PixDLM: A Dual-Path Multimodal Language Model for UAV Reasoning Segmentation

Supplementary Material

In this supplementary material, Sec. A provides additional visualizations of the DRSeg dataset. Sec. B discusses the design of the loss functions. Sec. C presents extended experimental results, covering the baselines of GeoPix and Seg-Zero across the three reasoning dimensions, along with the performance of PixDLM on the ReferSeg benchmark. Sec. D presents ablation studies on alternative Multi-Path Encoder fusion strategies and a one-layer decoder variant.

A. DRSeg

A.1. Prompt Design

Figure A.1 illustrates the instruction template used to construct the reasoning annotations in DRSeg. To ensure consistent and high-quality supervision,we design a unified prompt format that guides the model to generate three types of reasoning signals: Spatial, Attribute, and Scene-level reasoning. For each annotated instance, the prompt requires: (1) a reasoning-oriented question that uniquely locates the target; (2) a structured reasoning chain that analyzes contextual cues such as environmental elements, functional roles, spatial relations, and appearance features; and (3) a concise, verifiable answer. The output is constrained to a strict JSON format, enabling robust automatic parsing and ensuring annotation consistency across all 10,000 samples.

A.2. Dataset Statistics

We provide extended visualizations of DRSeg's key statistics in Table A.2, including scene category distribution, altitude levels, day/night proportions, and object scale diversity. These plots offer a more comprehensive view of the dataset's geometric, semantic, and environmental variability.

A.3. Human Verification of DRSeg Annotations

To systematically assess the annotation quality of **DRSeg**, we invited remote sensing researchers to conduct a human verification study on a stratified random sample of 2,000 images. The evaluation covers three dimensions: reasoning question uniqueness, CoT reasoning consistency, and Mask–text alignment. Each sample was independently reviewed and quantified based on the following criteria.

Uniquely Identifiable Rate: This metric measures whether a reasoning question can uniquely refer to a single target entity in the image. We compute the proportion of samples for which both annotators agree that the question is uniquely identifiable.

CoT Reasoning Consistency: This metric evaluates whether the chain-of-thought reasoning is consistent with the visual evidence and the final target. Each CoT explanation is scored according to the following three criteria, with 1 point awarded for each satisfied condition (score range: 0–3):

- The entities mentioned in the CoT exist in the image;
- The reasoning steps clearly describe the localization process;
- The final predicted target matches the image semantics.

The average CoT score is defined as:

$$\text{CoT-Score} = \frac{1}{N}\sum_{i=1}^{N} s_i,$$

where s_i denotes the CoT score of the i-th sample.

MaskAlign-Rate: This metric measures whether the segmentation mask aligns with the textual description. A sample is considered aligned if (1) the target region coverage is $\geq 95\%$, and (2) the non-target miscoverage is $\leq 5\%$. We report the proportion of samples for which both annotators judged the mask to be aligned.

The results show that the Uniquely Identifiable Rate reaches **95%**, the CoT-Score achieves **2.83** out of 3, and the MaskAlign-Rate is **97%**. These findings demonstrate that DRSeg exhibits highly reliable annotation quality in terms of reasoning target identification, reasoning-chain soundness, and visual-text alignment.

B. Training Configuration

B.1. Loss Design

The overall objective $\mathcal{L}$ constitutes the weighted sum of these losses, calibrated by λ_{ref} and λ_{dice}:

$$\mathcal{L} = \mathcal{L}_{\text{txt}} + \lambda_{\text{ref}}\,\mathcal{L}_{\text{ref}} + \lambda_{\text{dice}}\,\mathcal{L}_{\text{dice}}. \tag{1}$$

L denotes the overall training objective, L_{txt} is the text cross-entropy loss, L_{ref} is the binary cross-entropy loss supervising the referred target mask, and L_{dice} is the Dice loss measuring region-level overlap. Following LISA [1], PixDLM keeps the same CE and Dice loss weights ($\lambda_{\text{ref}} = 2.0$ and $\lambda_{\text{dice}} = 0.5$)

B.2. Trainable Parameters

To preserve the knowledge of the pre-trained multimodal LLM, PixDLM adopts LoRA for efficient fine-tuning while

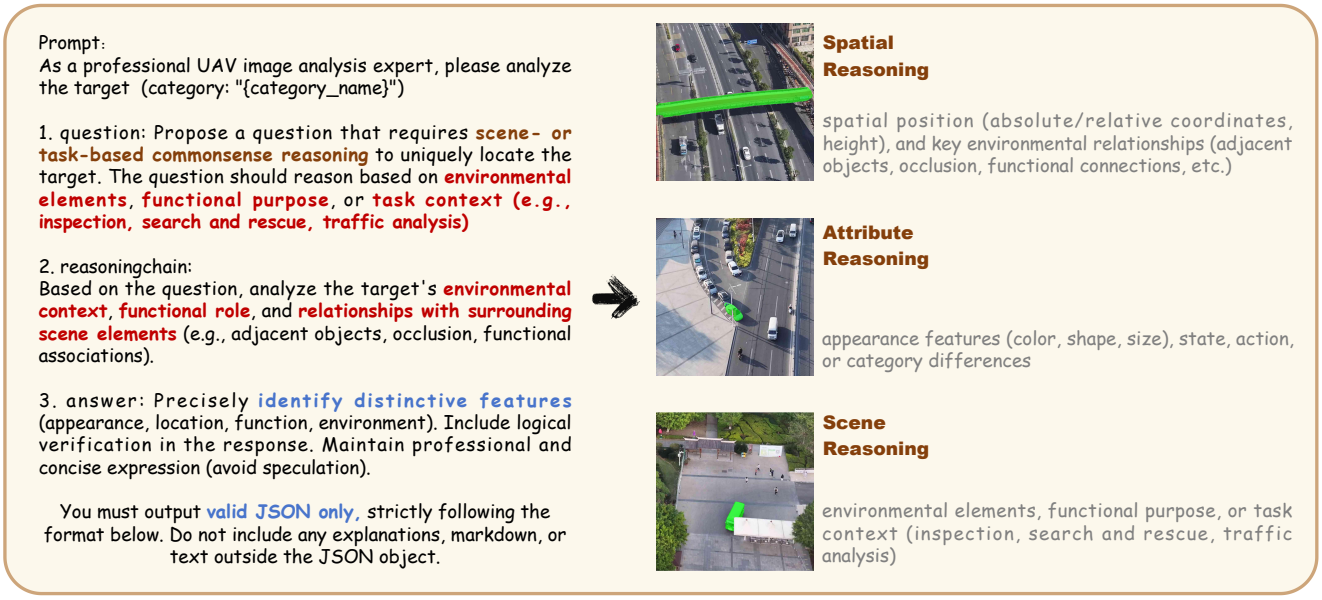

Figure A.1. Instruction template used for constructing reasoning annotations in DRSeg. It includes structured prompts for generating Spatial, Attribute, and Scene-level reasoning, each guiding the model to produce questions, reasoning chains, and answers in a unified JSON format.

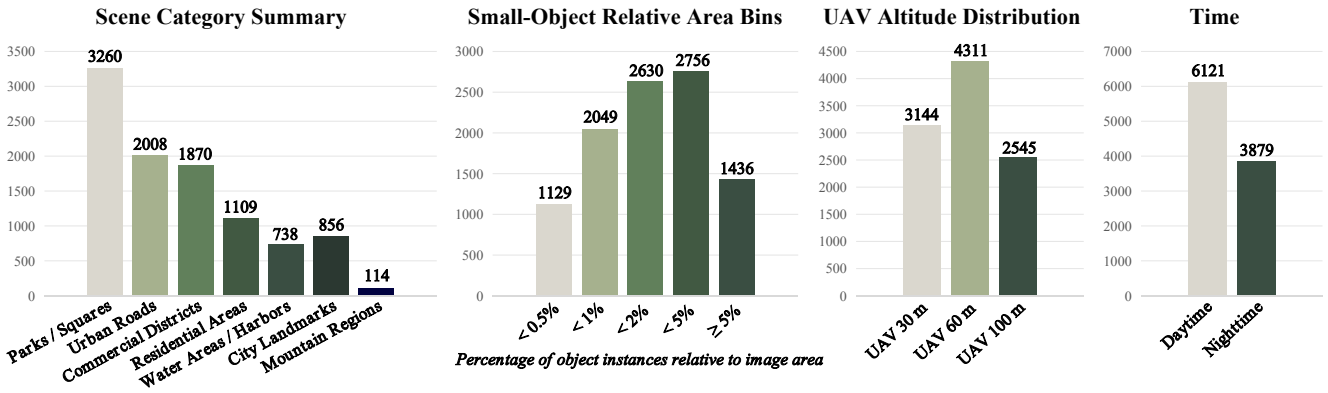

Figure A.2. Dataset statistics of DRSeg, including small-object relative area distribution, scene category summary, UAV altitude distribution, and day/night proportions.

completely freezing the CLIP vision encoder, the SAM vision encoder, and the base LLaMA/LLaVA language backbone (only the inserted LoRA adapters carry gradients). The trainable components include the MultiPath Alignment modules, the hierarchical reasoning decoder, the text-hidden projection MLP, the SAM-to-LLM bridging convolution , the prompt encoder, and the image-feature neck adapter.

B.3. Inference Efficiency and Trainable Parameter Budget

Baseline evaluations on an RTX 3090 GPU show that PixDLM requires an average of 1.12 seconds (0.89 FPS) to process a single 1024×1024 UAV image. This measurement is obtained under the full model configuration, including the multi-level segmentation decoder and the dual-path semanticstructural alignment mechanism. The results indicate that PixDLM maintains stable throughput even with fine-grained structural modeling and cross-path interaction enabled, making it suitable for offline analysis and large-scale batch inference. In addition, the model contains 4,194,304 trainable parameters (4.19M) out of a total of 7,303,624,675 parameters (7.3B), corresponding to a trainable ratio of 0.0574

Table A.1. Zero-shot comparison on the **DRSeg** benchmark across three reasoning dimensions.

Setting	Model Name	Attribute Reasoning		Scene Reasoning		Spatial Reasoning	
		$gIoU \uparrow$	$cIoU \uparrow$	$gIoU \uparrow$	$cIoU \uparrow$	$gIoU \uparrow$	$cIoU \uparrow$
Zero-shot	Seg-Zero-7B [2]	49.32	50.58	25.99	29.56	31.12	36.25
	GeoPix-7B [3]	42.96	46.51	36.84	41.67	36.79	39.73

Table A.2. Results on the referring segmentation benchmark.

Model Name	refCOCO			refCOCO+			refCOCOg	
	val	testA	testB	val	testA	testB	val	test
GSVA-7B [6]	76.4	77.4	72.8	64.5	67.7	58.6	71.1	72.0
LaSagnA-7B [5]	**76.8**	78.7	**73.8**	66.4	70.6	60.1	70.6	71.9
LISA-7B [1]	74.1	76.5	71.1	62.4	67.4	56.5	64.5	66.7
PixelLM [4]	73.0	76.5	68.2	66.3	71.7	58.3	69.3	70.5
Ours	75.2	**80.2**	70.5	**68.5**	**73.3**	**60.6**	**73.3**	**72.8**

C. Extended Experimental Results

C.1. Additional SOTA Comparisons

To further strengthen the comprehensiveness of our evaluation, we additionally benchmark two recently proposed multimodal segmentation models **Seg-Zero-7B** [2] and **GeoPix-7B** [3] under the zero-shot setting in Table A.1 These models have demonstrated strong generalization capabilities in open-world or geospatial reasoning tasks, making them relevant candidates for comparison on the DRSeg benchmark.

C.2. ReferSeg Results

Table A.2 shows that PixDLM, despite being designed primarily for UAV reasoning segmentation, transfers well to standard referring segmentation tasks. The model achieves leading accuracy on multiple splits of refCOCO and refCOCO+, and remains highly competitive on refCOCOg. These results highlight the versatility and robustness of our dual-path design and Multi-Path Alignment module.

D. Ablation Studies

D.1. Ablations on Multi-Path Fusion Schemes

Table B.3 reports an ablation study on the fusion directions of the Multi-Path Alignment module. We compare three alternatives: (*i*) directly summing features from both paths (SAM+CLIP), (*ii*) injecting SAM features into the CLIP branch (CLIP→SAM), and (*iii*) injecting CLIP features into the SAM branch (SAM→CLIP). The results show that the SAM→CLIP direction consistently yields the best performance across all three reasoning dimensions. This confirms that guiding the low-resolution semantic path with high-resolution structural cues is more effective than the reverse integration strategy.

D.2. Ablations on Hierarchical Reasoning Decoder

To better understand the contribution of hierarchical multi-level reasoning, we conduct ablations on non-hierarchical decoder variants, as shown in Table B.4. Specifically, we compare two single-layer decoders: (1) a **Single-Layer Decoder** that performs one-step reasoning without any cross-level fusion, and (2) a **SAM2.1-L Decoder** baseline that replaces our decoder with the mask prediction head from SAM2.1-Hiera-L.

Both variants operate without hierarchical fusion, thus isolating the effect of multi-stage reasoning depth. The results demonstrate that all single-layer designs exhibit a clear performance drop across attribute, scene, and spatial reasoning dimensions. In particular, our Single-Layer Decoder achieves higher scores than the SAM decoder baseline, indicating that even without hierarchical fusion, a reasoning-oriented decoder provides better cross-modality alignment than a purely segmentation-oriented head.

However, both single-layer variants lag significantly behind our full hierarchical decoder (Table 4 in the main paper), highlighting the importance of progressive multi-level reasoning for capturing complex UAV semantics, long-range relations, and fine-grained spatial cues. These results verify that hierarchical fusion is essential for robust reasoning segmentation under UAV-specific challenges.

References

[1] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: reasoning segmentation via

Table B.3. Ablation study of the **Multi-Path Alignment Fusion Strategy** on the **DRSeg** benchmark. We compare three fusion directions: **SAM→ CLIP**, **SAM+CLIP**, and **CLIP→ SAM**. Metrics (%) are reported as $gIoU \uparrow$ and $cIoU \uparrow$.

Fusion Strategy	**Attribute Reasoning**		**Scene Reasoning**		**Spatial Reasoning**	
	$gIoU \uparrow$	$cIoU \uparrow$	$gIoU \uparrow$	$cIoU \uparrow$	$gIoU \uparrow$	$cIoU \uparrow$
SAM + CLIP	60.11	58.73	56.80	56.23	59.34	60.32
CLIP → SAM	<u>60.24</u>	<u>60.28</u>	<u>57.12</u>	<u>57.84</u>	<u>59.36</u>	<u>60.97</u>
SAM → CLIP	**62.80**	**62.84**	**61.75**	**64.03**	**62.51**	**62.80**

Table B.4. Ablation study of single-layer decoders on the DRSeg benchmark. We compare our Single-Layer Decoder (no fusion) with the SAM Decoder baseline.Metrics (%) are reported as gIoU ↑ and cIoU ↑.

Decoder Layer	**Attribute Reasoning**		**Scene Reasoning**		**Spatial Reasoning**	
	$gIoU \uparrow$	$cIoU \uparrow$	$gIoU \uparrow$	$cIoU \uparrow$	$gIoU \uparrow$	$cIoU \uparrow$
1-layer	47.06	50.07	45.16	51.35	45.81	53.27
sam2.1-l	46.43	48.80	44.78	49.90	46.16	51.95

large language model. In *CVPR*, pages 9579–9589, 2024. 1, 3

[2] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *CoRR*, abs/2503.06520, 2025. 3

[3] Ruizhe Ou, Yuan Hu, Fan Zhang, Jiaxin Chen, and Yu Liu. Geopix: Multi-modal large language model for pixel-level image understanding in remote sensing. *CoRR*, abs/2501.06828, 2025. 3

[4] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *CVPR*, pages 26364–26373, 2024. 3

[5] Cong Wei, Haoxian Tan, Yujie Zhong, Yujiu Yang, and Lin Ma. Lasagna: Language-based segmentation assistant for complex queries. *CoRR*, abs/2404.08506, 2024. 3

[6] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. GSVA: generalized segmentation via multimodal large language models. *CoRR*, https://doi.org/10.48550/arXiv.2312.10103, 2023. 3