Interpretable Open-Vocabulary Referring Object Detection with Reverse Contrast Attention

Drandreb Earl O. Juanico[1], Rowel O. Atienza[1,2], Jeffrey Kenneth Go[3]
[1]AI Graduate Program, University of the Philippines, Diliman, Quezon City
[2]EEEI, University of the Philippines, Diliman, Quezon City
[3]Samsung R&D Institute Philippines
dojuanico@up.edu.ph, earl.juanico@gmail.com, rowel@eee.upd.edu.ph, jeff.go@samsung.com

Abstract

We propose Reverse Contrast Attention (RCA), a plug-in method that enhances object localization in vision-language transformers without retraining. RCA reweights final-layer attention by suppressing extremes and amplifying mid-level activations to let semantically relevant but subdued tokens guide predictions. We evaluate it on Open Vocabulary Referring Object Detection (OV-RefOD), introducing FitAP, a confidence-free average precision metric based on IoU and box area. RCA improves FitAP in 11 out of 15 open-source VLMs, with gains up to $+26.6\%$*. Effectiveness aligns with attention sharpness and fusion timing; while late-fusion models benefit consistently, models like* `DeepSeek-VL2` *also improve, pointing to capacity and disentanglement as key factors. RCA offers both interpretability and performance gains for multimodal transformers. Codes and data set in* https://github.com/earl-juanico/rca

1. Introduction

Vision-language transformers are widely considered as effective computational models for studying how natural language reasoning interfaces with visual perception. These models align and integrate information across image and text with multi-modal attention mechanisms. However, the interpretability of outputs by a vision-language model (VLM), especially in explaining how cross-modal attention pathways selectively propagate visual features in response to linguistic cues, remains a critical and ongoing research challenge. Underscoring this challenge is a debate sparked by two highly influential work. Jain and Wallace's "*Attention is Not Explanation*" [11] demonstrated how significant modifications to the attention weights did not change the model output or responses. This conclusion undermines previous findings supporting the argument that attention reliably represents model reasoning. On the contrary, in their equally influential work, "*Attention is Not Not Explanation*", Wiegreffe and Pinter [26] argued that in particular methodically crafted, constrained scenarios, attention may provide insightful interpretation.

The central debate in both influential papers concerns whether or not attention weights uniquely explain a VLM output. As Jain and Wallace demonstrated, multiple distinct attention distributions can often yield the same model response, indicating that no single configuration of attention weights acts as a definitive trace of the model's internal reasoning. In contrast, Wiegreffe and Pinter contended that attention can still serve as a useful interpretive tool, provided that modifications to the attention distributions are made deliberately, under constraints that preserve output fidelity and align with known model behavior. Taken together, these perspectives suggest that attention flexibility (the capacity to vary without altering predictions) can be productively used for interpretability. This functional plasticity of attention may thus be viewed not as a limitation but as a resource for probing the model's internal decision structure.

In this study, we propose to improve VLM performance in a computer vision task without additional training or fine-tuning by leveraging attention functional plasticity to directly manipulate the attention weights used to compute logits during inference. This manipulation preserves the gains from extensive pretraining on large-scale data sets while enabling inference-time adaptation to task-specific objectives. Building on previous eXCV studies focused on mapping and visualizing VLM attention, our method capitalizes on this attention-derived guidance for performance improvement, thus linking interpretability to functional enhancement.

We establish this improvement by defining the task and the probing mechanism to achieve it in such a task. We then discuss how the manner of attention manipulation relates to VLM explainability. In particular, we draw parallels to the

relevance propagation idea to interpret bi-modal transformers [1] and the tracer scoring approach with deep Taylor decomposition [2], which recognize the potential of attention traces not just for explanation, but also for actionable model intervention during inference.

1.1. Related Work

Reverse Contrast Attention (RCA) generalizes the focus shifting principle introduced by Chen *et al*. [4], in which confident predictions are erased to help recover missed object regions in a top-down manner. Similarly, Huang *et al*. [9] employed reverse attention to suppress incorrect class predictions in confusing image regions. This suppression effectively redirected the model's attention where it underperforms. Li *et al*. used a "reverse-and-distill" strategy to disentangle attribute and object representations, where reverse attention is used to guide the learning of less visible semantic components by masking confident attribute/object features [15]. Although it does not use reverse attention per se, Hyeon-Woo *et al*. [10] tackle the bias of peaked softmax attention by injecting uniform attention to support denser token interactions. Following these works, RCA indirectly enhances mid-level attention activation by suppressing extremes in the transformer attention matrix, allowing focus redistribution across insufficiently attended yet semantically relevant image tokens.

Our approach is also informed by the hierarchical design by Wang *et al*. [25], where stacked attention modules progressively emphasize important image regions. We abstract this concept by applying contrastive modulation not on spatial features, but rather, indirectly to attention weights with a parameter-based flooring of the final layer hidden states (Eq. 3). Furthermore, akin to the work "*Self-Attention with Relative Position Representations*" [19], which adjusts attention based on position, our method promotes a more balanced and structured distribution of attention across tokens, not by relying on positional information (*e.g*. positional encoding) but rather by contrast enhancement to emphasize moderately attended regions and suppress overly dominant or neglected ones.

Recent works such as RA-Net [25], RTA-Former [16], and SRaNet [13] confirm the growing value of reverse attention mechanisms in guiding residual learning, boundary recovery, and transformer attention refinement. Sun *et al*. [22] introduced both reverse and boundary attention units in a residual refinement module to gradually refine road segmentation by focusing on previously missed regions and road edges. Xie *et al*. [28] introduced learnable bidirectional attention maps, including reverse attention, which suppresses known regions so the U-net can focus solely on reconstructing missing parts. Our RCA offers a general-purpose formulation for this class of methods, operating on attention scores through hidden states to improve detection, grounding, and interpretability across vision-language tasks.

A recent study by Venhoff *et al*. [23] introduces a controlled framework for analyzing how a trainable adapter maps visual representations into the feature space of a frozen LLM. With tools based on the sparse autoencoder (SAE), they demonstrate that vision-language alignment predominantly emerges in the middle-to-late transformer layers. This discovery reinforces the rationale behind RCA, which operates directly on the final transformer layer's attention matrix to boost weakly attended but semantically relevant tokens. The SAE-based observations indicate that attentional reweighting strategies like RCA are most effective when applied to the layers where visual features have already begun to resemble the internal language representations, specifically, the mid-to-late layers of the transformer where cross-modal alignment naturally occurs.

2. Methodology

2.1. Computer Vision Task

Various VLMs were subjected to the task of identifying and localizing (with bounding boxes) all objects in an image matching a free-form natural language prompt, despite the possible absence of object categories in a fixed label set. We refer to this task as *Open Vocabulary Referring Object Detection* (OV-RefOD), which has recently emerged at the interface between open-vocabulary object detection (OVD) and referring expression comprehension (REC), such as in visual grounding [24] and attribute recognition [3].

OV-RefOD is open-ended localization guided by natural language, where:

- **Input:** image and an arbitrary text prompt
- **Output:** bounding box coordinates `[x1, y1, x2, y2]` (parsed) of at least one of all visible instances in the image matching the description

The concept behind the task has been gaining traction in discussions around VLM evaluation. Notable developments include the introduction of OV-VG, a benchmark for open-vocabulary visual grounding [24] and phrase localization [32]; GroundVLP [20], which exploits zero-shot visual grounding from vision-language pretraining and OVD; and grounded spatial reasoning in VLMs [5], which released the Open Spatial Dataset with five million open-vocabulary boxes and masks to test grounding capabilities. More recent efforts include LED [32], which augments OV detectors by integrating hidden states from LLMs to enhance grounding on OmniLabel benchmarks, and zPROD [21], which introduces a zero-shot framework for OVD, segmentation, and grounding in challenging autonomous-driving contexts. Collectively, these studies reflect a broader trend toward systematically evaluating VLM grounding under open-vocabulary and zero-shot regimes.

2.2. Reverse Contrast Attention

This paper introduces Reverse Contrast Attention (RCA), a novel method inspired by treating the transformer's attention matrix as an image. In this analogy, the attention matrix reveals visual patterns and contrast adjustment can be used to emphasize certain features. Traditional contrast enhancement amplifies extremes, making high values brighter and low values darker, effectively emphasizing dominant patterns. However, reverse contrast enhancement suppresses extremes and brings out mid-range features that might otherwise be overlooked. RCA applies this principle to the final-layer attention maps of VLMs, effectively establishing a "floor" on the hidden states. This adjustment improves the model's sensitivity to moderately activated but semantically relevant visual tokens, thereby enhancing the average precision of its responses in OV-RefOD tasks.

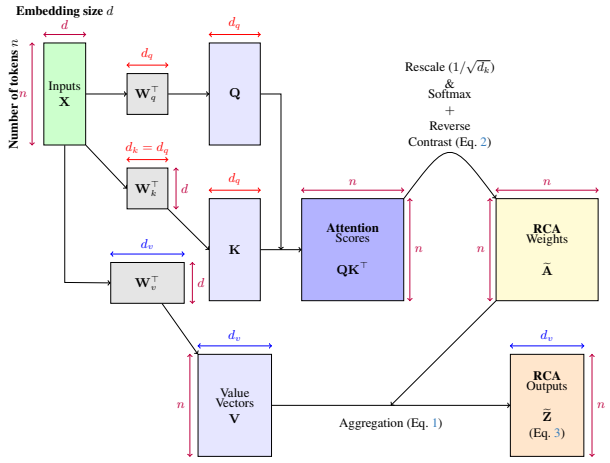


Figure 1. Illustration of the RCA mechanism.

In a standard transformer layer, the hidden state z_i is a vector that represents the token i as a superposition of the value vectors $v_j \in \mathbf{V}$ of the text prompt and the image regions according to the attention distribution $\mathbf{A} = \{\alpha_{ij}\}$ (Figure 1):

$$z_i = \sum_{j=1}^{n} \alpha_{ij} v_j \in \mathbf{Z}. \tag{1}$$

RCA restructures $\mathbf{A}$ by amplifying α_{ij} that are near some central value m, while inhibiting α_{ij} that are far above or below m. This restructuring can be accomplished by a nonmonotonic reweighting that can particularly take either form:

- **Inverse distance from m:**

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

- **Gaussian peaking around m:**

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

The free, but possibly optimizable, parameter γ regulates *reverse contrast*, or how sharply large deviations from m are penalized. Without loss of generality, we take $\gamma = 1$ whereas m, which is another potentially optimizable free parameter, is manually selected.

The reverse contrast weights α' then go through a renormalization:

$$\widetilde{\alpha_{ij}} = \frac{\max\left(\alpha'_{ij}, 0\right)}{\sum_{k=1}^{n} \max\left(\alpha'_{ij}, 0\right)}, \tag{2}$$

which are $\in \widetilde{\mathbf{A}}$ in Figure 1. We claim that RCA implies the element-wise flooring operation applied to the final-layer hidden states such that:

$$\widetilde{\mathbf{Z}} = \left\{ \max\left(z_i, \vartheta\right) | z_i \in \mathbf{Z}, \vartheta \in \mathbb{R} \right\}, \tag{3}$$

in which ϑ is a free parameter linked to m and γ. Adjusting ϑ implies tuning of m or γ.

2.3. Inference

During inference, a model receives an input pair, $\mathbf{X} = (I, Q)$, where I is an image, and Q is the free-form natural language query:

```
'Give the normalized bounding box coordinates
in the format [x1, y1, x2, y2] of all
instances of {cls} in the image.'
```

in which `{cls}` refers to the category or the descriptive phrase of the object, and `x1`, `x2`, `y1`, and `y2` are ideally floating point values $\in [0, 1]$. This prompt was applied in all VLM to generate one or more parsed bounding boxes $\{B_k\}$ that supposedly align with the object(s) referenced in Q. Some models returned B_k in pixel coordinates rather than in normalized format. To ensure consistency, all outputs were standardized to a common format using regular expression parsing.

2.4. VLM Selection

We considered open-source VLMs from the OpenCompass Multi-Modal Academic Leaderboard (OC-MMAL) at https://rank.opencompass.org.cn/leaderboard-multimodal due to their publicly benchmarked performance on image-text reasoning tasks and other multi-modal capabilities. The VLM must satisfy the following criteria:

- LLM size $<$ 35B parameters, for efficiency reasons;
- Provides bounding box coordinates in the form B_k;
- Checkpoints available in HuggingFace

The list samples top- and middle-ranking open-source VLM in OC-MMAL: Although we attempted as broad a coverage as possible, some models like `Ola-7b` and `Intern-VL` do not respond to the prompt Q with parsable bounding box information or were designed to interpret the prompt as a REC rather than OV-RefOD. Other models are earlier versions or have a version with a higher rank in OC-MMAL; hence, we opt for the updated or higher ranked version.

Model	Params	LLM	Vision
Ovis2-34B	34.9B	Qwen2.5-32B	AIMv2-1B
SAIL-VL-1.6-8B	8.33B	Qwen2.5-7B	AIMv2 Huge
WeThink-Qwen2.5VL-7B	8.29B	Qwen2.5-7B	QwenViT
Qwen2.5-VL-7B	8.29B	Qwen2.5-7B	QwenViT
MiniCPM-o-2.6	8.67B	Qwen2.5-7B	SigLIP-400M
valley2_dpo	8.88B	Qwen2.5-7B	SigLIP-400M
Kimi-VL-A3B	16.4B	Moonlight -16B-A3B	MoonViT
Ristretto-3B	3.84B	Qwen2.5-3B	SigLIP-400M
POINTS1.5-Qwen2.5-7B	8.3B	Qwen-2.5-7B	NaViT
Valley-Eagle	8.9B	Qwen2.5-7B	SigLIP-400M
Gemma3-27B	27.4B	Gemma3-27B	SigLIP-400M
VARCO-VISION-14B	15.2B	Qwen2.5-14B	SigLIP-400M
DeepSeek-VL2	27.5B	DeepSeekMoE -27B	SigLIP-400M
PaliGemma2-3B-mix-448	3B	Gemma2-2B	SigLIP-400M
Moondream2	1.9B	Phi-1.5	SigLIP-400M

Table 1. Selected VLMs from OC-MMAL in ascending rank.

Model	**FitAP** (↑)		
	pre-RCA	post-RCA	% Change
Ovis2-34B	3.23869	3.52222	**+8.75**
SAIL-VL-1.6-8B	4.84873	5.67149	**+17.0**
WeThink-Qwen2.5VL-7B	39.9640	37.7606	−5.51
Qwen2.5-VL-7B	37.0005	46.8535	**+26.6**
MiniCPM-o-2.6	0.03064	0.07334	**+139**
valley2_dpo	11.5145	11.6927	**+1.55**
Kimi-VL-A3B	30.7194	32.2176	**+4.88**
Ristretto-3B	9.12887	7.94552	−13.0
POINTS1.5-Qwen2.5-7B	9.75203	9.45686	−3.03
Valley-Eagle	11.7736	11.2598	−4.36
Gemma3-27B	2.74179	3.01913	**+10.1**
VARCO-VISION-14B	27.3592	28.7003	**+4.90**
DeepSeek-VL2	3.38530	3.99586	**+18.0**
PaliGemma2-3B-mix-448	38.7982	41.1179	**+5.98**
Moondream2	47.0039	47.0819	**+0.17**

Table 2. FitAP of VLM before and after applying RCA.

2.5. Evaluation

In this paper, we propose **FitAP** (supplementary section S5), a modified evaluation metric derived from Average Precision (AP), commonly used in object detection tasks [6]. Unlike standard AP, which ranks detections by confidence scores typically produced by region proposal networks [18], FitAP is designed to evaluate OV-RefOD in VLMs that lack explicit confidence output. FitAP ranks predicted bounding boxes according to the product of their normalized area A_{box} and their intersection-over-union (IoU) with ground-truth annotations. This alternative ranking strategy preserves the precision-recall structure of AP while enabling evaluation in settings where traditional confidence-based sorting is not available or unreliable. Thus, FitAP provides a practical and interpretable measure for assessing detection quality in VLMs operating under weakly supervised or prompt-based regimes [12, 14].

We evaluated the models in Table 1 on the novel split of `COCO val 2017` [7, 31] consisting of 2064 (I, Q) pairs. The mean FitAP is the average value at different IoU thresholds $\in [0.5 : 0.05 : 0.95]$.

3. Results and Discussion

RCA improved six of the top seven and, in general, 11 of 15 VLM in Table 1, despite the absence of systematic optimization of the ϑ parameter in Eq. 3. The authors of `WeThink-Qwen2.5VL-7B` [29] observed that additional training stages, especially supervised fine-tuning and chain-of-thought, degraded grounding precision and latent object detection capability of `Qwen2.5-VL-7B`, which could partially explain the negative effect of RCA on this model. For `MiniCPM-o-2.6`, the low FitAP is probably due to internal randomization imposed on response generation through sampling decoding [30].

In the bottom half of the list (Table 2) is `PaliGemma2-3B-mix-448`, which, like `Gemma3-27B` garnered a positive RCA effect. The more significant OV-RefOD improvement of `PaliGemma2-3B-mix-448` than `Gemma3-27B` can be explained by its integration of image and text at the model level, enabling deep cross-modal reasoning; hence, naturally suitable for object detection and visual grounding. `Gemma3` entirely lacks this modality fusion.

Figures 2 and 3, in which TP and FP represent true positive and false positive detection, respectively, illustrate the improved recall and precision that corroborate positive FitAP changes due to RCA (Table 2). Green boxes (solid) belong to ground-truth annotations from `COCO val 2017`, while red boxes (dashed) correspond to parsed VLM detections. In Fig. 2, RCA enhanced the detection of multiple instances of objects `bus` and `elephant`, while increasing its precision in detecting small objects such as `snowboard` and `sink` (higher IoU). In Fig. 3, RCA sharpened box precision leading to higher IoU in multiple instances of `elephant`, and single instances of `cup`, `umbrella`, and `airplane` in their respective images. These improvements are explainable with RCA enforcing a shift in attention focus toward subdued image tokens (Figure 4).

In the original attention matrix (Fig. 4, left), the highlighted patch indices exhibit strong attention scores that extend downward to the final rows. However, several tokens corresponding to distinct image regions remain visually indistinct, suggesting insufficient attention. After applying RCA, these previously subdued tokens are amplified in the transformed attention matrix (right), making their associated image patches more prominent. This redistribution of attention likely allowed the model to detect an additional object instance, namely the `kite`, which was not distinguished in the pre-RCA outputs.

The RCA propagates its influence through a structured transformation of $\mathbf{A}$ centered on a chosen mid-value m, which is the *mean of the column-wise (i.e., along the* $\mathbf{K}$*-dimension of* $\mathbf{A}$*) maximum attention across multiple heads*. This restructuring inhibits extremes, that is, tokens or image regions receiving disproportionately high or low atten-

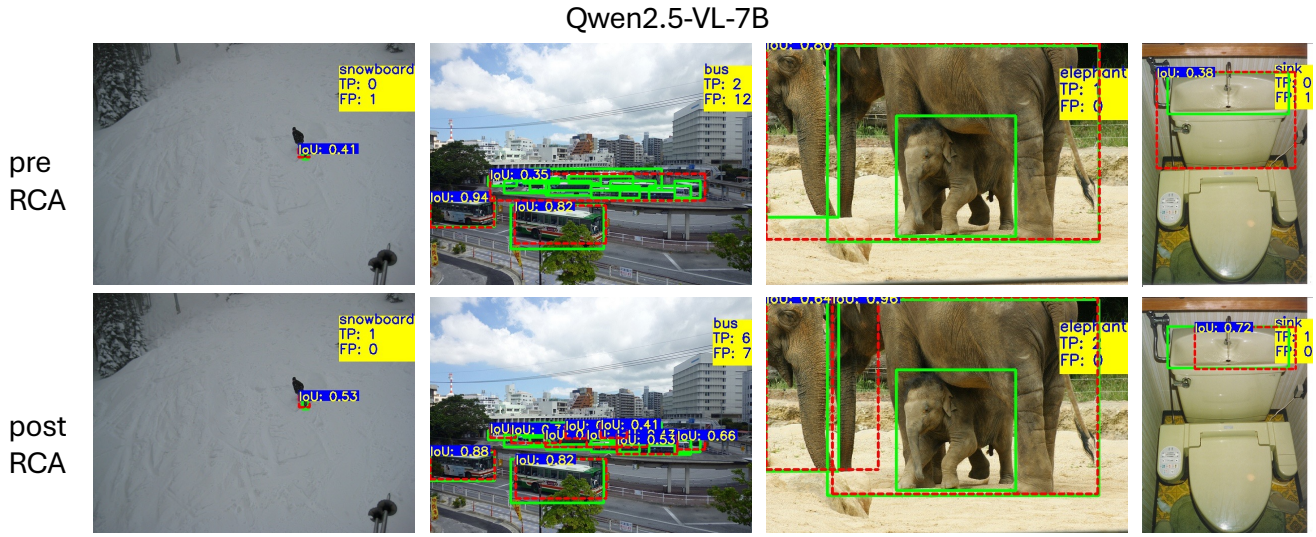


Figure 2. Selected examples suggestive of RCA's positive impact on `Qwen2.5-VL-7B`: solid, green boxes (ground truth); dashed, red boxes (parsed detections). These cases illustrate improved object localization and precision following the application of RCA.

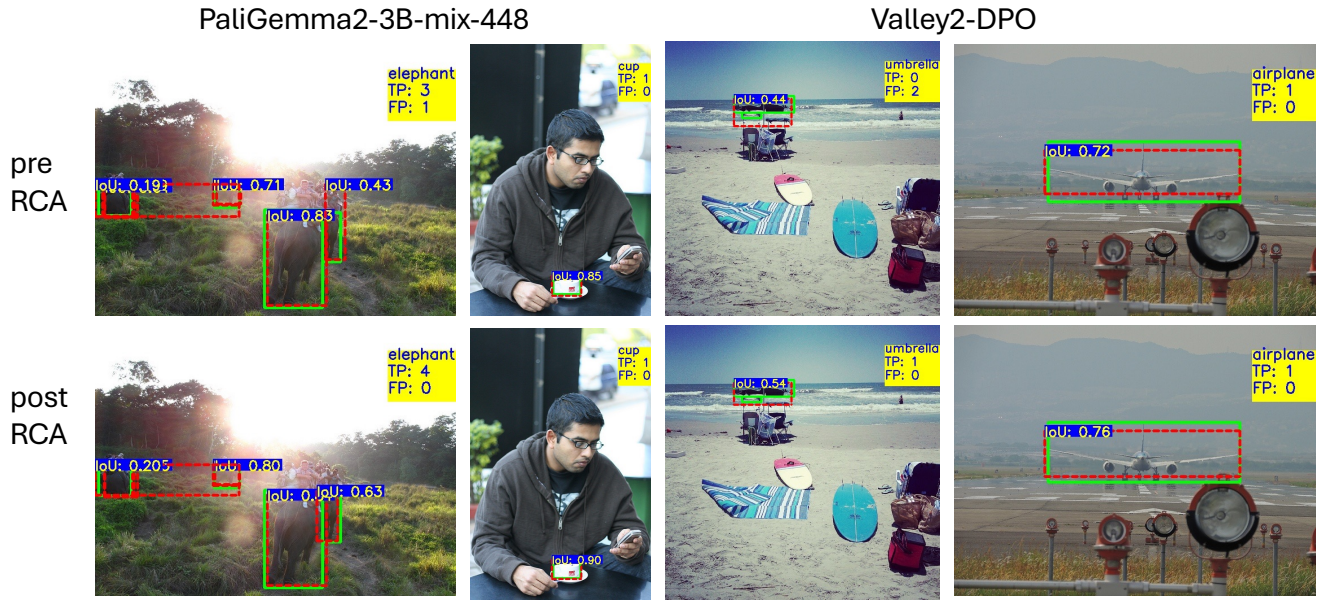


Figure 3. Selected examples from `PaliGemma2-3B-mix-448` and `valley2_dpo`, qualitatively illustrating the observed improvements in detection after RCA is applied: solid green boxes (ground truth); dashed red boxes (detections)

tion, while it amplifies contributions closer to the midpoint. Thus, $\widetilde{\mathbf{A}}$ is a more equalized attention profile, implying more stable and bounded hidden states from the final transformer layer. As extreme attentions are suppressed, the resultant vectors from applying the superposition (Eq. 1) on $\widetilde{\mathbf{A}}$ rather than on $\mathbf{A}$ are less likely to fall below a threshold ϑ (supplementary Fig. S9) as outlined in the following argument. Suppose that

$$\widetilde{z}_i = \sum_{j=1}^{n} \widetilde{\alpha_{ij}} v_j \tag{4}$$

represents the transformed hidden states from the final transformer layer, where $\widetilde{\alpha_{ij}} \in \widetilde{A}$ is the renormalized attention weight (Eq. 2) for the query token i over "key" token j, and $v_j \in \mathbb{R}^{d_v \times 1}$ is the value vector for the token j. Con-

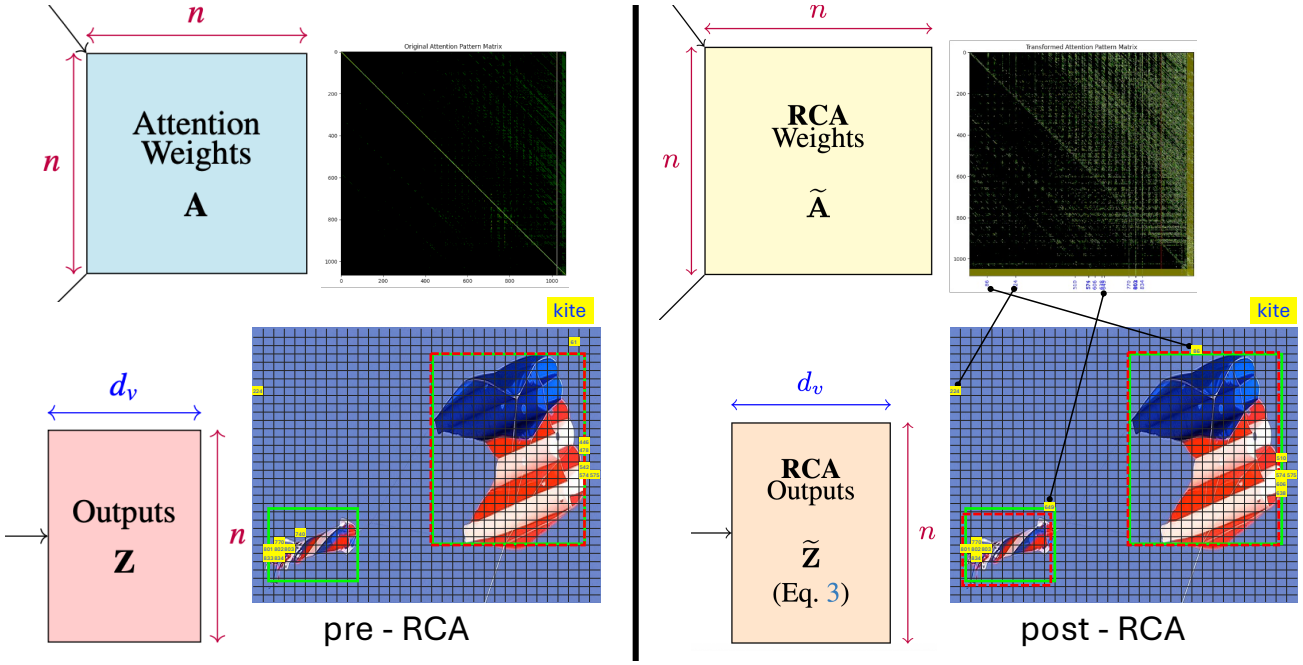


Figure 4. Visualizing the link between RCA and OV-RefOD in `PaliGemma2-3B-mix-448` for a sample of `kite` from `COCO val 2017` base vocabulary subset: (*left*) before applying RCA; (*right*) after applying RCA. Attention matrix images are beside their corresponding diagram; and patched image directly below the corresponding attention matrix image. The outputs generate bounding box information that can be drawn on the image. Solid green boxes are ground truths; dashed red boxes are the parsed VLM detections of the class `kite`. The highlighted patches (yellow with blue font) correspond to sufficiently attended image tokens. The indexes of these tokens are shown in the post-RCA attention matrix. Some patches that emerged post-RCA are linked to their corresponding positions in the image for emphasis.

sider in further detail a partition of tokens into two sets of disjoint token index.

- **Subthreshold contributors:** are tokens that dip below ϑ in dimension d.

$$\mathcal{J}_{\downarrow} = \{j | v_j(d) < \vartheta\}$$

- **Suprathreshold contributors:** are tokens of which value vector is at least ϑ in dimension d.

$$\mathcal{J}_{\uparrow} = \{j | v_j(d) \geq \vartheta\}$$

where the inequalities hold element-wise in all dimensions. Applying this partition to each component d, which is the d^{th} scalar element representing one channel of the model's internal feature space, in Eq. 4, then

$$\widetilde{z}_i(d) = \sum_{j \in \mathcal{J}_{\downarrow}} \widetilde{\alpha_{ij}} v_j(d) + \sum_{j \in \mathcal{J}_{\uparrow}} \widetilde{\alpha_{ij}} v_j(d) \tag{5}$$

Performing the partition (Section 3.1) ultimately leads to the following inequality:

$$\widetilde{z}_i(d) \geq \vartheta + \underbrace{(v^- - \vartheta)}_{\text{negative}} \underbrace{\left(\sum_{j \in \mathcal{J}_{\downarrow}} \widetilde{\alpha_{ij}} \right)}_{\text{small if penalized}}. \tag{6}$$

When the penalty term is sufficiently small, $\widetilde{z}_i(d)$ ends up "close enough" to ϑ, resulting in the RCA outputs defined in Eq. 3. The term $\sum_{j \in \mathcal{J}_{\downarrow}} \widetilde{\alpha_{ij}}$, the renormalized mass assigned to the subthreshold contributors, must be small. The guarantee that RCA increases performance relies on this term being close to zero, which could explain why it improved OV-RefOD performance in some models but not in others (Table 2).

3.1. Soft Guarantee to RCA Flooring

Assume that among the subthreshold contributors to the hidden state of the final transformer layer, the minimal (or near-minimal) component is $v^- < \vartheta$. Then, for $j \in \mathcal{J}_{\downarrow}$, $v_j(d) \geq v^-$, while for $j \in \mathcal{J}_{\uparrow}$, $v_j(d) \geq \vartheta$. Hence, Eq. 5 simplifies into an inequality:

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

Rewrite by factoring out constants:

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

But by virtue of normalization,

$$\sum_{j\in\mathcal{J}_\uparrow} \widetilde{\alpha_{ij}} + \sum_{j\in\mathcal{J}_\downarrow} \widetilde{\alpha_{ij}} = 1,$$

such that the preceding inequality becomes

$$\widetilde{z_i}(d) \geq \vartheta \left(1 - \sum_{j\in\mathcal{J}_\downarrow} \widetilde{\alpha_{ij}}\right) + v^- \left(\sum_{j\in\mathcal{J}_\downarrow} \widetilde{\alpha_{ij}}\right).$$

Thus,

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

which is the inequality (6).

If $\sum_{j\in\mathcal{J}_\downarrow} \widetilde{\alpha_{ij}}$ is very small, say $\delta \ll 1$, then

$$\widetilde{z_i}(d) \geq \vartheta + \left(v^- - \vartheta\right)\delta \approx \vartheta.$$

Of course, $\widetilde{z_i} \geq \vartheta$ is not a strict guarantee, but is based on the original distribution $\mathbf{A}$ of attention weights, the central value m, and the subthreshold range, $(v^- - \vartheta)$. The parameter ϑ could possibly be derived from γ and m and could be optimized based on the magnitude of the improvement in the VLM performance in OV-RefOD due to RCA.

3.2. Empirical Test of RCA Effect

To empirically verify the assumption underpinning condition (6), we investigated the correlation between the mean attention weights (a suitable central value) and the number of subthreshold contributions. Specifically, we defined the index set $S = \{i : \widetilde{z_i} < \vartheta\}$, the cardinality $|S|$ of which represents the number of subthreshold contributions. The mean attention weight is denoted as m.

Figure 5 plots the $|S|$ against m for `Qwen2.5-VL-7B`, `DeepSeek-VL2`, and `WeThink`. Each data point represents the VLM response to Q on `COCO val 2017` (all categories included). In the first two cases, a moderately negative but statistically significant Pearson correlation was observed ($r = -0.09$ and $r = -0.73$, respectively). This consistent inverse relationship expresses that when the central value m is lower (that is, attention is more diffusely allocated across tokens), the subthreshold contributions are higher in number. However, the correlation coefficient $r = -0.02$ in `WeThink` is not significant, corroborating its non-improvement with RCA (Table 2).

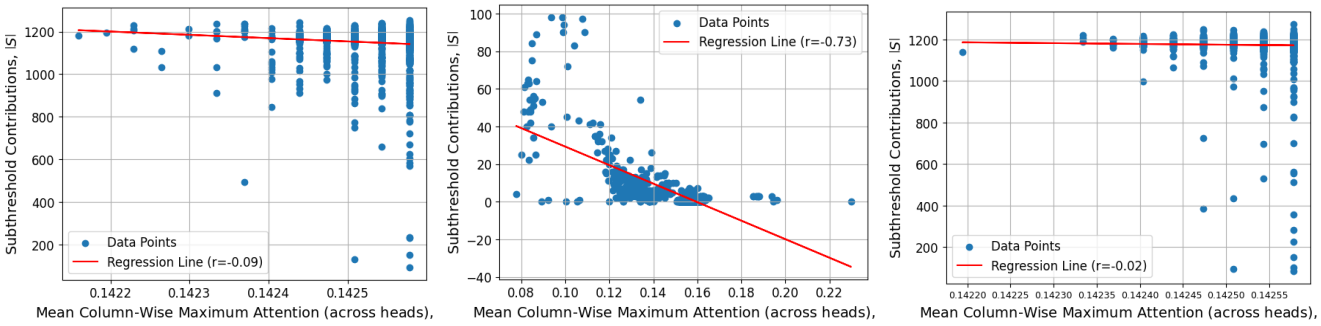
These observations empirically support the underlying assumption of Eq. 3 that the subthreshold factor found in condition (6) diminishes with increasing attention sharpness (supplementary section S6). Hence, by elevating the relevant token attention in these models, the RCA mechanism suppresses the contribution of low-activation (subthreshold) tokens in the final transformer layer. The negative correlation coefficient supports the hypothesis that RCA enhances model performance by reducing the influence of tokens receiving low attention, effectively reducing noise in the final hidden representations. This selective emphasis endows the model with more precise object localization and response generation.

These theoretical conditions align with the findings of Venhoff *et al*. [23], who showed that meaningful alignment of vision and language modalities only emerges in the middle to late layers of the transformer. Their use of SAEs revealed that visual tokens become semantically integrated within the LLM's internal representation only in later layers, which is precisely where RCA operates. This observation supports RCA's core premise: modifying attention in late layers is most effective, as the visual features have already been mapped into language-relevant embeddings. In particular, even early-fusion models such as `DeepSeek-VL2` can benefit from RCA, likely because their high-capacity architectures allow latent modality disentanglement to emerge in deeper layers. In such cases, RCA amplifies meaningful but under-attended visual cues while suppressing irrelevant or overly dominant ones, sharpening the model's focus during object localization.

3.3. Impact of Vision-Language Fusion Timing

Our empirical findings highlight a clear pattern: the *timing and structure of modality fusion* in transformer-based VLMs (Figure 6) critically determine whether RCA enhances or degrades OV-RefOD performance. The theoretical foundation of RCA (condition 6) requires the suppression of subthreshold visual tokens, which is only possible if visual and linguistic information remain sufficiently separable in the attention mechanism.

While most models that improved with RCA, such as `Qwen2.5-VL-7B`, `SAIL-VL-1.6-8B`, `MiniCPM-o-2.6`, `Gemma3-27B`, and `PaliGemma2-3B`, share a modular late-stage fusion architecture (Fig. 6), the case of `DeepSeek-VL2` presents a notable counterexample. Despite employing a tight early fusion strategy using Mixture-of-Experts across modalities, `DeepSeek-VL2` still benefited from RCA ($+18.0\%$ FitAP). This outcome suggests that tight early fusion is not inherently incompatible with RCA, as long as the model has compensating properties, such as large representational capacity, redundancy-aware design, or latent gating structures that allow attention distributions to evolve separately across layers. This observation is contrary to previous findings that advocate delayed or bottlenecked cross-modal fusion to maintain modality separability and improve alignment and interpretability [8, 17]. In particular, Hori *et al*. [8] demonstrate that selective application of attention between modalities during decoding improves output quality, while Liang *et al*. [17] propose delaying fusion until each modality is internally encoded to minimize distri-



(a) `Qwen2.5-VL-7B` ($\vartheta = -1.5$) (b) `DeepSeek-VL2` ($\vartheta = -13$) (c) `WeThink` ($\vartheta = -1.5$)

Figure 5. Correlations between the number $|S|$ of subthreshold contributions and mean cross-head max m of attention weights evaluated on 2064 (I, Q) pairs with Pearson correlations and p-value: (a) $r = -0.09$, $p = 0.00004$, (b) $r = -0.73$, $p < 0.00001$, (c) $r = -0.02$, $p = 0.32$. At the 0.05 confidence level, the correlation coefficient in (c) is not statistically significant.

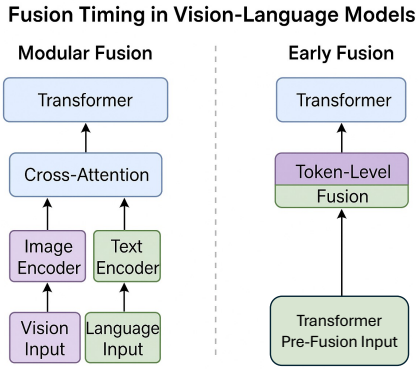


Figure 6. Modular vs. early fusion strategies in VLM

butional mismatch. `DeepSeek-VL2`'s RCA compatibility suggests that learned internal disentanglement, enabled by high-capacity attention pathways, may offer an alternative path to attention reweighting success, even in early fusion architectures.

In contrast, models that degraded post-RCA, such as `Valley-Eagle`, `WeThink`, and `POINTS1.5`, share characteristics that limit the effectiveness of RCA. These include early token-level fusion without structural constraints or training objectives, such as Chain-of-Thought supervision or instruction tuning, that emphasize abstract reasoning over spatial grounding. In the case of `Valley-Eagle`, for example, the Eagle module directly integrates image tokens into the transformer embedding space early in the architecture [27]. `WeThink-Qwen2.5-VL`, while built on a strong modular foundation, was subject to reinforcement learning and chain of thought supervision that deemphasized precise visual grounding [29]. `POINTS1.5`, similarly, adopts a modality fusion scheme designed for efficiency and general reasoning rather than for attention interpretability and separability. In these cases, the distribution of attention on the visual tokens remains too diffuse or semantically entangled, violating the assumption in condition (6) that subthreshold contributors can be isolated and suppressed (Fig. S9). These findings reinforce the need to consider the fusion strategy when designing VLMs for explainability or plug-in interpretability methods such as RCA. Models with sufficient capacity and flexible attention dynamics can still benefit from RCA even with an early stage fusion strategy, as demonstrated by `DeepSeek-VL2`.

While we focused on architectural factors to explain RCA's effectiveness, other elements—like dataset composition, prompt structure, and implementation choices—may also influence outcomes. These factors could interact with architecture in subtle ways, warranting future research to explore their interplay and uncover broader principles behind attention-based inference-time interventions.

4. Conclusion

This work proposed Reverse Contrast Attention (RCA), a novel method that reformulates attention maps at inference time to enhance object localization in vision-language transformers without altering model parameters. Through both theoretical analysis and empirical evaluation on the OV-RefOD task, we demonstrate that RCA selectively boosts the influence of semantically relevant but neglected tokens, thereby improving interpretability and performance. The introduction of FitAP allows meaningful benchmarking in the absence of explicit confidence scores, and our findings highlight the importance of architectural factors such as late-stage modality fusion to make RCA effective. Beyond performance gains, RCA offers a diagnostic lens into the internal workings of VLMs, showing that attention plasticity, when deliberately guided, can serve as a tool not only for eXCV but for functional enhancement. These insights lay the groundwork for future research into adaptive attention reweighting and post hoc interpretability methods across multimodal models.

References

[1] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *ICCV*, pages 397–406, 2021. 2

[2] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *CVPR*, pages 782–791, 2021. 2

[3] Keyan Chen, Xiaolong Jiang, Yao Hu, Xu Tang, Yan Gao, Jianqi Chen, and Weidi Xie. OvarNet: Towards open-vocabulary object attribute recognition. In *CVPR*, pages 23518–23527, 2023. 2

[4] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *ECCV*, pages 234–250, 2018. 2

[5] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatial-rgpt: Grounded spatial reasoning in vision-language models. *NeurIPS*, 37:135062–135093, 2025. 2

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) challenge. *IJCV*, 88:303–338, 2010. 4

[7] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 4

[8] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In *ICCV*, pages 4193–4202, 2017. 7

[9] Qin Huang, Chunyang Xia, Chihao Wu, Siyang Li, Ye Wang, Yuhang Song, and C-C Jay Kuo. Semantic segmentation with reverse attention. In *BMVC*, pages 18.1–18.13, 2017. 2

[10] Nam Hyeon-Woo, Kim Yu-Ji, Byeongho Heo, Dongyoon Han, Seong Joon Oh, and Tae-Hyun Oh. Scratching visual transformer's back with uniform attention. In *ICCV*, pages 5807–5818, 2023. 2

[11] Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, 2019. 1

[12] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR-modulated detection for end-to-end multi-modal understanding. In *ICCV*, pages 1780–1790, 2021. 4

[13] Go-Eun Lee, Jungchan Cho, and Sang-II Choi. Shallow and reverse attention network for colon polyp segmentation. *Scientific Reports*, 13(1):15243, 2023. 2

[14] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, pages 10965–10975, 2022. 4

[15] Yun Li, Zhe Liu, Saurav Jha, and Lina Yao. Distilled reverse attention network for open-world compositional zero-shot learning. In *ICCV*, pages 1782–1791, 2023. 2

[16] Zhikai Li, Murong Yi, Ali Uneri, Sihan Niu, and Craig Jones. Rta-former: Reverse transformer attention for polyp segmentation. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–5. IEEE, 2024. 2

[17] Tao Liang, Guosheng Lin, Lei Feng, Yan Zhang, and Fengmao Lv. Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion. In *ICCV*, pages 8148–8156, 2021. 7

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015. 4

[19] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, 2018. 2

[20] Haozhan Shen, Tiancheng Zhao, Mingwei Zhu, and Jianwei Yin. GroundVLP: Harnessing zero-shot visual grounding from vision-language pre-training and open-vocabulary object detection. In *AAAI*, pages 4766–4775, 2024. 2

[21] Poulami Sinhamahapatra, Shirsha Bose, Karsten Roscher, and Stephan Günnemann. Zero-shot open-vocabulary OOD object detection and grounding using vision language models. In *Northern Lights Deep Learning Conference*, pages 230–238. PMLR, 2025. 2

[22] Jee-Young Sun, Seung-Wook Kim, Sang-Won Lee, Ye-Won Kim, and Sung-Jea Ko. Reverse and boundary attention network for road segmentation. In *ICCV*, pages 0–0, 2019. 2

[23] Constantin Venhoff, Ashkan Khakzar, Sonia Joseph, Philip Torr, and Neel Nanda. How visual representations map to language feature space in multimodal llms. *arXiv preprint arXiv:2506.11976*, 2025. 2, 7

[24] Chunlei Wang, Wenquan Feng, Xiangtai Li, Guangliang Cheng, Shuchang Lyu, Binghao Liu, Lijiang Chen, and Qi Zhao. OV-VG: A benchmark for open-vocabulary visual grounding. *Neurocomputing*, 591:127738, 2024. 2

[25] Zhenyuan Wang, Xuemei Xie, Jianxiu Yang, and Xiaodan Song. Ra-net: reverse attention for generalizing residual learning. *Scientific Reports*, 14(1):12771, 2024. 2

[26] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 11–20. Association for Computational Linguistics, 2019. 1

[27] Ziheng Wu, Zhenghao Chen, Ruipu Luo, Can Zhang, Yuan Gao, Zhentao He, Xian Wang, Haoran Lin, and Minghui Qiu. Valley2: Exploring multimodal models with scalable vision-language design. *arXiv preprint arXiv:2501.05901*, 2025. 8

[28] Chaohao Xie, Shaohui Liu, Chao Li, Ming-Ming Cheng, Wangmeng Zuo, Xiao Liu, Shilei Wen, and Errui Ding. Image inpainting with learnable bidirectional attention maps. In *ICCV*, pages 8858–8867, 2019. 2

[29] Jie Yang, Feipeng Ma, Zitian Wang, Dacheng Yin, Kang Rong, Fengyun Rao, and Ruimao Zhang. WeThink: To-

ward general-purpose vision-language reasoning via reinforcement learning. *arXiv preprint arXiv:2506.07905*, 2025. 4, 8

[30] Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, et al. RLAIF-V: Open-source AI feedback leads to super GPT-4V trustworthiness. In *CVPR*, pages 19985–19995, 2025. 4

[31] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, pages 14393–14402, 2021. 4

[32] Yang Zhou, Shiyu Zhao, Yuxiao Chen, Zhenting Wang, and Dimitris N Metaxas. LED: LLM enhanced open-vocabulary object detection without human curated data generation. *arXiv preprint arXiv:2503.13794*, 2025. 2

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

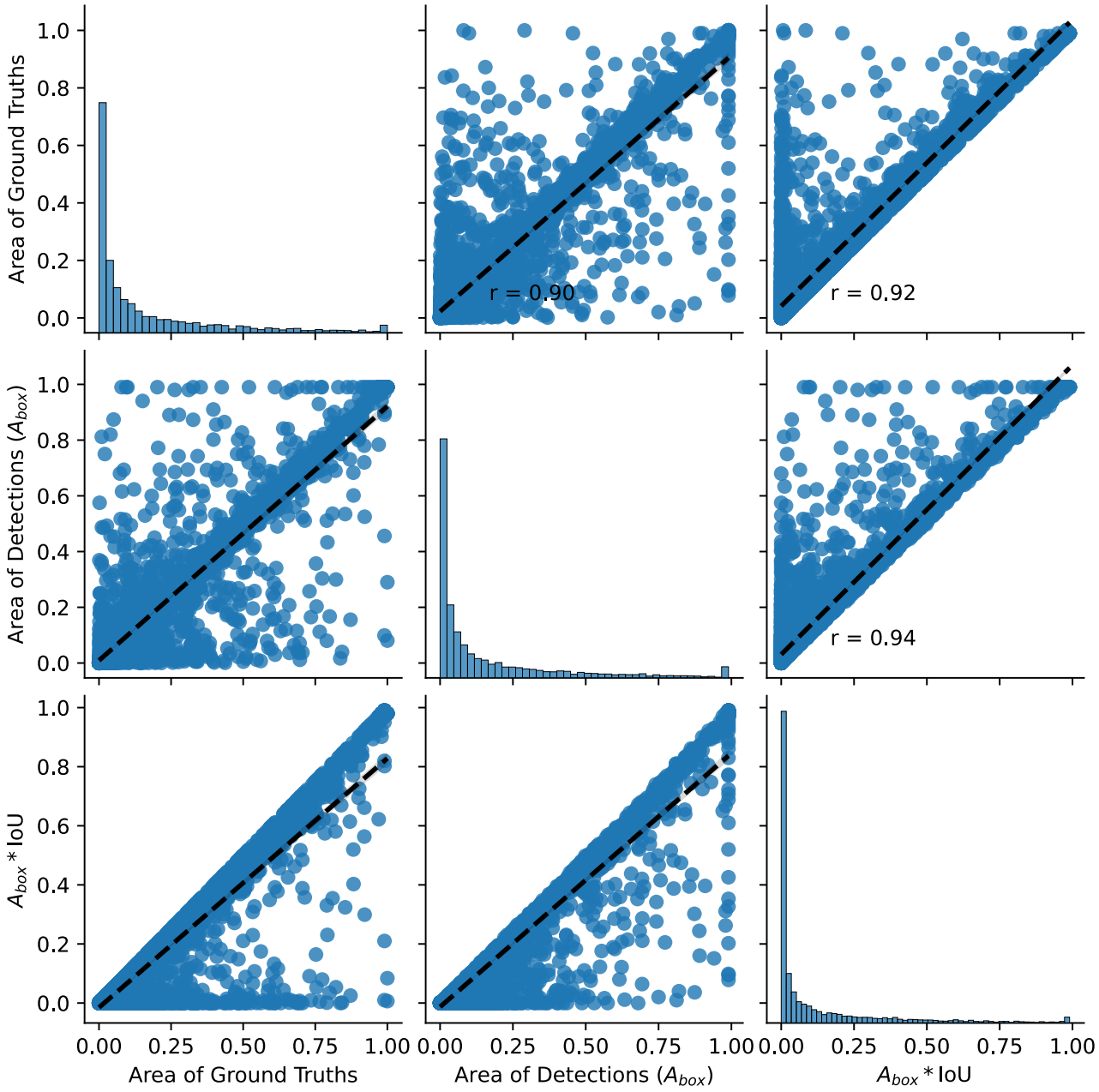
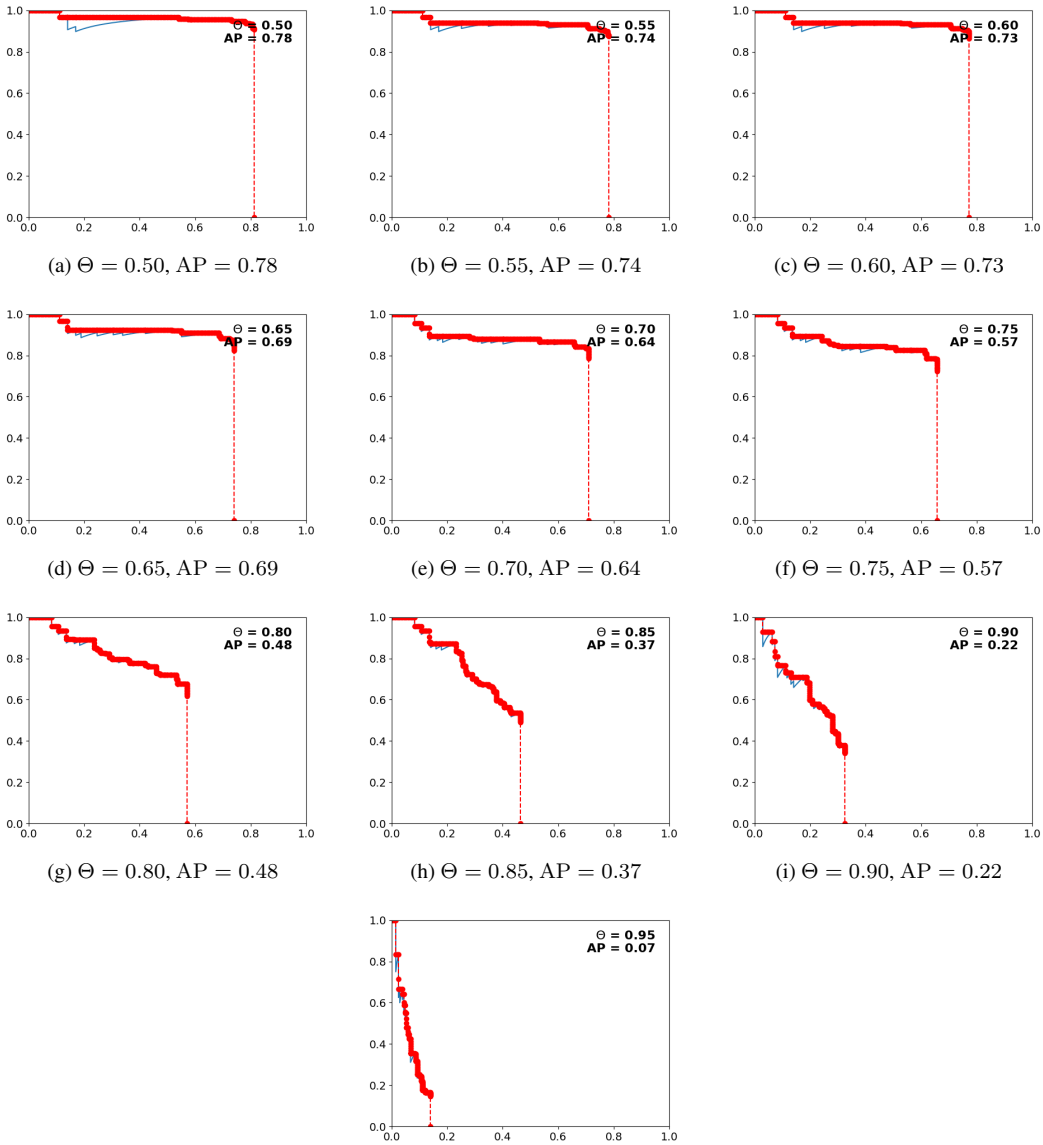


Figure S7. Correlations of ground-truth box area with A_{box} and $A_{\text{box}} \times \text{IoU}$ from the output of VLM queried with p_5. The plots include results from all object categories taken at IoU threshold, $\Theta = 0.50$. Diagonals are the univariate histograms of ground-truth area, A_{box}, 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 2.2. 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.

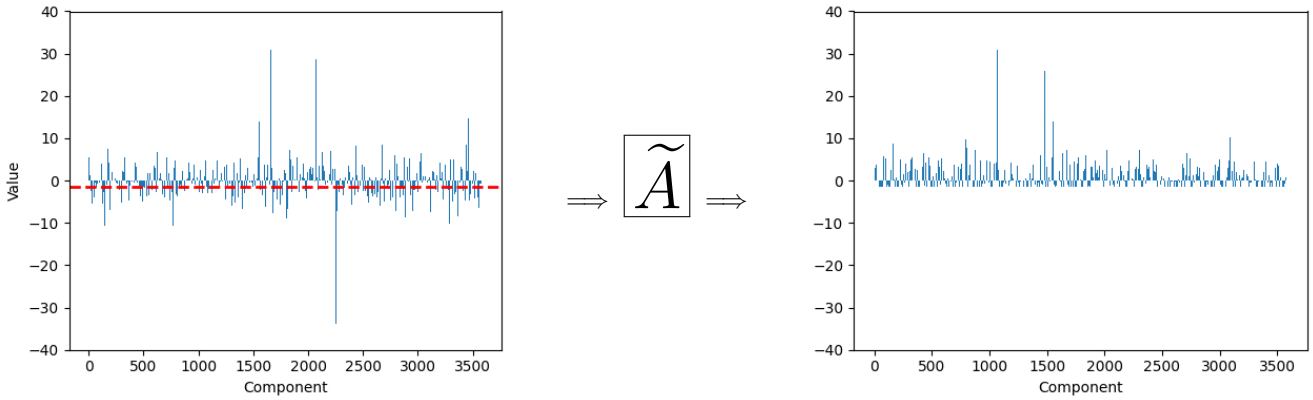
From the paper:

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

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:

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

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.

S7. Online Repository

The codes and data sets used by this study are accessible from https://github.com/earl-juanico/rca