Supplementary Material

1. Related work

1.1. Multimodal Large Language Model

Multimodal large language models (MLLMs) extend large language models (LLMs) by integrating visual and textual modalities for unified understanding and reasoning across vision and language. A typical MLLM employs a vision encoder (e.g., CLIP[11]) to extract image features, which are projected into the language embedding space and processed jointly with text tokens through the LLM backbone. This unified architecture enables a wide range of tasks such as visual question answering, image captioning, and referring expression comprehension (REC). MLLMs are commonly trained in two stages: large-scale image–text pre-training for cross-modal alignment, followed by instruction tuning to enhance multimodal reasoning and task generalization. Recent works further improve efficiency and scalability through visual token compression, modality adapters, and alignment-optimized training strategies.

Representative open-source MLLMs include LLaVA [9], Qwen-VL [1], InternVL [5], and InstructBLIP [6]. Most adopt Rotary Position Embedding (RoPE)[12] to encode relative positional information within the LLM backbone. While effective for relational reasoning, RoPE[12] provides no explicit mechanism for representing absolute coordinates in visual space, limiting the spatial precision required for localization-oriented tasks such as REC.

1.2. Visual Grounding

Visual grounding, also known as referring expression comprehension (REC), aims to localize image regions described by natural language. Unlike object detection, it requires fine-grained reasoning over attributes, spatial relations, and context to distinguish similar targets. Early two-stage methods (e.g., MAttNet[17]) generate region proposals and rank them based on multimodal similarity, while one-stage approaches (e.g.,SeqTR[18], GroundingDINO[10]) directly predict coordinates in an end-to-end manner. Recent vision–language foundation models (e.g., ONE-PEACE[14]) improve representation alignment through large-scale contrastive pretraining but remain discriminative rather than generative. The latest MLLM-based frameworks (e.g., Qwen-VL[1], Shikra[4], InternVL[5], VPP[13], Ferret[15], Att-grounding[7]) further reformulate grounding as a generative process, leveraging large language models to directly produce coordinates in text form. This transition enhances reasoning and generalization but still suffers from limited spatial precision due to the absence of explicit positional modeling.

However, traditional methods struggle with complex language reasoning, while MLLM-based approaches still lag behind them at higher IoU thresholds. Our LocBoost bridges this gap through explicit spatial grounding and inference-time refinement.

1.3. MLLMs for Visual Grounding

Recent advances in multimodal large language models (MLLMs) for visual grounding can be grouped into two paradigms: *external model–dependent* and *intrinsic end-to-end* approaches. External model–dependent methods, such as VPP [13] and Att-grounding [7], incorporate pretrained visual experts (e.g., DETR [2], SAM [8]) to enhance spatial perception. Intrinsic

end-to-end approaches, represented by Shikra [4] and Ferret [15], jointly model visual and textual tokens to learn grounding directly. While the former improves localization precision but increases system complexity, the latter maintains a unified architecture yet struggles with fine-grained spatial accuracy.

A key limitation persists in both paradigms. Most MLLMs adopt Rotary Position Embedding (RoPE) [12] to encode relative geometric priors, which offer only implicit positional cues. Visual grounding, however, requires explicit 2D coordinates, and bridging this gap demands complex mappings from token-level relations to absolute positions. This mismatch fundamentally constrains localization precision and motivates our approach to inject explicit positional awareness into MLLMs.

2. Experiments

2.1. Datasets

We conduct experiments on standard REC benchmarks, including the RefCOCO family and the Ref-L4 dataset. **RefCOCO Series.** The RefCOCO family of datasets, including RefCOCO [16], RefCOCO+, and RefCOCOg, provides comprehensive benchmarks for referring expression comprehension. RefCOCO allows free-form descriptions, while RefCOCO+ prohibits location words to encourage appearance-based reasoning. RefCOCOg, in contrast, emphasizes longer and more descriptive expressions. Together, they comprise over 360k referring expressions across approximately 26k images, covering diverse linguistic and visual grounding challenges.

Ref-L4. Ref-L4 [3] is a recently introduced benchmark designed for more comprehensive REC evaluation. It contains 45k expressions over 9.7k images, spanning 365 categories with a larger vocabulary and longer descriptions (average length 24.2 words), making it more challenging than RefCOCO-style datasets.

In our experiments, the model is trained exclusively on the RefCOCO training split, while evaluation is conducted on RefCOCO validation/test sets and on Ref-L4. This setup ensures that performance on Ref-L4 reflects genuine generalization ability rather than exposure to its training data.

2.2. Performance Comparision

2.2.1. Full Localization Results Across All IoU Thresholds

In the main paper, we report IoU@0.5, IoU@0.85, and mIoU for clarity and space efficiency. Here, we provide the full breakdown of localization accuracy across all IoU thresholds from **0.50 to 0.95** (step size 0.10) on the RefCOCO, RefCOCO+, RefCOCOg, and Ref-L4 benchmarks. These extended results enable a more detailed examination of localization precision under varying strictness levels and further substantiate our observations regarding the effectiveness of **LocBoost**.

To complement the main paper, we provide in Tables 3 and 1 the complete performance curves over IoU thresholds 0.50–0.95 on all RefCOCO splits and the Ref-L4 benchmark. These results allow a fine-grained analysis of localization precision and further highlight the strengths of **LocBoost** under increasingly strict spatial alignment requirements.

2.2.2. RefCOCO/+/g: Consistent and Increasing Gains Under Higher Strictness

Across all RefCOCO, RefCOCO+, and RefCOCOg splits, LocBoost-3B and LocBoost-7B deliver **universal improvements at every IoU threshold**, confirming the trend reported in the main paper. The advantage becomes even more striking under high-IoU evaluation:

- **Large high-IoU margins.** At $\text{IoU}_{0.85}$ and $\text{IoU}_{0.95}$, LocBoost surpasses all MLLM-based baselines by dramatic margins, frequently improving accuracy by **10–20+ points** compared with Ferret-7B and Shikra-7B. These thresholds capture fine-grained spatial grounding, where existing MLLMs notoriously break down. LocBoost, in contrast, maintains strong performance even as the IoU becomes extremely strict.
- **Outperforming detector-based models without using detectors.** Although ONE-PEACE relies on explicit region proposals, LocBoost-7B exceeds its performance at nearly all thresholds,

including difficult settings such as RefCOCO val at $\text{IoU}_{0.85}$ (79.15 vs. 74.33), despite operating with a fully end-to-end MLLM pipeline.

- **Superior mIoU across all datasets.** mIoU gains further confirm that LocBoost does not merely improve at a single threshold, but reshapes the entire localization curve, elevating both coarse recall ($\text{IoU}_{0.5}$) and precise alignment ($\text{IoU}_{0.85}$–0.95).s

Overall, these full-IoU results reinforce a key takeaway of the main paper: **LocBoost turns general-purpose MLLMs into reliable high-precision localizers**, a capability that has remained elusive in prior MLLM grounding research.

2.2.3. Ref-L4: High-Precision Localization at a New Level

Table 1 reports exhaustive comparisons on Ref-L4. This benchmark requires reasoning over objects across varying scales and evaluating models at extremely strict thresholds (e.g., $\text{IoU}_{0.9}$).

Several observations stand out:

- **New state of the art at all strict thresholds.** LocBoost-7B achieves **75.86 IoU@0.75** and **53.72 IoU@0.9**, surpassing CogVLM-Grounding (a 17B model) by a large margin at the high-IoU region. These results highlight LocBoost's ability to deliver precise spatial alignment beyond what significantly larger models achieve.
- **Strong cross-dataset generalization without Ref-L4 training.** Even without training on Ref-L4, LocBoost-3B reaches **84.28 IoU@0.5** and **68.71 mAcc**, outperforming CogVLM-17B by 2.6 points in mAcc while using **82% fewer parameters**. This demonstrates that LocBoost's coordinate refinement module generalizes across datasets and object distributions, bridging the gap between semantic reasoning and geometric precision.
- **Robust improvements across object scales.** LocBoost enhances grounding for small, medium, and large objects alike. The 7B variant achieves **94.45 IoU@0.5** on large objects and strong gains on small ones, where prior MLLMs struggle most.

2.2.4. Conclusion: High-Precision Grounding Is Now Possible for MLLMs

The complete IoU sweeps confirm and further strengthen the central claim of our work: **LocBoost fundamentally elevates the localization capability of MLLMs**, enabling them to match, and in many cases surpass, detector-based systems under strict localization metrics, while still retaining the strong semantic reasoning abilities of end-to-end language–vision models.

These results demonstrate that the combination of explicit coordinate modeling and inference-time refinement provides a powerful and broadly applicable solution to the long-standing challenge of achieving precise visual grounding in MLLMs.

2.3. Additional Analysis

2.3.1. Inference-Time Expansion (ITE) Training and Inference Cost

For completeness, we detail the training and inference cost introduced by our Inference-Time Enhancement (ITE). During training, ITE is supervised with a lightweight denoising objective that perturbs the bounding box and asks the model to recover the correct coordinates. This denoise training does not require any extra data and is applied on top of standard grounding supervision, resulting in only a small increase in training computation compared with the base MLLM training. No additional parameters or modules are introduced beyond the existing prediction pipeline.

At inference time, each ITE refinement step feeds the updated bounding box back into the MLLM and performs an additional forward pass. Therefore, increasing the iteration count T leads to a roughly linear cost: each extra refinement step adds approximately the cost of one more MLLM forward. In practice, however, most of the performance improvement occurs within the first one or two iterations, so the overall latency increase is modest.

Overall, ITE introduces minimal extra training cost and a controllable inference overhead, while offering disproportionately large gains in strict localization accuracy under high-IoU evaluation.

Table 1. Full Comparison on RefL4 benchmark (Val+Test).The symbol † denotes models that outputs segmentation masks.

Method	Val+Test				Val	Test	Small		Medium		Large	
	$IoU_{0.5}$	$IoU_{0.75}$	$IoU_{0.9}$	mAcc	mAcc	mAcc	$IoU_{0.5}$	mAcc	$IoU_{0.5}$	mAcc	$IoU_{0.5}$	mAcc
ONE-PEACE	70.82	60.09	36.12	55.07	55.49	54.89	22.18	13.98	83.26	63.39	83.81	70.04
Ferret-7B	57.54	42.44	21.01	40.29	40.31	40.28	30.93	14.57	62.40	43.72	68.18	52.92
Ferret-13B	64.44	49.04	27.46	46.88	47.31	46.71	36.46	17.88	70.50	51.86	73.92	59.09
Shikra-7B	65.06	39.62	10.45	38.60	38.91	38.47	43.91	18.50	75.98	46.27	60.60	39.34
Qwen-VL-Chat	73.80	58.05	37.16	55.94	56.18	55.83	47.66	26.26	79.80	61.06	82.01	68.37
CogVLM-Grounding	81.70	70.77	48.35	66.09	66.25	66.02	**75.06**	**52.85**	86.43	71.31	77.91	66.25
Qwen2.5-VL-3B-SFT	80.11	54.29	38.03	56.47	56.68	55.94	61.10	49.48	79.75	53.62	91.19	68.91
PixelLM-13B†	49.89	35.37	18.42	34.10	34.52	33.92	17.05	8.54	53.40	35.48	67.59	50.34
LISA-Explanatory†	65.12	52.35	38.26	50.77	50.89	50.72	39.11	27.16	70.03	54.61	75.25	61.09
LISA†	66.23	54.02	39.73	52.18	52.44	52.07	39.24	27.49	71.17	56.05	77.01	63.22
GlaMM†	71.90	60.27	45.15	57.89	58.16	57.78	47.07	34.36	77.17	62.28	80.50	67.14
LocBoost-3B (ours)	84.28	74.21	51.02	68.71	69.05	68.49	62.49	42.72	88.85	72.61	92.98	82.55
LocBoost-7B (ours)	**84.71**	**75.86**	**53.72**	**69.95**	**70.39**	**69.63**	63.27	44.08	**89.57**	**74.30**	**94.45**	**84.48**

2.3.2. Training Stability

To verify that our method does not rely on a lucky run, we trained the LocBoost-3B model twice independently. Table 2 reports the mean and standard deviation (std) of the two runs. The differences between the runs are very small across all IoU thresholds, indicating that the training process is stable and the improvements brought by LocBoost are consistent. The main paper reports the better result of the two runs.

Table 2. Mean and standard deviation (std) over two independent training runs of LocBoost-3B on RefCOCO val.

Metric	Mean	Std
$IoU_{0.5}$	92.55	0.06
$IoU_{0.65}$	89.92	0.27
$IoU_{0.75}$	86.44	0.52
$IoU_{0.85}$	77.93	1.23
$IoU_{0.95}$	33.17	0.53

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 1

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, page 213–229, Berlin, Heidelberg, 2020. Springer-Verlag. 1

[3] Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S-H Gary Chan, and Hongyang Zhang. Revisiting referring expression comprehension evaluation in the era of large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 513–524, 2025. 2

[4] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic, 2023. 1, 2

[5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024. 1

[6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2023. Curran Associates Inc. 1

[7] Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9339–9350, 2025. 1

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 1

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1

[10] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino withnbsp;grounded pre-training fornbsp;open-set object detection. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLVII*, page 38–55, Berlin, Heidelberg, 2024. Springer-Verlag. 1

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[12] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C), 2024. 1, 2

[13] Wei Tang, Yanpeng Sun, Qinying Gu, and Zechao Li. Visual position prompt for mllm based visual grounding. *arXiv preprint arXiv:2503.15426*, 2025. 1

[14] Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities, 2023. 1

[15] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 1, 2

[16] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pages 69–85. Springer, 2016. 2

[17] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315, 2018. 1

[18] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer, 2022. 1

Table 3. Overall comparison on RefCOCO, RefCOCO+, and RefCOCOg splits. mIoU denotes the mean localization accuracy averaged over IoU thresholds from 0.50 to 0.95. Only methods that we were able to reproduce are included in this table.

		Traditional Methods		**MLLM-Based Methods**							
Dataset	**Metric**	Grounding DINO	ONE PEACE	Qwen VL-Chat	Qwen2.5-VL 3B-SFT	Qwen2.5-VL 7B-SFT	InternVL 3.5-2B	Shikra 7B	Ferret 7B	**LocBoost 3B(ours)**	**LocBoost 7B(ours)**
RefCOCO val	$IoU_{0.5}$	84.47	**92.51**	88.55	89.39	90.10	88.70	87.01	87.49	92.49	**92.60**
	$IoU_{0.65}$	82.08	**89.32**	80.64	83.12	86.07	73.63	78.61	77.24	89.65	**90.18**
	$IoU_{0.75}$	79.10	**85.24**	69.83	71.14	72.96	65.36	66.65	67.00	85.92	**86.95**
	$IoU_{0.85}$	74.28	**74.33**	59.64	62.21	63.38	50.01	41.70	39.99	77.88	**79.15**
	$IoU_{0.95}$	**45.47**	24.66	19.75	21.38	22.32	15.85	5.85	7.95	32.24	**33.69**
	mIoU	73.08	**73.21**	63.68	65.45	66.97	58.71	55.96	55.93	75.64	**76.51**
RefCOCO testA	$IoU_{0.5}$	88.47	**94.18**	92.27	91.97	92.87	91.60	90.61	91.35	94.29	**95.09**
	$IoU_{0.65}$	85.88	**91.92**	85.50	88.03	89.13	78.33	83.88	84.61	92.47	**92.70**
	$IoU_{0.75}$	83.35	**88.14**	78.38	81.19	81.84	69.14	74.05	71.71	87.99	**89.34**
	$IoU_{0.85}$	77.66	**77.73**	65.65	70.08	71.13	53.81	49.20	45.99	80.28	**81.63**
	$IoU_{0.95}$	**47.55**	26.55	22.09	21.34	22.69	17.50	7.12	6.89	34.04	**35.64**
	mIoU	**76.58**	75.70	68.78	70.52	71.53	62.08	60.97	60.11	77.81	**78.88**
RefCOCO testB	$IoU_{0.5}$	80.24	**89.38**	84.51	84.73	85.92	84.80	81.81	82.45	88.89	**89.40**
	$IoU_{0.65}$	76.57	**84.46**	73.33	75.67	76.79	66.38	70.03	71.02	83.99	**85.38**
	$IoU_{0.75}$	73.19	**79.04**	65.63	68.08	70.39	57.80	57.45	54.25	80.10	**81.63**
	$IoU_{0.85}$	**68.05**	67.67	52.40	56.73	57.86	45.40	32.62	33.82	71.85	**73.46**
	$IoU_{0.95}$	**38.69**	21.28	17.35	20.78	21.17	13.82	3.57	5.84	30.62	**31.68**
	mIoU	67.34	**68.37**	58.64	61.20	62.43	53.64	49.10	49.48	71.09	**72.31**
RefCOCO+ val	$IoU_{0.5}$	73.54	**82.36**	82.82	82.57	84.78	82.70	81.60	80.78	87.67	**87.95**
	$IoU_{0.65}$	71.17	**79.33**	74.67	76.37	77.43	67.90	73.84	73.34	84.45	**85.56**
	$IoU_{0.75}$	68.35	**75.87**	68.27	68.12	69.57	60.58	63.24	62.21	81.82	**82.33**
	$IoU_{0.85}$	64.35	**65.82**	55.70	57.22	58.78	46.47	40.03	39.16	74.05	**75.15**
	$IoU_{0.95}$	**39.74**	22.45	20.09	21.59	22.76	14.84	5.76	6.91	31.98	**32.97**
	mIoU	63.43	**65.16**	60.31	61.17	62.66	54.50	52.89	52.48	71.99	**72.79**
RefCOCO+ testA	$IoU_{0.5}$	81.86	**88.11**	88.59	88.26	89.78	88.40	87.36	87.38	91.32	**92.18**
	$IoU_{0.65}$	79.58	**85.99**	81.96	83.32	83.60	74.55	80.13	80.84	89.68	**89.68**
	$IoU_{0.75}$	76.95	**82.26**	75.06	78.14	79.28	65.82	71.20	69.87	86.27	**86.81**
	$IoU_{0.85}$	71.25	**71.87**	62.17	64.27	65.36	51.48	46.25	43.58	78.20	**79.51**
	$IoU_{0.95}$	**43.29**	26.06	21.79	23.26	24.17	17.03	6.69	6.93	33.83	**35.00**
	mIoU	70.59	**70.86**	65.91	67.45	68.44	59.46	58.33	57.72	75.86	**76.64**
RefCOCO+ testB	$IoU_{0.5}$	63.80	**73.63**	76.79	76.45	78.03	76.60	72.12	73.14	81.41	**82.18**
	$IoU_{0.65}$	60.97	**69.81**	65.47	67.01	68.13	59.42	62.61	61.93	78.52	**78.75**
	$IoU_{0.75}$	58.01	**64.94**	59.07	59.09	60.37	51.97	51.50	48.29	74.86	**75.21**
	$IoU_{0.85}$	54.10	**55.90**	47.13	51.12	52.31	40.70	29.52	29.19	66.91	**68.42**
	$IoU_{0.95}$	**31.21**	18.20	17.32	19.19	19.76	12.97	3.29	2.86	30.11	**31.36**
	mIoU	53.62	**56.50**	53.16	54.57	55.72	48.33	43.81	43.08	66.36	**67.18**
RefCOCOg val	$IoU_{0.5}$	78.25	**85.33**	85.96	85.61	87.78	85.60	82.27	83.93	88.40	**89.24**
	$IoU_{0.65}$	74.47	**81.84**	73.73	75.04	76.30	70.30	63.46	64.58	84.44	**85.99**
	$IoU_{0.75}$	71.49	**77.88**	67.16	68.16	69.70	61.23	52.10	50.07	80.78	**81.92**
	$IoU_{0.85}$	65.89	**69.32**	56.58	60.49	61.38	46.16	30.94	31.60	71.97	**73.39**
	$IoU_{0.95}$	**36.44**	32.07	28.08	25.85	26.12	13.62	3.41	4.97	32.11	**33.44**
	mIoU	65.31	**69.29**	62.30	63.03	64.25	55.38	46.44	47.03	71.54	**72.80**
RefCOCOg test	$IoU_{0.5}$	77.94	**85.20**	86.32	86.12	87.67	85.50	82.19	84.76	88.85	**88.92**
	$IoU_{0.65}$	74.83	**82.02**	74.35	77.31	77.33	69.50	64.24	66.36	85.46	**86.09**
	$IoU_{0.75}$	71.77	**78.63**	67.49	68.24	69.35	61.24	53.47	54.34	81.37	**82.93**
	$IoU_{0.85}$	67.07	**70.45**	56.22	59.56	60.01	47.24	31.50	32.80	74.04	**75.12**
	$IoU_{0.95}$	**38.78**	34.47	27.61	29.15	29.84	15.96	3.45	4.91	33.68	**34.83**