Part-Aware Bottom-Up Group Reasoning for Fine-Grained Social Interaction Detection

Dongkeun Kim, Minsu Cho, Suha Kwak

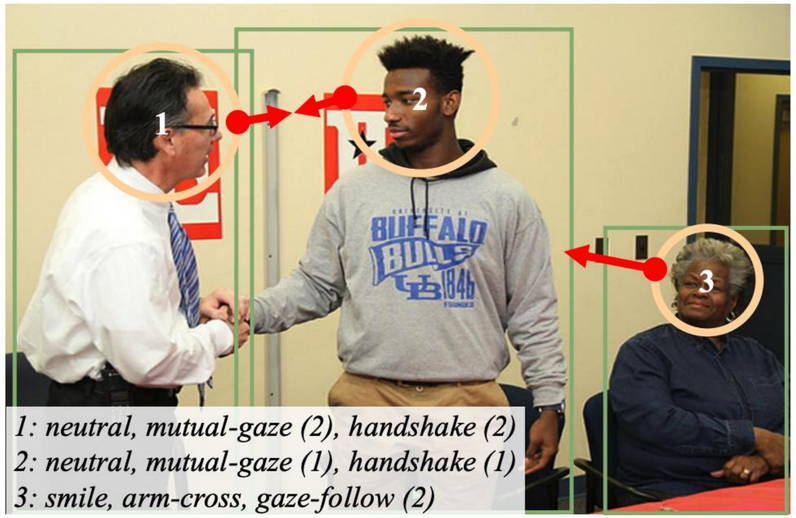
Pohang University of Science and Technology (POSTECH)

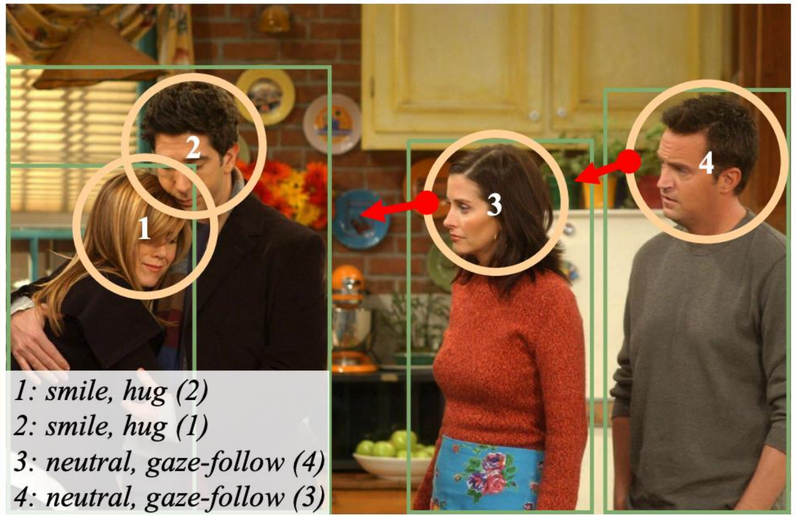
Understanding Social Interactions

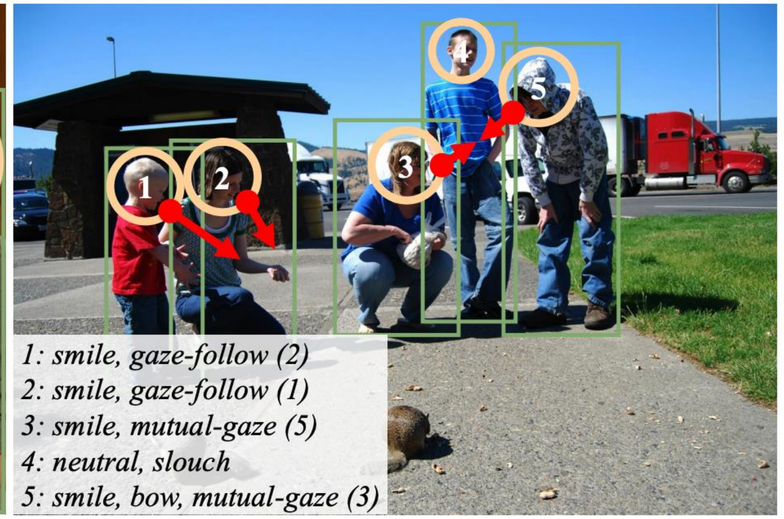
Perceiving subtle visual cues beyond global appearance

Fine-Grained Social Interaction Detection

Predict <individual, group, interaction> triplets



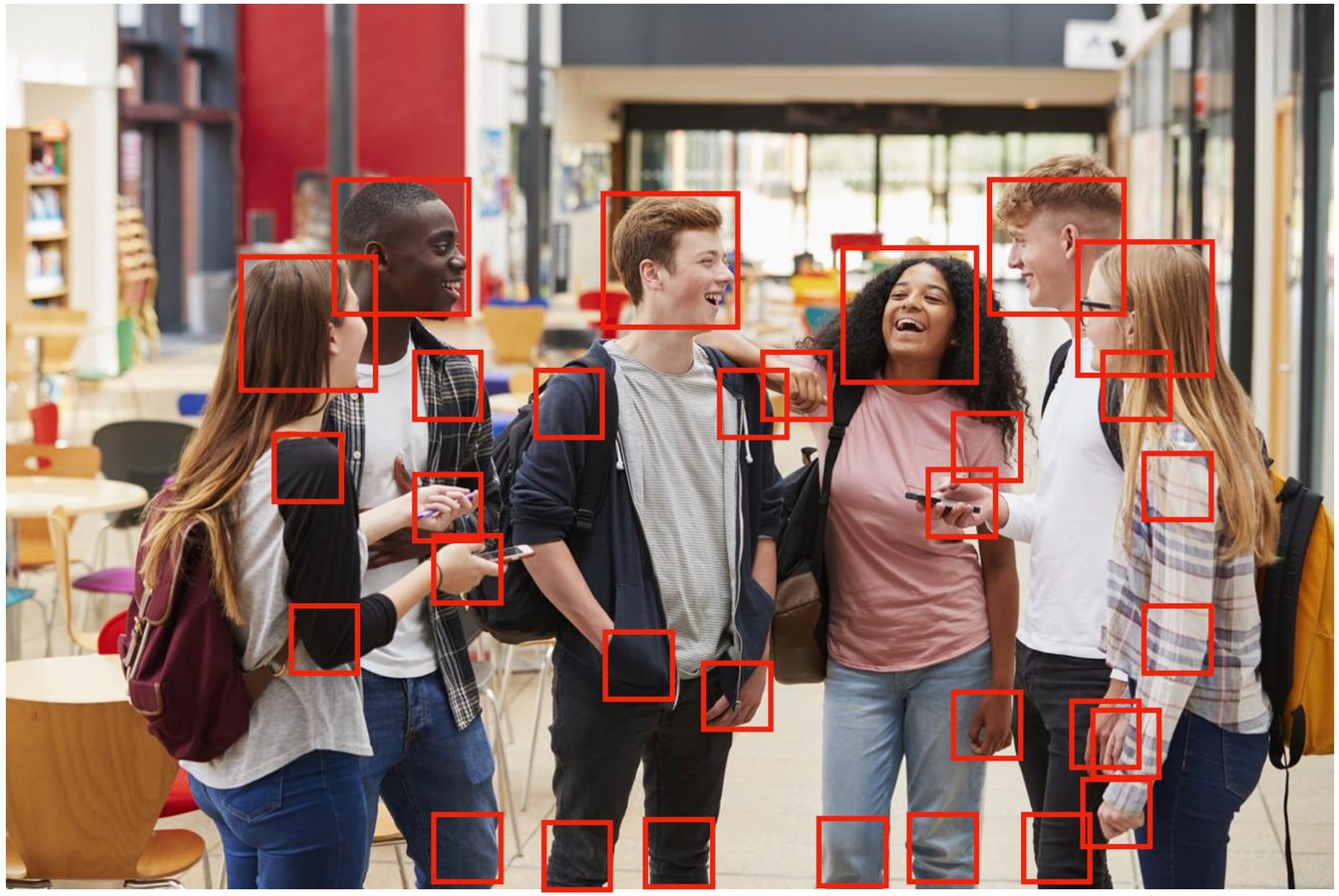


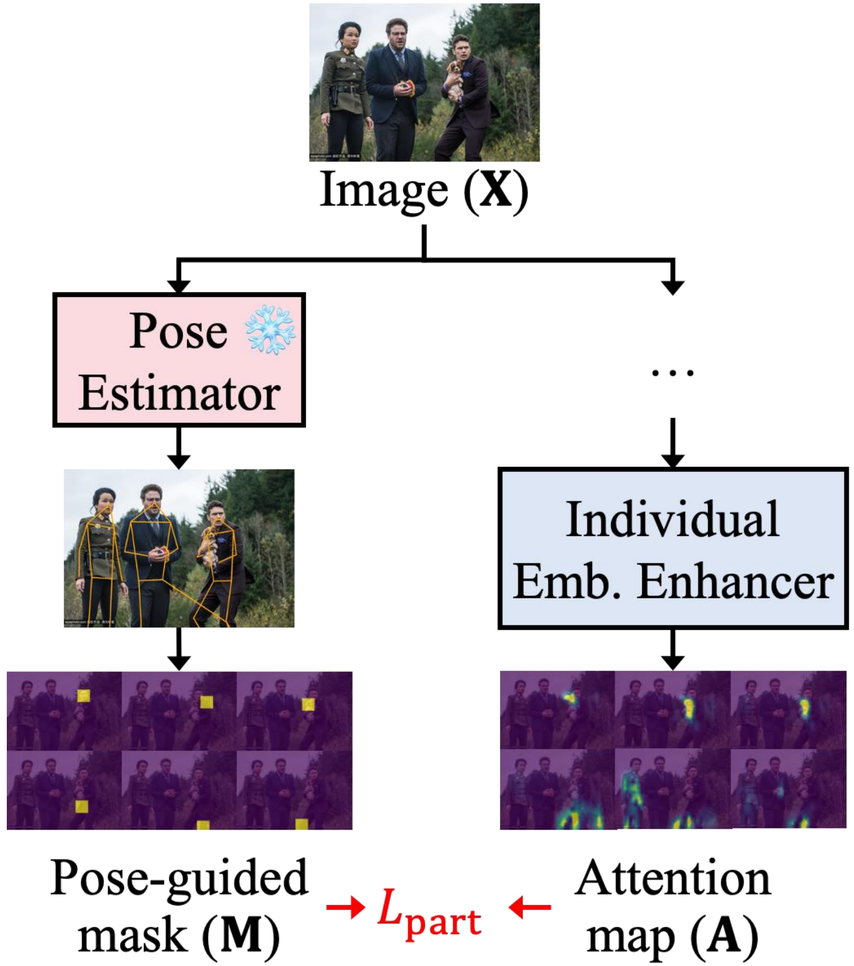


Wei et al., Nonverbal Interaction Detection, ECCV 2024

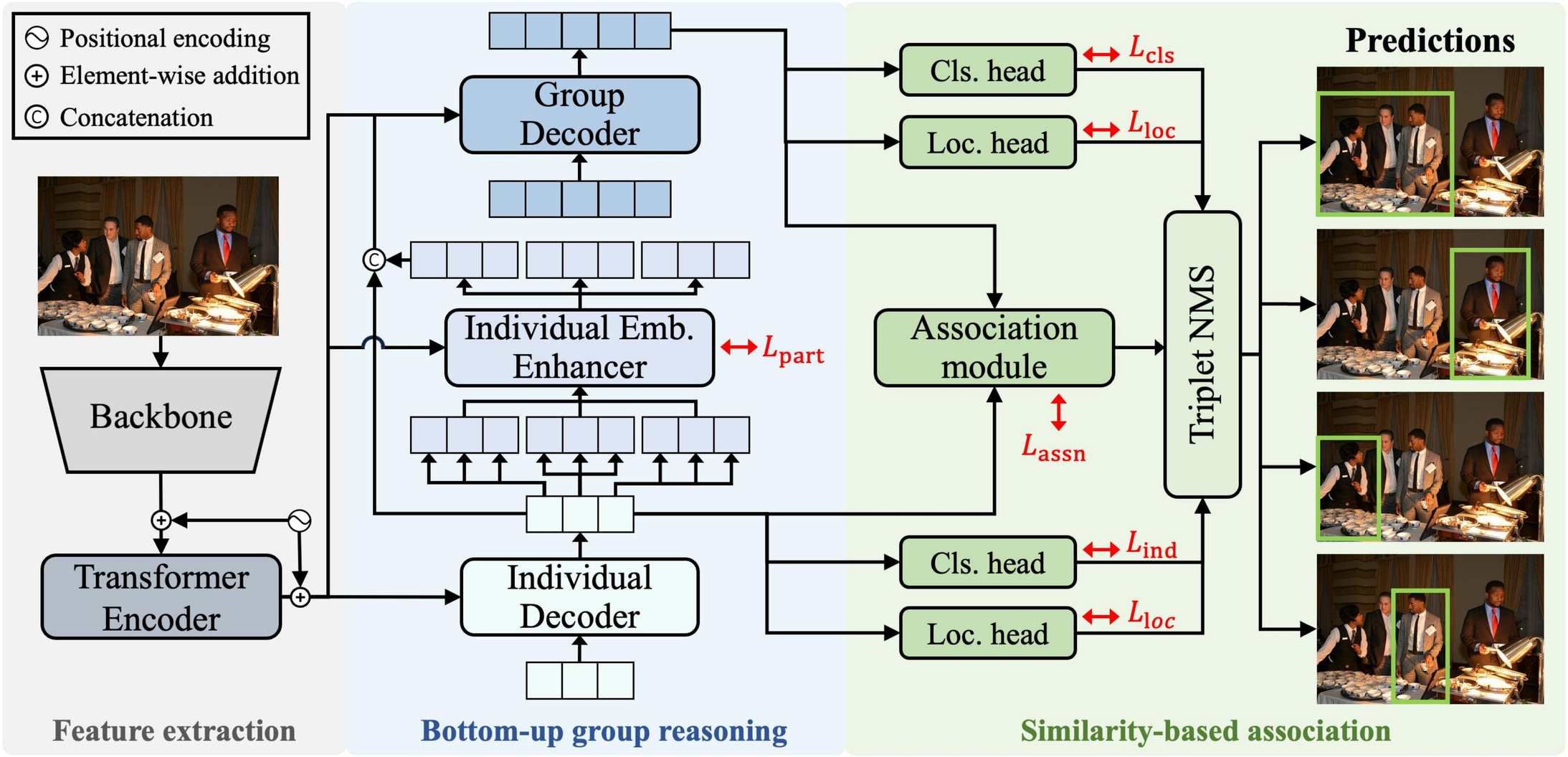
Motivation

Leveraging pose as privileged information to capture part cues

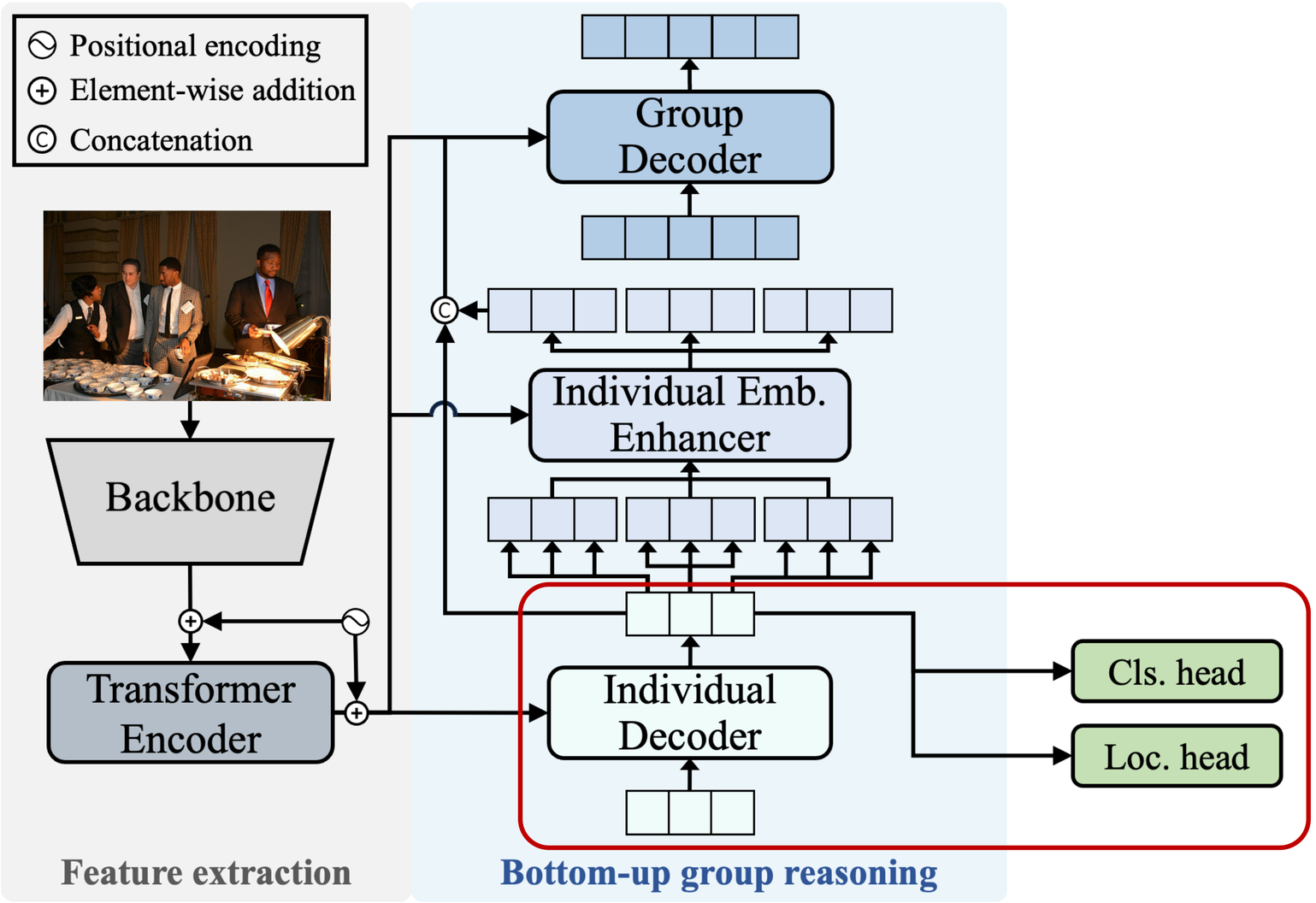




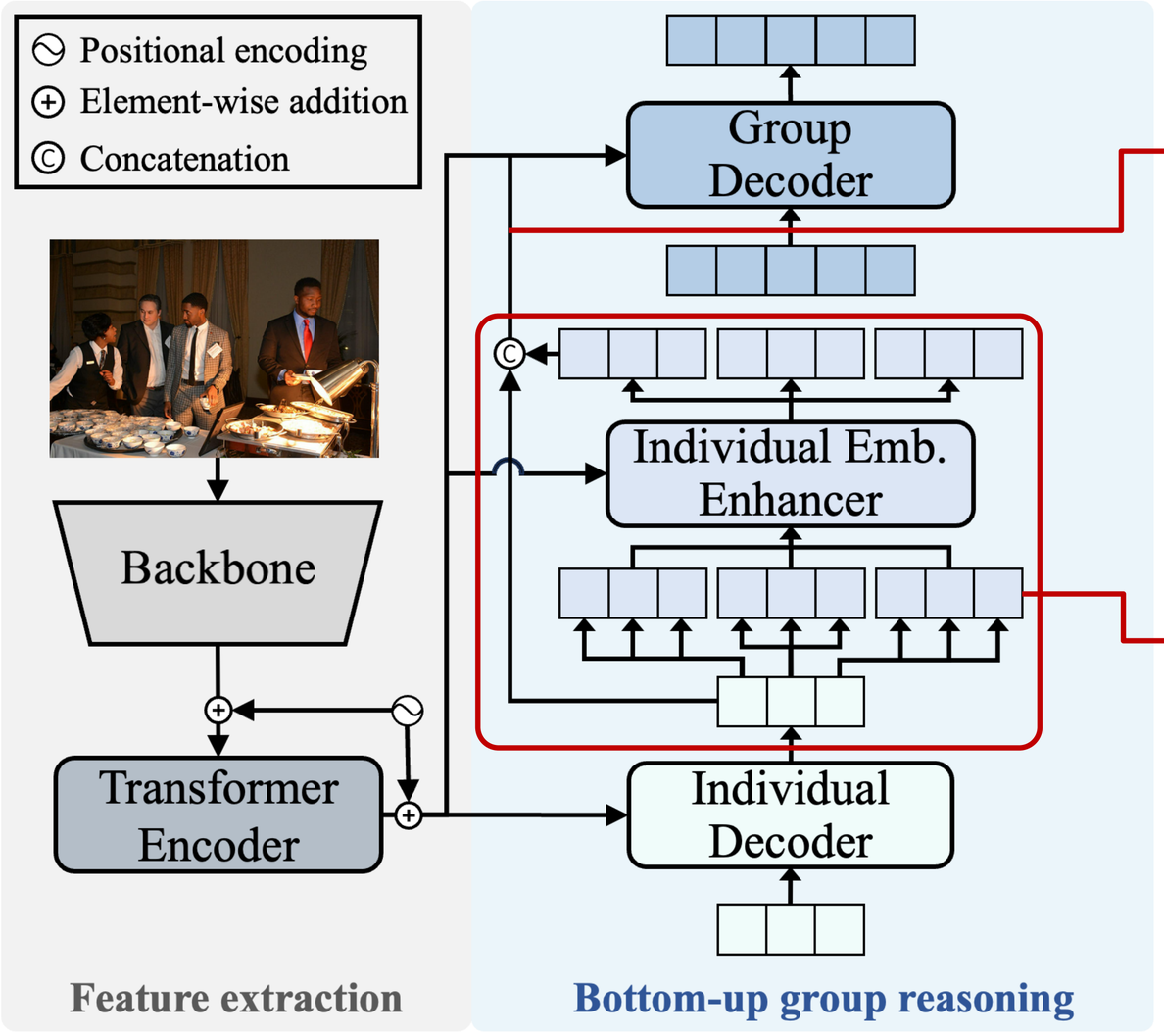
Part-Aware Bottom-Up Group Reasoning



Part-Aware Bottom-Up Group Reasoning



Part-Aware Bottom-Up Group Reasoning



Part-aware individual features

$$E_A = [E_I, E_P^1, \cdots, E_P^P] \cdot W_{fuse} \in \mathbb{R}^{N_I \times D}$$

Individual embedding

Part embedding

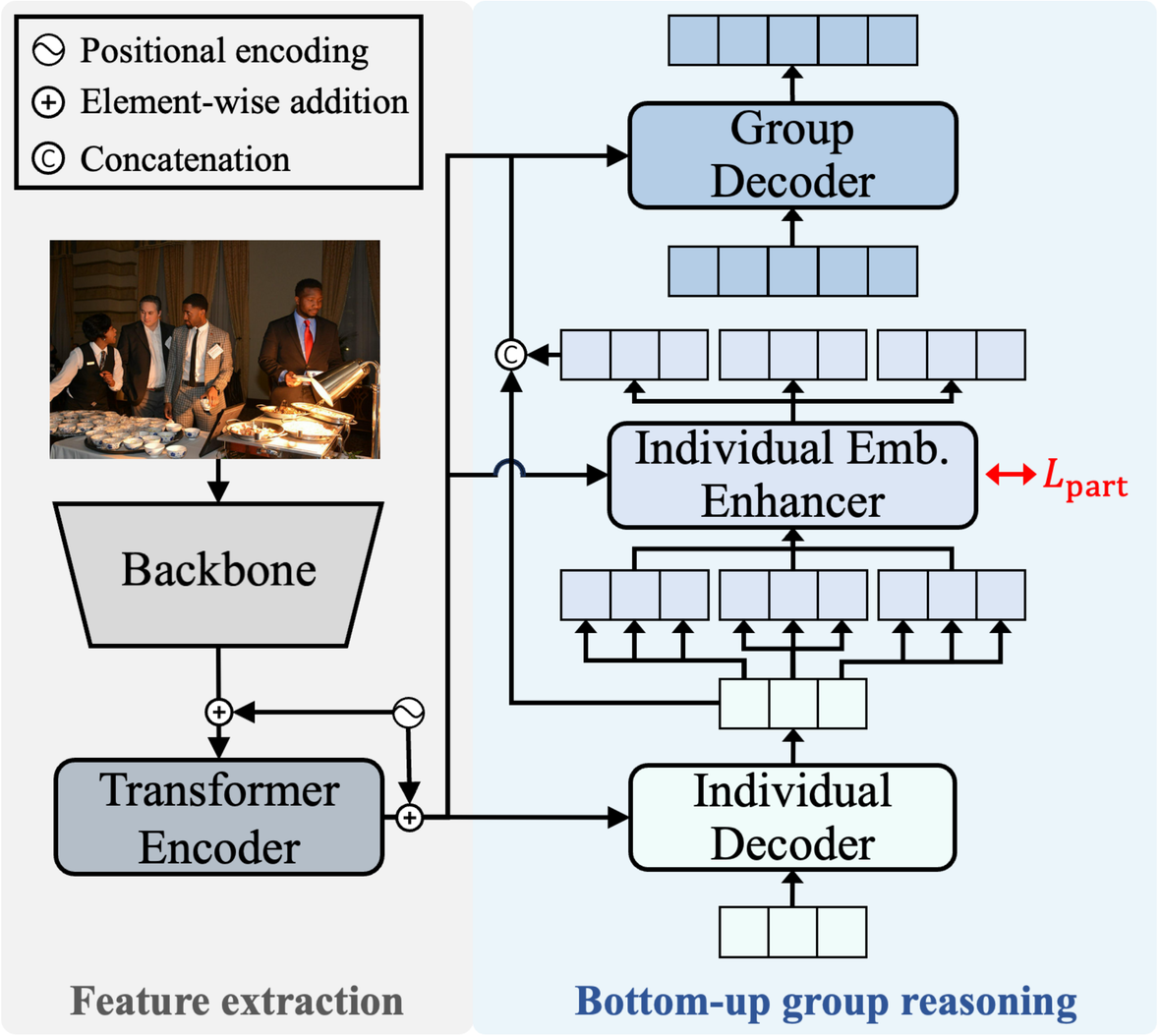
Part-specific queries

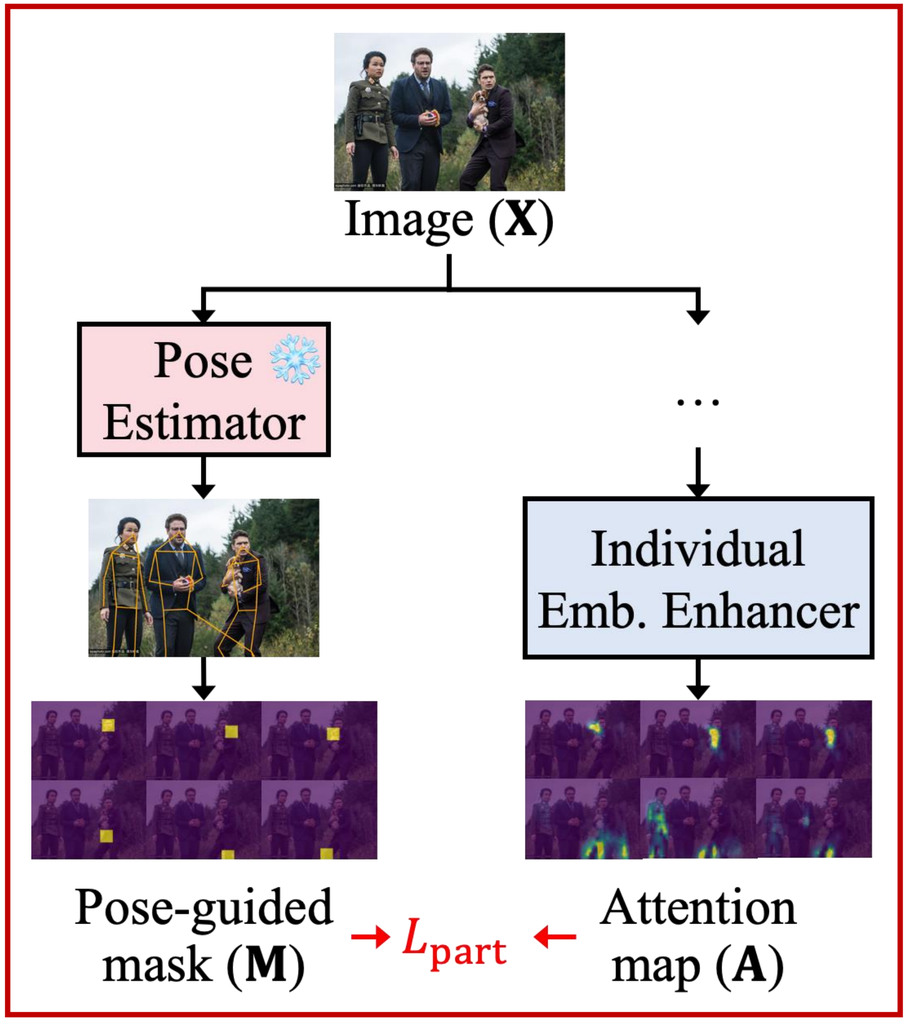
$$Q_P = E_I \cdot [W_1, W_2, \cdots, W_P] \in \mathbb{R}^{N_I \times P \times D}$$

Individual embedding

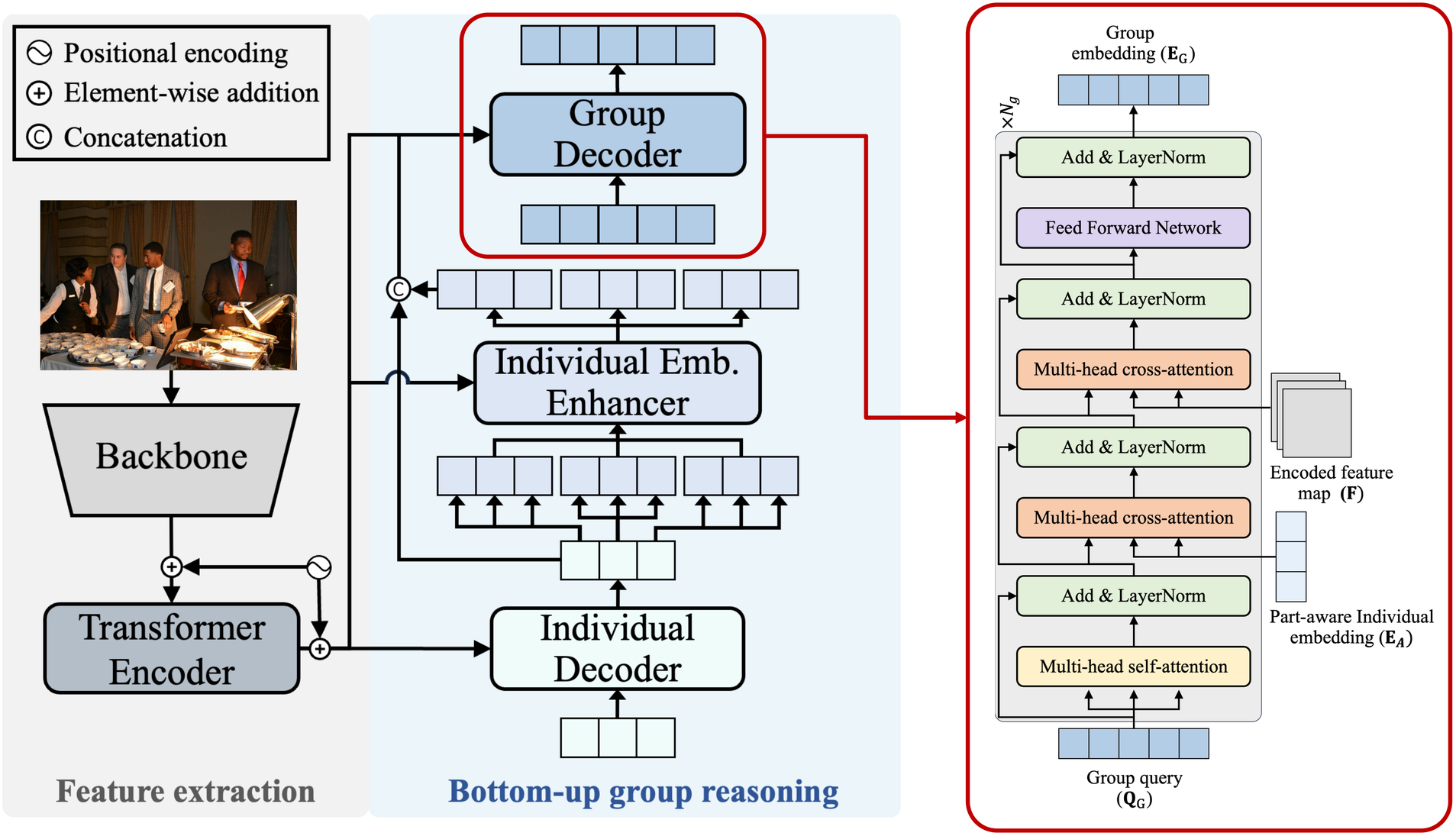
Part projection

Part-Aware Bottom-Up Group Reasoning

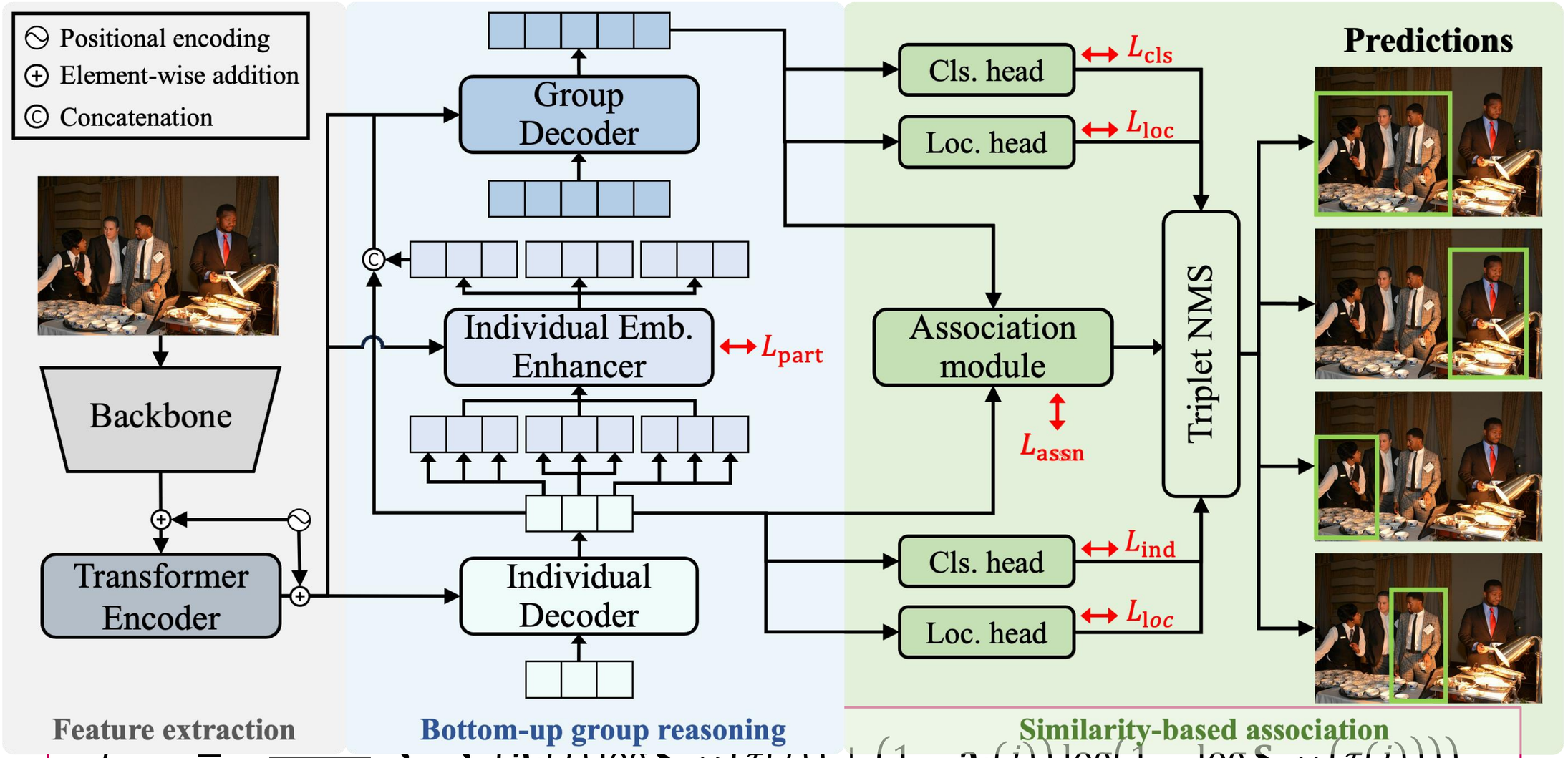




Part-Aware Bottom-Up Group Reasoning



Part-Aware Bottom-Up Group Reasoning



$$\mathcal{L}_{assn} = -\frac{1}{|\mathcal{I}||\mathcal{J}_i|}\sum_i \sum_{j\in\mathcal{J}_i} \left(\mathbf{a}_i(j)\log \mathbf{S}_{\sigma(i)}(\tau(j)) + \left(1-\mathbf{a}_i(j)\right)\log\left(1-\log \mathbf{S}_{\sigma(i)}(\tau(j))\right)\right)$$

Quantitative Result

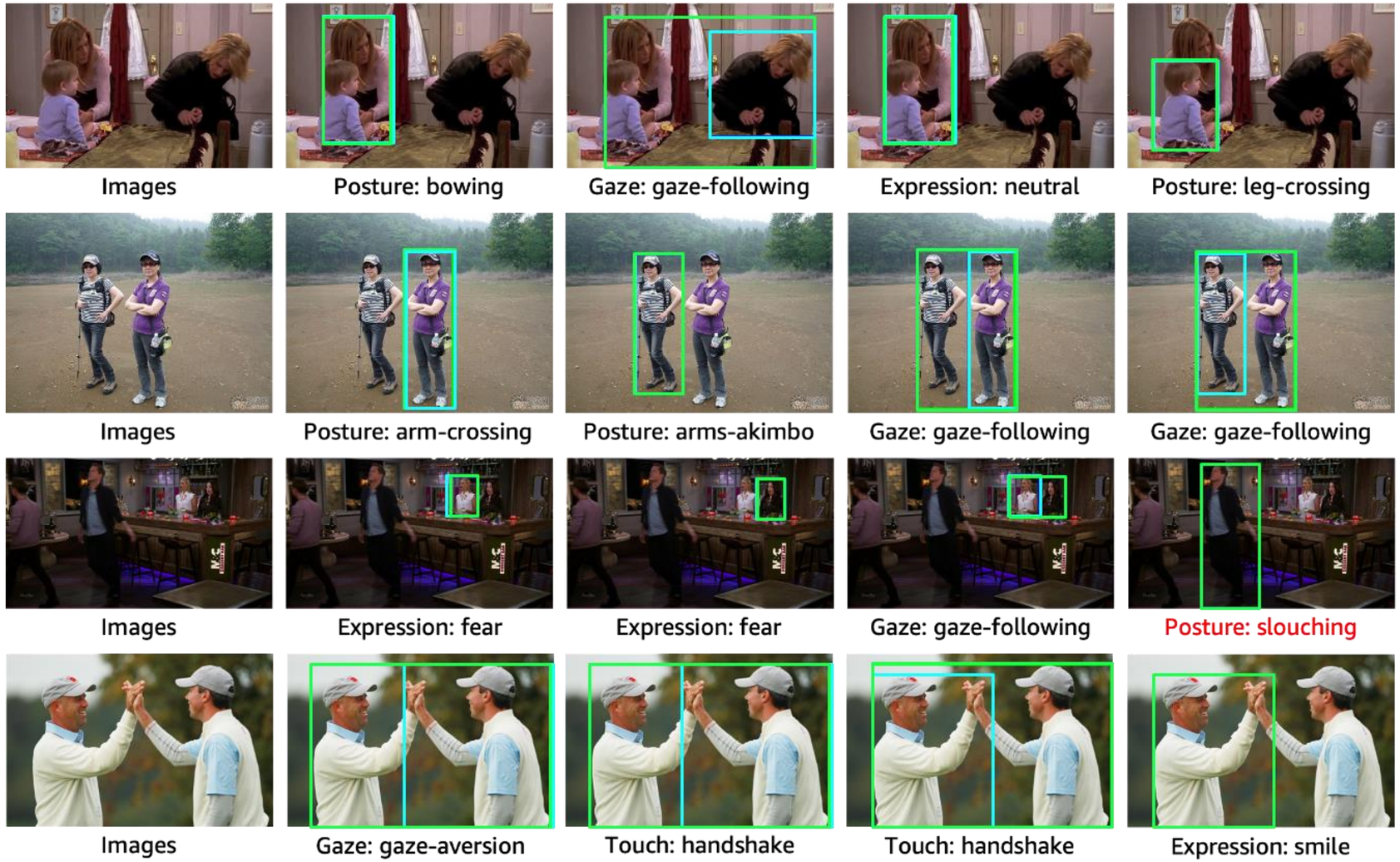
Method	val				test			
	mR@25	mR@50	mR@100	AR	mR@25	mR@50	mR@100	AR
m-QPIC	56.89	69.52	78.36	68.26	59.44	71.46	80.07	70.32
m-CDN	55.57	71.06	78.81	68.48	59.01	72.94	82.61	71.52
m-GEN-VLKT	50.59	70.87	80.08	67.18	56.68	74.32	84.18	71.72
NVI-DEHR	54.85	73.42	85.33	71.20	59.46	76.01	88.52	74.67
Ours	**59.43**	**76.62**	**87.43**	**74.49**	**63.59**	**80.62**	**91.34**	**78.52**

Comparison on NVI dataset

Method	Split by view			Split by place		
	Group mAP 1.0	Group mAP 0.5	Outlier mIoU	Group mAP 1.0	Group mAP 0.5	Outlier mIoU
Joint	9.14	31.83	42.93	6.08	18.43	2.83
JRDB-base	12.63	35.53	31.85	8.15	22.68	33.03
HGC	6.77	31.08	57.65	4.27	24.97	57.70
Café-base	14.36	37.52	63.70	8.29	28.72	59.60
Ours	**18.23**	**46.88**	**67.62**	**10.65**	**39.03**	**63.60**

Comparison on Café dataset

Qualitative Result



Conclusion

- Leverage human pose as privileged information to obtain part-aware representation.
- Propose a bottom-up group reasoning framework.
- Achieve superior performance on NVI and Café datasets.

Poster Session #1

December 3 (Wednesday)

11:00 ~ 14:00

In San Diego

Project Page

Paper