Reasoning-based Spatial Prior (RSP): Learning Spatial Priors from Multimodal LLMs for Object Detection

Cagri Gungor* Meta cagrigungor@meta.com

Qingshuang Chen **Hongda Mao** **Chi Zhang** **Yelin Kim** Amazon {cicichen, hongdam, zhanbchi, kimyelin}@amazon.com

Abstract

Grounding natural language queries to target objects in images requires both high-level reasoning about intentions and precise spatial localization. Multimodal large language models (MLLMs) excel at reasoning about which object is desired given an implicit intention query but turning this understanding into precise bounding box predictions remains challenging. Conversely, specialized detectors such as GroundingDINO offer robust localization capabilities but lack the inherent capacity to understand implicit intentions. Existing hybrid methods couple these components only at the semantic level, such as by passing object names or a single global token to the detector, thereby discarding the rich, dense spatial signals encoded in MLLM attention maps. We propose **R**easoning-based **S**patial **P**rior (**RSP**), a reasoning-based detection framework that converts the MLLM's internal attention into a learnable spatial prior and injects it directly into the detector. Concretely, we learn input-adaptive gating over attention heads, supervise the fused spatial prior with a SoftIoU localization loss, and integrate it into GroundingDINO via a prior-guided deformable fusion module that biases sampling toward regions highlighted by the reasoning process. Our method consistently outperforms strong MLLM-only, detector-only, and hybrid baselines on the EgoIntention and RIO benchmarks, achieving state-of-the-art performance with substantial gains across both frequent and uncommon categories.

1 Introduction

Conventional object detection [4, 10, 12, 15, 29] has long relied on recognizing objects from a fixed set of explicit category names. However, recent work from robotic assistants [5, 6, 16] to egocentric systems [8, 25, 28, 31] has increasingly focused on identifying objects based on a user's underlying *need* or *intention*. For example, an assistant may need to locate not a "hammer" by name, but "something to secure a loose piece of wood," forcing the system to infer the target from context. These scenarios demand that the model first infer *what* type of object satisfies the intention (semantic reasoning) and then localize the correct instance among visually similar distractors (spatial localization).

Multimodal large language models (MLLMs) [2, 19, 33] excel at the former, inferring *which* object a user might need even when the target is not explicitly named, but their ability to pinpoint the correct instance remains weaker than specialized detectors. Conversely, detectors such as GroundingDINO [21] and YOLO-World [7] produce precise boxes when the query explicitly names the object or its attributes, but lack the reasoning needed to interpret implicit intentions. Hybrid approaches attempt to bridge this gap but remain shallow: sequential pipelines like DETGPT [26] pass only the predicted

*This work was done while working at Amazon.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

category to the detector, and methods like Lenna [35] pass a single global semantic token (e.g., <DET>), both discarding the dense spatial structure encoded in MLLM attention maps.

We propose a reasoning-based detection framework that directly exploits these spatial cues. Central to our approach is the **R**easoning-based **S**patial **P**rior (**RSP**), a learnable spatial prior derived from MLLM attention maps, which often exhibit structured patterns that correlate with relevant object regions. Prior work [17] has shown in zero-shot settings that certain attention heads concentrate on semantically meaningful parts of the image. Instead of relying on heuristic head selection or treating these maps as static signals, we extract all decoder attention maps over visual tokens from InternVL3.5-8B [33], learn adaptive per-head weights via a lightweight spatial-aware gating module, and refine the fused map with a differentiable SoftIoU loss that aligns the spatial prior with ground-truth object boxes. This supervision transforms diffuse attention patterns into task-aligned spatial signals. We then introduce a probability-level fusion mechanism that injects the refined prior into the sampling distribution of GroundingDINO's deformable attention, guiding the detector toward regions favored by the MLLM while still allowing it to rely on its own learned cues when the prior is uncertain.

We evaluate our approach on two intention-oriented grounding datasets, EgoIntention [31] and RIO [27], where our method consistently surpasses strong MLLM-only, detector-only, and hybrid baselines, establishing new state-of-the-art performance. Beyond intention grounding, we further evaluate on referring expression datasets (RefCOCO/+/g [18, 23]) and observe consistent gains, indicating that our framework generalizes beyond implicit intentions to explicit referring-expression grounding. We summarize our major contributions as follows:

- We introduce the **RSP** framework, which transforms MLLM attention maps into a task-aligned spatial prior via adaptive head weighting and SoftIoU supervision.
- We propose a probability-level fusion mechanism that incorporates this spatial prior into GroundingDINO's deformable attention, guiding the detector toward semantically relevant regions.
- Extensive experiments on EgoIntention, RIO, and the RefCOCO family show that RSP delivers strong performance across both implicit and explicit grounding tasks.

2 Related Work

2.1 Open-vocabulary and reasoning-based object detection.

Referring expression comprehension approaches localize objects described by explicit natural-language phrases by matching textual descriptions to visual regions [9, 11, 30, 32, 37]. Open-vocabulary detectors such as GLIP [20], OWL-ViT [24], GroundingDINO [21], YOLO-World [7], and LLMDet [12] extend this idea beyond fixed label sets by aligning visual features with large vision–language models, but typically assume that the query explicitly names the target category or its attributes, limiting their ability to handle intention-driven goals where the object must be inferred from context. This gap has motivated reasoning-oriented grounding benchmarks and datasets such as RIO [27], EgoIntention [31], SIGAR [13], and IntentionVG [34], which emphasize contextual ambiguity and multi-step inference. Our work adopts this reasoning-based perspective but differs by converting MLLM attention maps into a learnable spatial prior and injecting it into an open-vocabulary detector, moving beyond approaches that rely only on explicit text queries or global semantic tokens.

2.2 Attention maps in MLLMs.

Recent work has explored attention maps in large multimodal models as a source of visual grounding signals. Visualization tools such as VL-InterpreT [1] and LVLM-Interpret [3] show that cross-modal attention often highlights semantically meaningful regions without task-specific supervision, and several approaches exploit this property with frozen MLLMs: Prompt Highlighter [39] steers internal attention via prompt manipulation, [38] uses attention maps to surface small discriminative details in a training-free manner, F-LMM [36] refines them with a lightweight decoding module for grounding without updating the base model, and [17] selects inherently strong localization heads to produce pseudo-masks. Although these methods demonstrate that MLLM attention contains useful spatial

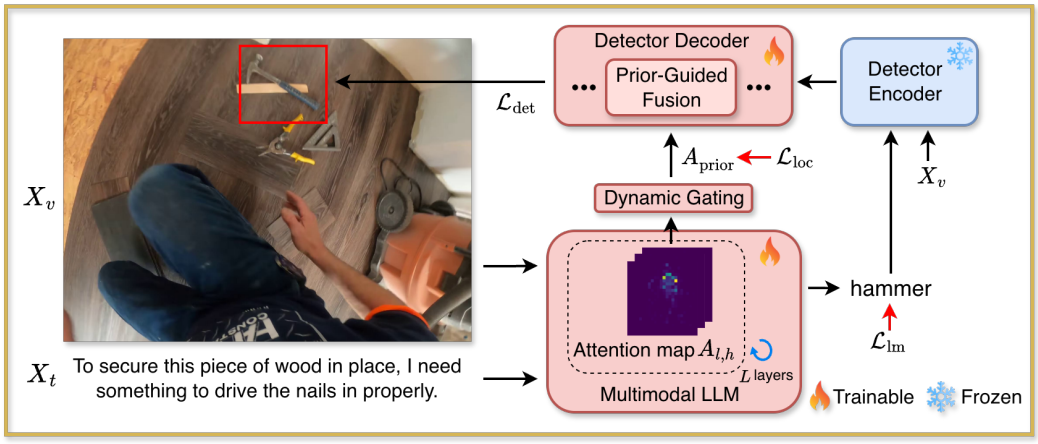


Figure 1: Overview of our **R**easoning-based **S**patial **P**rior (**RSP**) framework. Given an input image X_v and an implicit intention query X_t, the multimodal LLM predicts the object name *hammer* that is best aligned with intention and produces attention maps $A_{\ell,h}$. A dynamic gating module aggregates these maps into a spatial prior A_{prior}, which guides the GroundingDINO decoder through prior-guided deformable fusion.

structure, they treat attention maps as static cues. In contrast, RSP explicitly trains these maps with a SoftIoU localization loss and integrates the resulting prior into an open-vocabulary detector, enabling the attention signal to be sharpened, task-aligned, and directly exploited for precise localization.

3 Method

In this section, we present our RSP framework, which combines a multimodal large language model (MLLM) with a GroundingDINO-based detector. The MLLM jointly reasons over the intention query and image, and its multimodal decoder attention maps are converted into a learnable spatial prior using the gating and SoftIoU supervision described in Sec. 3.1. This prior, together with predicted object-name tokens, guides the detector through prior-guided deformable fusion in Sec. 3.2. The joint training objectives are described in Sec. 3.3.

3.1 Reasoning-Conditioned Spatial Prior from the MLLM

Given an input image X_v and an implicit intention query X_t, the multimodal LLM encodes the visual tokens $Z_v = \{v_i\}_{i=1}^{P^2}$, where $P \times P$ is the spatial layout of the visual token grid, and the textual tokens $Z_t = \{t_j\}_{j=1}^{T}$, where T denotes the length of the input text tokens, jointly through L transformer decoder layers, each equipped with H attention heads. During decoding, the model autoregressively generates a sequence that expresses its inferred object prediction. We use the final generated token y_{gen} as the query for attention extraction, as it reflects the context and semantics of the full generated token sequence and provides the most stable reference point for revealing the model's reasoning-dependent visual focus. This choice ensures that the resulting spatial attention map is conditioned on the model's causal reasoning process and reflects where the model attends while forming its prediction, rather than being tied to the final input text token. The anchor must be a generated token for intention queries, where the target object is not named in the input and the object name appears only in the model's output (e.g., "hammer" in Fig. 1): its attention over visual tokens shows where the model looks while predicting that object, whereas attention from the implicit input text is not tied to any specific object. Among the generated tokens, the final one is the only position conditioned on the full context: the image, the intention, and all previously generated tokens. For referring expressions, where the object is named in the input, input-token anchoring is also viable, and in qualitative analyses during development we observed closely similar attention patterns under both choices; we therefore use the final generated token as the single anchor valid in both settings, which keeps extraction uniform under our joint training over both query types.

3.1.1 Multimodal decoder attention.

Consider layer $\ell \in \{1, \ldots, L\}$ and head $h \in \{1, \ldots, H\}$. Let $q_{\text{gen}}^{(\ell,h)} \in \mathbb{R}^{d_h}$ denote the query projection of the final generated token y_{gen}, where d_h is the dimensionality of each attention head, and let $K^{(\ell,h)} \in \mathbb{R}^{(P^2+T)\times d_h}$ be the key matrix containing both visual and textual keys. Following the standard multi-head attention formulation:

$$\text{Attn}_{\ell,h}(q_{\text{gen}}) = \text{softmax}\left(\frac{q_{\text{gen}}^{(\ell,h)} {K^{(\ell,h)}}^\top}{\sqrt{d_h}}\right) \tag{1}$$

Since we are interested in spatial grounding, we take only the first P^2 entries of this attention vector, corresponding to visual tokens, and denote them as $A_{\ell,h} = \text{Attn}_{\ell,h}(q_{\text{gen}})[\,1 : P^2\,]$. Reshaping these values into a $P \times P$ map yields a reasoning-conditioned attention map.

3.1.2 Dynamic head weighting via learned gating.

Prior work has shown that large vision-language models naturally exhibit a small subset of attention heads that behave as effective *localizers* in zero-shot settings [17]. These localization heads are typically identified through per-head statistics, such as overall image-attention mass and spatial entropy, after which the top–K heads that most consistently focus on compact regions are selected. However, because this procedure is entirely heuristic and fixed at inference time, the selected heads cannot adjust to individual inputs and cannot benefit from learning signals during training.

Our goal is to obtain a spatial prior that adapts dynamically to each image–query pair. For this purpose, we treat all attention maps as candidates and learn their relative importance on a per-instance basis. We stack all maps into a tensor $A = [A_{1,1}; \ldots; A_{L,H}] \in \mathbb{R}^{(LH)\times P\times P}$, treating each map as an input channel, and process them in a single forward pass with a lightweight gating network F: a depthwise 3×3 convolution (BN, GELU) extracts a spatial-pattern descriptor for each map, a pointwise 1×1 convolution exchanges information across heads, and global average pooling followed by a two-layer MLP produces one logit per head, $s = F(A) \in \mathbb{R}^{LH}$. The gating weights are obtained by a softmax over this logit vector, $g_{\ell,h} = \exp(s_{\ell,h}) / \sum_{\ell',h'} \exp(s_{\ell',h'})$, and the final fused spatial prior is then defined as follows:

$$A_{\text{prior}} = \sum_{\ell,h} g_{\ell,h}\, A_{\ell,h} \tag{2}$$

Because the pointwise convolution and the shared MLP mix information across all maps, each gating weight is estimated with full cross-head context rather than from a single map in isolation. The model thus learns to emphasize heads whose spatial patterns align with the relevant region while suppressing diffuse or noisy activations.

3.1.3 SoftIoU loss for localization.

Although MLLMs exhibit strong zero-shot grounding behavior, their attention maps are not explicitly optimized for localization. As a result, their spatial attention patterns can be informative but are not explicitly optimized to match object extents. We therefore supervise the spatial prior with a differentiable localization loss.

Because A_{prior} is a continuous $P \times P$ heatmap while supervision is given as a bounding box, a standard IoU loss would require non-differentiable binarization. Instead, we introduce a ***SoftIoU*** formulation that treats the attention map as a continuous spatial density. Let $B \in \{0,1\}^{P\times P}$ denote the projected ground-truth mask. Concretely, B is obtained by mapping the ground-truth box from image coordinates onto the $P \times P$ visual-token grid, where each cell corresponds to one 28×28-pixel patch of the 448×448 input, and setting to 1 every cell the box overlaps, i.e., taking the floor and ceiling of the box extents in grid coordinates so that partially covered boundary cells are included. We apply no dilation or smoothing: the differentiability of the SoftIoU loss comes entirely from the continuous prediction side, with the binary target playing the same role as a ground-truth mask in standard segmentation losses. We define:

$$\text{SoftIoU}(A_{\text{prior}}, B) = \frac{\sum A_{\text{prior}} \odot B}{\sum A_{\text{prior}} + \sum B - \sum A_{\text{prior}} \odot B} \tag{3}$$

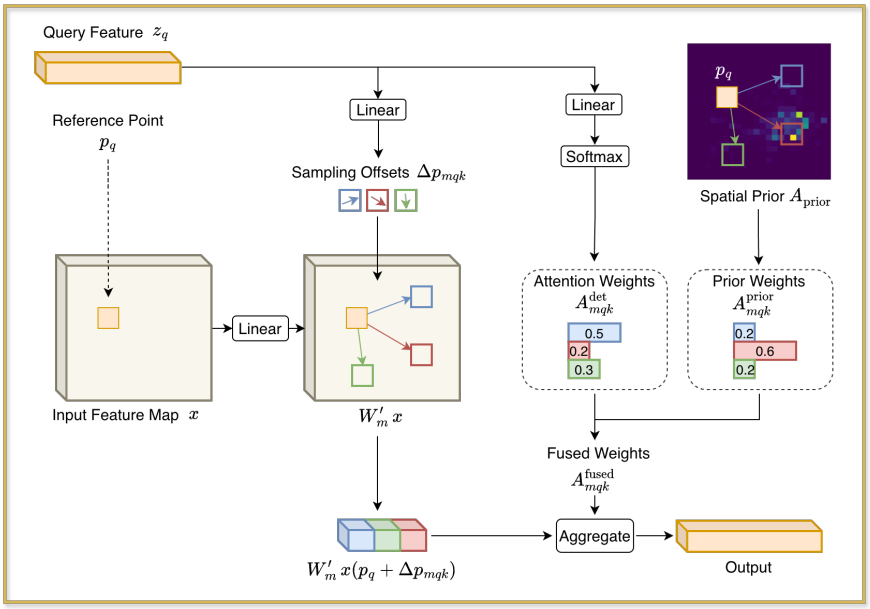


Figure 2: Prior guided deformable attention. For each query, deformable attention samples features around a reference point, while our MLLM supplies a spatial prior that is sampled at the same locations and fused with the detector weights to form A^{fused}_{mqk}.

Here, $\odot$ denotes elementwise multiplication and $\sum$ denotes summation over all spatial locations. The localization loss is then defined as

$$\mathcal{L}_{\text{loc}} = 1 - \text{SoftIoU}(A_{\text{prior}}, B) \tag{4}$$

which encourages the spatial prior to concentrate mass inside the ground-truth region while remaining fully differentiable with respect to the attention maps and all upstream MLLM parameters. This supervision improves the model's internal grounding behavior, transforming its zero-shot localization capability into a more reliable, task-adapted spatial prior within the RSP framework.

We arrived at this formulation empirically, comparing three localization objectives during development (Appendix C): direct box regression from the attention map failed to converge, an energy-ratio loss never improved over the zero-shot prior, and the SoftIoU formulation, which aligns the training objective with the inference-time IoU metric, trained stably. The box supervision it requires adds no annotation cost: boxes are the native annotation of all benchmarks we use, and every fine-tuned method we compare against trains on exactly the same labels; $\mathcal{L}_{\text{loc}}$ simply reuses them to additionally supervise the MLLM's internal attention.

3.2 Prior-Guided Deformable Fusion in GroundingDINO

The spatial prior $A_{\text{prior}} \in \mathbb{R}^{P\times P}$ is a purely spatial signal, whereas GroundingDINO [21] operates over high-dimensional multi-scale features. Injecting A_{prior} as a feature-level input would require a learned projection that dilutes the signal and adds extra parameters. Instead, we inject it into the image cross-attention block of the GroundingDINO decoder, which is implemented as multi-scale deformable attention in the style of Deformable DETR [40] and explicitly controls *where* the detector samples evidence for each object query, making it the most natural site to introduce a spatial prior.

Let $z_q \in \mathbb{R}^C$ denote the feature of query q, where C is the shared feature dimension of queries and image features. Following the deformable attention formulation of [40], and omitting multi-scale indices for clarity, the image cross-attention output for query q can be written as

$$\text{DeformAttn}(z_q, p_q, x) = \sum_{m=1}^{M} W_m \left[\sum_{k=1}^{K} A^{\text{det}}_{mqk}\, W'_m\, x\big(p_q + \Delta p_{mqk}\big) \right], \tag{5}$$

where M is the number of attention heads, $m \in \{1, \ldots, M\}$ indexes the heads, and $k \in \{1, \ldots, K\}$ indexes one of the K sampling points around the reference point p_q. The offsets $\Delta p_{mqk} \in \mathbb{R}^2$ are learned from z_q, and bilinear interpolation is used to evaluate the feature map x at fractional coordinates. The coefficients A^{det}_{mqk} are the standard deformable attention weights obtained from per-head logits by a softmax over k for each (m, q) pair.

We adopt GroundingDINO's reference points unchanged: language-guided query selection picks the top-K encoder proposals as initial reference points p_q, which are then refined layer by layer in the decoder, as in Deformable DETR [40]. RSP leaves both the reference points and the sampling offsets untouched and modifies only the attention distribution over the sampled locations, as follows. To inject the reasoning-conditioned spatial prior, we keep the sampling locations $p_q + \Delta p_{mqk}$ unchanged and modify only the distribution over them. We first resize A_{prior} to the resolution of the feature map x and sample it at the same locations:

$$r_{mqk} = A_{\text{prior}}\big(p_q + \Delta p_{mqk}\big). \tag{6}$$

Although A_{prior} is normalized over the full $P \times P$ grid, the subset of values $\{r_{mqk}\}_{k=1}^{K}$ for a given (m, q) is not guaranteed to sum to one. We therefore renormalize them to obtain a valid prior distribution over the sampled points:

$$A^{\text{prior}}_{mqk} = \frac{r_{mqk}}{\sum_{k'=1}^{K} r_{mqk'} + \varepsilon}, \tag{7}$$

with a small constant $\varepsilon > 0$ for numerical stability.

Finally, we fuse the detector's native deformable attention distribution A^{det}_{mqk} with the MLLM-induced prior distribution A^{prior}_{mqk}. For each head m we maintain a learnable scalar parameter γ_m and define $\alpha_m = \sigma(\gamma_m)$ using a sigmoid so that $\alpha_m \in (0, 1)$. The fused attention weights are given by

$$A^{\text{fused}}_{mqk} = (1 - \alpha_m)\, A^{\text{det}}_{mqk} + \alpha_m\, A^{\text{prior}}_{mqk}, \tag{8}$$

which corresponds to probability-level mixing between the normalized deformable attention weights and the normalized samples from A_{prior}. We obtain our prior-guided deformable attention module by replacing A^{det}_{mqk} with A^{fused}_{mqk} in Eq. (5). Intuitively, A^{det} captures bottom-up detector cues while A^{prior} injects top-down MLLM reasoning, so the fused weights bias sampling toward regions supported by both signals while preserving the detector's default behavior when the prior is uncertain.

A natural alternative to the convex combination of Sec. 3.1 would be to feed all attention maps to a network that predicts the spatial prior directly. We deliberately keep the prior a convex combination of the maps: it cannot place mass on regions the MLLM never attends to, so it remains a faithful summary of the model's reasoning rather than an independently learned saliency head; and the system as a whole is not confined to the subspace spanned by the maps, since Eq. (8) mixes the prior with the detector's native attention under a learned per-head weight.

3.3 Training Objectives

We jointly optimize three complementary objectives. (i) A standard autoregressive language modeling loss $\mathcal{L}_{\text{lm}}$ on the ground-truth object-label tokens $\mathbf{y}^* = \{y^*_1, \ldots, y^*_{T_{\text{obj}}}\}$ conditioned on X_v and X_t, which encourages the MLLM to generate object-name tokens that are semantically consistent with the intention and visual context. (ii) The SoftIoU localization loss $\mathcal{L}_{\text{loc}}$ defined in Eq. (4), which encourages the spatial prior to place most of its mass inside the ground-truth region. (iii) The standard GroundingDINO detection loss $\mathcal{L}_{\text{det}} = \lambda_{L_1} L_1(\textit{pred}, \textit{gt}) + \lambda_{\text{GIoU}} \,\text{GIoU}(\textit{pred}, \textit{gt}) + \lambda_{\text{con}} \,\text{Contrast}(\textit{pred}, \textit{gt})$, which combines bounding-box regression with a contrastive term that aligns text phrases with their corresponding object queries. We optimize the weighted sum

$$\mathcal{L} = \lambda_{\text{lm}}\, \mathcal{L}_{\text{lm}} + \lambda_{\text{loc}}\, \mathcal{L}_{\text{loc}} + \lambda_{\text{det}}\, \mathcal{L}_{\text{det}}. \tag{9}$$

Gradients from $\mathcal{L}_{\text{lm}}$ and $\mathcal{L}_{\text{loc}}$ update the MLLM and the gating module that produces A_{prior}, while $\mathcal{L}_{\text{det}}$ trains the detector and the learnable fusion weights α_m. Because the deformable fusion operates directly on probability distributions derived from the MLLM attention maps, the two components are trained in a coordinated way, with the MLLM learning to provide sharper priors and the detector learning to exploit them.

4 Experiments

In this section, we first describe our experimental setup, including the network architecture, implementation details, datasets, and evaluation metrics (Sec. 4.1). We then present quantitative comparisons with competitive baselines on EgoIntention, RIO, and RefCOCO (Sec. 4.2). Next, we analyze the contribution of each component of the spatial prior through ablation studies (Sec. 4.3), and finally we provide qualitative visualizations of the learned priors and detections (Sec. 4.4).

4.1 Experimental Setup

4.1.1 Network architecture.

We instantiate RSP with InternVL3.5-8B [33] as the MLLM and GroundingDINO with a Swin-base backbone [21] as the detector, coupled through the prior-guided deformable fusion of Sec. 3.2. The MLLM is adapted via LoRA [14] adapters in its attention and feed-forward layers; the detector's backbone and encoder are frozen, and we fine-tune only the decoder layers together with the fusion parameters $\{\alpha_m\}$.

4.1.2 Implementation details.

We train the model end to end using AdamW [22] with weight decay 0.0 and $\beta_2{=}0.999$. The learning rate is scheduled with linear warmup followed by cosine decay, starting from $1\mathrm{e}{-4}$ and decaying to $1\mathrm{e}{-6}$, with 200 warmup iterations. We set the effective global batch size to 64, achieved through data-parallel training on $4\times$ NVIDIA L40S (48 GB) GPUs combined with gradient accumulation. For the MLLM, we freeze the base weights and train only the LoRA [14] adapters, using rank $r{=}8$, scaling factor $\alpha_{\mathrm{LoRA}}{=}16$, and LoRA dropout 0.05 in the projection layers of attention as well as the feed forward blocks. On the detector side, optimization is restricted to the GroundingDINO decoder layers and the fusion parameters $\{\alpha_m\}$. The loss weights are set to $\lambda_{\mathrm{lm}}{=}0.5$, $\lambda_{\mathrm{loc}}{=}1.0$, and $\lambda_{\mathrm{det}}{=}1.0$. InternVL3.5-8B takes 448×448 input images and produces a 16×16 grid of 256 visual tokens, so the spatial prior of Sec. 3.1 has $P{=}16$.

4.1.3 Datasets.

We evaluate on three benchmarks. **EgoIntention** [31] and **RIO** [27] are intention-driven grounding benchmarks (egocentric and third-person, respectively), where each sample pairs an image with an intention sentence and one or more target-object boxes; both provide *content*/*common* and *uncommon* splits covering frequent and rare objects. **RefCOCO / RefCOCO+ / RefCOCOg** [18, 23] are standard referring expression benchmarks where queries are explicit descriptive phrases rather than intentions. We use the official training/validation/test splits of all datasets. We collect no new data: a single model is trained jointly on the union of the official training splits of EgoIntention, RIO, and RefCOCO/+/g, and evaluated on each benchmark's official evaluation split. All fine-tuned baselines in Tables 1 and 2 are trained under the identical setup with the same backbones, so the comparisons isolate our method's contribution.

4.1.4 Evaluation metrics.

On EgoIntention and RIO, we follow prior work and report Precision@0.5: a prediction is deemed correct if its highest IoU with the annotated boxes for the target object exceeds 0.5, and results are reported separately for common and uncommon categories. On the RefCOCO benchmarks, we follow the standard protocol and report top-1 detection accuracy with IoU > 0.5 on the validation and test splits.

4.1.5 Baselines.

All baselines share the same backbones (InternVL3.5-8B [33] and GroundingDINO with Swin-base [21]) to isolate the effect of localization supervision and fusion. **R-D** is a hybrid reasoner-detector baseline in two settings: zero-shot (ZS, both frozen) and fine-tuned (FT, trained with $\mathcal{L}_{\mathrm{lm}}$ and $\mathcal{L}_{\mathrm{det}}$ but no prior fusion or localization loss). **RoG** [31] fine-tunes the MLLM alone with reasoning and referring objectives, without a detector branch. **Lenna** [35] is a hybrid method that

Method	EgoIntention		RIO	
	content	uncommon	common	uncommon
InternVL (ZS)	24.6	20.2	37.7	20.3
R-D (ZS)	44.3	29.7	51.3	28.5
RoG [31]	58.3	52.9	66.3	43.2
R-D (FT)	61.3	56.5	67.6	45.8
Lenna [35]	62.4	57.9	69.1	47.2
RSP (ours)	**65.3**	**61.9**	**72.3**	**53.2**

Table 1: Comparison with baselines on EgoIntention and RIO datasets. We report Precision@0.5 for content and uncommon categories on EgoIntention, and common/uncommon splits on RIO. Our method achieves the best results across all settings.

Method	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val	test
GroundingDINO	89.2	91.9	85.0	81.1	87.4	74.8	84.2	85.0
Lenna [35]	89.7	92.4	86.4	86.1	89.8	83.7	**90.4**	90.4
RSP (ours)	**91.3**	**93.6**	**89.3**	**87.9**	**91.7**	**83.4**	90.2	**90.6**

Table 2: Results on RefCOCO, RefCOCO+, and RefCOCOg benchmarks. We report top-1 detection accuracy (IoU > 0.5) on the validation and test splits.

injects an additional global semantic token into the detector. We reimplement RoG and Lenna with InternVL3.5-8B for fair comparison, as no official implementations are available.

4.2 Comparison with Baselines

4.2.1 Intention grounding on EgoIntention and RIO.

Table 1 shows a clear progression across MLLM-only, detector-only, and hybrid models. In the zero-shot setting, coupling InternVL3.5-8B with GroundingDINO (R-D ZS) yields **36–80%** relative Precision@0.5 gains over the MLLM-only baseline (InternVL ZS), indicating that once the object name is known, a detector trained with box-level losses localizes targets more effectively. Fine-tuning further boosts performance: RoG [31] achieves **76–162%** relative gains over InternVL (ZS) through language-modeling supervision alone, but remains **2–7%** behind R-D (FT) since it lacks an explicit detection branch. Lenna [35] adds **2–3%** over R-D (FT) by passing a global <DET> token from the MLLM to the detector, yet this token transfers only coarse semantic cues. In contrast, our RSP injects a dense attention-based spatial prior into deformable attention, sharing both fine-grained spatial structure and semantics with the detector, and surpasses Lenna by **5–6%** on EgoIntention and **5–12%** on RIO. To make the source of these gains explicit, we note both decompositions of the improvement over the zero-shot cascade R-D (ZS). Fine-tuning accounts for the larger share, as expected, since it adapts both models to the task; we therefore treat R-D (FT) as the common starting point. On top of this identical fine-tuned floor, Lenna improves over R-D (FT) by only $+1.1/+1.4/+1.5/+1.4$ P@0.5 across the four splits, whereas RSP improves by $+4.0/+5.4/+4.7/+7.4$, roughly a fourfold difference attributable to the transfer mechanism rather than to fine-tuning itself.

4.2.2 Referring expressions on RefCOCO.

Table 2 reports results on RefCOCO, RefCOCO+, and RefCOCOg, where queries are explicit referring expressions rather than implicit intentions. In this setting, object names and attributes are directly specified, and the MLLM is used only to provide a spatial prior. Even so, transferring information from the MLLM to the detector remains beneficial. Lenna improves over the strong GroundingDINO baseline by **1–12%** relative across the eight splits by injecting a semantic token into the detector. Our RSP framework instead transfers an attention-based spatial prior and achieves additional gains over Lenna on most splits, with **1–3%** relative improvements. These results show

Exp	*Fusion*	*Gating*	$\mathcal{L}_{\text{loc}}$	EgoIntention	RIO
E1				61.3	67.6
E2	✓			62.3	68.2
E3	✓	✓		62.5	68.8
E4	✓		✓	64.4	70.6
E5	✓	✓	✓	**65.3**	**72.3**

Table 3: Ablation on the components of our spatial prior. ***Fusion*** uses MLLM attention maps to form a spatial prior and injects it into deformable attention; ***Gating*** denotes learned, dynamic head weighting (otherwise top–3 heads are uniformly averaged following [17]); and $\mathcal{L}_{\text{loc}}$ is the localization loss applied to this prior. EgoIntention and RIO columns report P@0.5 on the content and common splits, respectively.

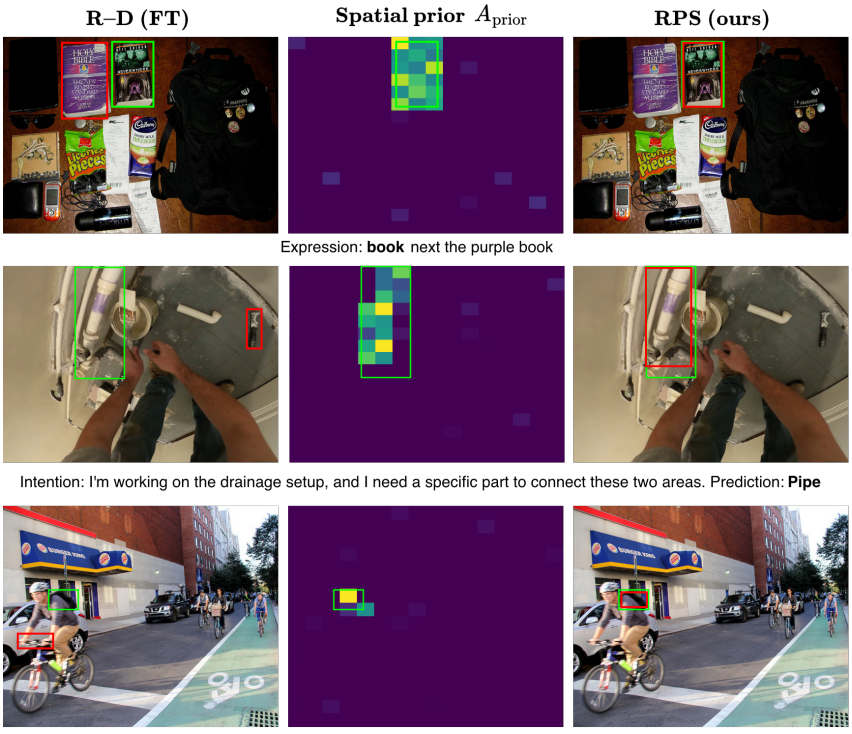


Figure 3: Qualitative examples of our reasoning-based prior. For each triplet, the left image shows the **R–D (FT)** prediction without fusion, the middle image shows the MLLM spatial prior A_{prior} (upsampled to image resolution to show its spatial correlation with the target), and the right image shows the prediction of our **RSP** model after prior-guided fusion. Red boxes denote predictions and green boxes denote GT boxes.

that spatial priors from MLLM attention generalize beyond intention-driven grounding to standard referring expression benchmarks.

4.3 Ablation on Spatial Prior Components

Table 3 analyzes five experiments (E1–E5), each progressively adding components of our spatial prior to measure their impact. E1 corresponds to the fine-tuned R–D (FT) baseline, where InternVL3.5-8B is optimized with the language modeling loss $\mathcal{L}_{\text{lm}}$ and GroundingDINO with the detection loss $\mathcal{L}_{\text{det}}$, without using MLLM attention maps. In E2 we enable ***Fusion***, injecting attention maps into deformable attention using the heuristic top–3 head selection and uniform averaging strategy of [17]. E3 adds ***Gating***, allowing the model to learn input-dependent head weights, which consistently improves over E2 and shows the benefit of end-to-end head weighting. In E4 we introduce the

localization loss $\mathcal{L}_{\text{loc}}$ to supervise the spatial prior. Even without gating, this improves over E1 by around **4.5–5%** relative on both EgoIntention and RIO, confirming that $\mathcal{L}_{\text{loc}}$ enhances spatial precision. Finally, combining all components in E5 yields the full RSP model, achieving the best results and an overall **6.5–7%** relative improvement over E1 across both benchmarks.

4.4 Qualitative Analysis of Reasoning-based Priors

Figure 3 illustrates how the MLLM prior guides the detector on both referring-expression and intention-driven cases. In all three examples, the left image shows that the R–D (FT) baseline drifts to a visually similar distractor such as another book or a wrong tool. In contrast, the middle image visualizes the MLLM prior A_{prior}, which focuses on the correct region (the book next to the purple one, the pipe and the backpack implied by the intention). When this prior is injected into the decoder via our prior-guided deformable fusion (Sec. 3.2) within the RSP framework, the deformable sampling weights are biased toward locations favored by the MLLM, which suppresses nearby distractors and pulls the final box onto the intended object. The right images show that this coupling reliably shifts the prediction onto the intended object, indicating that the MLLM learns a strong reasoning-based spatial signal and that our fusion module effectively turns this signal into better localization.

We close with a note on multi-object grounding. Existing intention-grounding benchmarks, including EgoIntention and RIO, effectively evaluate a single target per query, and no established benchmark quantifies multi-target intention grounding, so we could not measure this setting. Architecturally, RSP supports it: GroundingDINO decodes a full set of scored object queries, and the spatial prior is a distribution over the entire token grid that can be multimodal; in qualitative analyses during development we observed the learned prior placing mass on multiple regions when an intention is satisfiable by several instances (e.g., multiple cups for "something to drink from"). We see multi-target intention grounding as a valuable direction for future benchmarks.

5 Conclusion

In this work, we introduced the Reasoning-based Spatial Prior (RSP) framework, a novel approach that bridges the gap between high-level MLLM reasoning and precise detector localization. Unlike existing hybrid methods that rely on sparse semantic tokens, we demonstrate that the MLLM's internal cross-modal attention maps contain rich, dense spatial signals that can be effectively harvested. By refining these signals into a learnable prior via dynamic gating and SoftIoU supervision, and injecting them directly into the detector through our prior-guided deformable fusion, we enable the detector to look exactly where the MLLM reasons. Extensive experiments on EgoIntention, RIO, and RefCOCO benchmarks confirm that this deep spatial coupling yields state-of-the-art performance, with particularly substantial gains on challenging uncommon categories.

6 Limitations

Our framework is instantiated with a specific MLLM-detector pair (InternVL3.5-8B and GroundingDINO). The learnable spatial-prior stage transfers across MLLMs (Appendix A) and the fusion interface assumes only multi-scale deformable attention; nevertheless, the full framework has not been retrained with other pairs, so pair-level generality remains to be demonstrated end to end. The current instantiation is not optimized for real-time deployment, as the 8B-parameter MLLM dominates the compute cost, and lighter backbones remain future work. Finally, current intention-grounding benchmarks assess a single target per query, so multi-target intention grounding (Sec. 4.4) remains unquantified.

Acknowledgements and Disclosures

This work was done by Cagri Gungor while working at Amazon as an Applied Scientist Intern. No external funding supported this work, and the authors declare no competing interests.

References

[1] Aflalo, E., Du, M., Tseng, S.Y., Liu, Y., Wu, C., Duan, N., Lal, V.: VL-InterpreT: An interactive visualization tool for interpreting vision-language transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21406–21415 (2022)

[2] Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)

[3] Ben Melech Stan, G., Aflalo, E., Rohekar, R.Y., Bhiwandiwalla, A., Tseng, S.Y., Olson, M.L., Gurwicz, Y., Wu, C., Duan, N., Lal, V.: LVLM-Intrepret: An interpretability tool for large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops). pp. 8182–8187 (2024)

[4] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science, vol. 12346, pp. 213–229. Springer (2020)

[5] Chen, Y., Gu, K., Wen, Y., Zhao, Y., Wang, T., Nie, L.: IntentionVLA: Generalizable and efficient embodied intention reasoning for human-robot interaction. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). pp. 13494–13501 (2026)

[6] Chen, Z., Luo, X., Li, D.: VisRL: Intention-driven visual perception via reinforced reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2545–2555 (2025)

[7] Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: YOLO-World: Real-time open-vocabulary object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16901–16911 (2024)

[8] Chu, Q., Zhang, H., Liu, M., Feng, Y., Shi, H., Nie, L.: Intention-guided cognitive reasoning for egocentric long-term action anticipation. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 40, pp. 17436–17444 (2026)

[9] Dang, R., Feng, J., Zhang, H., Ge, C., Song, L., Gong, L., Liu, C., Chen, Q., Zhu, F., Zhao, R., Song, Y.: InstructDET: Diversifying referring object detection with generalized instructions. In: International Conference on Learning Representations (ICLR) (2024)

[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)

[11] FitzGerald, N., Artzi, Y., Zettlemoyer, L.: Learning distributions over logical forms for referring expression generation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1914–1925 (2013)

[12] Fu, S., Yang, Q., Mo, Q., Yan, J., Wei, X., Meng, J., Xie, X., Zheng, W.S.: LLMDet: Learning strong open-vocabulary object detectors under the supervision of large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14987–14997 (2025)

[13] Guo, H., Zhu, J., Fan, W., Yi, C., Jiang, F.: Beyond object categories: Multi-attribute reference understanding for visual grounding. arXiv preprint arXiv:2503.19240 (2025)

[14] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)

[15] Hussain, M.: YOLOv1 to v8: Unveiling each variant–a comprehensive review of YOLO. IEEE Access **12**, 42816–42833 (2024)

[16] Jin, S., Xu, J., Lei, Y., Zhang, L.: Reasoning grasping via multimodal large language model. In: Proceedings of the Conference on Robot Learning (CoRL). Proceedings of Machine Learning Research, vol. 270, pp. 3809–3827. PMLR (2025)

[17] Kang, S., Kim, J., Kim, J., Hwang, S.J.: Your large vision-language model only needs a few attention heads for visual grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9339–9350 (2025)

[18] Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: ReferItGame: Referring to objects in photographs of natural scenes. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 787–798 (2014)

[19] Kimi Team, Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., et al.: Kimi-VL technical report. arXiv preprint arXiv:2504.07491 (2025)

[20] Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10965–10975 (2022)

[21] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science, vol. 15105, pp. 38–55. Springer (2024)

[22] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)

[23] Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11–20 (2016)

[24] Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., Wang, X., Zhai, X., Kipf, T., Houlsby, N.: Simple open-vocabulary object detection with vision transformers. In: European Conference on Computer Vision (ECCV). pp. 728–755. Springer (2022)

[25] Pei, B., Huang, Y., Xu, J., He, Y., Chen, G., Wu, F., Qiao, Y., Pang, J.: EgoThinker: Unveiling egocentric reasoning with spatio-temporal CoT. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 38, pp. 49361–49389 (2025)

[26] Pi, R., Gao, J., Diao, S., Pan, R., Dong, H., Zhang, J., Yao, L., Han, J., Xu, H., Kong, L., Zhang, T.: DetGPT: Detect what you need via reasoning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 14172–14189 (2023)

[27] Qu, M., Wu, Y., Liu, W., Liang, X., Song, J., Zhao, Y., Wei, Y.: RIO: A benchmark for reasoning intention-oriented objects in open environments. Advances in Neural Information Processing Systems (NeurIPS) **36**, 43041–43056 (2023)

[28] Ravi, S., Sarch, G.H., Vineet, V., Wilson, A.D., Kumaravel, B.T.: Out of sight, not out of context? egocentric spatial reasoning in VLMs across disjoint frames. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 16135–16150 (2025)

[29] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems (NeurIPS) **28**, 91–99 (2015)

[30] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: VL-BERT: Pre-training of generic visual-linguistic representations. In: International Conference on Learning Representations (ICLR) (2020)

[31] Sun, P., Xiao, J., Tse, T.H.E., Li, Y., Akula, A., Yao, A.: Visual intention grounding for egocentric assistants. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2512–2522 (2025)

[32] Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Chen, K., Xu, B., Li, J., Dong, Y., Ding, M., Tang, J.: CogVLM: Visual expert for pretrained language models. Advances in Neural Information Processing Systems (NeurIPS) **37**, 121475–121499 (2024)

[33] Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)

[34] Wang, W., Zhang, Y., He, X., Yan, Y., Zhao, Z., Wang, X., Liu, J.: Beyond literal descriptions: Understanding and locating open-world objects aligned with human intentions. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 762–776 (2024)

[35] Wei, F., Zhang, X., Zhang, A., Zhang, B., Chu, X.: Lenna: Language enhanced reasoning detection assistant. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)

[36] Wu, S., Jin, S., Zhang, W., Xu, L., Liu, W., Li, W., Loy, C.C.: F-LMM: Grounding frozen large multimodal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24710–24721 (2025)

[37] Yang, Z., Gan, Z., Wang, J., Hu, X., Ahmed, F., Liu, Z., Lu, Y., Wang, L.: UniTAB: Unifying text and box outputs for grounded vision-language modeling. In: European Conference on Computer Vision (ECCV). pp. 521–539. Springer (2022)

[38] Zhang, J., Khayatkhoei, M., Chhikara, P., Ilievski, F.: MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In: International Conference on Learning Representations (ICLR) (2025)

[39] Zhang, Y., Qian, S., Peng, B., Liu, S., Jia, J.: Prompt highlighter: Interactive control for multi-modal LLMs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13215–13224 (2024)

[40] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (ICLR) (2021)

Appendix

A Generalization of the Learnable Prior Across MLLMs

During method development, we validated the learnable spatial-prior stage across three MLLMs from two model families on RefCOCO. These experiments used an earlier variant of the localization objective, input-text-token anchoring, and boxes derived directly from the trained attention maps, without the detector. Table 4 reports the attention-stage Precision@0.5 in two settings: *zero-shot*, where attention maps are read directly from the frozen MLLM and aggregated with the strongest available heuristic (uniform averaging of the top-3 localization heads, following [17]), and *trained prior*, where the maps are adapted with localization supervision as in RSP. Training the prior yields large, consistent gains for all three MLLMs (16–22 points), so the mechanism is not specific to InternVL3.5-8B; we selected the strongest model for the full framework. The same comparison answers the frozen-maps question directly: removing adaptation degrades the prior severely and consistently, confirming that localization-supervised adaptation of the MLLM is necessary for its attention maps to function as an effective spatial prior. These are preliminary development-stage results, not full-RSP numbers, and are not comparable to Table 2.

MLLM	Zero-shot P@0.5	Trained prior P@0.5
LLaVA-1.5-7B	26.1	44.7
InternVL2.5-8B	39.4	55.5
InternVL3.5-8B	33.9	56.1

Table 4: Attention-stage results on RefCOCO for three MLLMs, without the detector. Zero-shot uses frozen attention maps with top-3 head averaging [17]; trained prior applies localization supervision to the maps. Development-stage results with an earlier objective variant; not comparable to Table 2.

B Sensitivity to the Visual Token Grid

The spatial prior is defined on the MLLM's visual token grid, which for InternVL3.5-8B at 448×448 input is 16×16 (P=16, 256 tokens). Development experiments give a direct sensitivity signal: raising the effective resolution to a 32×32 grid via a 2×2 tiling of the input further improved attention-stage results on RefCOCO ($56.1 \rightarrow 56.6$ P@0.5) and qualitatively helped small and dense objects, so results are moderately sensitive to grid resolution, with headroom at higher resolutions. In the full framework this sensitivity is reduced by design: the prior only biases *where* deformable attention samples, while precise box geometry is carried by the detector's multi-scale features, so the prior needs to be directionally correct rather than pixel-precise. We nevertheless kept the default single-view configuration in the final model as a deliberate cost-benefit choice: the 2×2 tiling quadruples the number of visual tokens ($256 \rightarrow 1024$), which substantially increases MLLM training and inference cost for a modest attention-stage gain ($+0.5$ P@0.5), and the default configuration relies on the MLLM's standard preprocessing, keeping the framework simpler and easier to reproduce.

C Choice of the Localization Objective

During development we compared three localization objectives for supervising the attention maps on LLaVA-1.5/RefCOCO. (i) *Direct box regression*: a lightweight CNN+MLP head regressing box coordinates from the attention map, trained with GIoU and L1 losses, failed to converge across learning rates and loss weightings, with and without LoRA adaptation. (ii) *Energy-ratio loss*: maximizing the fraction of attention mass inside the ground-truth box never improved over the zero-shot prior; it optimizes an energy fraction rather than spatial overlap. (iii) *SoftIoU* (Sec. 3.1): aligns the training objective with the inference-time IoU metric and trained stably. We adopted (iii). The same development program informed the design of the prior itself: a free-form head predicting localization directly from the attention maps behaved like variant (i), whereas learning only the reweighting of the maps (our convex combination with learned gating) trained stably, which is why RSP builds on the head-selection-and-averaging mechanism of [17] and makes it learnable and supervised.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state three main contributions: (i) the RSP framework that transforms MLLM attention maps into a task-aligned spatial prior via adaptive gating and SoftIoU supervision, (ii) a probability-level fusion mechanism for GroundingDINO's deformable attention, and (iii) consistent state-of-the-art results on EgoIntention, RIO, and RefCOCO. All three claims are substantiated by the quantitative and qualitative experiments in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in a dedicated Limitations section (Sec. 6), where we note that our framework is instantiated with a specific MLLM-detector pair (InternVL3.5-8B and GroundingDINO), is not yet optimized for real-time deployment with lighter backbones, and is evaluated on benchmarks that assess a single target per query. Extensions to other MLLM-detector combinations and to weakly-supervised settings are identified as directions for future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include theoretical results. All contributions are empirical and architectural: the RSP framework, the SoftIoU formulation, and the prior-guided deformable fusion are practical components without formal theorems or proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4.1 provides the full network architecture (InternVL3.5-8B with LoRA adapters and GroundingDINO with a Swin-base backbone), training hyperparameters (AdamW optimizer, learning rate schedule, batch size, warmup iterations, LoRA rank and scaling factor, dropout), loss weights, dataset splits, and evaluation metrics needed to reproduce our main results. The RSP method itself is fully described in Section 3 with all equations for the gating module, SoftIoU loss, and prior-guided deformable fusion.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: This work was conducted as part of an industrial research collaboration, and the implementation cannot be released due to institutional and intellectual property constraints. However, all datasets used (EgoIntention, RIO, RefCOCO/+/g) are publicly available under their original licenses, and the backbone models (InternVL3.5-8B and GroundingDINO) are open-source. Section 4.1 provides sufficient architectural, training, and hyperparameter details for independent reimplementation.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 specifies all training and test details necessary to understand the results, including the optimizer (AdamW with weight decay 0.0 and β_2=0.999), learning rate schedule (cosine decay from $1\mathrm{e}{-4}$ to $1\mathrm{e}{-6}$ with 200 warmup iterations), batch size (64), LoRA configuration (r=8, α_{LoRA}=16, dropout 0.05), loss weights, dataset splits, and evaluation metrics.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report single-run results following the standard evaluation protocols of EgoIntention, RIO, and RefCOCO, which is the convention in closely related prior work on open-vocabulary and reasoning-based grounding (e.g., GroundingDINO, Lenna, RoG), none of which report error bars either. Multiple-seed experiments were not feasible given the compute cost of end-to-end fine-tuning an 8B-parameter MLLM jointly with a transformer-based detector across multiple benchmarks.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources are specified in Section 4.1 (Implementation details): all experiments are run on 4$\times$ NVIDIA L40S GPUs (48 GB each) with data-parallel training and gradient accumulation to achieve an effective global batch size of 64.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work conforms to the NeurIPS Code of Ethics. It uses publicly available research datasets under their standard licenses, does not involve human subjects or personally identifiable information, and does not raise concerns related to fairness, privacy, or security.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Our work is a foundational contribution to open-vocabulary and reasoning-based object detection, without direct deployment implications. Improvements in visual grounding could support beneficial applications such as assistive robotics or accessibility tools, and could in principle be misused in surveillance contexts, but these impacts depend entirely on downstream deployment choices rather than on our methodological contribution.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release pre-trained models, scraped datasets, or generative systems that pose a high risk of misuse. All experiments are performed on existing public research datasets using publicly released backbone models.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets used in this work, including the InternVL3.5-8B and GroundingDINO backbone models, LoRA, the EgoIntention, RIO, and RefCOCO/+/g benchmark datasets, and the Deformable DETR formulation, are properly cited in the paper. These assets are released by their original authors under standard research-use licenses and are used in accordance with their intended research purposes.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release any new assets (datasets, models, or code). Our contribution is a training framework applied to existing pre-trained models and public benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects. All data comes from pre-existing public research benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects, so IRB approval or an equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as a core methodological component in the sense intended by this question. InternVL3.5-8B is used as a multimodal model within the proposed framework and is fully described as a standard, publicly available component in Sections 3 and 4. LLMs were not used for research ideation or experiment design.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.