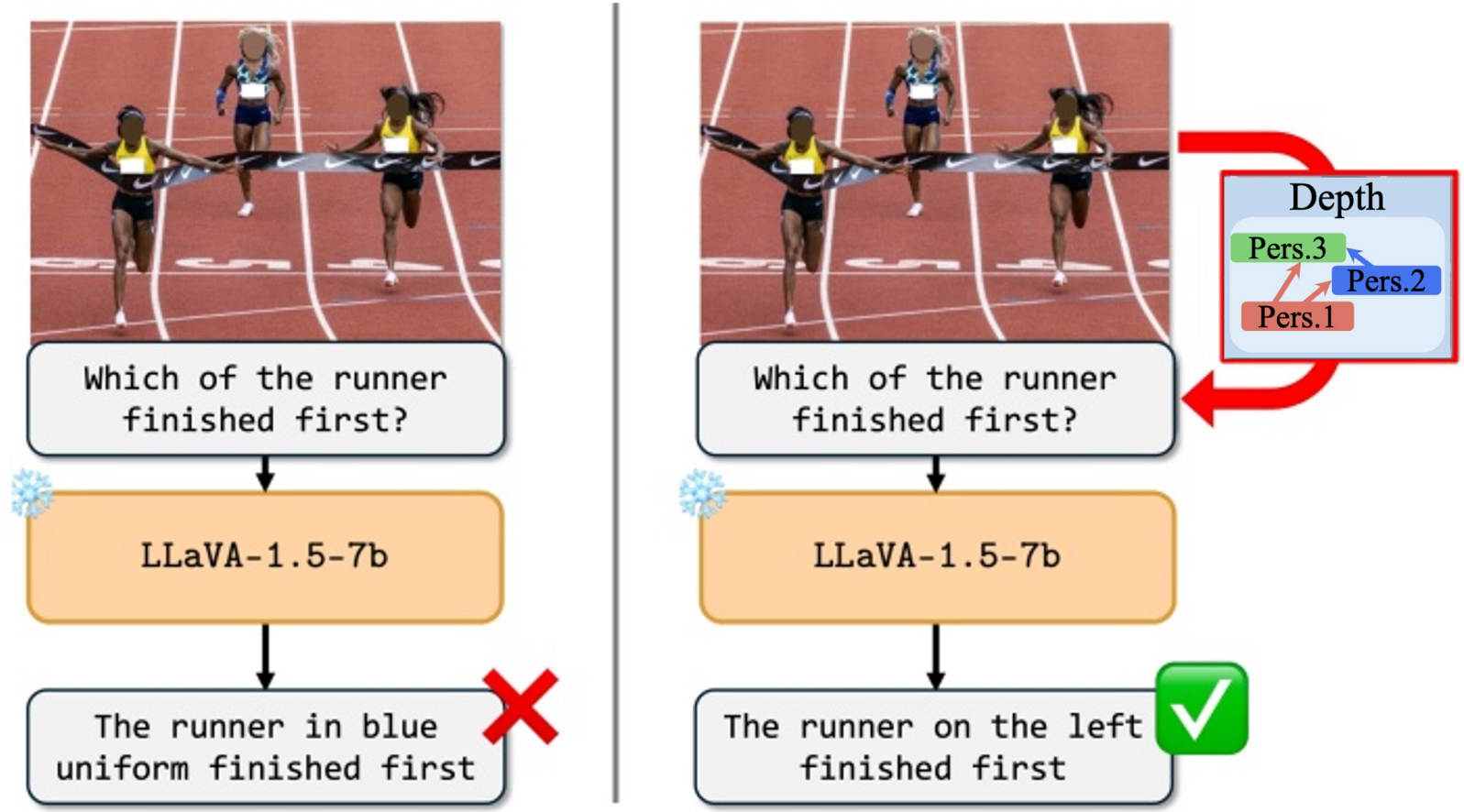


Visual Language Model Prompting

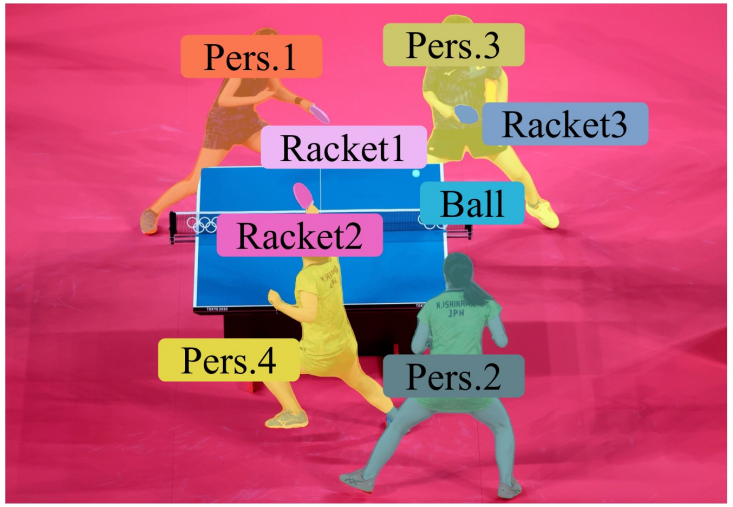


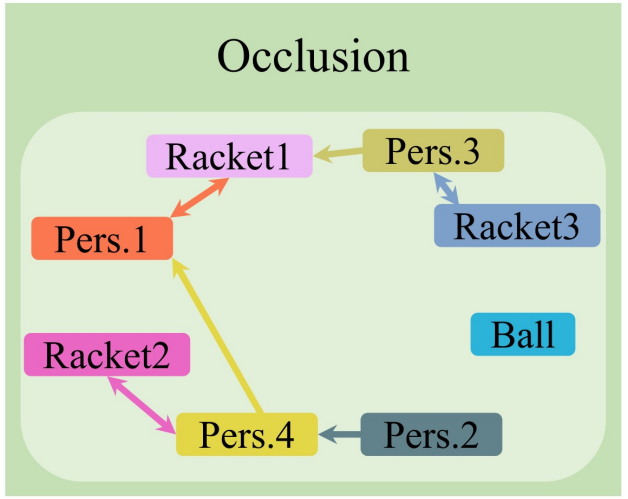
Which of the runner finished first ?

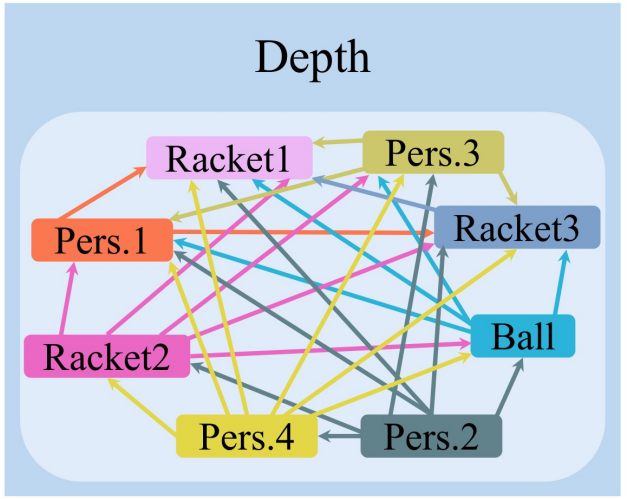
Visual Language Model: A Caveat



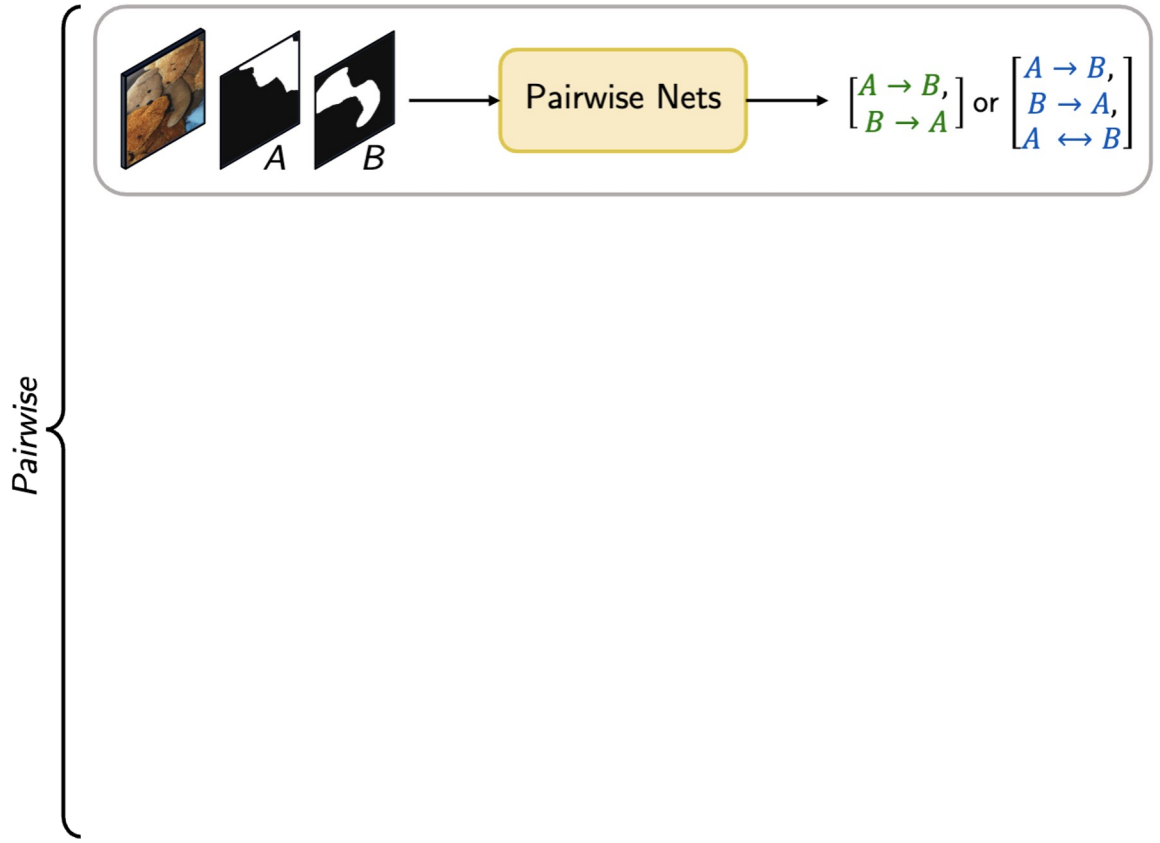
Order Prediction?





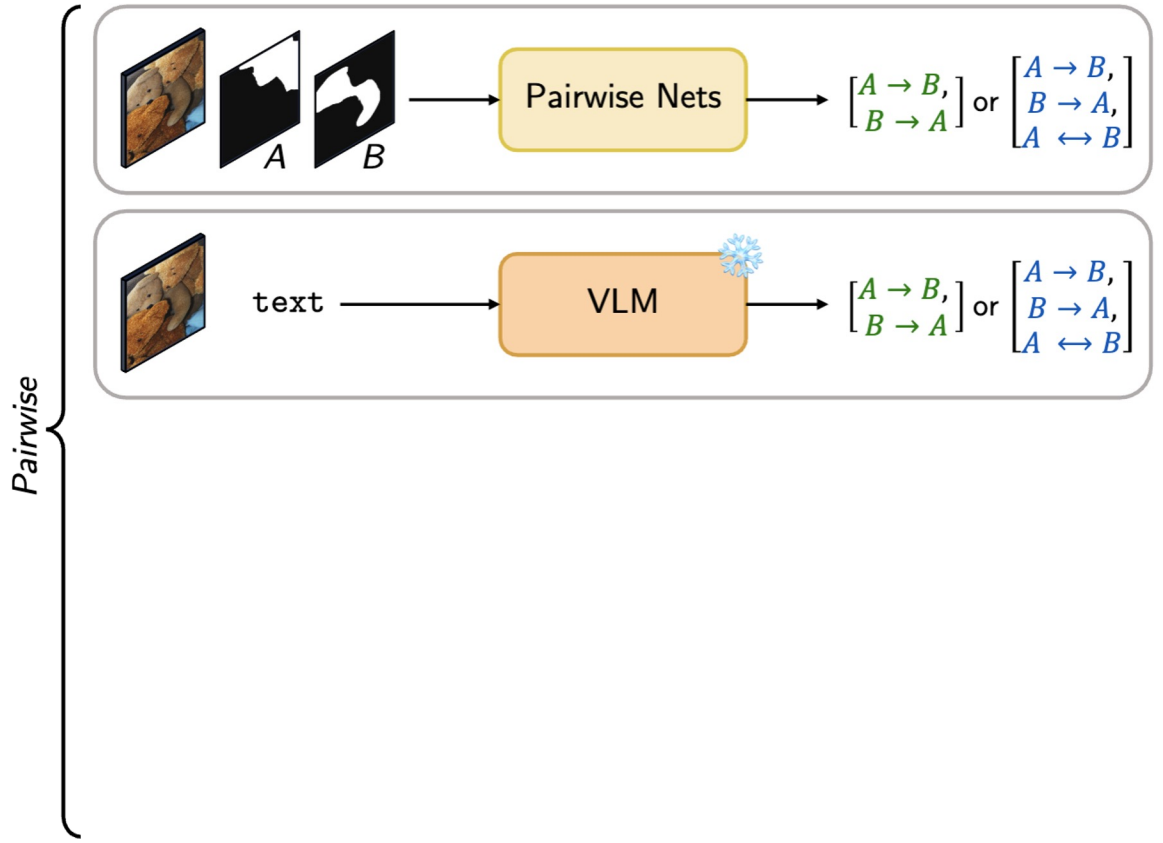


Previous Work



ConvNet Based,
Requires GT Masks
Pairwise Inferences

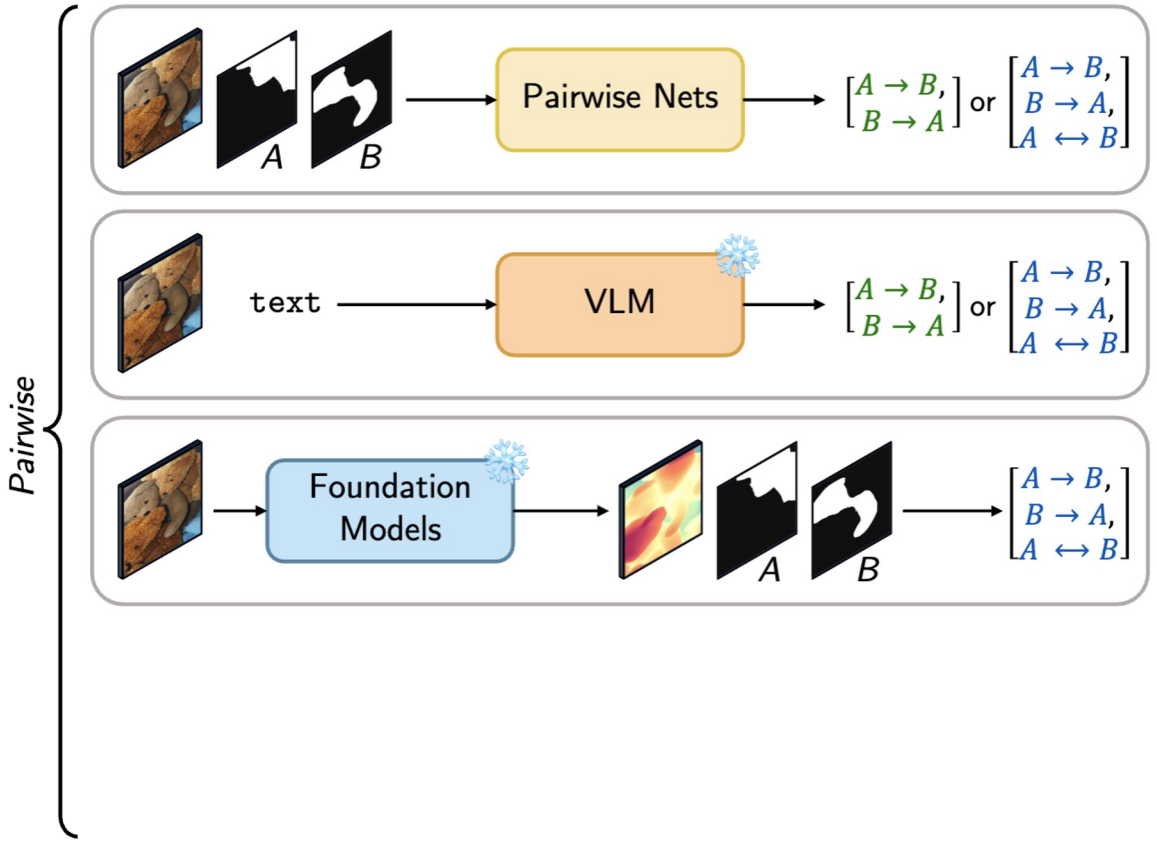
Previous Work



ConvNet Based,
Requires GT Masks
Pairwise Inferences

Limited
improvement

Previous Work

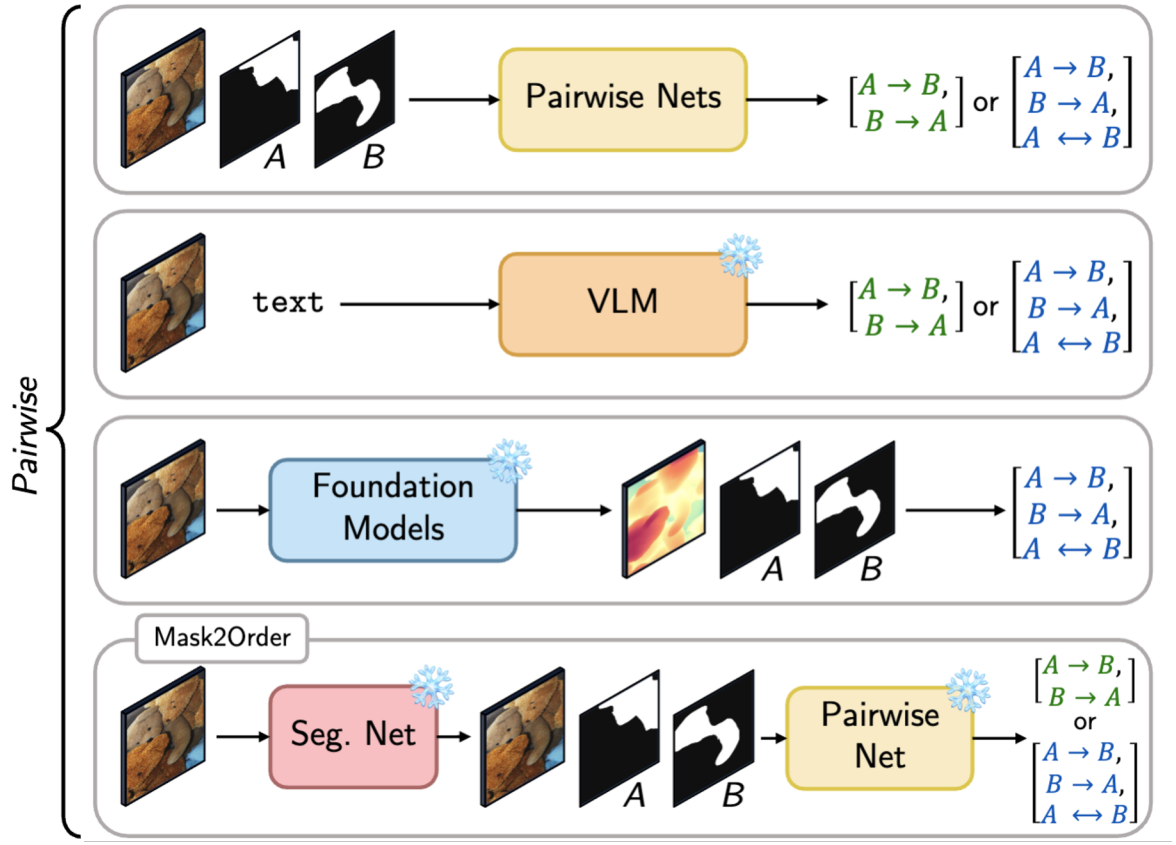


ConvNet Based,
Requires GT Masks
Pairwise Inferences

Limited reasoning

Pairwise Inferences

Previous Work



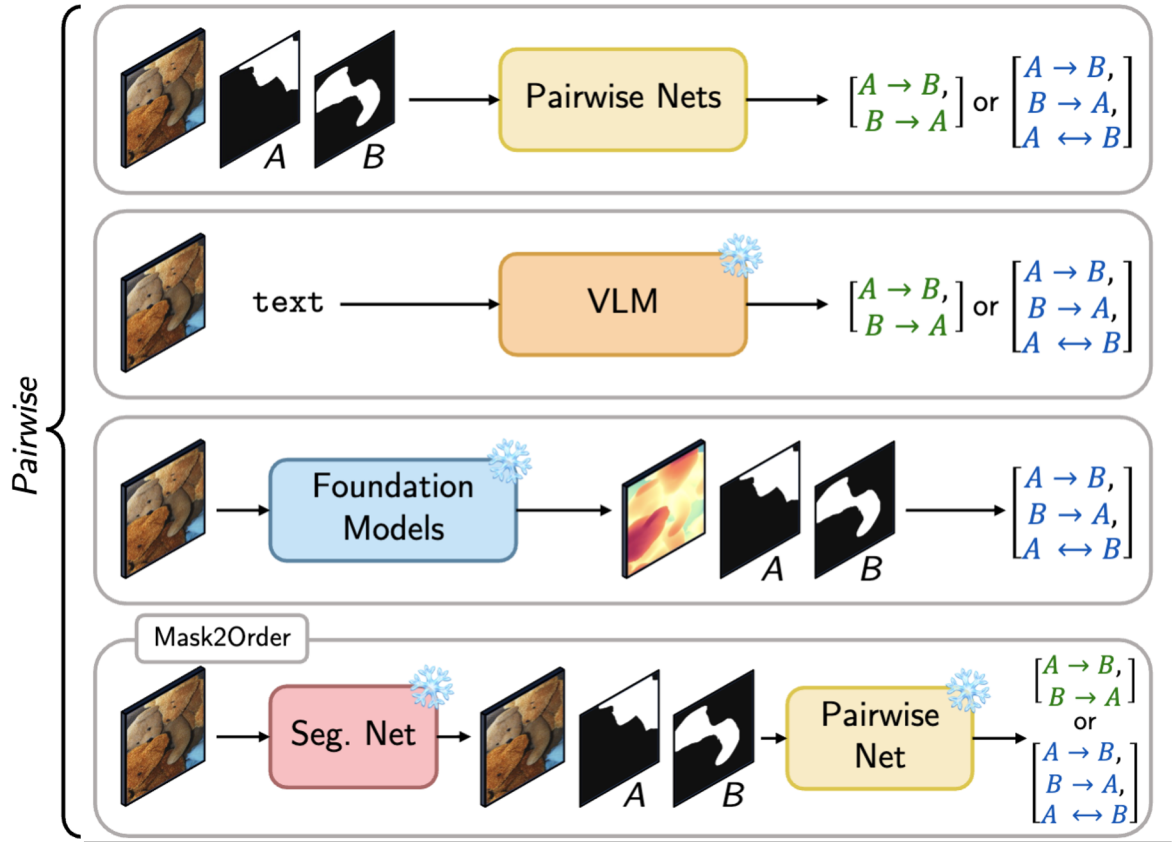
ConvNet Based,
Requires GT Masks
Pairwise Inferences

Limited
improvement

Pairwise Inferences

Cascade
Pairwise Inferences

Previous Work



ConvNet Based,
Requires GT Masks
Pairwise Inferences

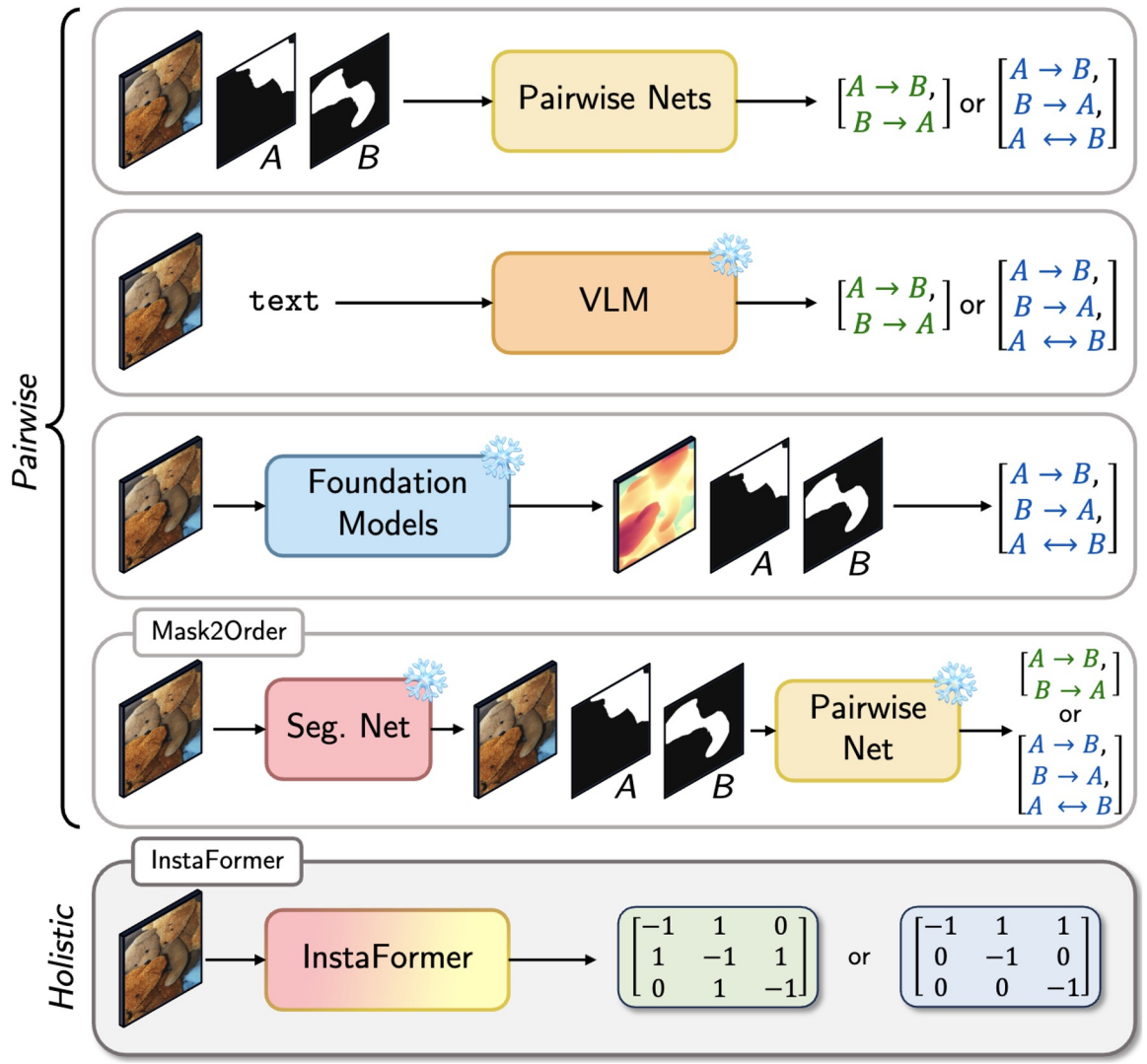
Limited
improvement

Pairwise Inferences

Cascade
Pairwise In

Can we do this better?

Holistic Order Prediction (ours)



ConvNet Based,
Requires GT Masks
Pairwise Inferences

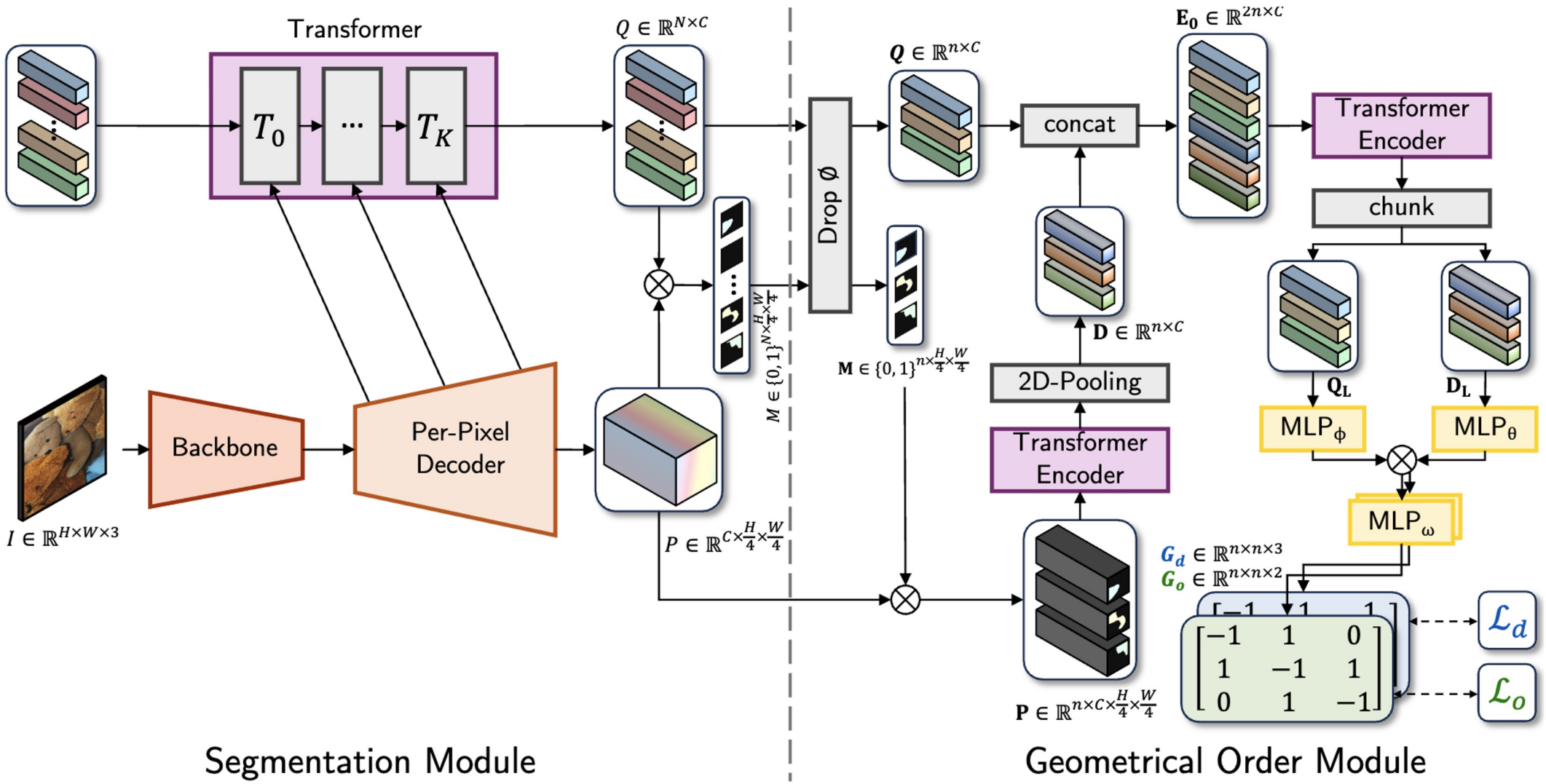
Limited
improvement

Pairwise Inferences

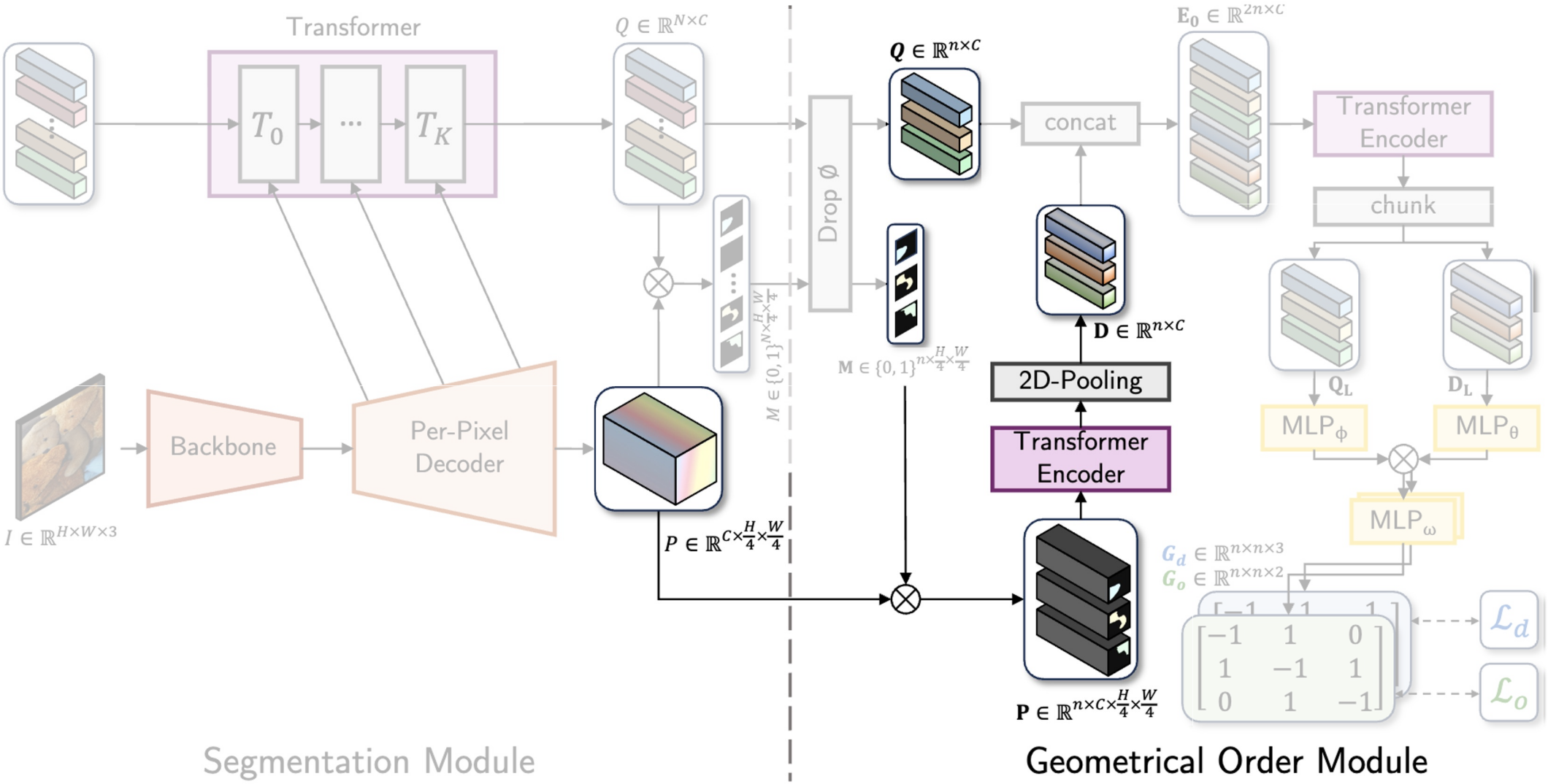
Cascade
Pairwise Inferences

All pairs comparison from latent instance-wise representations

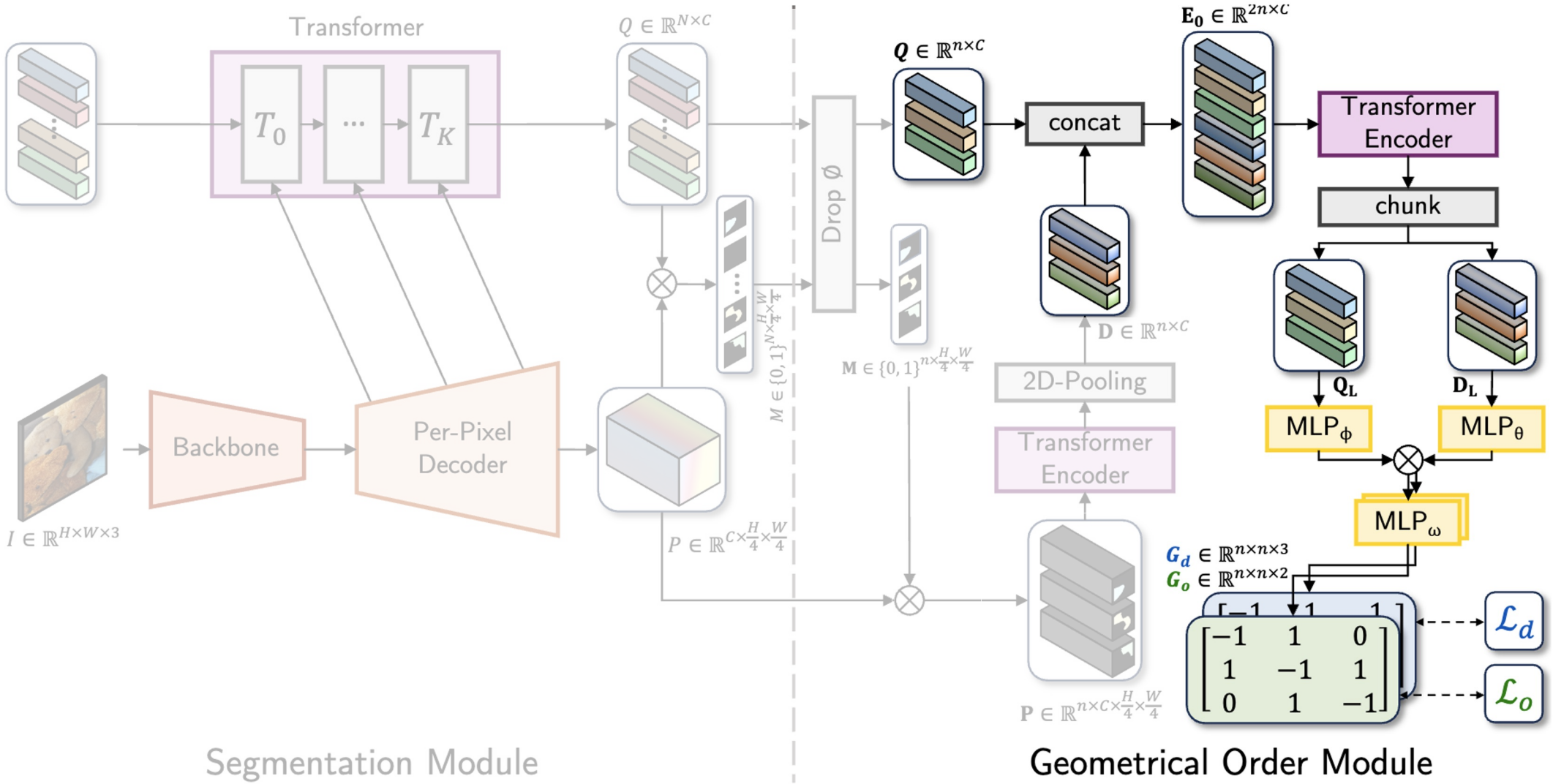
Method


Transformer
T_0
T_K
$Q \in \mathbb{R}^{N \times C}$
Drop ∅
$Q \in \mathbb{R}^{n \times C}$
concat
$\mathbf{E_0} \in \mathbb{R}^{2n \times C}$
Transformer Encoder
chunk
$\mathbf{Q_L}$
$\mathbf{D_L}$
MLP$_\phi$
MLP$_\theta$
MLP$_\omega$
$\mathbf{D} \in \mathbb{R}^{n \times C}$
2D-Pooling
Transformer Encoder
$M \in \{0,1\}^{N \times \frac{H}{4} \times \frac{W}{4}}$
$\mathbf{M} \in \{0,1\}^{n \times \frac{H}{4} \times \frac{W}{4}}$
Backbone
Per-Pixel Decoder
$I \in \mathbb{R}^{H \times W \times 3}$
$P \in \mathbb{R}^{C \times \frac{H}{4} \times \frac{W}{4}}$
$\mathbf{P} \in \mathbb{R}^{n \times C \times \frac{H}{4} \times \frac{W}{4}}$
$G_d \in \mathbb{R}^{n \times n \times 3}$
$G_o \in \mathbb{R}^{n \times n \times 2}$
$\mathcal{L}_d$
$\mathcal{L}_o$
Segmentation Module
Geometrical Order Module

Method (order module)

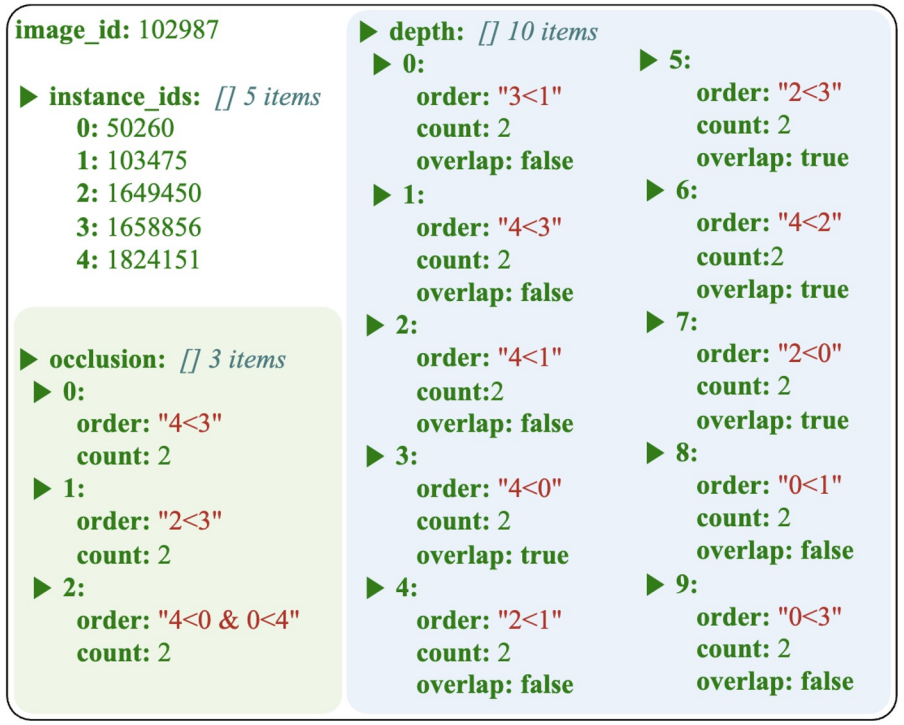


Method (order module)

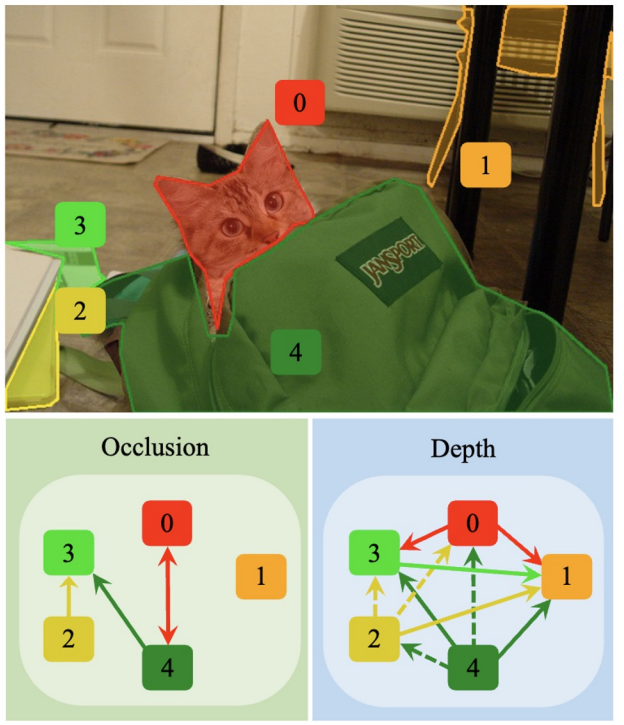


InstaOrder Dataset

- 100K images, 503K instances
- 2.8M instance-wise occlusion and depth orders



(a) `json` file provided by INSTAORDER dataset



(b) Instance-wise occlusion and depth orders

Instance-wise Occlusion and Depth Orders in Natural Scenes – H. Lee & J. Park, CVPR 2022

Metrics

- Occlusion Order

$$\text{Recall} = \frac{\sum_{AB} \mathbb{1}(\hat{o}_{AB} = 1 \text{ and } o_{AB} = 1)}{\sum_{AB} \mathbb{1}(o_{AB} = 1)},$$

$$\text{Precision} = \frac{\sum_{AB} \mathbb{1}(\hat{o}_{AB} = 1 \text{ and } o_{AB} = 1)}{\sum_{AB} \mathbb{1}(\hat{o}_{AB} = 1)},$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},$$

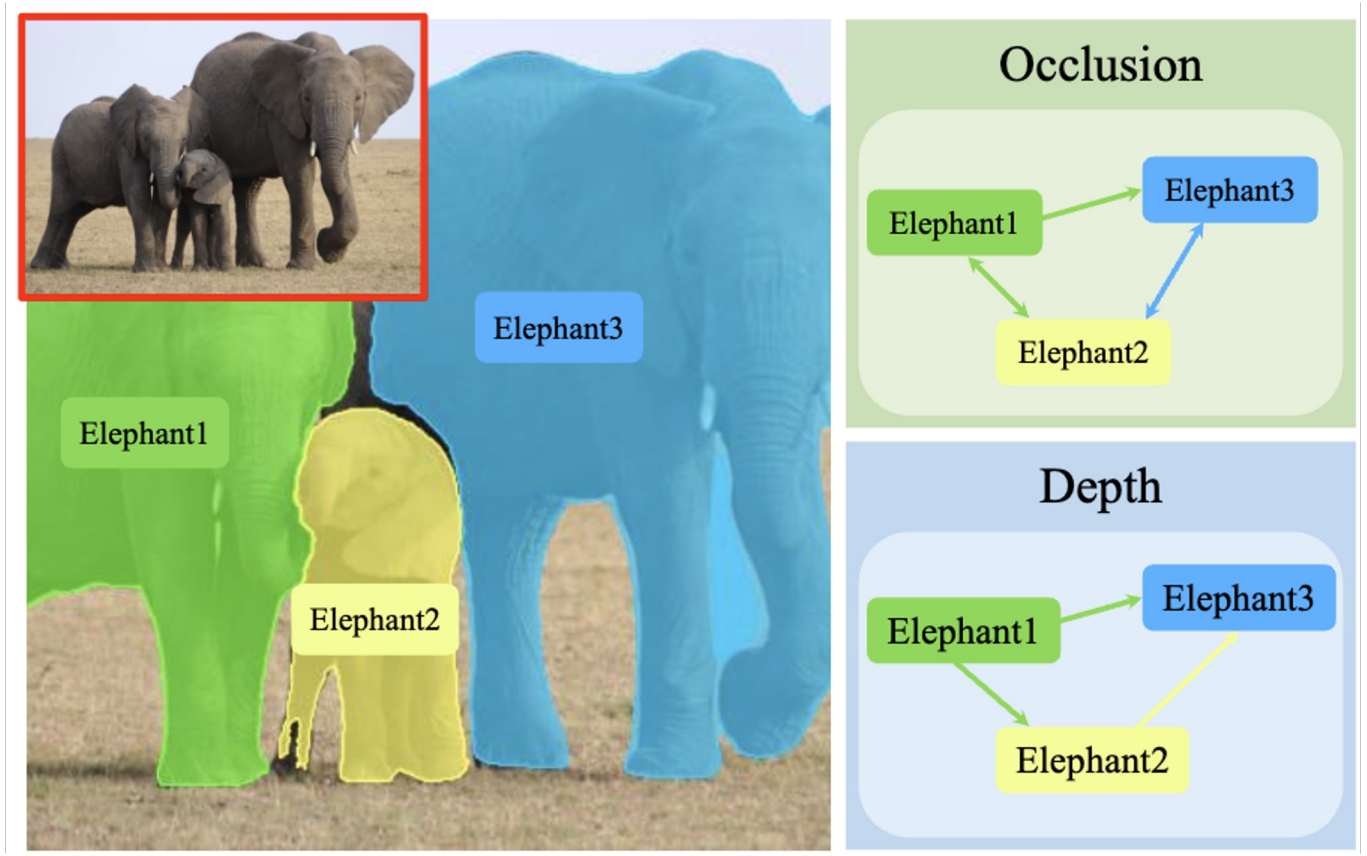
- Depth Order (Weighted Human Disagreement Rate)

$$\text{WHDR} = \frac{\sum_{AB} w_{AB} \cdot \mathbb{1}(\hat{d}_{AB} \neq d_{AB})}{\sum_{AB} w_{AB}},$$

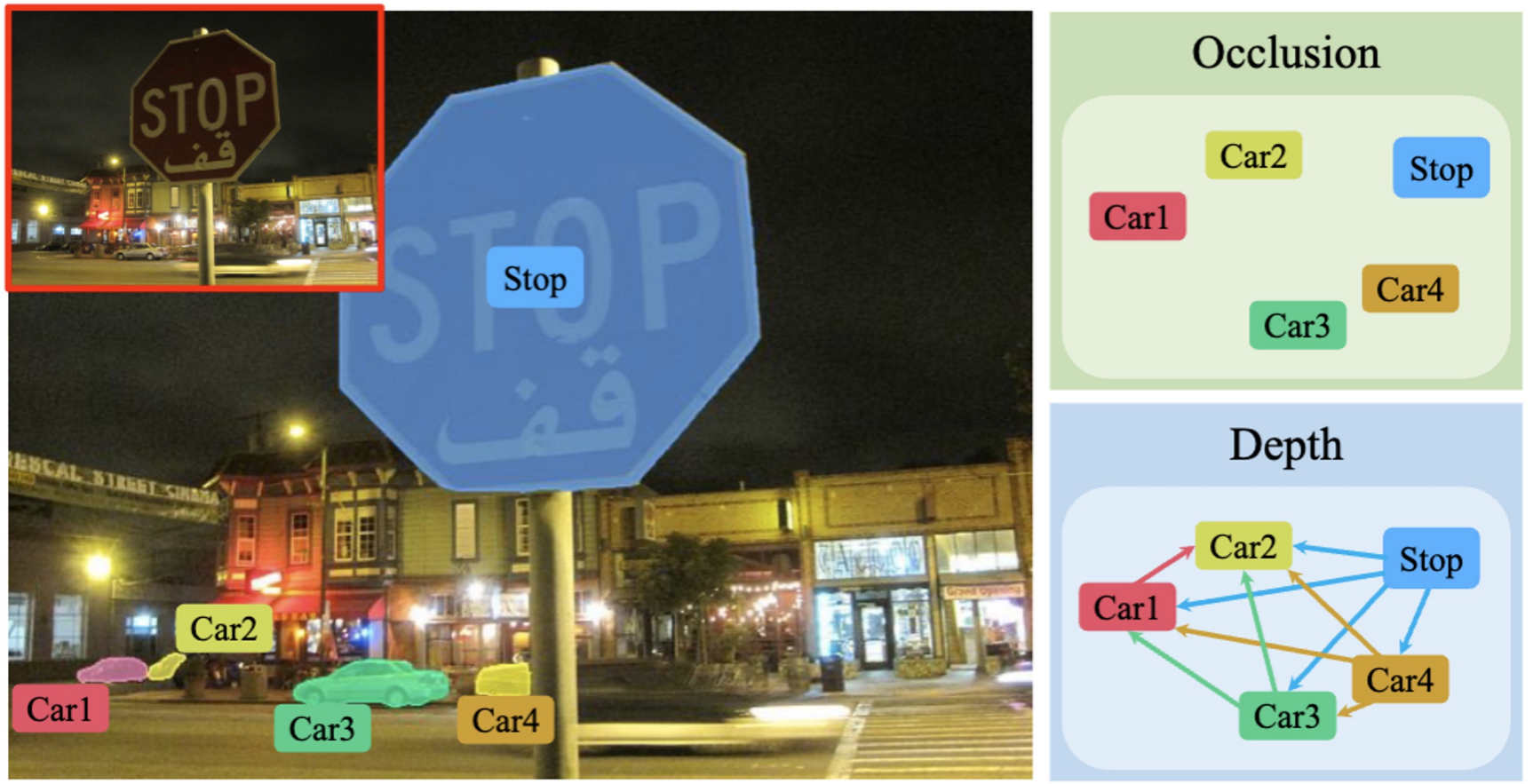
Comparison

Dataset	Method	Input				Output			Occlusion acc. ↑			WHDR ↓		
		GT Masks	Image	Category	Text	Occ. order	Depth order	Seg.	Recall	Prec.	F1	Distinct	Overlap	All
VQA	LLaVA [27]	–	✓	–	✓	✓	✓	–	49.98	48.28	34.10	37.37	44.56	39.52
	LLaVA‡ [27]	–	✓	–	✓	✓	✓	–	85.41	55.10	60.41	15.95	27.87	25.98
INSTAORDER [24]	Area	✓	–	–	–	✓	–	–	56.33	71.55	59.67	30.90	35.66	32.19
	Y-axis	✓	–	–	–	✓	–	–	44.84	57.34	47.30	22.19	39.04	29.20
	PCNet-M [47]	✓	✓	–	–	✓	–	–	59.19	76.42	63.02	–	–	–
	OrderNet^{M+I}(ext.) [32]	✓	✓	–	–	✓	–	–	84.93	78.21	77.51	–	–	–
	InstaOrderNet^{o}(M)	✓	–	–	–	✓	–	–	87.35	79.07	78.98	–	–	–
	InstaOrderNet^{o}(MC)	✓	–	✓	–	✓	–	–	88.70	78.21	79.18	–	–	–
	InstaOrderNet^{o}(MIC)	✓	✓	✓	–	✓	–	–	89.38	79.00	79.98	–	–	–
	InstaOrderNet^{o}	✓	✓	–	–	✓	–	–	89.39	79.83	80.65	–	–	–
	InstaOrderNet^{d}(M)	✓	–	–	–	–	✓	–	–	–	–	22.96	30.46	25.23
	InstaOrderNet^{d}(MC)	✓	–	✓	–	–	✓	–	–	–	–	23.19	28.56	36.45
	InstaOrderNet^{d}(MIC)	✓	✓	✓	–	–	✓	–	–	–	–	13.33	26.60	17.89
	InstaOrderNet^{d}	✓	✓	–	–	–	✓	–	–	–	–	12.95	25.96	17.51
	InstaOrderNet^{o,d} [24]	✓	✓	–	–	✓	✓	–	82.37	**88.67**	81.86	11.51	25.22	15.99
	MiDaS(Mean) [34]	✓	✓	–	–	–	✓	–	–	–	–	10.42	37.67	21.70
	MiDaS(Median) [34]	✓	✓	–	–	–	✓	–	–	–	–	10.31	36.08	20.92
	Foundation (Min-Max)	–	✓	–	–	–	✓	✓	–	–	–	21.71	42.89	29.85
	Foundation (Mean)	–	✓	–	–	–	✓	✓	–	–	–	10.70	39.22	22.46
	Foundation (Median)	–	✓	–	–	–	✓	✓	–	–	–	10.80	39.11	22.36
	Mask2Order^{o}	–	✓	–	–	✓	–	✓	84.37	78.48	77.49	–	–	–
	Mask2Order^{o‡}	–	✓	–	–	✓	–	✓	77.92	88.58	79.1	–	–	–
	Mask2Order^{d}	–	✓	–	–	–	✓	✓	–	–	–	14.16	27.15	18.52
	Mask2Order^{d‡}	–	✓	–	–	–	✓	✓	–	–	–	12.96	28.6	18.40
	Mask2Order^{o,d}	–	✓	–	–	✓	✓	✓	77.51	85.50	77.17	12.29	27.03	17.09
	Mask2Order^{o,d‡}	–	✓	–	–	✓	✓	✓	79.81	86.86	79.09	12.44	28.44	18.19
	InstaFormer^{o}	–	✓	–	–	✓	–	✓	**89.82**	78.10	**81.89**	–	–	–
	InstaFormer^{d}	–	✓	–	–	–	✓	✓	–	–	–	8.47	24.91	13.73
	InstaFormer^{o,d}	–	✓	–	–	✓	✓	✓	89.57	78.07	81.37	**7.90**	**24.68**	**13.30**

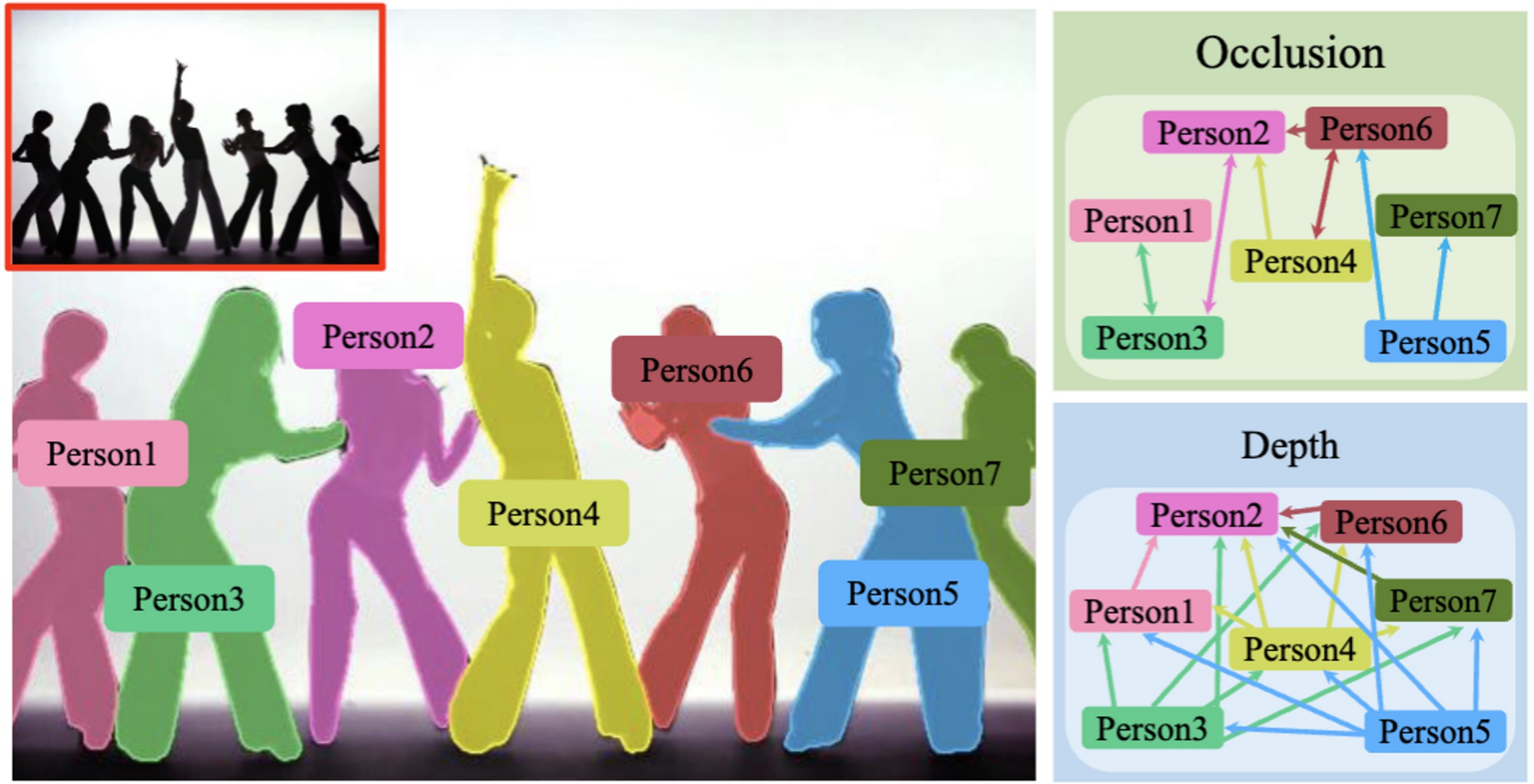
Results (OOD)


Elephant3
Elephant1
Elephant2
Occlusion
Elephant1
Elephant3
Elephant2
Depth
Elephant1
Elephant3
Elephant2

Results (validation)

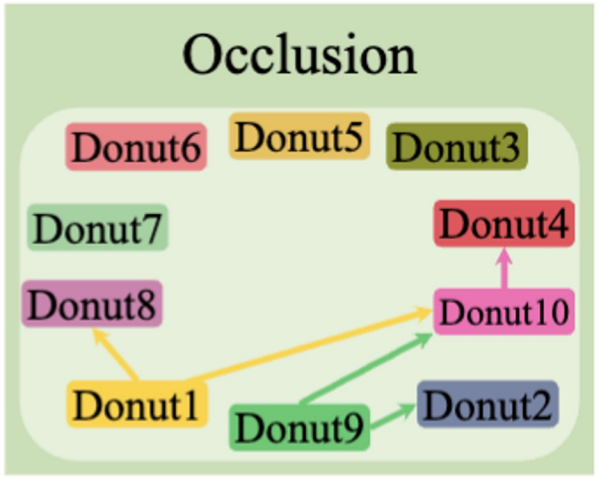


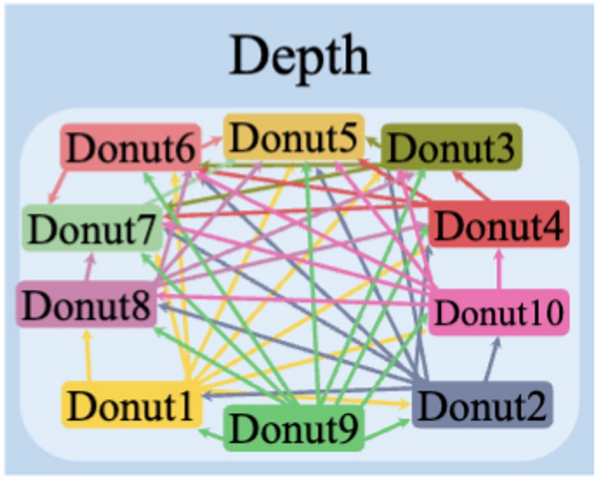
Results (OOD)



Robust under very sparse order clues!

Results (validation)





Holistic Order Prediction in Natural Scenes

NeurIPS 2025

Paper arXiv link

Pierre Musacchio[1], Hyunmin Lee[2], Jaesik Park[1]

[1]Seoul National University
[2]LG AI Research