Heuristic-inspired Reasoning Priors Facilitate Data-Efficient Referring Object Detection

Supplementary Material

7. Additional Ablation Studies

7.1. Data-Rich Setting.

Although our primary focus is label-scarce regimes, we also benchmark HeROD under full supervision to assess whether prior injection harms performance when abundant annotations are available. Results are shown in Tab. 5.

Compared with classical ROD models such as MAttNet [50], TransVG [7], and RefTR [22], foundation-style detectors like MDETR [14], DQ-DETR [25], and Grounding DINO [26] achieve substantially higher accuracy across RefCOCO/+/g. Our HeROD variants, which simply inject reasoning priors into these backbones, match or surpass the strongest baselines. For example, HeROD-G improves over Grounding DINO by +0.7 to +1.0 points across the three datasets, while HeROD-U achieves consistent gains over UNINEXT.

Although these margins are smaller than those in scarce-data regimes (likely reflecting the near-saturation of current detectors in full-data training) they confirm two important points: (1) reasoning priors provide measurable benefits even when abundant labels are available, and (2) our integration does not compromise performance in the data-rich case. This is important because explicit priors might restrict model flexibility under full supervision. Instead, HeROD maintains or slightly improves accuracy, confirming that the injected reasoning priors are complementary to learned features rather than competing with them.

7.2. Impact of Pre-trained CLIPSeg

To further examine the role of CLIPSeg, we compare our approach with a direct fusion baseline where CLIPSeg embeddings are simply added to Grounding DINO (see Tab. 6). This naive strategy yields only a marginal gain of **+0.33** points. By contrast, our full HeROD framework achieves a much larger improvement of **+14.25** points under the same setting.

This comparison highlights that CLIPSeg alone is insufficient for improving data-efficient ROD: its relevance maps are often coarse and do not directly influence the learning dynamics of the detector. In addition, further fusing spatial priors yields a modest gain of +0.77 over the baseline, confirming the priors are informative. The much larger gains of HeROD show that the benefit mainly comes from structured integration rather than naive fusion. In HeROD, however, CLIPSeg is reinterpreted as a visual reasoning prior and integrated systematically into the pipeline, affecting proposal ranking, prediction fusion, and the training objective. This principled integration allows CLIPSeg signals to guide both learning and inference, demonstrating that our gains come not from plugging in a pretrained model, but from explicitly embedding heuristic-inspired priors into the detection process.

7.3. Overall Impact of Spatial and Visual Priors

To examine the contributions of spatial (H_s) and visual (H_v) reasoning priors, we conduct an ablation study over the full HeROD pipeline (Tab. 7). Without any priors, the performance is 63.66. Adding only the spatial prior raises the score to 73.53, and integrating both spatial and visual priors further boosts performance to **77.91**. These results show that each prior contributes positively, and together they provide a complementary mechanism that substantially enhances data-efficient ROD, reinforcing the overall effectiveness of the HeROD framework.

7.4. Loss Weight Selection

Our final loss (Eq. (9)) combines three components: classification loss, bounding box loss, and heuristic confidence loss. Following standard DETR practice, we use a 1:5 ratio between classification and bounding box losses. Since the heuristic confidence term aligns more closely with classification than with box regression, we assign it the same weight as the classification loss. We validate this choice in Tab. 8, which shows that varying the relative weights has minimal impact on final performance. This suggests that HeROD's improvements are robust to the exact weighting scheme.

7.5. Results on Cleaned Datasets

Recent work [3] identified annotation noise in RefCOCO and released a cleaned version of the dataset. While the original RefCOCO remains the most widely used ROD benchmark, we additionally evaluate HeROD on the cleaned split to test robustness. As shown in Tab. 9, HeROD achieves consistent improvements (e.g., +14.26 in the low-data regime), demonstrating that our approach is effective across both the original and cleaned datasets. This suggests that HeROD's gains are not sensitive to annotation noise and generalize well under varying data conditions.

Table 5. Top-1 accuracy comparison for the data-rich setting. * represents the reproduced results using only the ROD datasets with the official code.

Method	Fine-tuning	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
MAttNet [50]	w/	76.65	81.14	69.99	65.33	71.62	56.02	66.58	67.27
VGTR [9]	w/	79.20	82.32	73.78	63.91	70.09	56.51	65.73	67.23
TransVG [7]	w/	81.02	82.72	78.35	64.82	70.70	56.94	68.67	67.73
VILLA-L [10]	w/	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71
RefTR [22]	w/	85.65	88.73	81.16	77.55	82.26	68.99	79.25	80.01
MDETR [14]	w/	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89
DQ-DETR [25]	w/	88.63	91.04	83.51	81.66	86.15	73.21	82.76	83.44
UNINEXT* [47]	w/	78.61	80.86	73.72	63.87	68.98	55.33	64.18	64.56
HeROD-U (ours)	w/	81.93	84.87	75.70	68.78	74.73	58.01	69.06	69.19
Grounding DINO* [26]	w/o	50.61	57.43	44.75	51.46	57.06	46.06	60.38	59.51
Grounding DINO* [26]	w/	88.86	91.46	85.73	80.67	86.99	72.76	84.66	84.56
HeROD-G (ours)	w/	**89.57**	**91.46**	**86.06**	**81.79**	**87.67**	**74.11**	**85.40**	**84.78**

Table 6. Impact of CLIPSeg embedding on the RefCOCO dataset.

Fusion of CLIPSeg Embedding	Performance
fuse w/o HeROD pipeline	63.99
fuse w HeROD pipeline	**77.91**

Table 7. Overall impact of different heuristic-inspired reasoning priors.

H_s	H_v	Performance
✓	✓	**77.91**
✓	x	73.53
x	x	63.66

Table 8. Ablation studies on loss weights selection.

$L_{conf} : L_{cls}$	Performance
0.5	77.18
1.0	**77.91**
2.0	77.53

Table 9. Results on the cleaned RefCOCO datasets.

method	0.1% data	2% data
Grounding DINO	60.57	68.64
HeROD	**74.83**	**82.42**

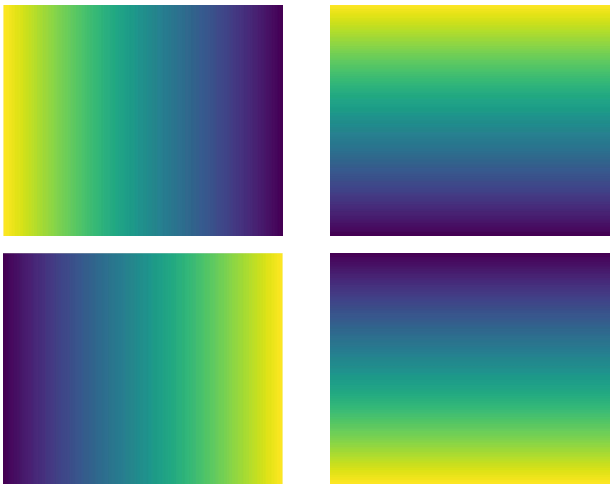

Figure 3. Visualization of four Spatial Heuristic-inspired Scoring Maps: left-related, top-related, right-related, and bottom-related (listed from the first column to the second column, from the first row to the second row). Brighter colors are used to indicate higher values, while darker colors represent lower values.

8. Visualizations

8.1. Visualization of Spatial Prior Maps

As described in Sec. 4, we visualize representative spatial prior maps in Fig. 3. For instance, a "left" map assigns higher scores to regions near the left edge of the image, while a "bottom" map highlights areas closer to the lower boundary. For compound descriptors such as "bottom left," we combine the corresponding maps (e.g., averaging the scores from "bottom" and "left"). These visualizations illustrate how simple positional priors provide interpretable spatial biases that can be directly injected into the detector.

8.2. Performance across Data Volumes

To illustrate the effect of HeROD under varying supervision, we plot performance curves of HeROD and baseline detectors (UNINEXT [47] and Grounding DINO [26]) across different training data volumes (Fig. 4, Fig. 5). The results show that HeROD yields significant improvements in low-data settings, highlighting its ability to enhance label efficiency and accelerate adaptation to the ROD task. Importantly, HeROD also generalizes well to the fully su-

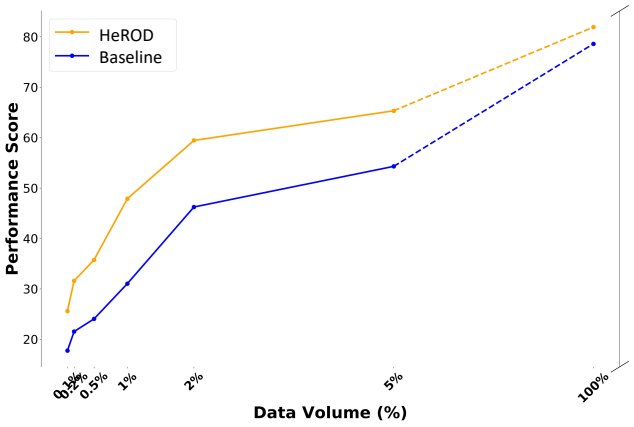

Figure 4. Visualization of performance comparison across varying data volumes between HeROD-U and the baseline UNINEXT.

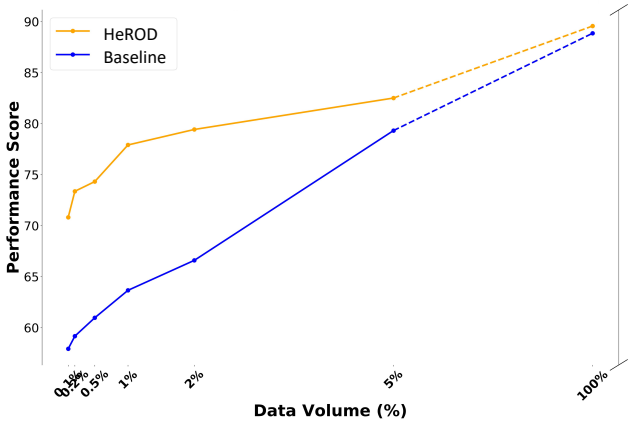

Figure 5. Visualization of performance comparison across varying data volumes between HeROD-G and the baseline Grounding DINO.

pervised regime, maintaining or improving accuracy even when sufficient annotations are available. This confirms that the injected reasoning priors are complementary to learned features across both scarce- and rich-data conditions.

8.3. Illustration of Pre-trained Detector Outputs

As discussed in Sec. 1, we observe that the pre-trained Grounding DINO [26] often fails to handle spatial descriptions in zero-shot evaluation. For example, when the referring phrase specifies "on the left," the model frequently ignores the spatial cue and predicts an incorrect region. Representative outputs are shown in Fig. 6. These errors highlight the model's insensitivity to spatial context and motivate our emphasis on injecting explicit spatial reasoning

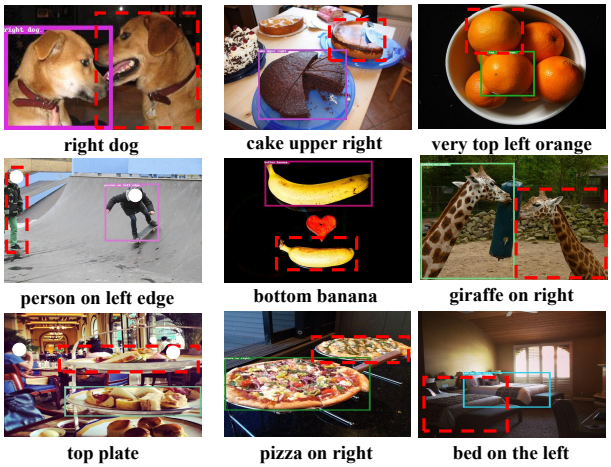

Figure 6. Predictions illustration of the pre-trained Grounding DINO. The ground truth bounding boxes are indicated by **dashed** red lines. It suggests that the vanilla grounding detection may overlook the contextual information in the text descriptions without heavy finetuning.

priors.

8.4. Qualitative Examples and Discussion

We show qualitative examples in Fig. 7 with different cases. In all cases, the baseline detectors fail, while our HeROD correctly localizes the target.

For potential misleading priors, in HeROD, priors provide *soft* guidance rather than hard constraints. Though absolute-direction priors can be imperfect for purely relative relations, their influence is adaptively modulated through stage-wise integration and learned fusion, allowing detector and visual evidence to down-weight misleading cues. As shown in Fig. 7 (e.g., "person right of umbrella"), the framework remains robust under some challenging expressions.

For complex relationships, it is true that our current spatial vocabulary is limited. This is a deliberate design choice: our goal is not to model all spatial relations, but to show that making even simple, explicit spatial priors actionable can substantially improve data efficiency under scarcity. Importantly, HeROD does not rely on spatial cues alone. Visual priors and stage-wise integration allow the framework to handle more compositional expressions in practice (e.g., "the white chair right above the dog" in Fig. 7). The framework is general: more expressive relational priors (e.g., relative, proximity cues) can be encoded as heatmaps and injected in the same pipeline. We view this work as establishing the foundation, with richer priors as a natural extension.

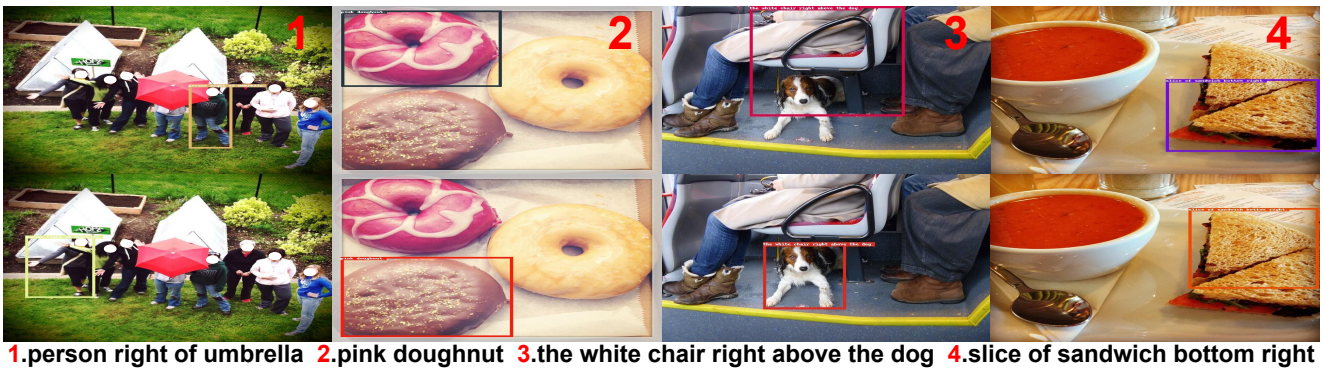

Figure 7. Qualitative examples. The top row shows correct prediction results by HeROD, the bottom row shows incorrect prediction results by the baseline. The responding referring expressions are illustrated below the figure.

9. Zero-Shot Setting

we also evaluate a zero-shot variant by disabling training while retaining prior injection. On RefCOCO, HeROD improves Grounding DINO from 50.61/57.43/44.75 to 59.24/64.96/53.35 (val/testA/testB), showing that it can also enhance zero-shot grounding.

10. Latency Discussion

Baseline (without CLIPSeg) 0.362 s/iter vs. HeROD 0.435 s/iter. GPU memory rise from 4742 to 5810 MB for batch size 8. Training-time overhead is similarly small since CLIPSeg is frozen, runtime remains dominated by detector.

11. Evaluation on Referring Image Segmentation

Although our primary focus is on referring object detection (ROD), we also evaluate the potential of HeROD for downstream referring image segmentation (RIS). We use the high-quality bounding boxes predicted by HeROD and apply HQ-SAM [17], a strong variant of SAM [19], to generate segmentation masks without fine-tuning.

This pipeline differs from conventional zero-shot RIS approaches (*e.g.*, Global-Local CLIP [51], TAS [39], IteRPrimE [42]), which directly predict masks from text without intermediate box proposals. Instead, our method first localizes the referred object via text and then applies segmentation within the localized region. Despite this difference, HeROD achieves strong RIS performance (Tab. 10), showing that accurate localization serves as an effective prompt for high-quality segmentation even without mask supervision.

While not directly comparable to zero-shot RIS methods, this experiment demonstrates the broader utility of data-efficient ROD. By providing robust localization under limited data, HeROD offers a practical stepping stone toward unified vision–language segmentation frameworks that integrate heuristic-driven localization with downstream mask generation.

12. Further Discussion of De-ROD and HeROD

De-ROD as a benchmark. Unlike convergence-acceleration methods such as Adam [18] that improve optimization techniques, De-ROD defines a new evaluation protocol focused on data efficiency in ROD. A successful approach should achieve strong performance in severely limited-data regimes while maintaining or slightly improving results under full supervision. By emphasizing intrinsic label efficiency, De-ROD provides a realistic and stringent test bed for future ROD research, especially in scenarios where large-scale annotations are infeasible. This benchmark highlights both the urgency and the value of designing ROD methods that require fewer labels yet deliver competitive accuracy.

Scope of spatial and visual priors. Our current spatial priors (H_s) are limited to cardinal and compositional directions, and thus cannot capture more complex contextual relations. This limitation is partially mitigated by the visual prior (H_v), which leverages CLIPSeg to align phrase semantics with image regions. Importantly, our approach goes well beyond simply adding CLIPSeg scores or ensembling with a grounding detector. CLIPSeg maps alone are coarse and do not influence detector learning; as shown in Sec. 7.2, naïve fusion yields only marginal gains. In HeROD, CLIPSeg signals are reinterpreted as visual reasoning priors and integrated systematically into proposal ranking, prediction fusion, and the training objective, which enables them to directly shape learning. While CLIPSeg is used here for convenience, other pretrained models that provide local alignment signals could be employed in the same role.

Scope and limitation. Our study targets data-scarce referring object detection on standard benchmarks, under

Table 10. Referring image segmentation results. * denotes that the paper's reported results were kept to one decimal place, and we directly replicated them here.

Method	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val	test
Global-Local CLIP [51]	26.70	24.99	26.48	28.22	26.54	27.86	33.02	33.12
TAS [39]	39.91	42.85	35.85	43.99	50.58	36.44	47.68	47.41
IteRPrimE* [42]	40.20	46.50	33.90	44.20	51.60	35.30	46.00	45.80
HeROD-G (ours)	**78.21**	**80.16**	**74.30**	**71.52**	**76.86**	**63.86**	**72.52**	**72.39**

the common assumption in modern vision–language research that pretrained foundation models are available. As such, HeROD is most applicable to real-world settings like robotics, where language-guided localization is needed but dense labeling is costly. A limitation is that our semantic prior relies on the availability of suitable pretrained models; in niche domains (e.g., medical imaging), effective deployment may require domain-specific foundation models/priors and datasets to produce reliable heatmaps. The framework itself is general, and can be adapted when appropriate domain foundations are established.

Novelty and extensibility. The novelty of HeROD lies in the systematic integration of spatial and visual reasoning priors into DETR-based detectors (Sec. 4.1). Unlike prior work that relies solely on implicit feature learning, HeROD explicitly injects interpretable priors into multiple stages of the pipeline, reducing dependence on large labeled datasets and improving cross-modal alignment. Although this work focuses on spatial and visual priors, the framework is extensible: additional priors, such as structural patterns, domain expertise, or medical knowledge, could be incorporated to further enhance data-efficient referring object detection. For instance, depth-based terms may be feasible within HeROD. Depth cues from a depth estimator can be converted into text-conditioned depth priors and injected as reasoning signals in the same pipeline.