Interpreting Object-level Foundation Models via Visual Precision Search

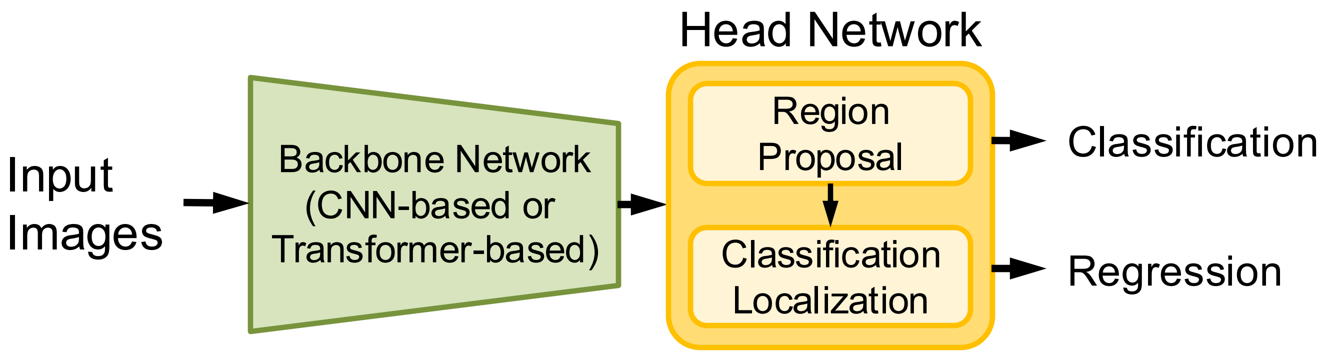
(Highlight Paper)

Ruoyu Chen, Siyuan Liang, Jingzhi Li, Shiming Liu, Maosen Li,
Zhen Huang, Hua Zhang, Xiaochun Cao

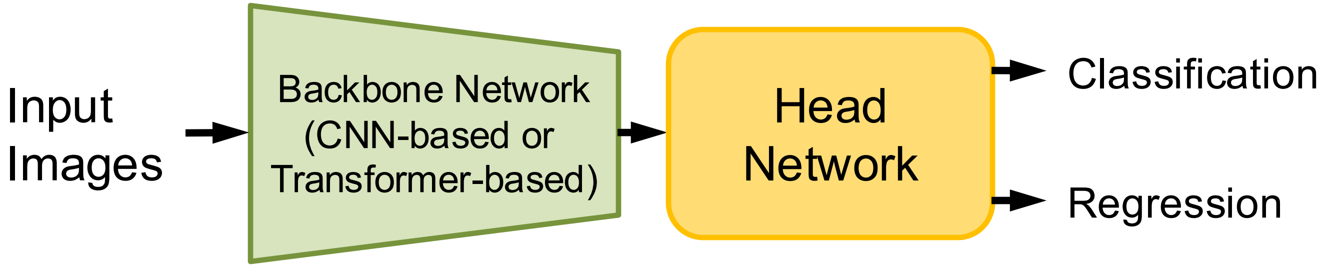
Code: https://github.com/RuoyuChen10/VPS

Related Work

Traditional Detector Types



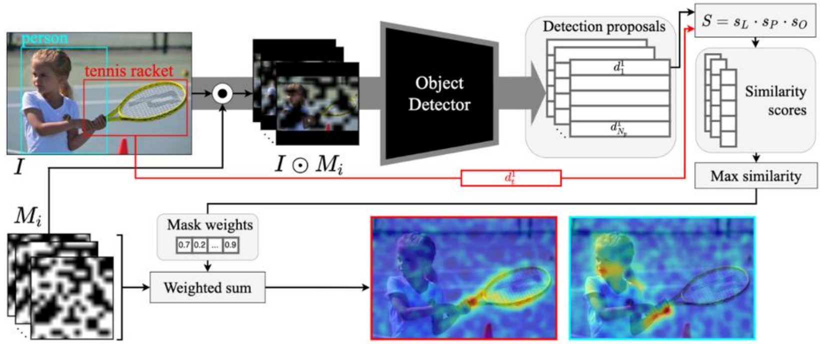
Two-stage Object Detector



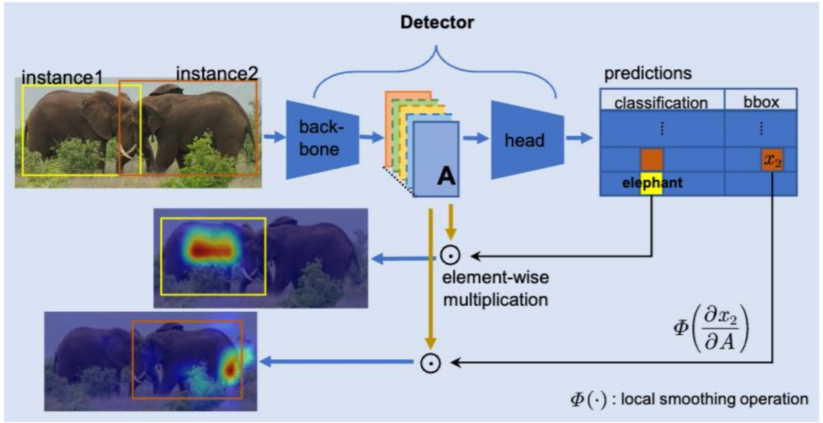
One-stage Object Detector

Corresponding Interpretation Methods

D-RISE[CVPR 21, Oral], a perturbation-based method.



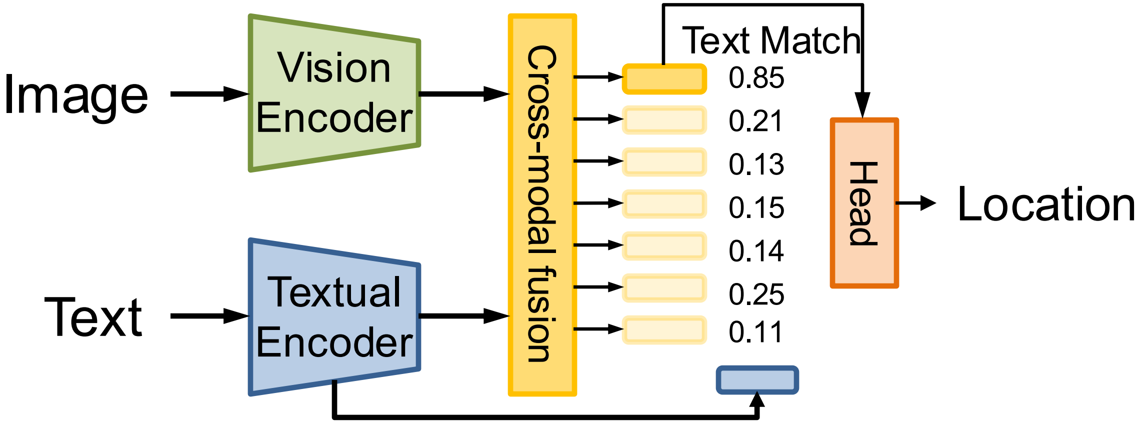
ODAM[TPAMI 24], a gradient-based method.



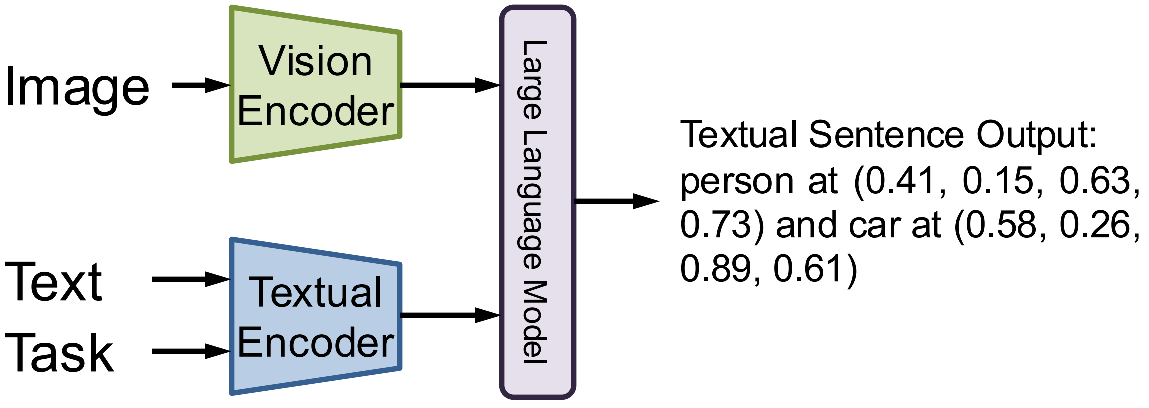
[1] Petsiuk, Vitali, et al. Black-Box Explanation of Object Detectors via Saliency Maps. CVPR 2021: 11443-11452
[2] Zhao, Chenyang, et al. Gradient-Based Instance-Specific Visual Explanations for Object Specification and Object Discrimination. TPAMI 46(9): 5967-5985 (2024)

From Traditional Detectors to Multimodal Foundation Model

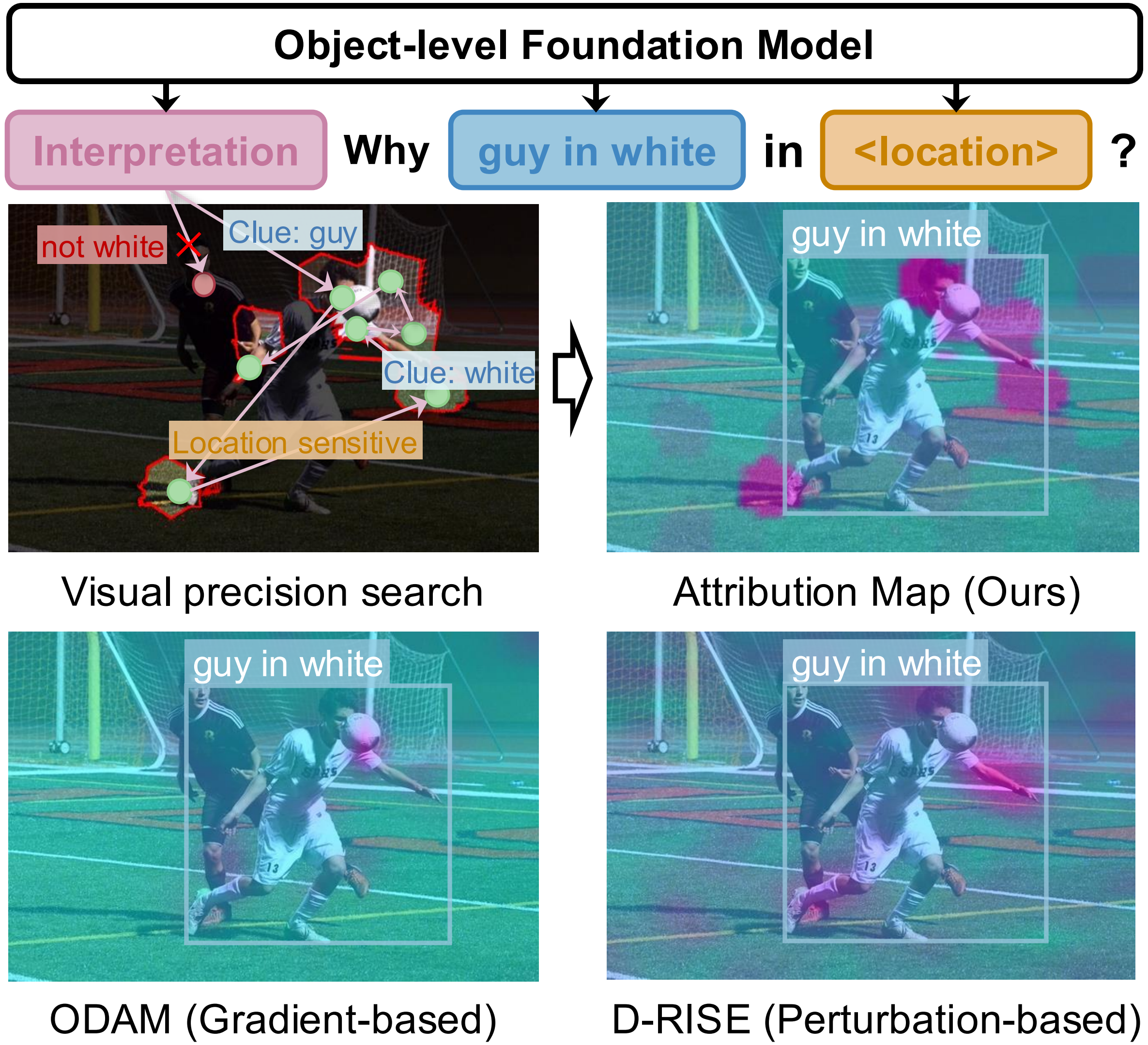
Key Problems: Early visual-text fusion in the object-level foundation model, which makes the gradient-based method unable to effectively attribute visual representations, and the perturbation-based method contains a lot of noise.



non-LLM architecture, *e.g.*, Grounding DINO



LLM architecture, *e.g.*, Florence-2



Our Motivation

Divide the image into a set of small sub-regions and ranking the sub-regions according to their importance.

Problem Formulation

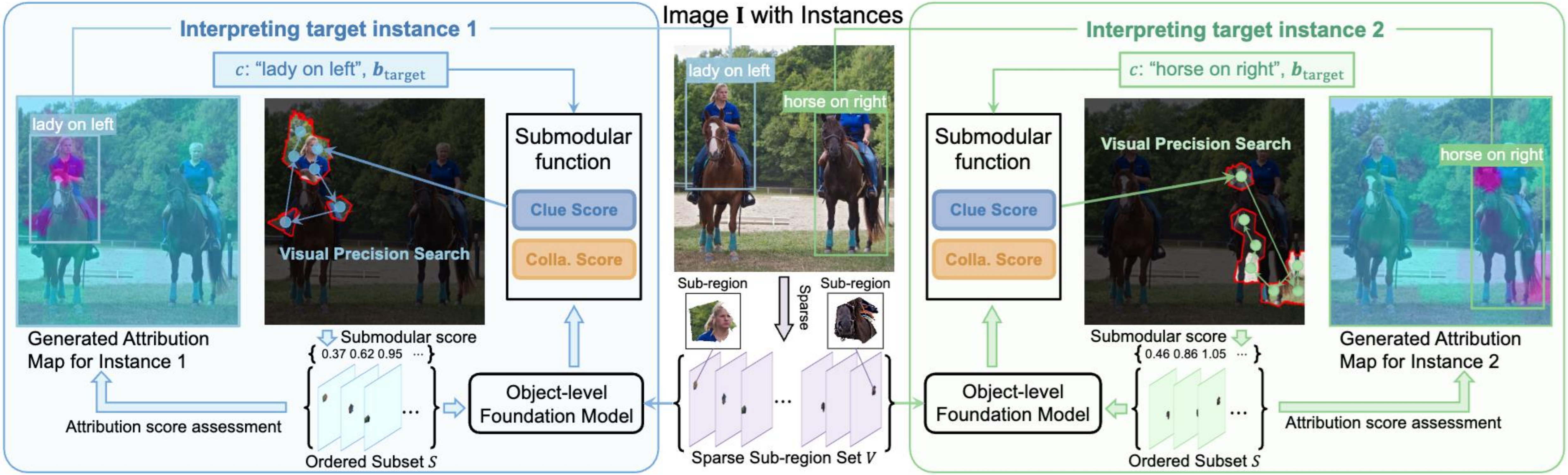
Given an image $\mathbf{I} \in \mathbb{R}^{h\times w\times 3}$ and an object-level foundation model $f\,(\cdot)$, the output can be represented as $f\,(\mathbf{I}) \ = \{(\boldsymbol{b}_i, c_i, s_i) \mid i \ = \ 1, 2, \ldots, N\}$.

Our goal is to generate a saliency map that explains the reasons behind the model's detection of a specific object.

To achieve this, we can sparsity the input region $V \ = \ \{\mathbf{I}_1^s, \ldots, \mathbf{I}_m^s\}$, where $\mathbf{I}_i^s$ represents the i-th sub-region. A set function $\mathcal{F}(\cdot)$ is defined to assess interpretability by determining whether a given region is a key factor in the model's decision. Then, the objectives are:

$$\max_{S\subseteq V, |S|<k} \mathcal{F}(S)$$

The Proposed VPS Method



Clue Score, accurately locate and identify objects while using fewer regions:

$$s_{\text{clue}}\left(S, \boldsymbol{b}_{\text{target}}, c\right) = \max_{(\boldsymbol{b}_i, c_i, s_i)\in f(S)} \text{IoU}\left(\boldsymbol{b}_{\text{target}}, \boldsymbol{b}_i\right)\cdot s_{c,i}$$

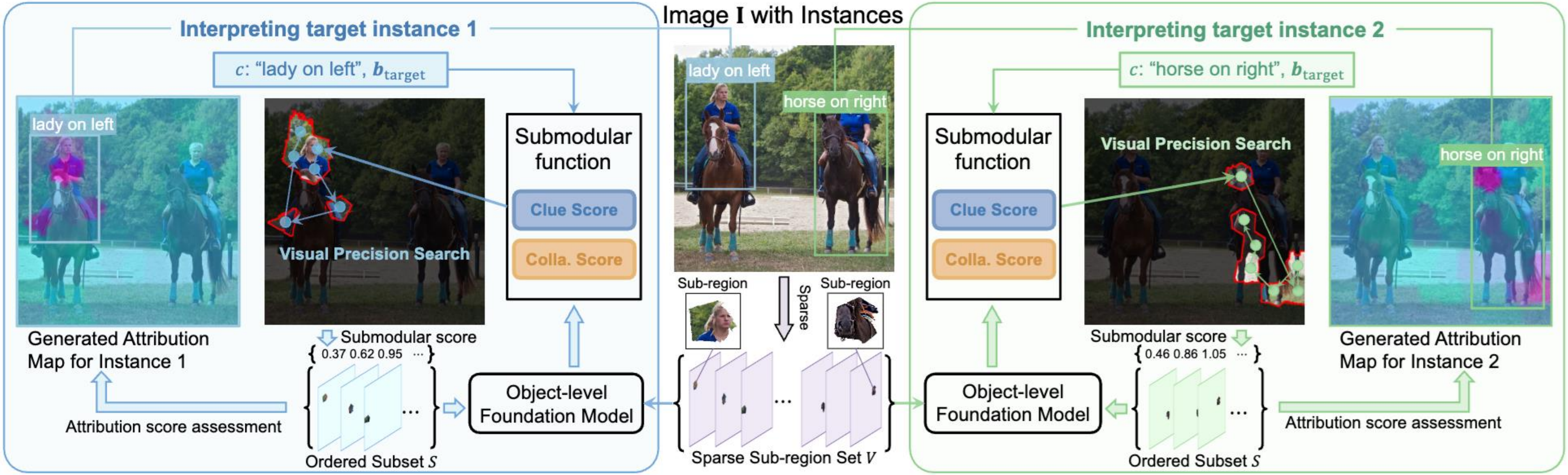
Collaboration Score, assess sub-regions with high sensitivity to decision outcomes:

$$s_{\text{colla.}}\left(S, \boldsymbol{b}_{\text{target}}, c\right) = 1 - \max_{(\boldsymbol{b}_i, c_i, s_i)\in f(V\setminus S)} \text{IoU}\left(\boldsymbol{b}_{\text{target}}, \boldsymbol{b}_i\right)\cdot s_{c,i}$$

Submodular Function:

$$\mathcal{F}\left(S, \boldsymbol{b}_{\text{target}}, c\right) = s_{\text{clue}}\left(S, \boldsymbol{b}_{\text{target}}, c\right) + s_{\text{colla.}}\left(S, \boldsymbol{b}_{\text{target}}, c\right)$$

The Proposed VPS Method



Scoring the sub-regions is necessary to better explain the importance of each sub-region, we evaluate the salient difference between the two sub-regions by the marginal effect. The attribution score:

$$\mathcal{A}_i = \begin{cases} b_{\text{base}} & \text{if } i = 1, \\ \mathcal{A}_{i-1} - \left|\mathcal{F}\left(S_{[i]}\right) - \mathcal{F}\left(S_{[i-1]}\right)\right| & \text{if } i > 1, \end{cases}$$

Experimental Results

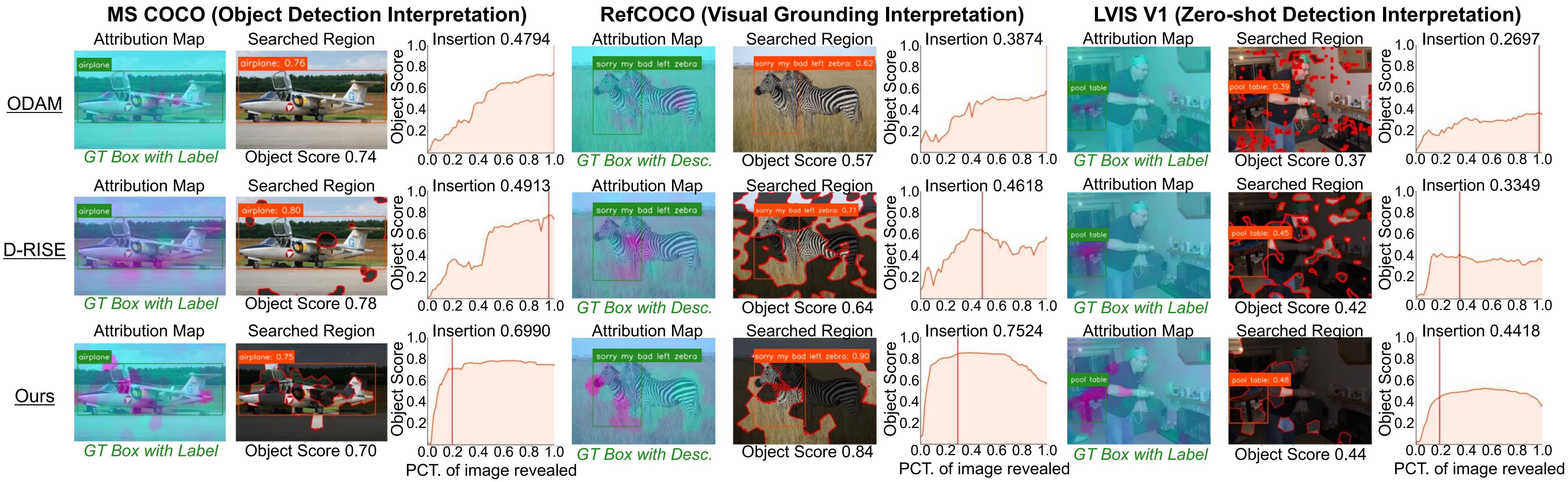
Faithfulness on Grounding DINO

Table 1. Evaluation of faithfulness metrics (Deletion, Insertion AUC scores, and average highest score) and location metrics (Point Game and Energy Point Game) on the MS-COCO, RefCOCO, and LVIS V1 (rare) validation sets for correctly detected or grounded samples using Grounding DINO.

Datasets	Methods	Faithfulness Metrics							Location Metrics	
		Ins. (↑)	Del. (↓)	Ins. (class) (↑)	Del. (class) (↓)	Ins. (IoU) (↑)	Del. (IoU) (↓)	Ave. high. score (↑)	Point Game (↑)	Energy PG (↑)
MS COCO [26] (Detection task)	Grad-CAM [40]	0.2436	0.1526	0.3064	0.2006	0.6229	0.5324	0.5904	0.1746	0.1463
	SSGrad-CAM++ [54]	0.2107	0.1778	0.2639	0.2314	0.5981	0.5511	0.5886	0.1905	0.1293
	D-RISE [35]	0.4412	0.0402	0.5081	0.0886	0.8396	0.3642	0.6215	0.9497	0.1850
	D-HSIC [33]	0.3776	0.0439	0.4382	0.0903	0.8301	0.3301	0.5862	0.7328	0.1861
	ODAM [59]	0.3103	0.0519	0.3655	0.0894	0.7869	0.3984	0.5865	0.5431	0.2034
	Ours	**0.5459**	**0.0375**	**0.6204**	**0.0882**	**0.8581**	**0.3300**	**0.6873**	**0.9894**	**0.2046**
RefCOCO [19] (REC task)	Grad-CAM [40]	0.3749	0.4237	0.4658	0.5194	0.7516	0.7685	0.7481	0.2380	0.2171
	SSGrad-CAM++ [54]	0.4113	0.3925	0.5008	0.4851	0.7700	0.7588	0.7561	0.2820	0.2262
	D-RISE [35]	0.6178	0.1605	0.7033	0.3396	0.8606	0.5164	0.8471	0.9400	0.2870
	D-HSIC [33]	0.5491	0.1846	0.6295	0.3509	0.8504	0.5120	0.7739	0.7900	0.3190
	ODAM [59]	0.4778	0.2718	0.5620	0.3757	0.8217	0.6641	0.7425	0.6320	0.3529
	Ours	**0.7419**	**0.1250**	**0.8080**	**0.2457**	**0.9050**	**0.5103**	**0.8842**	**0.9460**	**0.3566**
LVIS V1 (rare) [14] (Zero-shot det. task)	Grad-CAM [40]	0.1253	0.1294	0.1801	0.1814	0.5657	0.5910	0.3549	0.1151	0.0941
	SSGrad-CAM++ [54]	0.1253	0.1254	0.1765	0.1775	0.5800	0.5691	0.3504	0.1091	0.0931
	D-RISE [35]	0.2808	0.0289	0.3348	0.0835	0.8303	0.3174	0.4289	0.9697	0.1462
	D-HSIC [33]	0.2417	0.0353	0.2912	0.0928	0.8187	0.3550	0.4044	0.8303	0.1730
	ODAM [59]	0.2009	0.0410	0.2478	0.0844	0.7760	0.4082	0.3694	0.6061	**0.2050**
	Ours	**0.3695**	**0.0277**	**0.4275**	**0.0799**	**0.8479**	**0.3242**	**0.4969**	**0.9758**	0.1785

Experimental Results

Faithfulness on Grounding DINO



Experimental Results

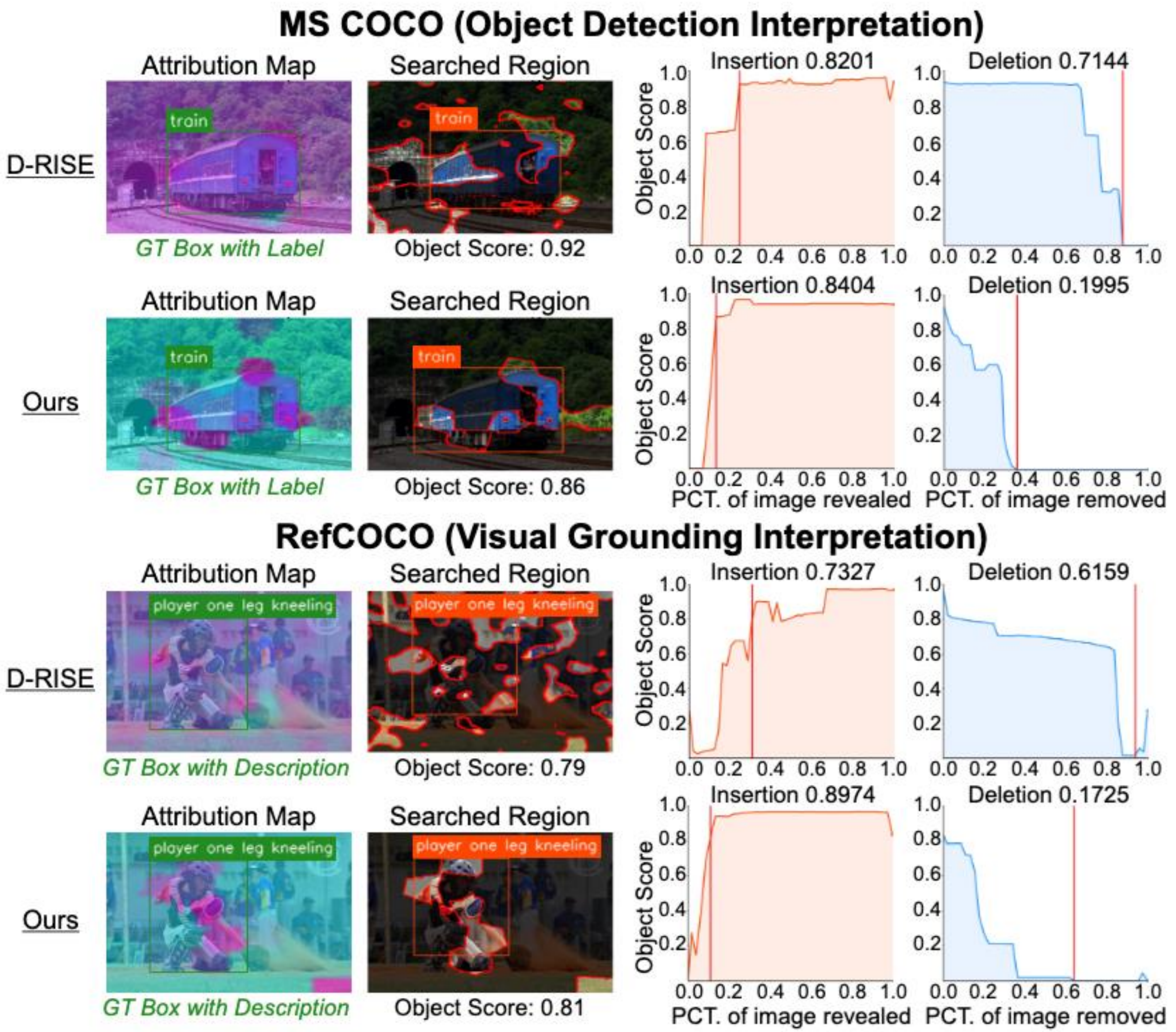
Faithfulness on Florence-2

Table 2. Evaluation of faithfulness metrics (Deletion and Insertion AUC scores) and location metrics (Point Game and Energy Point Game) on the MS COCO and RefCOCO validation sets for correctly detected and grounded samples using Florence-2.

Datasets	Methods	Faithfulness Metrics		Location Metrics	
		Insertion (↑)	Deletion (↓)	Point Game (↑)	Energy PG (↑)
MS COCO [26] (Detection task)	D-RISE [35]	0.7477	0.0972	0.8850	0.1568
	D-HSIC [33]	0.5345	0.2730	0.2925	0.0862
	Ours	**0.7759**	**0.0479**	**0.9583**	**0.2519**
RefCOCO [19] (REC task)	D-RISE [35]	0.7922	0.3505	0.8480	0.2464
	D-HSIC [33]	0.7639	0.3560	0.6980	0.2754
	Ours	**0.8409**	**0.1159**	**0.8660**	**0.3927**

Object detection interpretation: VPS outperforms D-RISE by 50.7% on the Deletion metric. Furthermore, VPS enhances the Point Game and Energy Point Game metrics by 8.3% and 60.7%, respectively.

Referring expression comprehension interpretation: VPS outperforming D-RISE by 66.9% in Deletion metrics, and also achieved SOTA localization results, with a 14.6% improvement in the Energy Point Game



Experimental Results

Interpreting REC Failures

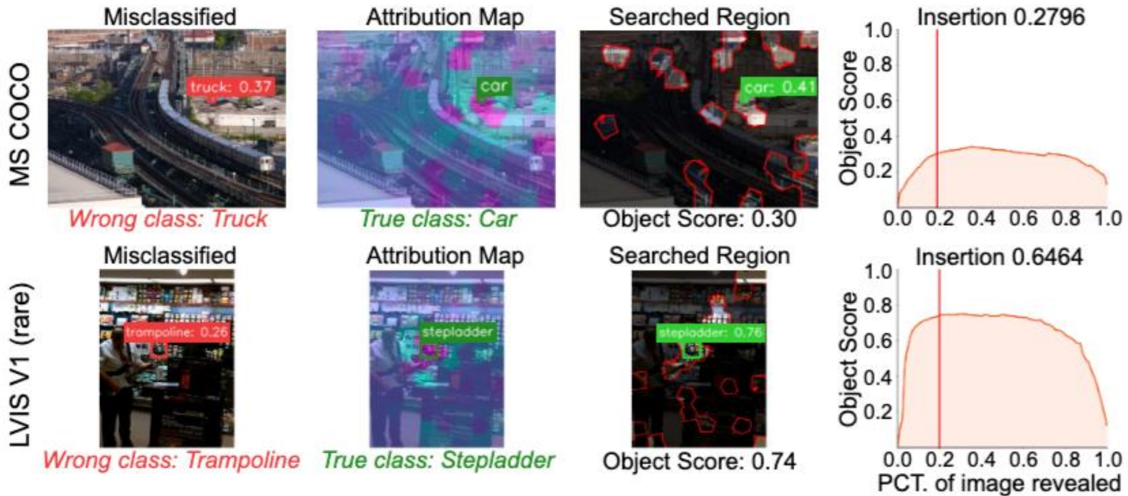
Table 3. Insertion AUC scores and the average highest score on the RefCOCO validation sets for or the samples with incorrect localization in visual grounding using Grounding DINO.

Datasets	Methods	Faithfulness Metrics		
		Ins. (↑)	Ins. (class) (↑)	Ave. high. score (↑)
RefCOCO [19] (REC task)	Grad-CAM [40]	0.1536	0.2794	0.3295
	SSGrad-CAM++ [54]	0.1590	0.2837	0.3266
	D-RISE [35]	0.3486	0.4787	0.6096
	D-HSIC [33]	0.2274	0.3488	0.4495
	ODAM [59]	0.1793	0.3001	0.3453
	Ours	**0.4981**	**0.5990**	**0.7007**

Interpreting Detection Failures (Misclassification)

Table 4. Insertion AUC scores, average highest score, and explaining successful rate (ESR) on the MS-COCO and the LVIS validation sets for misclassified samples using Grounding DINO.

Datasets	Methods	Faithfulness Metrics			
		Ins. (↑)	Ins. (class) (↑)	Ave. high. score (↑)	ESR (↑)
MS COCO [26] (Detection task)	Grad-CAM [40]	0.1091	0.1478	0.3102	38.38%
	SSGrad-CAM++ [54]	0.0960	0.1336	0.2952	33.51%
	D-RISE [35]	0.2170	0.2661	0.3603	50.26%
	D-HSIC [33]	0.1771	0.2161	0.3143	34.59%
	ODAM [59]	0.1129	0.1486	0.2869	32.97%
	Ours	**0.3357**	**0.3967**	**0.4591**	**69.73%**
LVIS V1 (rare) [14] (Zero-shot det. task)	Grad-CAM [40]	0.0503	0.0891	0.1564	12.50%
	SSGrad-CAM++ [54]	0.0574	0.0946	0.1580	11.84%
	D-RISE [35]	0.1245	0.1647	0.2088	28.95%
	D-HSIC [33]	0.0963	0.1247	0.1748	16.45%
	ODAM [59]	0.0575	0.0954	0.1520	9.21%
	Ours	**0.1776**	**0.2190**	**0.2606**	**53.29%**



Experimental Results

Interpreting Detection Failures (Undetected)

Table 5. Insertion, average highest score, and explaining successful rate (ESR) on the MS-COCO and the LVIS V1 (rare) validation sets for missed detection samples using Grounding DINO.

Datasets	Methods	Faithfulness Metrics			
		Ins. (↑)	Ins. (class) (↑)	Ave. high. score (↑)	ESR (↑)
MS COCO [26] (Detection task)	Grad-CAM [40]	0.0760	0.1321	0.2153	16.44%
	SSGrad-CAM++ [54]	0.0671	0.1151	0.2124	16.44%
	D-RISE [35]	0.1538	0.2260	0.2564	26.94%
	D-HSIC [33]	0.1101	0.1716	0.1945	13.56%
	ODAM [59]	0.0745	0.1350	0.2037	13.78%
	Ours	**0.2102**	**0.3011**	**0.3014**	**41.33%**
LVIS V1 (rare) [14] (Zero-shot det. task)	Grad-CAM [40]	0.0291	0.0689	0.0901	5.43%
	SSGrad-CAM++ [54]	0.0292	0.0680	0.0897	5.24%
	D-RISE [35]	0.0703	0.1184	0.1312	18.73%
	D-HSIC [33]	0.0516	0.0920	0.1168	13.48%
	ODAM [59]	0.0283	0.0716	0.0851	4.68%
	Ours	**0.1155**	**0.1886**	**0.1784**	**30.15%**

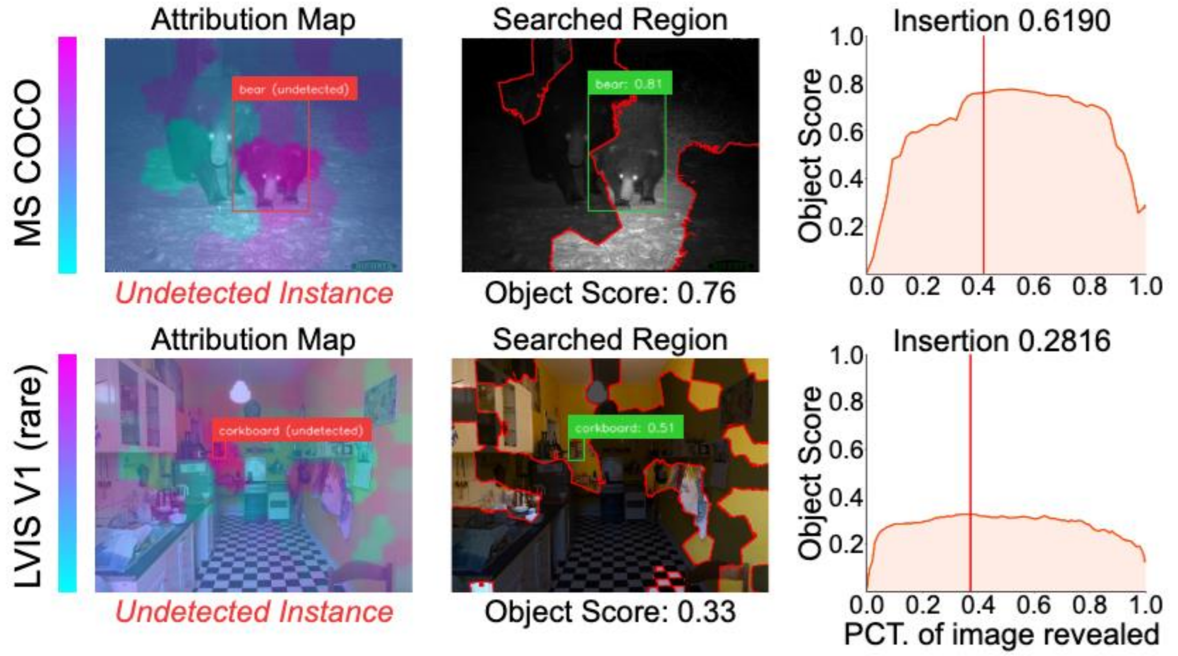


Figure 7. Visualization of our method reveals the causes of Grounding DINO undetected on MS COCO and LVIS. The cyan region in the saliency map highlights the regions responsible for the model's detection failure.

Ablation Study

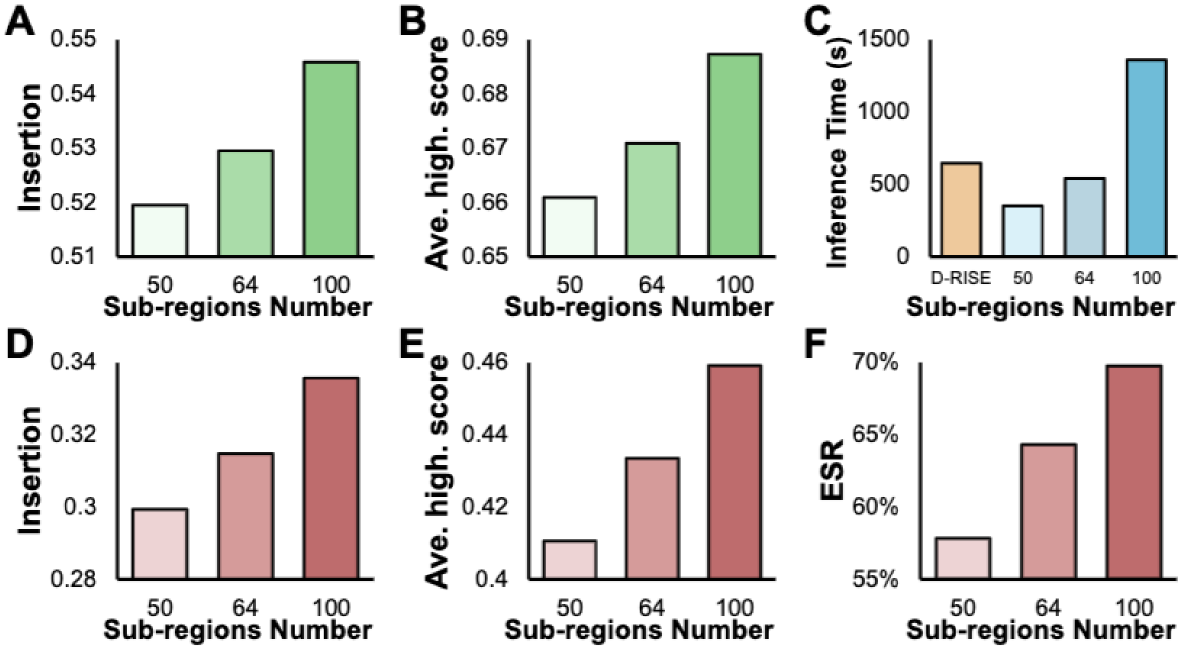
Ablation of the Submodular Function

Table 6. Ablation study on function score components for Grounding DINO on the MS COCO validation set.

Clue Score (Eq. 2)	Colla. Score (Eq. 3)	Faithfulness Metrics Insertion (↑)	Deletion (↓)	Ave. high. score (↑)
✗	✓	0.3632	0.0378	0.5967
✓	✗	0.5370	0.0799	0.6864
✓	✓	**0.5459**	**0.0375**	**0.6873**

Combining these scores enables our method to achieve optimal results across indicators, demonstrating the effectiveness of each score function within the submodular function.

Ablation on Divided Sub-region Number



Faithfulness: increasing the number of sub-regions leads to higher Insertion and average highest scores, indicating that finer divisions enhance the faithfulness of search results.

Computation time: Increasing sub-regions improves faithfulness but also rapidly increases inference time.

Thanks for Listening

Any Questions?