Supplementary Materials
Interpreting Object-level Foundation Models via Visual Precision Search

Ruoyu Chen[1,2], Siyuan Liang[3], Jingzhi Li[1,2,7], Shiming Liu[4], Maosen Li[5], Zhen Huang[6], Hua Zhang[1,2,*], and Xiaochun Cao[8,*]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China
[3]School of Computing, NUS [4]RAMS Lab, Huawei Technologies Co., Ltd. [5]IAS BU, Huawei Technologies Co., Ltd.
[6]College of Computer, NUDT [7]Key Lab. of Edu. Inf. for Nationalities (YNNU), Ministry of Education, Kunming, China
[8]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107, China

chenruoyu@iie.ac.cn pandaliang521@gmail.com {lijingzhi,zhanghua}@iie.ac.cn
{liushiming3,limaosen2}@huawei.com huangzhen@nudt.edu.cn caoxiaochun@mail.sysu.edu.cn

A. Proof of Theorem 1 (Submodular Properties)

Proof. Consider two sub-sets S_A and S_B in set V, where $S_A \subseteq S_B \subseteq V$. Given an element α, where $\alpha = V \setminus S_B$. Let $\alpha = V \setminus S_B$ represent an element not in S_B. For the function $\mathcal{F}(\cdot)$ to satisfy the submodular property, the following necessary and sufficient conditions must hold diminishing returns,

$$\mathcal{F}\left(S_A \cup \{\alpha\}\right) - \mathcal{F}\left(S_A\right) \geq \mathcal{F}\left(S_B \cup \{\alpha\}\right) - \mathcal{F}\left(S_B\right), \tag{S1}$$

and monotonic non-negative, $\mathcal{F}(S_A \cup \{\alpha\}) - \mathcal{F}(S) \geq 0$.

For Clue score (Eq. 2), let $f_{\text{cls}}(S) = \boldsymbol{s}_c$, $f_{\text{reg}}(S) = \boldsymbol{b}$, assuming that f_{cls} and $f_{\text{reg}}(S)$ is differentiable in S, the individual element α of the collection division is relatively small, according to the Taylor decomposition [6], we can locally approximate $f_{\text{cls}}\left(S_A + \alpha\right) = f_{\text{cls}}\left(S_A\right) + \nabla f_{\text{cls}}\left(S_A\right) \cdot \alpha$, and $f_{\text{reg}}\left(S_A + \alpha\right) = f_{\text{reg}}\left(S_A\right) + \nabla f_{\text{reg}}\left(S_A\right) \cdot \alpha$. Since object detection can output many candidate boxes, we can regard the model output with added sub-elements as incremental output, that is, $f_{\text{cls}}\left(S_A + \alpha\right) = \boldsymbol{s}_c + \nabla f_{\text{cls}}\left(S_A\right) \cdot \alpha = \boldsymbol{s}_c + \boldsymbol{s}_c^*$, $f_{\text{reg}}\left(S_A + \alpha\right) = \boldsymbol{b} + \nabla f_{\text{reg}}\left(S_A\right) \cdot \alpha = \boldsymbol{b} + \boldsymbol{b}^*$. Assuming that the searched α is valid, i.e., $\nabla f_{\text{reg}} > 0$ and $\nabla f_{\text{cls}} > 0$. Thus:

$$\begin{aligned}
&s_{\text{clue}}(S_A + \alpha, \boldsymbol{b}_{\text{target},c}) - s_{\text{clue}}(S_A, \boldsymbol{b}_{\text{target}}, c) \\
=&\max_{\boldsymbol{b}_i \in f_{\text{reg}}(S_A+\alpha), s_{c,i} \in f_{\text{cls}}(S_A+\alpha)} \text{IoU}(\boldsymbol{b}_{\text{target}}, \boldsymbol{b}_i) \cdot s_{c,i} \\
&- \max_{\boldsymbol{b}_i \in f_{\text{reg}}(S_A), s_{c,i} \in f_{\text{cls}}(S_A)} \text{IoU}(\boldsymbol{b}_{\text{target}}, \boldsymbol{b}_i) \cdot s_{c,i} \\
=&\max(s_{\text{clue}}(S_A, \boldsymbol{b}_{\text{target}}, c), \max_{\boldsymbol{b}_i \in \boldsymbol{b}^*, s_{c,i} \in \boldsymbol{s}_c^*} \text{IoU}(\boldsymbol{b}_{\text{target}}, \boldsymbol{b}_i) \cdot s_{c,i}) \\
&- s_{\text{clue}}(S_A, \boldsymbol{b}_{\text{target}}, c) \\
=&\max(0, \max_{\boldsymbol{b}_i \in \boldsymbol{b}^*, s_{c,i} \in \boldsymbol{s}_c^*} \nabla f_{\text{reg}}\left(S_A\right) \cdot \nabla f_{\text{cls}}\left(S_A\right) \cdot \alpha^2) \\
\geq&0,
\end{aligned} \tag{S2}$$

the clue score satisfies the monotonic non-negative property in the process of maximizing the marginal effect. Since $S_A \subseteq S_B \subseteq V$, for the model's candidate boxes, $f_{\text{reg}}(S_B) > f_{\text{reg}}(S_A)$, then the range of the gain candidate box $\boldsymbol{b}_A^*$ that can be generated is $f_{\text{reg}}(V) - f_{\text{reg}}(S_A)$. After introducing the new element α, a new candidate box with a gain $\boldsymbol{b}_A^* > \boldsymbol{b}_B^*$, closer to the target, can be generated. If both S_A and S_B contain positive subsets, then $\nabla f_{\text{cls}}\left(S_B\right)$ will become less severe or even disappear [12], thus, $\nabla f_{\text{cls}}\left(S_A\right) > \nabla f_{\text{cls}}\left(S_B\right)$. So we have:

$$\begin{aligned}
&\max_{\boldsymbol{b}_i \in \boldsymbol{b}_A^*, s_{c,i} \in \boldsymbol{s}_c^*} \nabla f_{\text{reg}}\left(S_A\right) \cdot \nabla f_{\text{cls}}\left(S_A\right) \cdot \alpha^2 > \\
&\max_{\boldsymbol{b}_i \in \boldsymbol{b}_B^*, s_{c,i} \in \boldsymbol{s}_c^*} \nabla f_{\text{reg}}\left(S_B\right) \cdot \nabla f_{\text{cls}}\left(S_B\right) \cdot \alpha^2,
\end{aligned} \tag{S3}$$

combining Eq. S2, we have:

$$\begin{aligned}
&s_{\text{clue}}(S_A + \alpha, \boldsymbol{b}_{\text{target},c}) - s_{\text{clue}}(S_A, \boldsymbol{b}_{\text{target}}, c) > \\
&s_{\text{clue}}(S_B + \alpha, \boldsymbol{b}_{\text{target},c}) - s_{\text{clue}}(S_B, \boldsymbol{b}_{\text{target}}, c).
\end{aligned} \tag{S4}$$

*Corresponding authors.

Table S1. 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 validation set for correctly detected or grounded samples using traditional object detectors.

Detectors	Methods	Faithfulness Metrics						Location Metric
		Ins. (↑)	Del. (↓)	Ins. (class) (↑)	Del. (class) (↓)	Ins. (IoU) (↑)	Del. (IoU) (↓)	Point Game (↑)
Mask R-CNN [4] (Two-stage)	Grad-CAM [9]	0.2657	0.2114	0.3746	0.3122	0.5348	0.4954	0.5554
	D-RISE [7]	0.6756	0.0814	0.7666	0.1570	0.8396	0.2987	0.8996
	ODAM [14]	0.6067	0.0787	0.7218	0.1860	0.7890	0.3188	0.9934
	Ours	**0.7991**	**0.0489**	**0.8678**	**0.1065**	**0.8968**	**0.2841**	**0.9987**
YOLO V3 [8] (One-stage)	Grad-CAM [9]	0.6283	0.2867	0.7961	0.4573	0.7271	0.5234	0.7268
	D-RISE [7]	0.7524	0.1889	0.8747	0.3629	0.8213	0.4587	0.8816
	ODAM [14]	0.7329	0.2766	0.8943	0.4707	0.7936	0.5283	0.9838
	Ours	**0.8674**	**0.1407**	**0.9490**	**0.3008**	**0.8984**	**0.3814**	**0.9900**
FCOS [13] (One-stage)	Grad-CAM [9]	0.2742	0.1417	0.3439	0.1845	0.6858	0.6176	0.5249
	D-RISE [7]	0.4421	0.0570	0.4968	0.1078	0.8578	0.3729	0.9193
	ODAM [14]	0.4266	0.0497	0.4742	0.0853	0.8713	0.4212	0.9935
	Ours	**0.5746**	**0.0414**	**0.6301**	**0.0815**	**0.8900**	**0.3698**	**0.9980**
SSD [5] (One-stage)	Grad-CAM [9]	0.3869	0.1466	0.4977	0.2022	0.6796	0.5366	0.7700
	D-RISE [7]	0.4882	0.0616	0.5722	0.0979	0.7852	0.4497	0.9243
	ODAM [14]	0.5117	0.0913	0.6072	0.1416	0.7900	0.4801	0.9778
	Ours	**0.6891**	**0.0483**	**0.7594**	**0.0835**	**0.8700**	**0.4366**	**0.9941**

Similar, for Collaboration score (Eq. 3), according to the Taylor decomposition [6], we can locally approximate $f_{\text{cls}}\left(V \setminus (S_A + \alpha)\right) = f_{\text{cls}}\left(V \setminus S_A\right) - \nabla f_{\text{cls}}\left(V \setminus S_A\right) \cdot \alpha$ and $f_{\text{reg}}\left(V \setminus (S_A + \alpha)\right) = f_{\text{reg}}\left(V \setminus S_A\right) - \nabla f_{\text{reg}}\left(V \setminus S_A\right) \cdot \alpha$. The model output can be viewed as a negative gain process, i.e., $f_{\text{reg}}\left(V \setminus (S_A + \alpha)\right) = \boldsymbol{b} - \boldsymbol{b}^*$. Assuming that the searched α is valid, i.e., $\nabla f_{\text{reg}} > 0$ and $\nabla f_{\text{cls}} > 0$. We have:

$$
\begin{aligned}
& s_{\text{colla.}}(S_A + \alpha, \boldsymbol{b}_{\text{target},c}) - s_{\text{colla.}}(S_A, \boldsymbol{b}_{\text{target}}, c) \\
= & \max_{\boldsymbol{b}_i \in f_{\text{reg}}(V \setminus S_A), s_{c,i} \in f_{\text{cls}}(V \setminus S_A)} \text{IoU}(\boldsymbol{b}_{\text{target}}, \boldsymbol{b}_i) \cdot s_{c,i} \\
& - \max_{\boldsymbol{b}_i \in f_{\text{reg}}(V \setminus (S_A + \alpha)), s_{c,i} \in f_{\text{cls}}(V \setminus (S_A + \alpha))} \text{IoU}(\boldsymbol{b}_{\text{target}}, \boldsymbol{b}_i) \cdot s_{c,i} \\
= & 1 - s_{\text{colla.}}(S_A, \boldsymbol{b}_{\text{target}}, c) \\
& - \min(1 - s_{\text{colla.}}(S_A, \boldsymbol{b}_{\text{target}}, c), \max_{\boldsymbol{b}_i \in \boldsymbol{b}^*, s_{c,i} \in \boldsymbol{s}_c^*} \text{IoU}(\boldsymbol{b}_{\text{target}}, \boldsymbol{b}_i) \cdot s_{c,i}) \\
= & \max(0, \max_{\boldsymbol{b}_i \in \boldsymbol{b}^*, s_{c,i} \in \boldsymbol{s}_c^*} \nabla f_{\text{reg}}\left(V \setminus S_A\right) \cdot \nabla f_{\text{cls}}\left(V \setminus S_A\right) \cdot \alpha^2) \\
\geq & 0,
\end{aligned}
\tag{S5}
$$

the collaboration score satisfies the monotonic non-negative property in the process of maximizing the marginal effect. Since $S_A \subseteq S_B \subseteq V$, more candidate boxes will be deleted, thus, $\boldsymbol{b}_A^* > \boldsymbol{b}_B^*$, and $\nabla f_{\text{reg}}\left(V \setminus S_A\right) > \nabla f_{\text{reg}}\left(V \setminus S_B\right)$. Since only a small number of candidate boxes $\boldsymbol{b}^*$ are removed, $\nabla f_{\text{cls}}(S_A) \cdot \alpha$ can be regarded as a tiny constant and can be ignored. Combining Eq. S5, we have:

$$
\begin{aligned}
& s_{\text{colla.}}(S_A + \alpha, \boldsymbol{b}_{\text{target},c}) - s_{\text{colla.}}(S_A, \boldsymbol{b}_{\text{target}}, c) > \\
& s_{\text{colla.}}(S_B + \alpha, \boldsymbol{b}_{\text{target},c}) - s_{\text{colla.}}(S_B, \boldsymbol{b}_{\text{target}}, c).
\end{aligned}
\tag{S6}
$$

We can prove that both the Clue Score and Collaboration Score satisfy submodularity under certain conditions. Since any linear combination of submodular functions is itself submodular [2], we have:

$$
\mathcal{F}\left(S_A \cup \{\alpha\}\right) - \mathcal{F}\left(S_A\right) \geq \mathcal{F}\left(S_B \cup \{\alpha\}\right) - \mathcal{F}\left(S_B\right), \tag{S7}
$$

and we can prove that Eq. 4 satisfies the submodular properties. □

From the above derivation, we find that the gain or negative gain condition of bounding boxes $\boldsymbol{b}^*$ is critical for satisfying submodularity. Thus, the detection model f, which can return relevant confidence candidate boxes for any combination of input sub-regions, is theoretically guaranteed to meet the required boundaries. Most detection models fulfill this condition by not filtering out low-confidence candidate boxes. However, multimodal large language model-based detection models, which may not directly output candidate boxes or confidence scores, do not fully satisfy these assumptions. This highlights room for improvement in explaining the results of such models.

B. Faithfulness in Traditional Detectors

We also validated the effectiveness of our interpretation method on traditional object detectors, including the two-stage detector Mask R-CNN (ResNet-50 backbone with Feature Pyramid Networks) [4] and the one-stage detectors YOLO v3 (DarkNet-53 backbone) [8], FCOS (ResNet-50 backbone with Feature Pyramid Networks) [13], and SSD [5]. We use the pre-trained models provided by

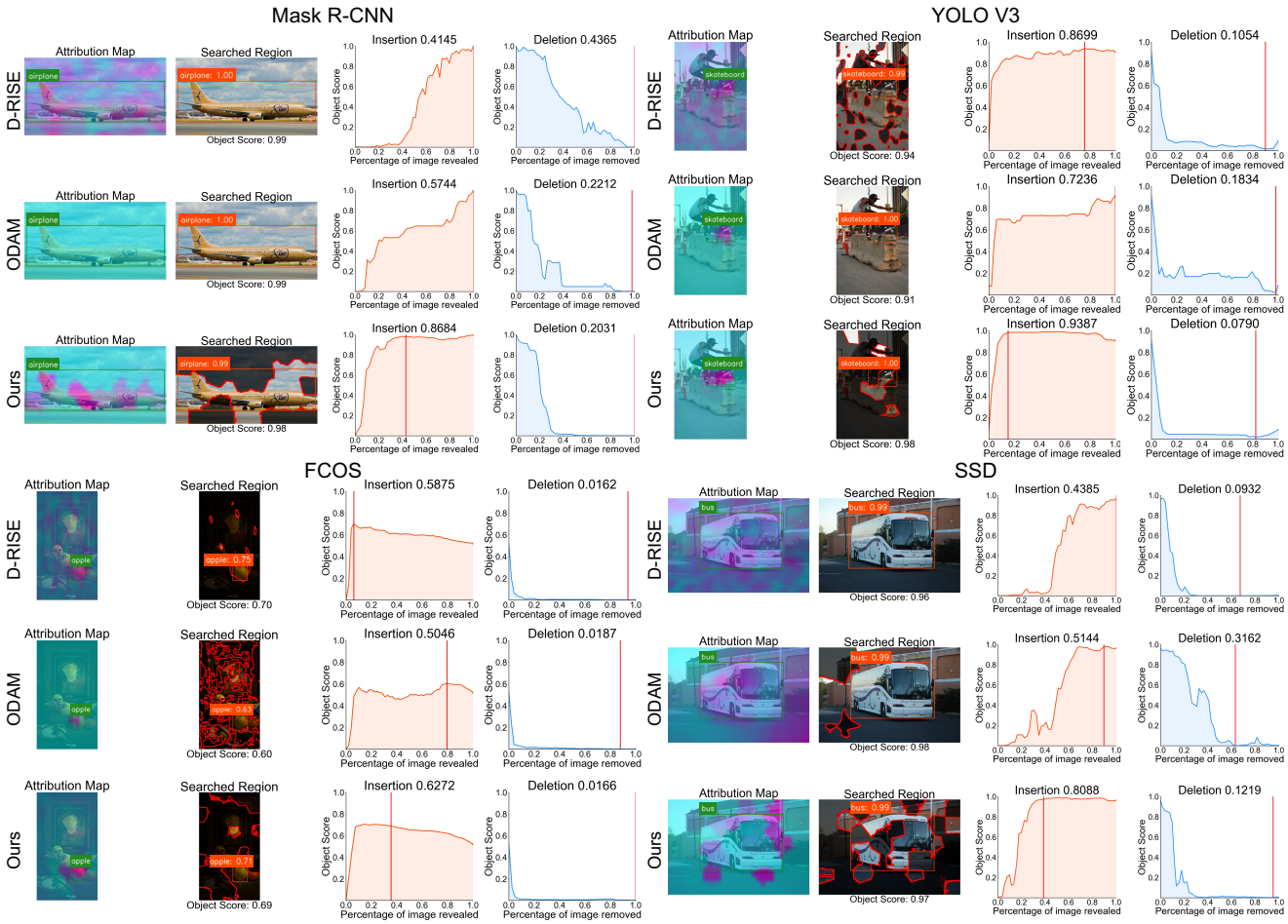

Figure S1. Visualization results of four object detectors for interpreting object detection task on the MS COCO dataset.

MMDetection 3.3.0[1] for interpretation. Following the evaluation settings of D-RISE [7] and ODAM [14], we selected samples for interpretation that were correctly predicted by the model with high confidence and precise localization.

Table S1 presents the results. We observe that, when considering both location and classification faithfulness metrics, D-RISE emphasizes location information more than ODAM, resulting in better performance in this aspect. In contrast, ODAM demonstrates higher faithfulness to classification scores on certain detectors, such as SSD and YOLO v3. Notably, ODAM outperforms D-RISE in the location metric Point Game. While ODAM and D-RISE have specific advantages across different metrics and models, our method consistently achieves state-of-the-art results across all models and metrics. On Mask R-CNN, our method outperforms D-RISE by 18.3% in Insertion and 31.7% in ODAM, as well as by 39.9% and 37.9% in Deletion. On YOLO V3, our method outperforms D-RISE by 15.3% in Insertion and 18.4% in ODAM and by 25.5% and 49.1% in Deletion. On FCOS, our method surpasses D-RISE by 30.0% in Insertion and 34.7% in ODAM, and by 27.4% and 16.7% in Deletion. On SSD, our method outperforms D-RISE by 41.2% in Insertion and 17.7% in ODAM, and by 21.6% and 47.1% in Deletion.

From the above results, we found that our method remains highly interpretable even on traditional object detection models, demonstrating its versatility in explaining both modern multimodal foundation models and traditional smaller detectors. Figure S1 presents the visualization results, demonstrating that our method maintains high faithfulness in explaining traditional detectors.

C. Semantic Interpretations

We apply our interpretation method to the visual grounding task using Grounding DINO, focusing on explaining the same location corresponding to different text expressions. As shown in Figure S2, although the important regions are similar when grounding the same object with different texts, the saliency map reveals distinct differences. For the inter-

[1] https://github.com/open-mmlab/mmdetection

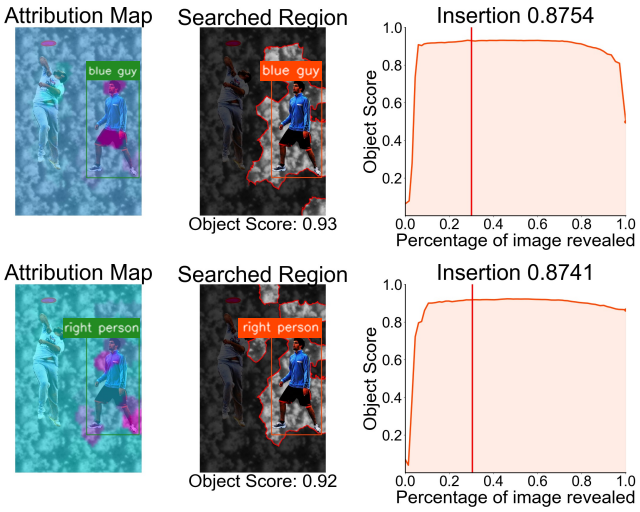

Figure S2. Interpretation of the same instance with different text expressions.

pretation of the text 'blue guy', the person wearing white clothes contributes less than the background region. In contrast, the interpretation of the 'right person' highlights only the correct object.

D. Computational complexity

Solving Eq. 1 is an $\mathcal{NP}$-hard problem, and the time complexity is $\mathcal{O}(2^{|V|})$. By employing a greedy search algorithm, we sort all subregions, resulting in a total number of inferences equal to $\frac{1}{2}|V|^2 + \frac{1}{2}|V|$, the algorithm's time complexity is $\mathcal{O}(\frac{1}{2}|V|^2 + \frac{1}{2}|V|)$.

E. Statistical Analysis

We utilize Grounding DINO's correct prediction attribution map and compute the numerical improvements of our method over the baseline for each sample. We then visualize the overall distribution of these improvements. As shown in Figure S3, the distributions on the MS COCO, RefCOCO, and LVIS V1 datasets illustrate that our method outperforms the baseline on the majority of samples, demonstrating its superior performance.

F. Evaluation Metrics

In this paper, we adopt 6 faithfulness metrics. Given the object location box information, $\boldsymbol{b}_{\text{target}}$, and the target category, c, that requires explanation. In this section we will formulate a description of faithfulness metrics.

For the **Deletion AUC score** [7], which quantifies the reduction in the model's ability of both location and classification when important regions are replaced with a baseline value. A sharp decline in performance indicates that

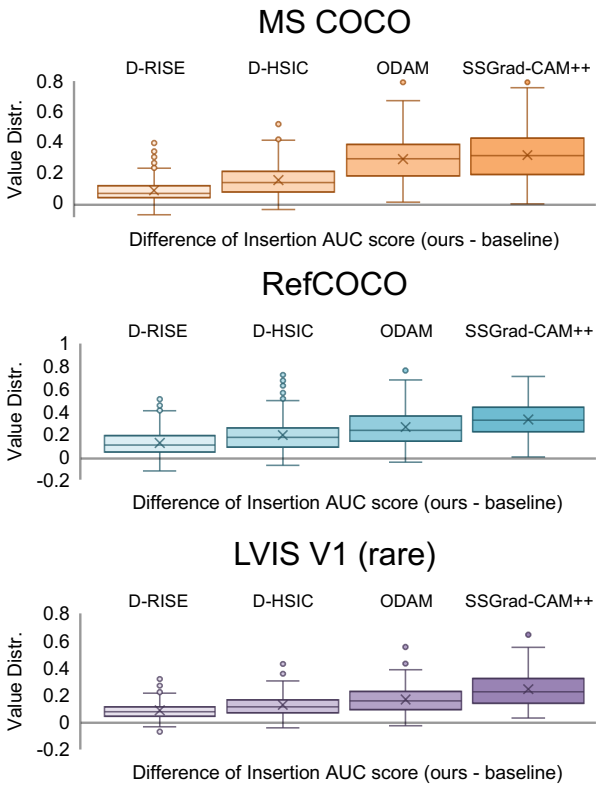

Figure S3. Distribution of our improvements over the baseline per sample on MS COCO, RefCOCO, and LVIS V1 datasets.

the explanation method effectively identifies the key variables influencing the decision. Let $\mathbf{x}_{[x_T=x_0]}$ denote the input where the T most important variables, according to the attribution map, are set to the baseline value $x_0 = 0$. Given a set $\mathcal{T} = \{T_0, T_1, \cdots, T_n\}$, where $T_0 = 0$ and T_n is the input size of $\mathbf{x}$, this set represents the selected numbers of the most important regions. Then, the Deletion AUC score is given by:

$$\text{Del.} = \sum_{i=1}^{n} \frac{\left(s_{\text{clue}}(\mathbf{x}_{[\mathbf{x}_{T_i}=x_0]}) + s_{\text{clue}}(\mathbf{x}_{[\mathbf{x}_{T_{i-1}}=x_0]})\right) \cdot (T_i - T_{i-1})}{2T_n}, \tag{S8}$$

the lower this metric, the better the attribution performance.

For the **Insertion AUC score** [7], which quantifies the increase in the model's output as important regions are progressively revealed. This metric is defined as follows:

$$\text{Ins.} = \sum_{i=1}^{n} \frac{\left(s_{\text{clue}}(\mathbf{x}_{[\mathbf{x}_{\bar{T}_i}=x_0]}) + s_{\text{clue}}(\mathbf{x}_{[\mathbf{x}_{\bar{T}_{i-1}}=x_0]})\right) \cdot (T_i - T_{i-1})}{2T_n}, \tag{S9}$$

where $\mathbf{x}_{[x_{\bar{T}}=x_0]}$ denotes the input where elements not belonging to the set T are set to the baseline value $x_0 = 0$. The

higher this metric, the better the attribution performance.

For the **Deletion AUC score (class)**, we first define s_{cls}, whose goal is to select the category score of the bounding box that is close to the explanation target:

$$s_{cls}(S) = \arg \max_{s_{c,i} \in f(S)} \text{IoU}(\boldsymbol{b}_{\text{target}}, \boldsymbol{b}_i) \cdot s_{c,i}, \tag{S10}$$

then,

$$\text{Del. (class)} = \sum_{i=1}^{n} \frac{\left(s_{cls}(\mathbf{x}_{[\mathbf{x}_{T_i}=x_0]}) + s_{cls}(\mathbf{x}_{[\mathbf{x}_{T_{i-1}}=x_0]})\right) \cdot (T_i - T_{i-1})}{2T_n}. \tag{S11}$$

Similar, for the **Insertion AUC score (class)**,

$$\text{Ins. (class)} = \sum_{i=1}^{n} \frac{\left(s_{cls}(\mathbf{x}_{[\mathbf{x}_{\bar{T}_i}=x_0]}) + s_{cls}(\mathbf{x}_{[\mathbf{x}_{\bar{T}_{i-1}}=x_0]})\right) \cdot (T_i - T_{i-1})}{2T_n}. \tag{S12}$$

For the **Deletion AUC score (IoU)**, we first define s_{iou}, whose goal is to select the IoU score of the bounding box that is close to the explanation target:

$$s_{iou}(S) = \text{IoU}\left(\arg \max_{\boldsymbol{b}_i \in f(S)} \text{IoU}(\boldsymbol{b}_{\text{target}}, \boldsymbol{b}_i) \cdot s_{c,i}, \boldsymbol{b}_{\text{target}}\right), \tag{S13}$$

then,

$$\text{Del. (IoU)} = \sum_{i=1}^{n} \frac{\left(s_{iou}(\mathbf{x}_{[\mathbf{x}_{T_i}=x_0]}) + s_{iou}(\mathbf{x}_{[\mathbf{x}_{T_{i-1}}=x_0]})\right) \cdot (T_i - T_{i-1})}{2T_n}. \tag{S14}$$

Similar, for the **Insertion AUC score (IoU)**,

$$\text{Ins. (IoU)} = \sum_{i=1}^{n} \frac{\left(s_{iou}(\mathbf{x}_{[\mathbf{x}_{\bar{T}_i}=x_0]}) + s_{iou}(\mathbf{x}_{[\mathbf{x}_{\bar{T}_{i-1}}=x_0]})\right) \cdot (T_i - T_{i-1})}{2T_n}. \tag{S15}$$

G. Limitation and Discussion

Limitations: The main limitation of our method is that (i) ***Sparse division*** impacts attribution faithfulness. When sub-regions mix positively and negatively contributing regions, attribution direction may be distorted. Refining sparse division strategies for different scenarios remains a region for improvement. (ii) A large number of sub-regions poses a challenge for ***attribution time***, as greedy search remains computationally demanding. Enhancing search efficiency or integrating external knowledge to filter unnecessary sub-regions can help accelerate attribution.

Why object score decrease sometimes: This phenomenon occurs because not all sub-regions contribute positively to the model's decision, underscoring a key advantage of our method: maximizing the decision response with fewest sub-regions. Exposing additional sub-regions may lower the object score, revealing that the remaining regions negatively impact the decision. This effect is more pronounced in incorrect decisions, where certain regions may cause errors. If such regions are excluded, the decision could potentially be corrected.

Future outlook: Our method primarily focuses on interpretable attribution at the input level of the model. There remains significant potential for attributing internal parameters, particularly in transformer-based models. This approach could extend to explaining additional tasks, such as instance segmentation. Future research could explore improving models based on this mechanism by identifying and correcting problematic parameters.

H. Actual Application

Attribution has numerous potential applications. Beyond aiding human understanding of model decisions, it can also help identify the causes of errors, enabling the analysis of potential hallucinations [1]. Some studies explore using attribution to guide model training and enhance performance [3], while others investigate detecting anomaly decisions by assessing whether the attribution distribution deviates from expected patterns during deployment [10, 11]. These diverse applications highlight the significant research value of attribution methods.

I. Additional Ablation

Ablation on thresholding the confidence scores: We discuss the impact of applying a confidence score threshold versus using all model predictions regardless of their confidence scores. As shown in Table S2, applying a threshold to filter boxes leads to a consistent decline in all faithfulness metrics as the threshold increases. Therefore, we recommend avoiding the use of thresholds.

Ablation on using confidence score: We discuss the impact of whether or not to use the Confidence score on the interpretation. In Table S3, without conf. score, both the attribution faithfulness for classes and the location will decrease, leading to imprecise attribution.

J. More Visualization

We present additional attribution visualizations for samples correctly predicted by Grounding DINO, including results from the MS COCO dataset to explain the object detection task (Figure S4) and from the RefCOCO dataset to illustrate the visual grounding task (Figure S5).

Table S2. Ablation on the confidence score threshold for Grounding DINO using the MS COCO validation set.

Threshold	Faithfulness Metrics						
	Ins. (↑)	Del. (↓)	Ins. (class) (↑)	Del. (class) (↓)	Ins. (IoU) (↑)	Del. (IoU) (↓)	Ave. high. score (↑)
None	**0.5459**	**0.0375**	**0.6204**	**0.0882**	**0.8581**	**0.3300**	**0.6873**
0.1	0.5267	0.0396	0.6007	0.0896	0.8498	0.3321	0.6660
0.2	0.5165	0.0423	0.5928	0.0916	0.8373	0.3408	0.6625
0.35	0.4862	0.0641	0.5638	0.1098	0.8023	0.3825	0.6519

Table S3. Ablation on the confidence score for Grounding DINO using the MS COCO validation set.

Submodular function	Faithfulness Metrics						
	Ins. (↑)	Del. (↓)	Ins. (class) (↑)	Del. (class)(↓)	Ins. (IoU) (↑)	Del. (IoU)(↓)	Ave. high. score (↑)
w/ conf. score	**0.5459**	**0.0375**	**0.6204**	**0.0882**	**0.8581**	**0.3300**	**0.6873**
w/o conf. score	0.3725	0.0917	0.4410	0.1622	0.8051	0.3421	0.5928

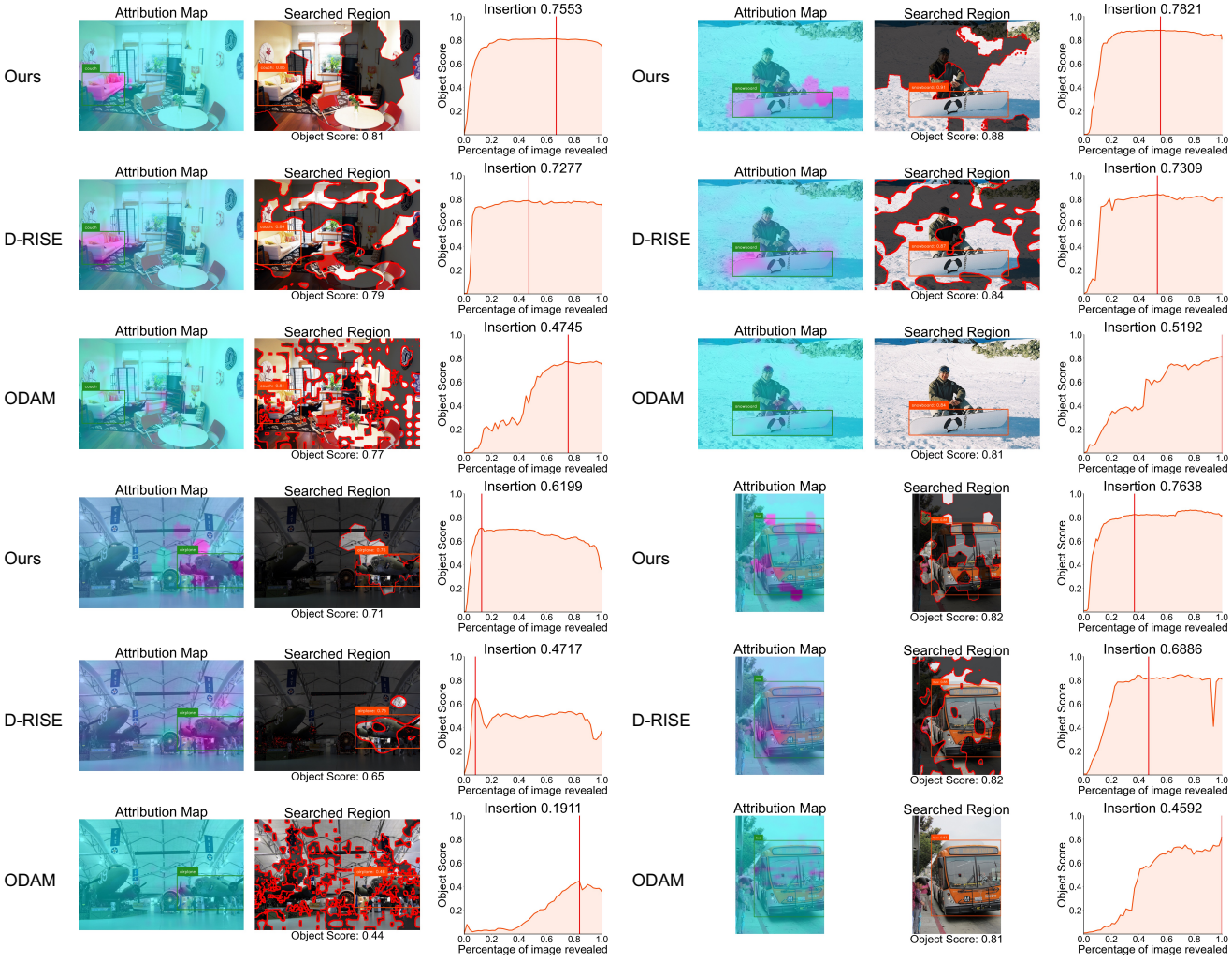

Figure S4. More visualization results of Grounding DINO for interpreting object detection task on the MS COCO dataset.

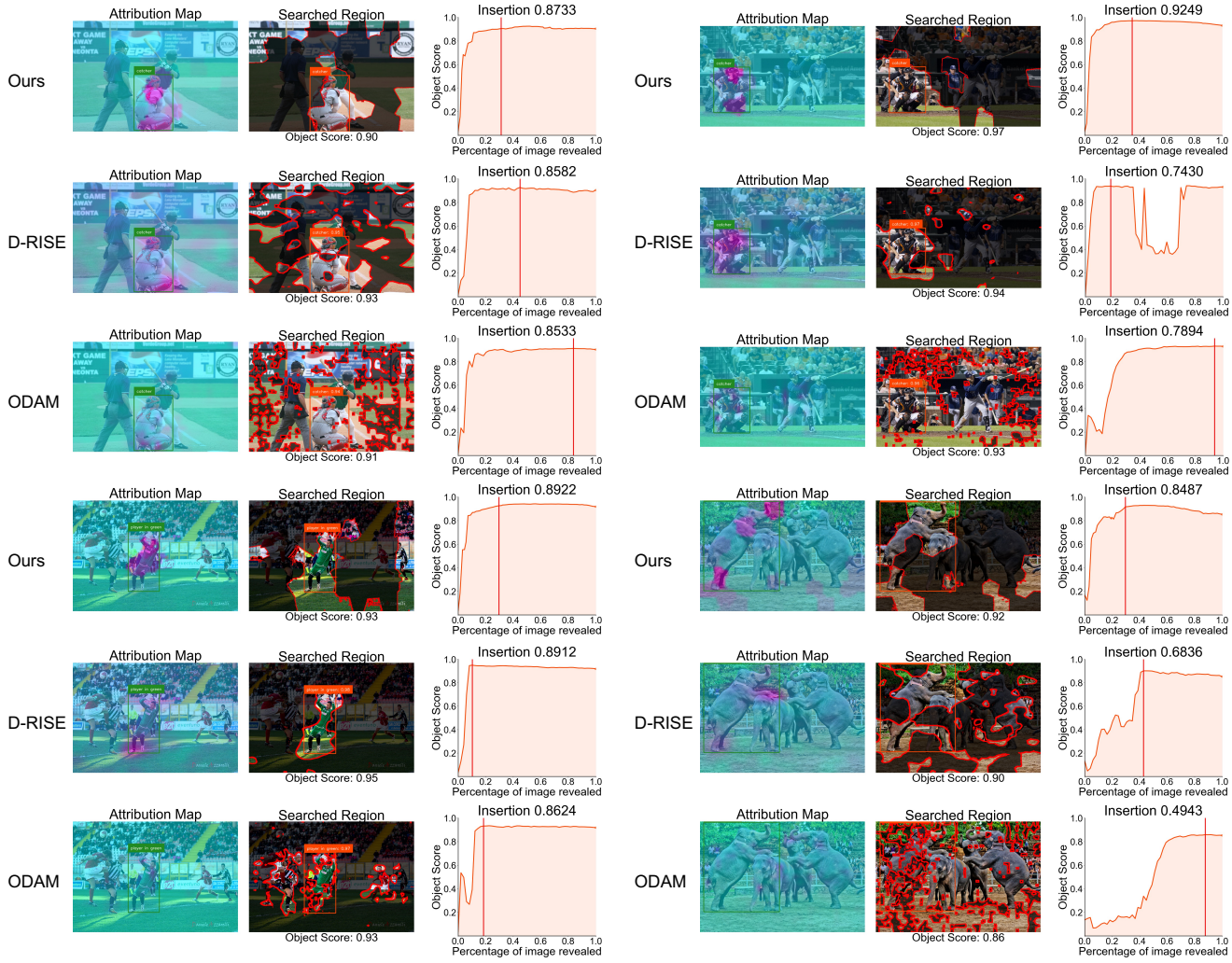

Figure S5. More visualization results of Grounding DINO for interpreting visual grounding task on the RefCOCO dataset.

References

[1] Ruoyu Chen, Hua Zhang, Siyuan Liang, Jingzhi Li, and Xiaochun Cao. Less is more: Fewer interpretable region via submodular subset selection. In *Int. Conf. Learn. Represent. (ICLR)*, 2024. 5

[2] Satoru Fujishige. *Submodular functions and optimization*. Elsevier, 2005. 2

[3] Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. Going beyond xai: A systematic survey for explanation-guided learning. *ACM Computing Surveys*, 56(7):1–39, 2024. 5

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):386–397, 2018. 2

[5] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 21–37, 2016. 2

[6] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017. 1, 2

[7] Vitali Petsiuk, Rajiv Jain, Varun Manjunatha, Vlad I Morariu, Ashutosh Mehra, Vicente Ordonez, and Kate Saenko. Black-box explanation of object detectors via saliency maps. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 11443–11452, 2021. 2, 3, 4

[8] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[9] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.*, 128:336–359, 2020. 2

[10] Shelley Zixin Shu, Aurélie Pahud de Mortanges, Alexander Poellinger, Dwarikanath Mahapatra, and Mauricio Reyes. Informer-interpretability founded monitoring of medical im-

age deep learning models. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 215–224, 2024. 5

[11] Andrea Stocco, Paulo J Nunes, Marcelo d'Amorim, and Paolo Tonella. Thirdeye: Attention maps for safe autonomous driving systems. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, pages 1–12, 2022. 5

[12] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Int. Conf. Mach. Learn. (ICML)*, pages 3319–3328, 2017. 1

[13] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: A simple and strong anchor-free object detector. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(4):1922–1933, 2020. 2

[14] Chenyang Zhao, Janet H Hsiao, and Antoni B Chan. Gradient-based instance-specific visual explanations for object specification and object discrimination. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024. 2, 3