RECS4R: Bridging Semantics and Geometry for Referring Remote Sensing Interpretation

Supplementary Material

6. Supplementary Experiments of RECS4R with VLM and MLLMs Methods on RECS

To further verify the effectiveness of RECS4R on RefDIOR [55], we extend the comparisons in Table 1 from RECS-specific multi-task methods to task-specific VLMs and MLLMs, as illustrated in Table 7. For RIS category methods, visual grounding scores are obtained by converting the predicted segmentation masks into tight bounding boxes, while for VG category methods we derive referring image segmentation results by running SAM2 [63] on the predicted bounding boxes. For MLLMs, we report zero-shot performance without any task-specific finetuning. Under this unified RECS protocol, RECS4R achieves the best results on VG, RIS, and RECS tasks. In visual grounding, RECS4R surpasses the strongest VLM baseline LQVG [33] by 13.33% oIoU and 12.00% mIoU, and in referring image segmentation, it outperforms the state-of-the-art MagNet [14] by 9.24% oIoU and 9.66% mIoU. It is worth noting that our method attains these gains with only 46.75G FLOPs, which is significantly lower than many high capacity RIS models, indicating that using contour vertices as a unified decoding target not only improves accuracy but also reduces computational cost. Moreover, RECS4R exceeds the best MLLM Qwen2-VL 72b [8] by 150.94% on the RECS Sum metric, revealing a large performance gap between current MLLMs and RECS4R and suggesting that ours proposed contour regressional paradigm can serve as a strong building block for future RECS capable MLLMs.

7. Supplementary Experiments of RECS4R with Other Methods on VG Task

To further validate the effectiveness and robustness of RECS4R for remote sensing visual grounding, we conduct additional experiments on the remote sensing OPT-RSVG [37] dataset with three visual backbones including Swin-Transformer [52], ConvNeXt [53], and VMamba [51], and compare our model with existing VG methods in Table 8. Among task-specific approaches, RECS representative method CCFormer [55] achieves the best scores with 86.57% oIoU and 76.47% mIoU, indicating that joint optimization of VG and RIS can bring mutual gains. On top of this strong baseline, RECS4R yields consistent improvements under all three backbones; in the Swin-Tiny [52] experiments, it surpasses CCFormer by 9.10% oIoU and 8.22% mIoU and pushes oIoU close to 95%, indicating that predicting dense contour vertices instead of only four box corners offers higher tolerance to localization errors and leads to more stable grounding. Similar margins are observed with ConvNeXt-Tiny [53] and VMamba-Tiny [51], where RECS4R remains the top-performing method. We attribute these gains to the residual coarse-to-fine encoding (RCE), which injects global priors into fine features, and the channel isolated multi-scale fusion (CIMF), which avoids confusion during the fusion of features at multi-scales and realizes lossless fusion.

8. Supplementary Experiments of RECS4R with Other Methods on RIS Task

To further assess the effectiveness and robustness of RECS4R for remote sensing referring image segmentation, we conduct additional experiments on the RRSIS-D [40] dataset using Swin-Transformer [52], ConvNeXt [53], and VMamba [51] as visual backbones, and compare our framework with representative RIS methods, as summarized in Table 9. Compared with LGCE [40], which first introduces the RRSIS-D [40] benchmark, RECS4R improves oIoU and mIoU by 7.58% and 11.67%, respectively. These gains mainly come from the proposed language-guided unified contour decoding paradigm (LUCDP), which uses contour-based regression to effectively avoid isolated and hollow region artifacts and from the gradient consistency loss (GCL), which adds explicit edge supervision and thus better fits thin, elongated, and topologically complex targets in remote sensing. Across all three backbones, RECS4R consistently achieves the best performance, indicating that the improvements are driven by our fourfold design in representation, refinement, reaggregation, and regularization, rather than by the capacity of any specific backbone.

9. Qualitative Result

9.1. Visual of VG Performance on Remote Sensing.

We provide qualitative comparisons of visual grounding results to further assess the robustness of RECS4R. In Fig. 4, we visualize a bridge example from RefDIOR [55] that represents a typical micro small object in remote sensing, where the target is heavily mixed with cluttered background structures and is easy to confuse with surrounding roads and rivers. Most existing methods completely fail on this case with IoU = 0.0%, and even the stronger task-specific models such as D-MDETR [44], LAVT [83], DMMI [92] and MagNet [14] only roughly hit the region of interest and still produce boxes that are largely mis-

Table 7. **Comparison with task-specific vision-language models and MLLMs methods on RefDIOR test [55] for RECS task.**

Category	Method	Pub.	#FLOPs	VG		RIS		RECS
				oIoU	mIoU	oIoU	mIoU	Sum
VG	TransVG [18]	ICCV'21	29.66G	67.10	46.69	61.42	46.30	221.51
	VLTVG [20]	CVPR'22	27.80G	70.22	51.74	64.34	51.40	237.70
	PseudoQ [60]	CVPR'22	30.49G	69.15	49.73	63.86	49.51	232.25
	QRNet [43]	CVPR'22	31.17G	69.53	50.05	63.41	49.85	232.84
	D–MDETR [44]	TPAMI'23	28.75G	69.94	52.53	63.78	52.14	238.39
	TransCP [75]	TPAMI'23	30.62G	59.56	33.36	55.31	32.73	180.96
	LQVG [36]	TGRS'24	58.60G	81.36	70.68	68.31	64.28	284.63
	LPVA [56]	TGRS'24	30.66G	71.07	52.65	64.65	52.25	240.62
RIS	LAVT [82]	CVPR'22	98.85G	74.38	62.03	79.02	60.81	276.24
	CGFormer [84]	CVPR'23	157.73G	75.18	63.34	79.28	61.31	279.11
	DMMI [92]	ICCV'23	89.84G	74.10	64.89	79.87	63.29	282.15
	CrossVLT [41]	TMM'23	98.26G	73.71	63.32	79.06	62.51	278.60
	LGCE [40]	TGRS'24	104.33G	76.64	64.41	79.55	62.31	282.91
	ReMamber [39]	ECCV'24	96.02G	78.34	66.10	80.22	63.86	288.52
	MagNet [10]	CVPR'24	104.35G	76.93	66.42	80.77	64.79	288.91
	RMSIN [78]	CVPR'24	102.14G	76.19	63.60	78.98	61.57	280.34
MLLMs	Qwen2–VL$_{2b}$ [2]	ArXiv'24	–	39.43	29.25	31.82	27.99	128.49
	Qwen2–VL$_{7b}$ [2]	ArXiv'24	–	44.97	43.08	38.28	42.40	168.73
	Qwen2–VL$_{72b}$ [2]	ArXiv'24	–	51.17	46.58	46.36	46.78	190.89
	InternV$L2.5_{1b}$ [8]	ArXiv'24	–	17.29	14.22	12.26	12.63	56.40
	InternV$L2.5_{2b}$ [8]	ArXiv'24	–	17.87	14.35	12.26	12.87	57.35
	InternV$L2.5_{4b}$ [8]	ArXiv'24	–	25.69	20.32	20.56	19.39	85.96
	InternV$L2.5_{8b}$ [8]	ArXiv'24	–	11.56	18.87	7.60	17.91	55.94
	DeepSeek–V$L2_t$ [54]	ArXiv'24	–	12.52	20.56	7.71	20.57	61.36
	GeoChat$_{7b}$ [93]	CVPR'24	–	43.26	27.61	41.19	27.41	139.47
RECS	RECS4R	–	46.75G	**94.69**	**82.68**	**90.01**	**74.45**	**341.83**

Table 8. **Performance comparison with existing methods on OPT-RSVG test [37] dataset for visual grounding task.**

Method	Pub.	#FLOPs	Swin-Tiny		ConvNeXt-Tiny		VMamba-Tiny	
			oIoU	mIoU	oIoU	mIoU	oIoU	mIoU
TransVG [18]	ICCV'21	29.66G	76.77	61.85	75.49	61.35	79.18	65.52
VLTVG [20]	CVPR'22	27.80G	75.80	61.47	76.05	62.16	78.55	64.84
PseudoQ [60]	CVPR'22	30.49G	76.96	62.03	75.93	61.56	80.28	67.14
QRNet [43]	CVPR'22	31.17G	75.17	60.87	74.31	59.56	78.94	67.35
D-MDETR [44]	TPAMI'23	28.75G	75.51	60.17	74.66	61.93	78.70	66.53
TransCP [75]	TPAMI'23	30.62G	58.28	39.77	56.52	39.36	65.09	44.99
LQVG [36]	TGRS'24	58.60G	84.40	74.81	82.98	74.39	85.26	75.64
LPVA [56]	TGRS'24	30.66G	77.57	63.62	77.21	63.65	78.67	66.65
CCFormer [55]	GRSM'25	119.39G	86.57	76.47	85.49	76.35	87.28	76.74
RECS4R (Ours)	–	46.75G	**95.67**	**84.69**	**95.83**	**87.39**	**95.01**	**82.06**

aligned with the ground truth around 20% on IoU metric. Large multimodal language models including Qwen2-VL [2], InternV$L2.5$ [8], and DeepSeek-V$L2$ [54] tend to output over-sized boxes that cover large portions of the scene, indicating that they lack precise control over object scale. In contrast, RECS4R attains an IoU of 79%, improving the best competing method DMMI [92] by about 33%. The predicted box tightly matches both the position and ex-

Table 9. **Performance comparison with existing methods on RRSIS-D test [40] dataset for referring image segmentation task.**

Method	Pub.	#FLOPs	Swin-Tiny		ConvNeXt-Tiny		VMamba-Tiny	
			oIoU	mIoU	oIoU	mIoU	oIoU	mIoU
LAVT [83]	CVPR'22	98.85G	75.19	57.52	76.25	59.36	76.73	61.44
CGFormer [84]	CVPR'23	157.73G	76.26	59.86	76.54	60.83	76.60	60.59
DMMI [92]	ICCV'23	89.84G	75.82	60.47	76.22	62.18	75.66	60.60
CrossVLT [15]	TMM'23	98.26G	75.44	58.88	76.35	61.02	76.07	62.10
LGCE [40]	TGRS'24	104.33G	75.21	59.37	76.17	60.79	75.66	60.45
ReMamber [79]	ECCV'24	96.02G	77.28	62.35	77.19	63.04	76.79	63.49
MagNet [14]	CVPR'24	104.35G	76.19	62.43	76.51	63.28	76.53	64.21
RMSIN [48]	CVPR'24	102.14G	75.18	58.78	76.24	60.18	76.04	60.66
CCFormer [55]	GRSM'25	119.39G	<u>76.36</u>	<u>66.90</u>	<u>77.30</u>	<u>67.49</u>	<u>77.18</u>	<u>67.58</u>
RECS4R (Ours)	–	46.75G	**82.79**	**71.04**	**81.21**	**69.12**	**82.18**	**68.49**

tent of the bridge, which suggests that the proposed residual coarse-to-fine encoding supplies strong global priors for localization, while the contour-based coarse bounding box supervision further closes the loop between RIS and VG, leading to more accurate grounding on micro small objects.

9.2. Visual of RIS Performance on Remote Sensing.

We further present qualitative comparisons on the referring image segmentation task to assess performace of RECS4R. In Fig. 6, the target is a golf field with highly complex topology and blurred boundaries against the surrounding context, which makes both localization and contour recovery challenging. PolyFormer [45] achieves around 60% IoU, where its prediction roughly covers the correct center region, but the final mask severely distorts the true geometry. Moreover, mask-based methods such as CGFormer [84] and ReMamber [79] can recover a reasonable skeleton of the object, yet their dense mask decoding still tends to produce hole regions or isolated fragments around thin structures, leading to incomplete or noisy boundaries. For current multimodal large language models, this case is even more difficult and they rarely provide a usable referring segmentation mask. In contrast, RECS4R attains about 92% IoU, delivering both accurate region coverage and a complete, clean contour that closely matches the golf field layout. This visualization again confirms that our unified contour decoding refines feature representation for fine-grained structures, while the proposed gradient consistency loss effectively regularizes boundary geometry and suppresses artifacts such as holes and isolated components.

9.3. Visual of RECS Performance on Natural.

We further visualize the RECS4R predictions on the RefCOCO [86] dataset to assess its cross-domain reliability, as shown in Fig. 5. The visual grounding results are highly accurate, with most predicted boxes achieving over 95% IoU, demonstrating that replacing box regression with contour-based decoding in LUCDP effectively stabilizes localization. For referring image segmentation, RECS4R preserves fine structural details even for challenging objects with complex topology, such as the example in Fig. 5(c). This improvement is largely attributed to the gradient supervision introduced by GCL, which enforces contour-level consistency and leads to precise boundary reconstruction. Overall, the visualizations highlight that the unified contour decoding paradigm enables robust and geometry-aware RECS performance in natural domain scenes.

10. Ablation on Contour Vertices Number N_c

Contour vertices serve as the key carrier to resolve representation fragmentation in our unified contour decoding paradigm, so it is important to study the performance affectiveness of contour vertices number N_c . We therefore conduct an ablation by varying N_c from 50 to 100 with a step of 10 based on our proposed RECS4R, as shown in Fig. 7. When N_c increases from 50 to 80, both RIS and VG metrics improve steadily, indicating that more vertices help the model capture fine structures, such as thin roads and irregular building boundaries, and reduce the gap between semantic and geometric representations. However, when N_c exceeds 80, the gains become marginal while the computational cost grows linearly with sequence length and we also observe slightly slower convergence, which suggests that more dense contour sampling brings redundant points that do not provide new shape information but still increase memory burden. Considering both accuracy and efficiency, we set $N_c = 80$ in experiments, which achieves a balance between contour fidelity and computational overhead.

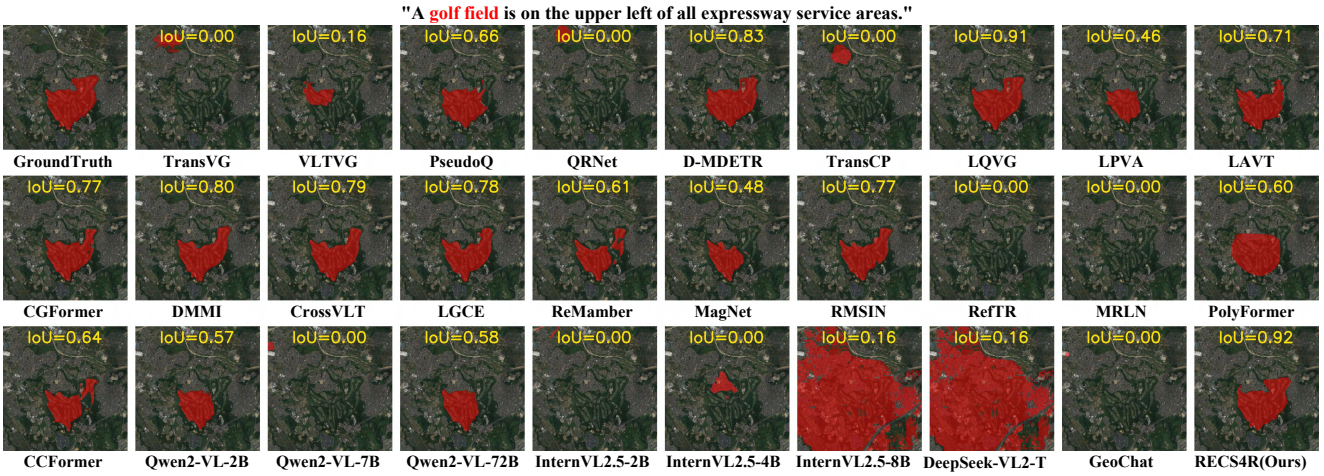

Figure 6. Performance comparison of RIS, VG, RECS and MLLMs methods for referring image segmentation on RefDIOR [55] test set.

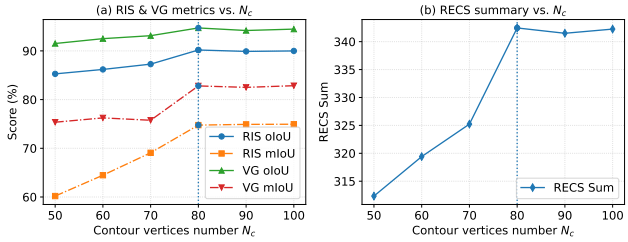

Figure 7. Ablation for contour vertices numberN_c of LUCDP.

Table 10. **Ablation of fine encoding on RefDIOR [55] test set.**

Encoding Type	RIS		VG		RECS
	oIoU	mIoU	oIoU	mIoU	Sum
Single Stage	86.40	60.71	90.33	72.91	310.35
C2F	87.98	67.07	92.24	78.68	325.97
C2F w RA	87.48	68.58	92.76	80.28	329.10
C2F w RCE	**90.01**	**74.45**	**94.69**	**82.68**	**341.83**

11. Analysis of Fine Feature Encoding

To better represent micro small objects in remote sensing scenes, we adopt a coarse-to-fine learning paradigm, where a coarse branch first captures global structure and a fine branch focuses on local refinement. In a standard coarse-to-fine design, the fine stage usually encodes the target from scratch and only weakly reuses the global information from the coarse stage, which easily leads to redundant computation and suboptimal guidance. To fully fuse the global information, we consider two fusion strategies: (i) residual addition (RA), which performs element-wise addition between F_{global} and F_{local} to inject global guidance and prior knowledge into local features; (ii) residual coarse-to-fine encoding (RCE), which sequentially applies channel modulation to reweight semantic channels according to coarse-stage context, spatial modulation to highlight fine-grained regions consistent with the global layout, and cross-modal attention modulation to inject language-guided global information into the fine-stage features, as illustrated in Fig. 2 (R2). Based on RECS4R model, we design the experiments in Table 10. We first note that the coarse-to-fine paradigm outperforms the single coarse processing stage by 6.36% and 5.77% on RIS and VG's mIoU, respectively, confirming the slight effectiveness of the fine processing stage in representing small objects. More importantly, based on coarse-to-fine paradigm, our RCE design surpasses RA in RIS and VG by 5.87% and 2.4% for mIoU metric, respectively. Therefore, we conclude that our RCE design is more effective, as it can provide prior features and global guidance for the fine stage and achieve the refinement.

Table 11. **Ablation of channel isolated multi-scale fusion.**

Fusion Type	w/o Channel Isolated			w/ Channel Isolated		
	RIS	VG	Sum	RIS	VG	Sum
FPN w/o Lang	58.52	42.35	242.94	60.48	44.36	258.48
FPN w/ Lang	59.21	44.24	254.60	**62.11**	**45.62**	**264.54**

12. Effectiveness of Channel Isolated and Languge Guidance for CIMF

To further verify the effectiveness of channel isolation and language guidance in the proposed CIMF module, we conduct a matrix style ablation that toggles these two factors based on baseline PolyFormer [45] using RefDIOR test [55] set, as summarized in Table 11. Comparing the two configurations with language guidance enabled, introducing channel isolated fusion yields clear gains: RIS oIoU and mIoU increase about 1.38% and 2.9%, and the overall RECS Sum improves from 254.60 to 264.54 that shows that assigning different scales to disjoint channel subspaces helps pre-

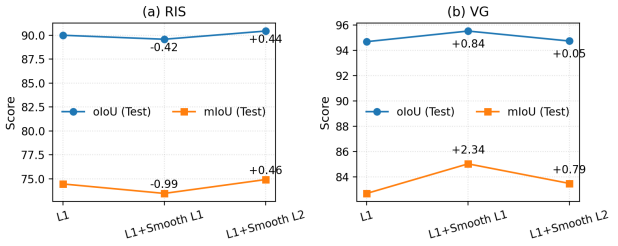

Figure 8. Performance comparison: Smooth-$L1$ vs Smooth-$L2$

serve their scale identity, avoids mixing statistics of tiny and large objects and thus makes multi-scale aggregation less destructive for both segmentation and grounding. At the same time, variants with language guidance consistently outperform their counterparts without text, indicating that the referring expression provides a powerful semantic prior to reweight scale-specific features and suppress distractor regions. In practice, the text-aware fusion allows CIMF to emphasize scales and locations that match the described attributes, which is crucial for correctly localizing micro small objects and visually similar targets.

13. Validation of Contour Sequence Loss

To strengthen the supervision on contour vertices, we build on the baseline that regresses the contour sequence with an $L1$ loss and design three variants in Fig. 8: pure $L1$, $L1$+Smooth $L1$, and $L1$+Smooth $L2$. All three keep the $L1$ term to provide sufficient driving force for large residuals, while the additional smooth terms reshape the penalty curve around zero. Among them, $L1$ + Smooth $L2$ yields the most balanced gains: compared with pure $L1$, it increases the average RIS mIoU from 74.45% to 75.91%, while keeping VG metrics competitive. This is consistent with its gentler penalty curve, which brings small but steady improvements on both RIS and VG. In contrast, $L1$ + Smooth $L1$ behaves like $L1$ in the large error region but approximates $L2$ in the small error region, weakening the strong correction effect around sharp corners and making polygon vertices converge to smoother shapes; as a result, region-level VG benefits significantly, whereas boundary-level RIS is slightly degraded about -0.99% in mIoU and -0.42% in oIoU. Overall, this ablation confirms that the proposed $L1$ + Smooth $L2$ loss achieves a better trade-off between stability and boundary fidelity and we adopt it as the default configuration.

14. Inference Latency and Scalability

Under the setting where images are resized to 512×512 and language sequence length is fixed to 20, RECS4R achieves 13 FPS, compared with 4 FPS for Mask2Former[12], indicating a clear advantage in inference efficiency. Furthermore, benefiting from its autoregressive design, RECS4R provides favorable scalability toward a wider spectrum of tasks and remote sensing multi-modal large language models, providing new ideas for unified multi-task optimization.

15. Analysis of GCL Differentiability

The gradient from M_p to P_f is enabled by SoftRas soft rasterization, which is specifically designed to back-propagate pixel-level losses to geometric vertices. Fig. 9 reports the gradient of P_f when optimizing with GCL only: the gradients are non-zero and gradually decrease during training, thus confirming that GCL effectively propagates to P_f.

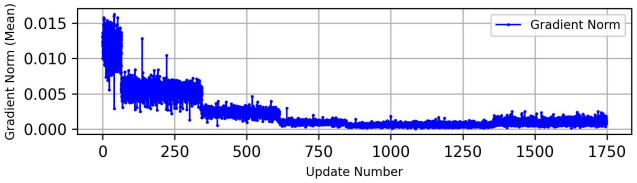

Figure 9. Visualization of GCL Differentiability