Interaction-aware Representation Modeling with Co-occurrence Consistency for Egocentric Hand-Object Parsing

Yuejiao Su, Yi Wang*, Lei Yao, Yawen Cui, Lap-Pui Chau*
Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong SAR

https://github.com/yuggiehk/InterFormer

Background

Hardware

GoPro [1]

Head-mounted Device [2]

Captured Views

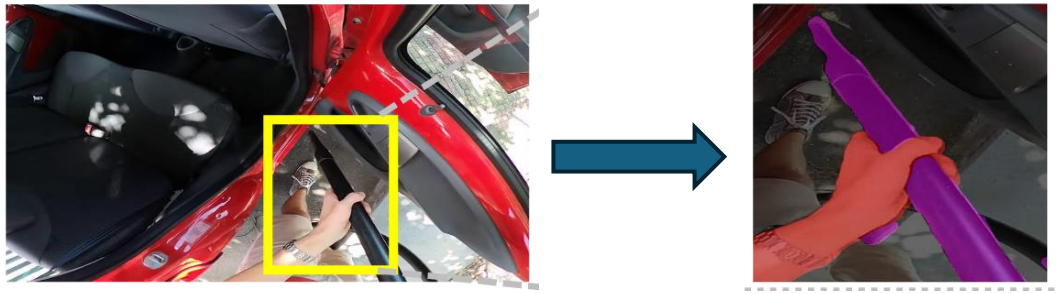
Egocentric [3]

Paired

Exocentric [4]

- Immersive visual information
- Details about interactions with the environment

Fundament Task

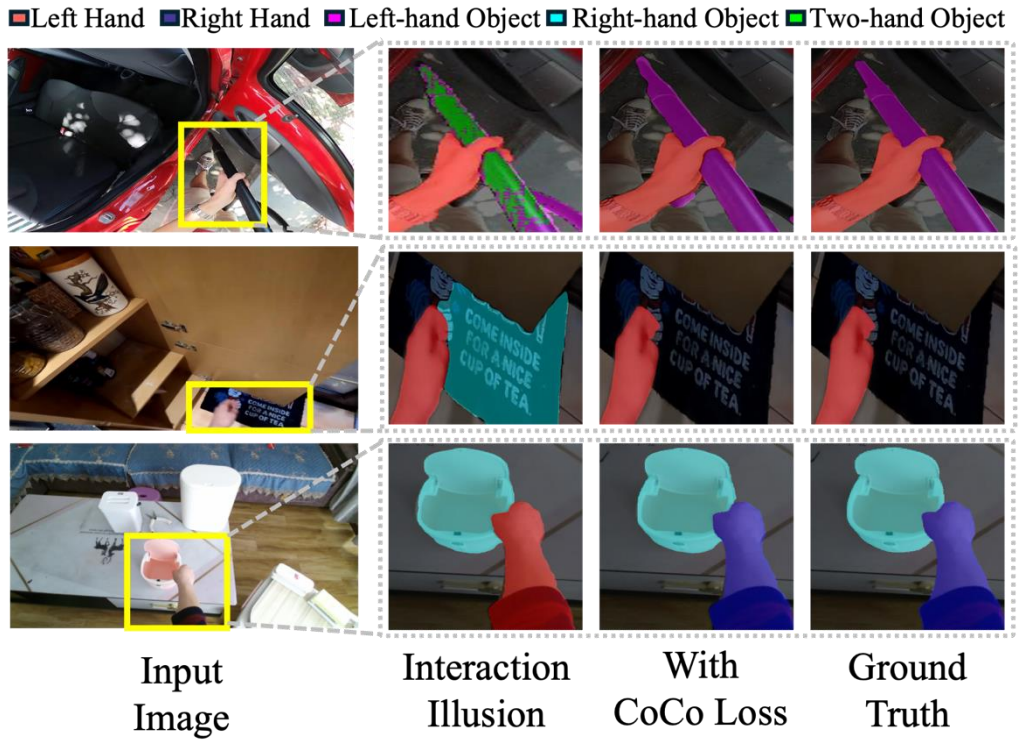


EgoHOS: Parsing the hands and objects involved in interaction.

Motivation

Several key limitations for transformer-based architectures when parsing egocentric hand-object interaction:

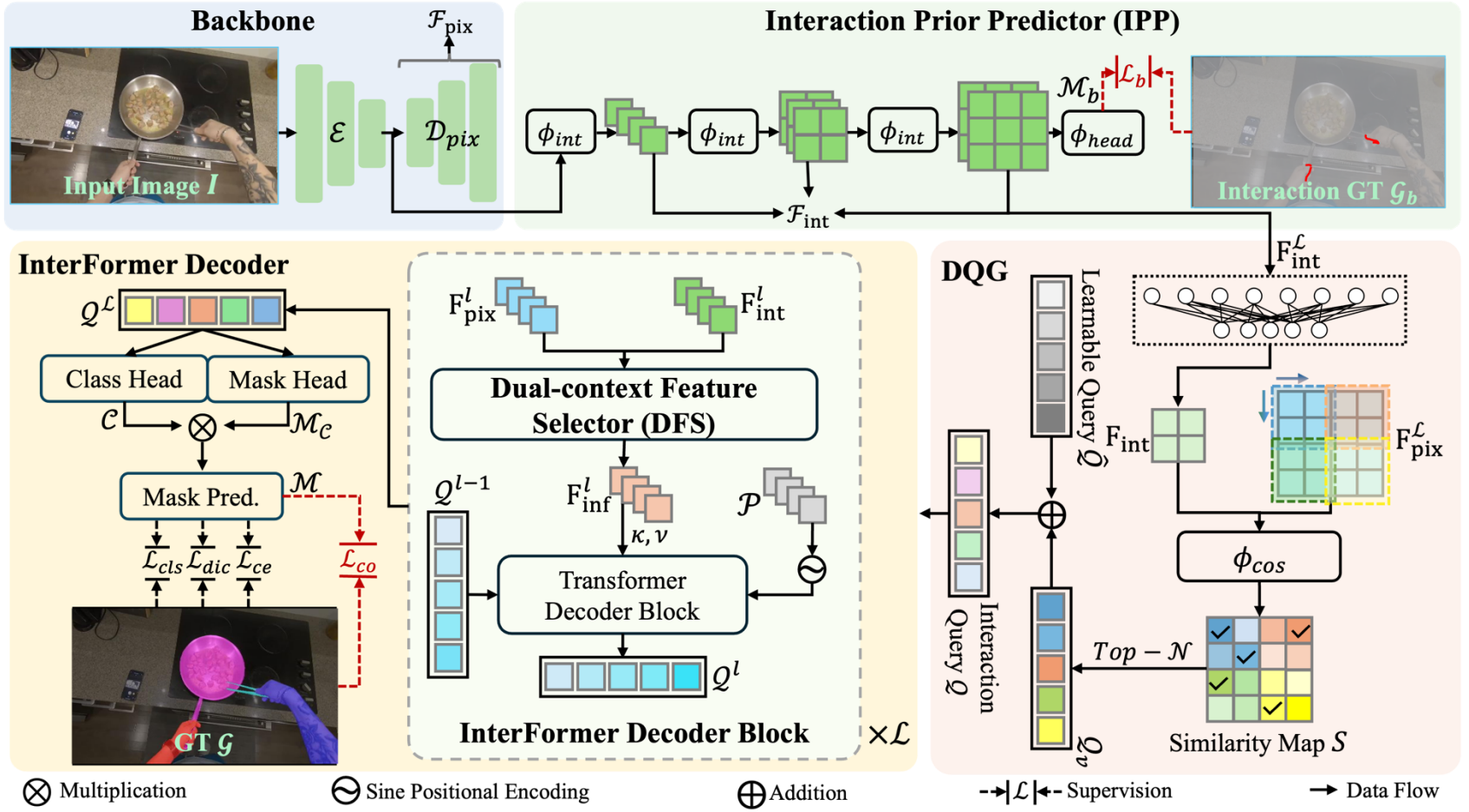
- Previous methods rely primarily on semantic cues or learnable parameters to initialize query, demonstrating **limited adaptability** to changing active objects across varying input scenes.
- Previous methods utilize pixel-level semantic features to iteratively refine queries during mask generation, which may **introduce interaction-irrelevant content** into the final embeddings.
- Prevailing models are susceptible to "**interaction illusion**", producing physically inconsistent outputs.



Contribution

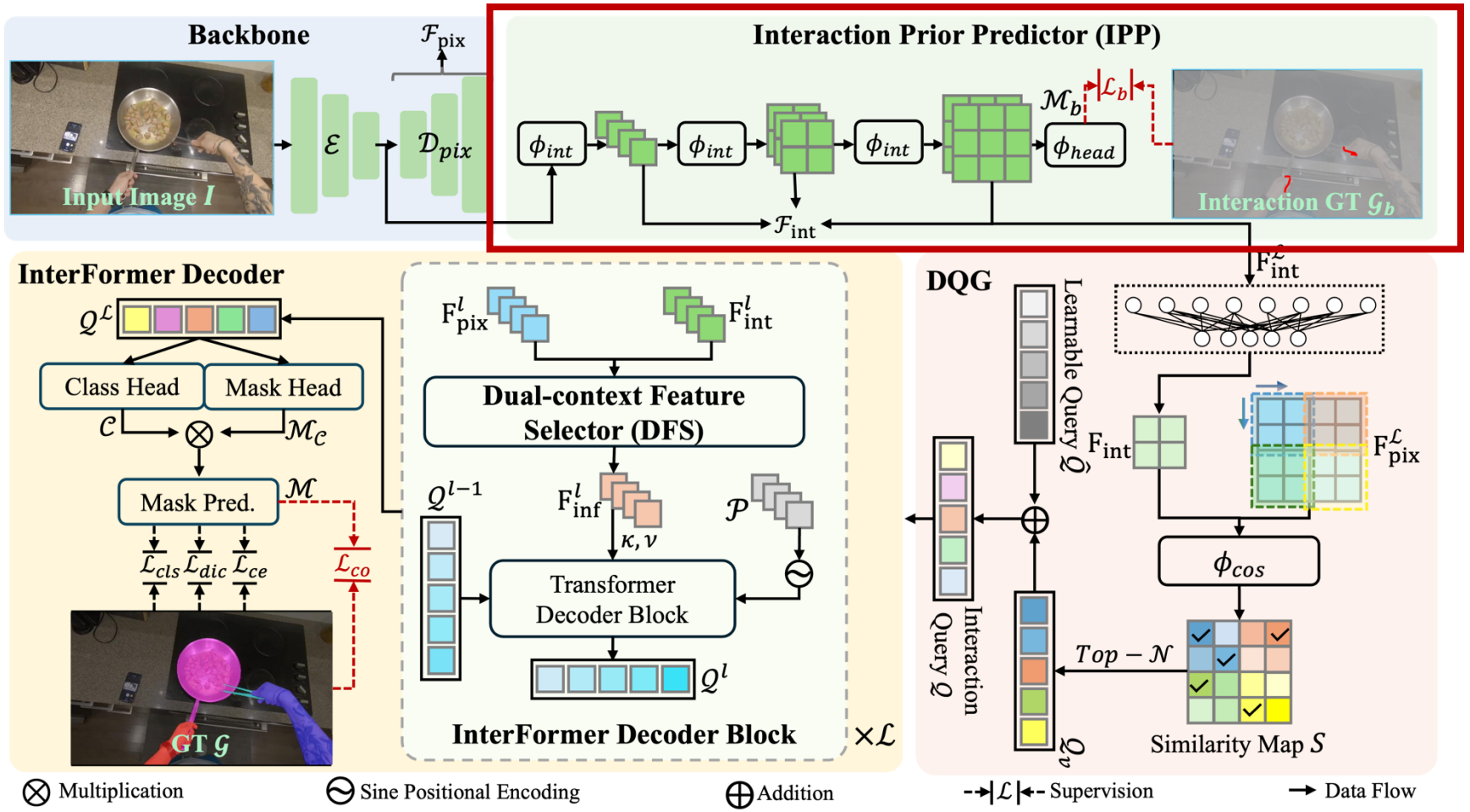
- We establish a novel query initialization paradigm, Dynamic Query Generator (DQG), which generates intrinsically **interaction-aware queries** by fusing coarse interaction-aligned semantic embeddings with learnable parameters, enabling dynamic adaptation to hands and diverse active objects across varying scenes.
- We propose Dual-context Feature Selector (DFS), which introduces an **interaction-centric refinement mechanism** that purifies semantic embeddings through boundary-guided feature fusion, effectively suppressing interaction-irrelevant noise and refocusing the model on contact relationships.
- We introduce a novel Conditional Co-occurrence (CoCo) loss that **encodes intuitive hand-object contact constraints into the learning process**. By penalizing physically implausible co-occurrences, the CoCo loss significantly mitigates the "interaction illusion" problem and improves segmentation consistency.

Framework



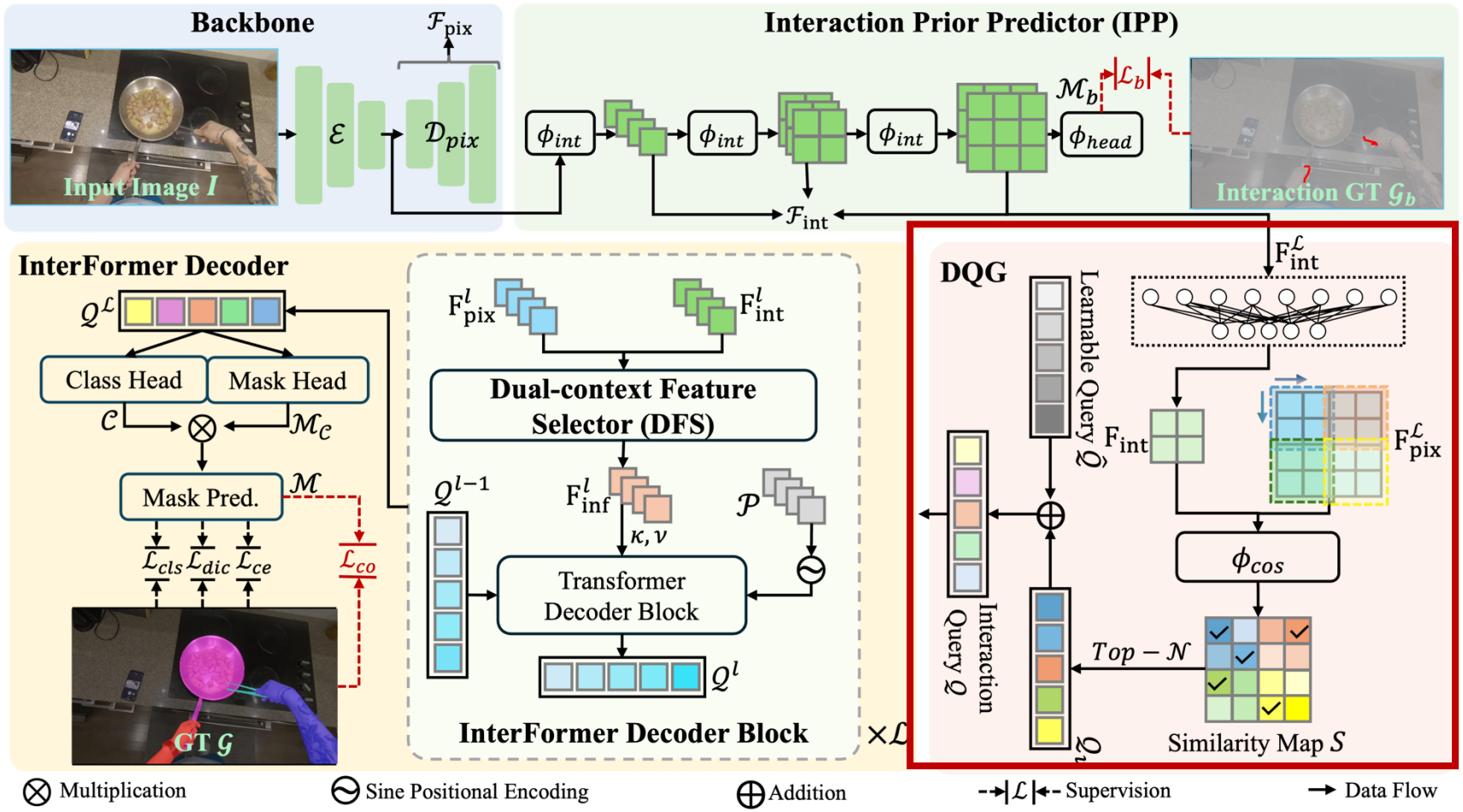
Given an input egocentric image, a backbone network first extracts global and multi-scale pixel-level features. We add an additional IPP branch to extract coarse boundary-guided representations that characterize the interaction. Subsequently, the DQG produces robust and dynamic queries by integrating interaction-relevant contextual information with learnable parameters. Finally, these queries and extracted features are fed into the InterFormer decoder, which employs the DFS to refine interaction-aware representations and generate the final segmentation masks.

IPP



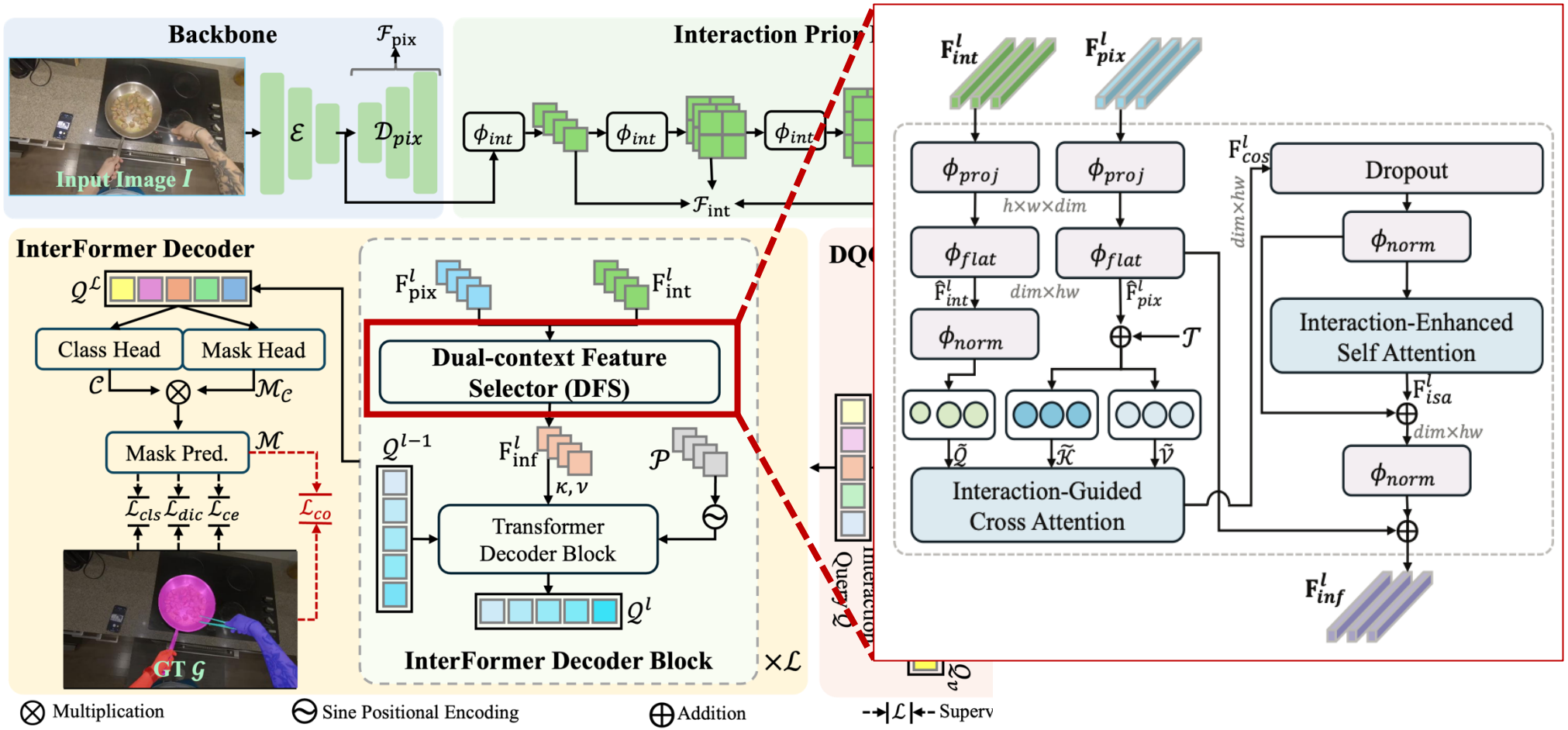
- Motivation: Since different actions involve different interacting objects, identifying active objects cannot rely solely on semantic information, but must depend on their relationship with the hand.
- Objective: To predict the interaction boundary, *i.e.*, the overlapping region between hands and interacting objects.
- Supervision: Interaction boundary ground truth.

DQG



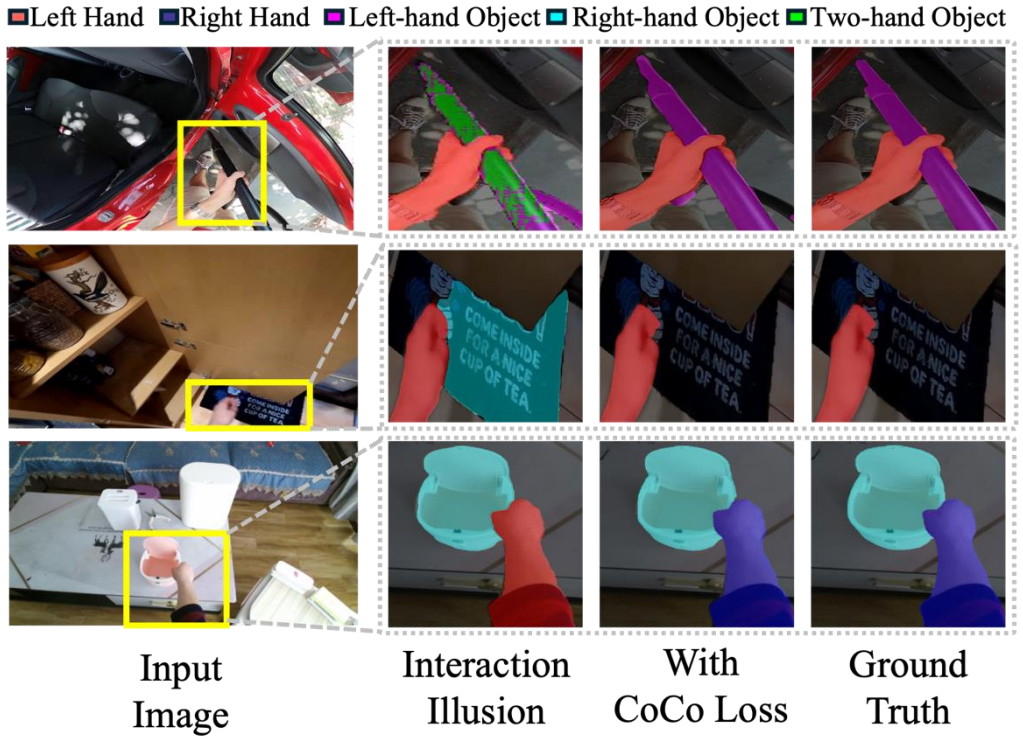
- Key innovation: Grounding query initialization in the dynamic spatial cues of interactions through a two-stage process.
- First stage: Extract interaction-relevant content by selecting semantic embeddings that demonstrate strong correspondence with boundary-guided features, ensuring the selection captures genuine contact relationships rather than relying solely on semantic information as in traditional feature-sampling methods
- Second stage: Synthesize selected features with learnable parameters to generate the final interaction-aware queries.

DFS



- This hierarchical attention mechanism enables progressive refinement of target localization through iterative feature alignment and interaction modeling. After L decoder layers, the final set of queries encodes rich target-specific representations, which are independently decoded into class predictions and mask reconstructions.

CoCo Loss



$$\mathcal{L}_{co}^{left} = (1 - \mathbb{I}_{\{\mathcal{N}_{lh}>\tau\}}) \cdot (\mathcal{N}_{lo} - \mathbb{I}_{\{\mathcal{N}_{lh}>\tau\}} \cdot \mathcal{N}_{lo}) = (1 - \mathbb{I}_{\{\mathcal{N}_{lh}>\tau\}}) \cdot \mathcal{N}_{lo},$$

$$\mathcal{L}_{co}^{right} = (1 - \mathbb{I}_{\{\mathcal{N}_{rh}>\tau\}}) \cdot (\mathcal{N}_{ro} - \mathbb{I}_{\{\mathcal{N}_{rh}>\tau\}} \cdot \mathcal{N}_{ro}) = (1 - \mathbb{I}_{\{\mathcal{N}_{rh}>\tau\}}) \cdot \mathcal{N}_{ro},$$

$$\mathcal{L}_{co}^{two} = (1 - \mathbb{I}_{\{N_{rh}>\tau \wedge N_{lh}>\tau\}}) \cdot (N_{to} - \mathbb{I}_{\{N_{rh}>\tau \wedge N_{lh}>\tau\}} \cdot N_{to}) = (1 - \mathbb{I}_{\{\mathcal{N}_{rh}>\tau \wedge \mathcal{N}_{lh}>\tau\}}) \cdot \mathcal{N}_{to},$$

- Principal: If the predicted mask for a given hand contains fewer pixels than a predefined threshold τ (indicating the absence of that hand), the loss penalizes any prediction of objects associated with that hand.
- This loss discourages implausible co-occurrence patterns, such as recognizing an object as the *left-hand object* when the left hand is not detected.

In-domain experimental results

Table 1: Comparison with SOTA methods on the EgoHOS in-domain test set measured by IoU ↑.

Method	Type	Left Hand	Right Hand	Left-hand Object	Right-hand Object	Two-hand Object	Overall
Segformer (Xie et al., 2021)	T	62.49	64.77	4.03	3.01	5.13	$27.89_{(+45.33)}$
SCTNet (Xu et al., 2024b)	T	81.94	82.12	17.77	16.60	21.74	$44.03_{(+29.19)}$
Para (Zhang et al., 2022)	T	69.08	73.50	48.67	36.21	37.46	$52.98_{(+20.24)}$
Segmenter (Strudel et al., 2021)	T	82.20	83.28	46.22	34.79	51.10	$59.52_{(+13.70)}$
UperNet (Xiao et al., 2018)	C	89.88	91.39	36.22	40.55	45.54	$60.71_{(+12.51)}$
Multi-UNet (Zhao et al., 2025)	C	86.35	87.64	44.80	45.29	46.72	$62.16_{(+11.06)}$
MaskFormer (Cheng et al., 2022a)	T	90.45	91.95	43.51	41.04	54.65	$64.32_{(+8.90)}$
OneFormer(Zhang et al., 2022)	T	90.38	91.95	43.88	44.37	52.64	$64.64_{(+8.58)}$
Mask2Former (Cheng et al., 2022a)	T	90.74	92.25	44.22	46.05	51.13	$64.88_{(+8.34)}$
Seq (Zhang et al., 2022)	T	87.70	88.79	**62.20**	44.40	52.77	$67.17_{(+6.05)}$
ANNEXE (Su et al., 2025b)	L	91.50	92.73	58.94	**57.32**	<u>56.41</u>	$71.38_{(+1.84)}$
Care-Ego (Su et al., 2025a)	T	<u>92.34</u>	**93.64**	60.07	56.69	54.73	$\underline{71.49}_{(+1.73)}$
InterFormer (Ours)	T	**92.51**	<u>93.50</u>	<u>60.86</u>	55.04	**64.17**	**73.22**

Out-of-domain experimental results

ICLR

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Table 2: Comparison results on the EgoHOS out-of-domain test set measured by IoU ↑.

Method	Type	Left Hand	Right Hand	Left-hand Object	Right-hand Object	Two-hand Object	Overall
Segformer(Xie et al., 2021)	T	71.97	71.44	7.60	5.00	4.91	$32.18_{(+40.64)}$
SCTNet(Xu et al., 2024b)	T	87.12	86.29	31.18	19.70	13.32	$47.52_{(+25.30)}$
UperNet(Xiao et al., 2018)	C	93.17	93.96	42.53	28.88	24.35	$56.58_{(+16.24)}$
Multi-UNet (Zhao et al., 2025)	C	92.76	83.08	44.31	39.07	37.15	$59.27_{(+13.55)}$
Maskformer(Cheng et al., 2022a)	T	92.69	94.02	51.81	39.84	39.43	$63.56_{(+9.26)}$
Mask2former(Cheng et al., 2022a)	T	91.46	93.04	53.41	44.90	35.61	$63.68_{(+9.14)}$
Segmenter(Strudel et al., 2021)	T	89.40	90.58	52.73	43.88	42.33	$63.78_{(+9.04)}$
Seq(Zhang et al., 2022)	T	81.77	78.82	46.93	26.40	42.38	$55.26_{(+17.56)}$
ANNEXE (Su et al., 2025b)	L	92.45	93.18	54.39	46.60	40.71	$65.36_{(+7.46)}$
CaRe-Ego (Su et al., 2025a)	T	**94.47**	94.41	51.56	36.80	41.84	$63.82_{(+9.00)}$
InerFormer (Ours)	T	94.38	**94.87**	**66.79**	**55.79**	**52.25**	**72.82**

Table 3: Comparison with SOTA methods on the mini-HOI4D dataset measured by IoU ↑.

Method	Type	Left Hand	Right Hand	Right-hand Object	Two-hand Object	Overall Object
Segformer(Xie et al., 2021)	T	30.16	56.44	5.17	12.02	$25.95_{(+40.12)}$
SCTNet(Xu et al., 2024b)	T	35.83	66.29	17.72	20.98	$35.21_{(+30.86)}$
Multi-UNet (Zhao et al., 2025)	C	52.15	83.64	25.70	41.60	$42.36_{(+23.71)}$
UperNet(Xiao et al., 2018)	C	54.82	84.43	20.34	29.34	$47.23_{(+18.84)}$
MaskFormer(Cheng et al., 2022a)	T	58.50	83.66	35.28	56.91	$58.59_{(+7.48)}$
Segmenter(Strudel et al., 2021)	T	**74.70**	85.58	22.38	58.67	$60.33_{(+5.74)}$
Seq(Zhang et al., 2022)	T	8.74	34.60	23.88	53.96	$30.30_{(+35.77)}$
Mask2Former(Cheng et al., 2022a)	T	70.13	88.57	32.37	55.72	$61.70_{(+4.37)}$
ANNEXE (Su et al., 2025b)	L	68.06	85.13	40.93	57.36	$62.87_{(+3.20)}$
CaRe-Ego (Su et al., 2025a)	T	70.39	**89.76**	27.56	60.08	$61.95_{(+4.12)}$
InterFormer (Ours)	T	66.44	87.07	**46.30**	**64.48**	**66.07**

Ablation study

Table 4: Ablation study results on the EgoHOS in-domain test set.

#	IPP	DQG	DFS	CoCo	Performance (%) mIoU ↑	mAcc ↑
1	–	–	–	–	70.72	77.48
2	–	–	–	✓	70.95	79.02
3	✓	–	–	–	71.23	79.97
4	✓	✓	–	–	71.50	79.68
5	✓	–	✓	–	71.26	79.11
6	✓	✓	✓	–	72.35	80.13
Ours	✓	✓	✓	✓	**73.22**	**80.68**

Table 5: Hyperparameter experiments of τ on the EgoHOS in-domain test set.

#	Hyper Parameter τ	Performance mIoU	mAcc
1	50	71.62	80.62
2	100	**73.22**	**80.68**
3	150	71.97	80.51
4	200	71.62	80.94
5	250	71.10	78.78
6	300	72.38	80.12

Visualization

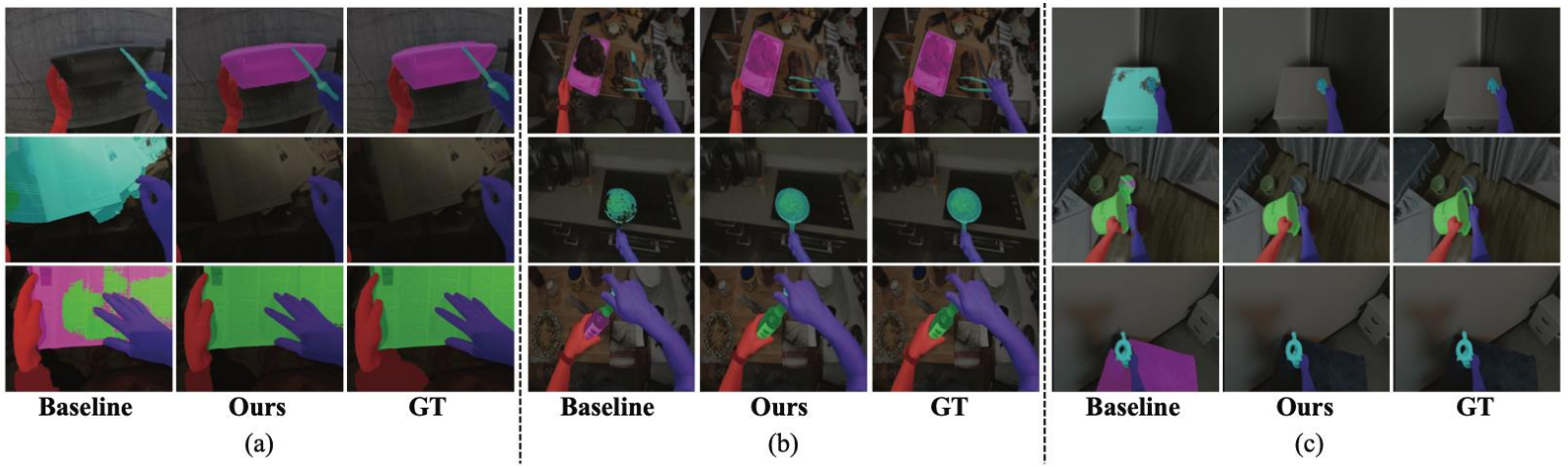


Figure 5: Visualization results on (a) EgoHOS in-domain test set, (b) EgoHOS out-of-domain test set, and (c) out-of-distribution mini-HOI4D dataset.

Conclusion

We propose a novel InerFormer approach for the EgoHOS task.

- Specifically, we introduce the DQG module to create robust queries that can adapt to various interacting objects in different images.
- We further design the DFS module to encourage the network to explicitly perceive interaction-aware features.
- Additionally, our CoCo loss guides the network to learn interaction relationships that are consistent with real-world logic.
- Experimental results on three public test sets demonstrate the remarkable effectiveness and generalization of InerFormer.