Interpretable Open-Vocabulary Referring Object Detection with Reverse Contrast Attention

Supplementary Material

S5. FitAP

Based on standard definitions, the FitAP, similar to the mean average precision in object detection, can be defined as

$$\text{FitAP} = \frac{1}{10}\sum_{i=1}^{10} \text{AP}(\Theta_i),$$

wherein $\boldsymbol{\Theta} = \{\Theta_i = 0.5 + (i-1)0.05 \mid i = 1, 2, \ldots, 10\}$.

In the absence of a confidence score from the parsed VLM detection results, we propose to use the product of the normalized box area of detection and the IoU, $A_{\text{box}} \times \text{IoU}$, for the quality ranking of detection against the ground truth data.

We establish this approach by first showing the correlations of the area of the ground truth boxes with A_{box} and $A_{\text{box}} \times \text{IoU}$. Then, we visualize samples of the precision-recall curves generated by this approach, pointing out how its features resemble those generated by vision-only models. From these, we confirm that $A_{\text{box}} \times \text{IoU}$ is a reliable substitute for traditional confidence scores in generating precision-recall curves and calculating AP.

The critical step is to ensure that the metric used reliably correlates with the probability of detection being *true positive*, which is crucial to accurately calculate the AP and understand the performance of the model.

S5.1. Area correlations

Here, we offer an empirical basis of $A_{\text{box}} \times \text{IoU}$ for the quality ranking of the detection boxes by VLM from the following:

1. The tendency of VLM's predictions to maintain proportional sizing with the actual object in the image, and
2. influence of the actual object's size on the detection accuracy.

Thus, we examine the correlation between the area of ground truths (normalized to image size) and those of detection A_{box}. Our results (Figure S7) confirm that this correlation is strong (Pearson $r = 0.90$), indicating that VLM tends to generate detections with areas similar to the actual objects, affirming VLM's sizing accuracy, which is an essential aspect of **objectness**. The size accuracy implies that VLM recognizes and localizes the actual object in the image.

We further establish the correlation between the ground truth box areas and the proposed metric $A_{\text{box}} \times \text{IoU}$. The results (Figure S7) also confirm that this correlation is strong (Pearson $r = 0.92$), suggesting that larger, more well-fitting boxes are more common when the model correctly detects objects. In fact, this metric captures both the size and the quality of fit of the detections.

S5.2. Sample precision-recall curves

The choice of a metric to replace confidence scores should ideally reflect the confidence in detections being true positives. By showing samples of the generated precision-recall curves, we empirically demonstrate that $A_{\text{box}} \times \text{IoU}$ correlates with actual detection performance and does not introduce bias or misrepresentation in model evaluation.

The first noticeable feature is the general decreasing trends shown in Figure S8. This trend expresses the expected trade-off between precision and recall. Attempting to fit tighter (more precise) boxes increases the tendency to miss actual objects (less recall). However, aiming for better recall comes at the expense of looser detections.

Another peculiar feature displayed in Figure S8(h), (i) is the zigzag pattern of the empirical curve. The zigzag is an artifact of deriving floating-point ratios, i.e. precision and recall, from counting. As we aim for better recall, more detections are necessary at the expense of some of these being false positives, which explains the abrupt vertical drops. Gradual recovery is attributed to the acquisition of true positives and the improvement in recall. Then, another peak is encountered, at which point the next drop-off starts. However, succeeding peaks are, nevertheless, getting lower such that the envelope maintains the downward trajectory of the curve.

Finally, notice how the AP correspondingly decreases as the IoU threshold Θ increases. This tradeoff is evident from the curve's displacement toward the plot's bottom-left corner. This displacement effectively reduces the area under the curve, hence reducing FitAP. Higher Θ expresses a stricter criterion to detect true positives, resulting in fewer correct detections.

The precision-recall curves for other categories display the same characteristics. Therefore, we have shown how well $A_{\text{box}} \times \text{IoU}$ performs in predicting true positives, making it applicable for evaluating the object detection capability of VLMs.

S6. Indicators of the RCA-driven improvement

Here we develop a formal mathematical argument discussion that shows how Condition (6) from the paper establishes a negative relationship between the number of sub-threshold contributions to the hidden state and the scaler

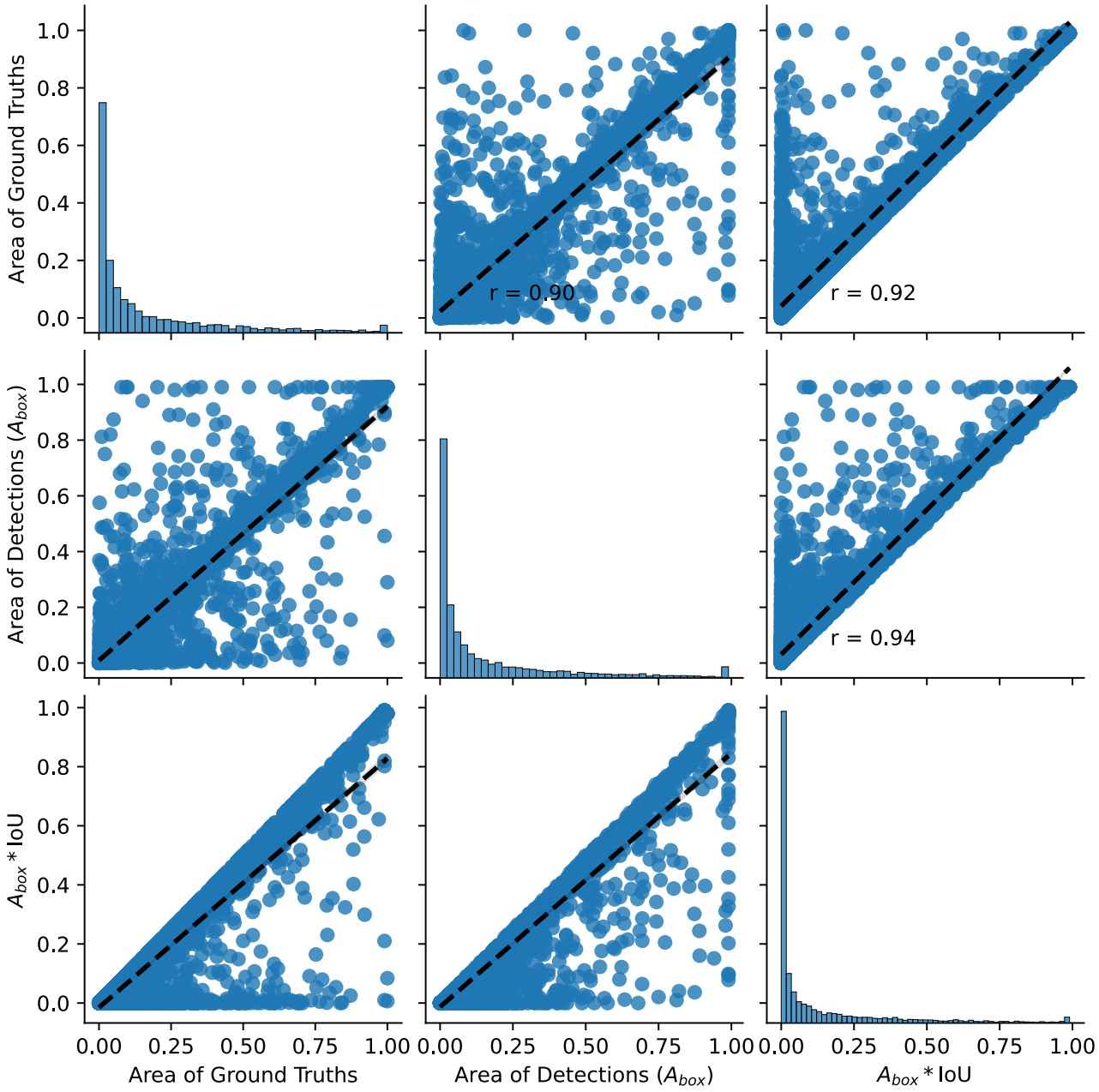
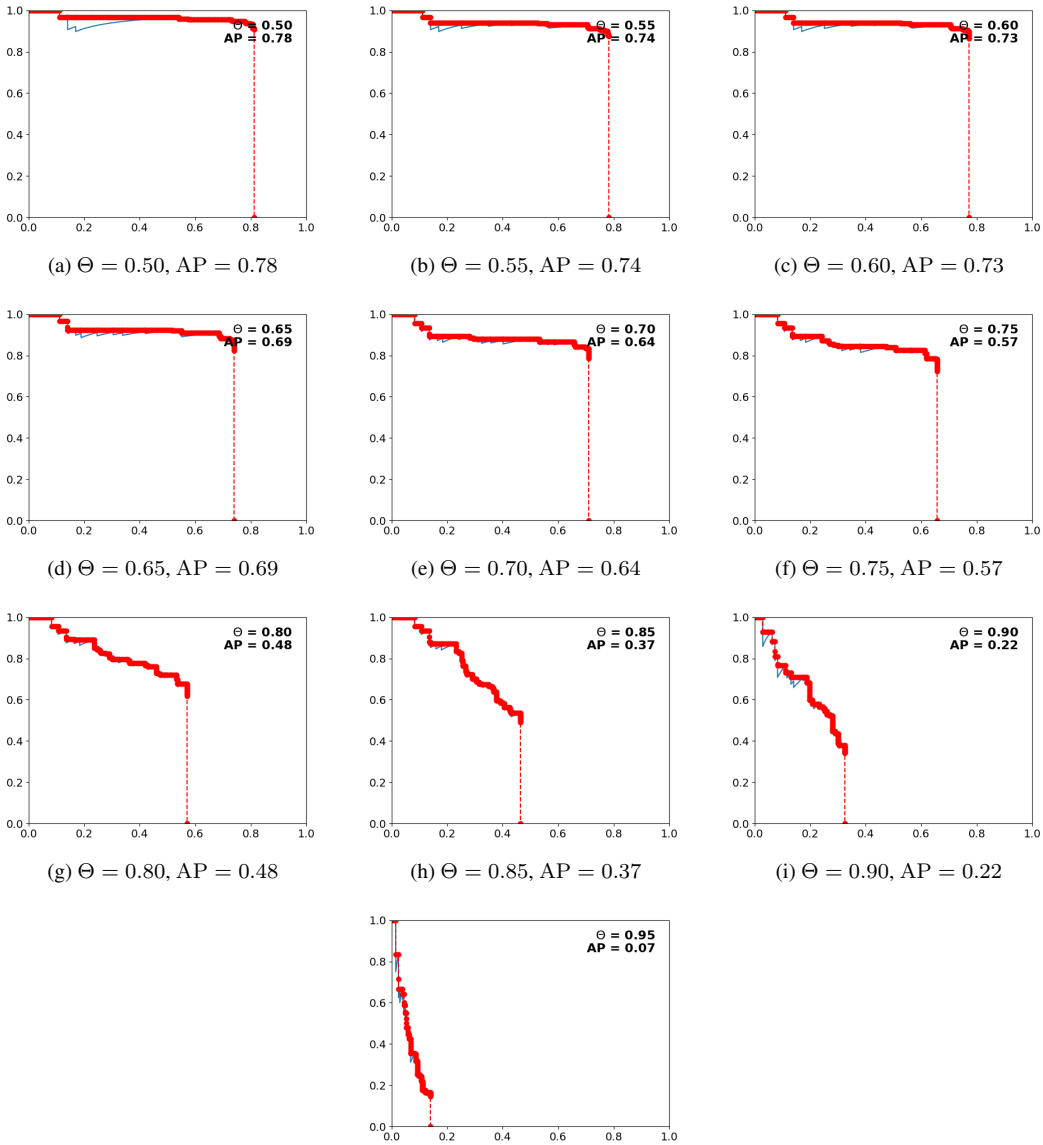


Figure S7. Correlations of ground-truth box area with A_{box} and $A_{\text{box}} \times \text{IoU}$ from the output of VLM queried with p_5. The plots include results from all object categories taken at IoU threshold, $\Theta = 0.50$. Diagonals are the univariate histograms of ground-truth area, A_{box}, and $A_{\text{box}} \times \text{IoU}$. The dashed lines represent linear regression fits with the Pearson correlation coefficient, r, shown only for the upper triangular plots.

m, which is the mean cross-head maximum of attention weights. We also prove that this inverse relationship is valid regardless of whether RCA uses inverse-distance or Gaussian peak reweighting, as defined in Section 1.2. For the preliminaries, let:

- $\alpha_{ij} \in [0, 1]$: the base attention weights from token i to j
- $A^{(h)} \in \mathbb{R}^{n \times n}$: attention map from head h, for $h = 1, \ldots, H$
- $A_{ij}^{\max} := \max_h A_{ij}^{(h)}$
- $m := \frac{1}{n} \sum_{j=1}^{n} \max_i A_{ij}^{\max}$: mean column-maximum of $A^{\max}$



(a) $\Theta = 0.50$, AP $= 0.78$ (b) $\Theta = 0.55$, AP $= 0.74$ (c) $\Theta = 0.60$, AP $= 0.73$

(d) $\Theta = 0.65$, AP $= 0.69$ (e) $\Theta = 0.70$, AP $= 0.64$ (f) $\Theta = 0.75$, AP $= 0.57$

(g) $\Theta = 0.80$, AP $= 0.48$ (h) $\Theta = 0.85$, AP $= 0.37$ (i) $\Theta = 0.90$, AP $= 0.22$

(j) $\Theta = 0.95$, AP $= 0.07$

Figure S8. Precision-recall curves for the category `giraffe` at different IoU thresholds Θ and the corresponding average precision, AP (area under the curve). Solid (blue) curves from actual data; dashed lines-points (red) represent envelopes from which FitAP is calculated as the average of AP for $\Theta \in [0.50 : 0.05 : 0.95]$.

The value of m quantifies the **global sharpness** in attention across all heads.

We want to show that:

$$m \uparrow \implies \sum_{j \in \mathcal{J}_\downarrow} \widetilde{\alpha_{ij}} \downarrow \implies \widetilde{z}_i(d) \uparrow$$

implying fewer subthreshold components, and thus, condition (6) implies a **negative relationship** between the number of subthreshold components and m.

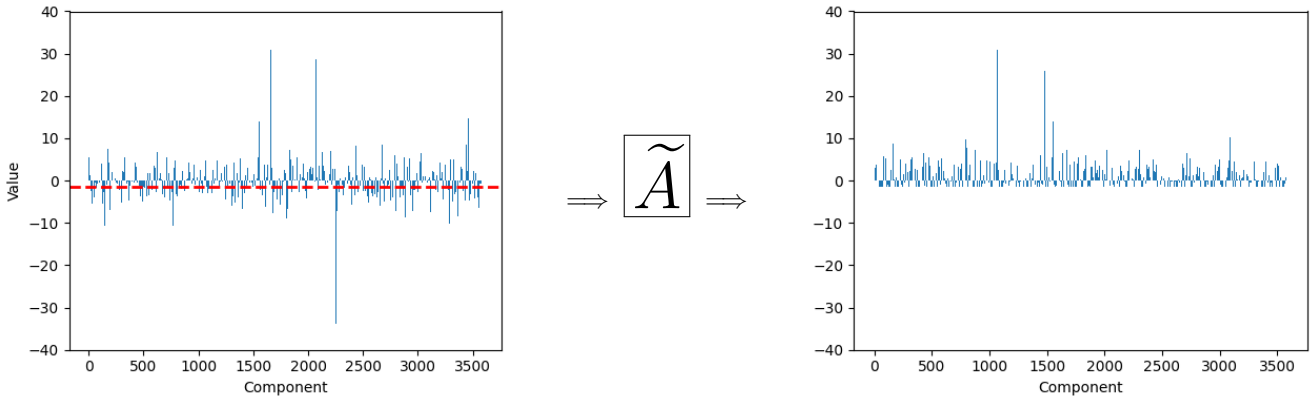
From the paper:

$$\widetilde{z}_i(d) \geq \vartheta + \left(v^- - \vartheta\right) \sum_{j \in \mathcal{J}_\downarrow} \widetilde{\alpha_{ij}},$$

where:

- ϑ: threshold value (floor)
- $v^- < \vartheta$: minimal value component of subthreshold tokens
- $\widetilde{\alpha_{ij}}$: RCA-transformed attention weights (depends on m)

Thus, minimizing $\sum_{j \in \mathcal{J}_\downarrow}$ tightens the bound so that $\widetilde{z}_i(d)$



(a) Pre-RCA hidden state z_i

(b) Post-RCA hidden state $\widetilde{z}_i$

Figure S9. Flooring the subthreshold contributions of a hidden state z_i implicitly implies $\widetilde{A}$ leading to $\widetilde{z}_i$. The red dashed horizontal line in (a) corresponds to $\vartheta = -1.5$. In this example, the embedding size is $d = 3584$

is closer to or above ϑ. For the key strategy, we show:

1. $m \uparrow \implies \widetilde{\alpha_{ij}}$ assigns less weight to $j \in \mathcal{J}_{\downarrow}$.
2. This assertion holds for both RCA schemes:
   - Inverse-distance from m
   - Gausian peaking around m

Therefore, the penalty term in condition (6) shrinks with increasing m, which increases $\widetilde{z}_i(d)$, decreasing the subthreshold count.

In the first case of inverse-distance reweighting:

$$\alpha'_{ij} = \frac{1}{1 + \gamma|\alpha_{ij} - m|},$$

which peaks at $\alpha_{ij} = m$ and decreases as α_{ij} deviates from m. Suppose m increases. Then for fixed α_{ij}, the distance $|\alpha_{ij} - m|$ increases unless α_{ij} tracks m. Thus, for subthreshold contributors $j \in \mathcal{J}_{\downarrow}$, which typically have $\alpha_{ij} < m$ and $v_j(d) < \vartheta$, we get:

$$\alpha'_{ij}(m) \downarrow \implies \widetilde{\alpha_{ij}} \downarrow \implies \sum_{j \in \mathcal{J}_{\downarrow}} \widetilde{\alpha_{ij}} \downarrow \implies \widetilde{z}_i(d),$$

implying that the subthreshold count decreases.

In the second case of Gaussian peak reweighting:

$$\alpha'_{ij} = \exp\left[-\gamma \left(\alpha_{ij} - m\right)^2\right],$$

which symmetrically peaks at $\alpha_{ij} = m$ and rapidly decays as α_{ij} moves away from m. Suppose m increases. For fixed α_{ij}, again $|\alpha_{ij} - m|$ increases and so α'_{ij} decreases and penalizes values further away from m. Thus, subthreshold tokens $j \in \mathcal{J}_{\downarrow}$ with mid- or low α_{ij}, get decreasing attention as m increases. So again,

$$\sum_{j \in \mathcal{J}_{\downarrow}} \widetilde{\alpha_{ij}} \downarrow \implies \widetilde{z}_i(d) \uparrow \implies \text{ subthreshold count } \downarrow .$$

From these arguments, we have shown that under both RCA reweighting strategies (inverse-distance and Gaussian peaking), as $m \uparrow$, subthreshold tokens $j \in \mathcal{J}_{\downarrow}$ receive less attention mass so $\sum_{j \in \mathcal{J}\downarrow} \widetilde{\alpha_{ij}} \downarrow$, which increases the lower bound of Condition (6). Thus, decreasing the number of components $\widetilde{z}_i$ that fall below ϑ, as visualized in Fig. S9

$$\boxed{\frac{d|S|(\widetilde{z}_i)}{dm} < 0 \text{ as implied by Condition (6)}}$$

where $|S|$ is the number of subthreshold contributors. This conclusion establishes that condition (6) supports a negative relationship between the subthreshold count and the attention sharpness measure m, regardless of RCA variant used.