GROUNDING BEYOND TOKENS: CONTINUOUS GEOMETRY FROM VISUAL EVIDENCE

Anonymous authors
Paper under double-blind review

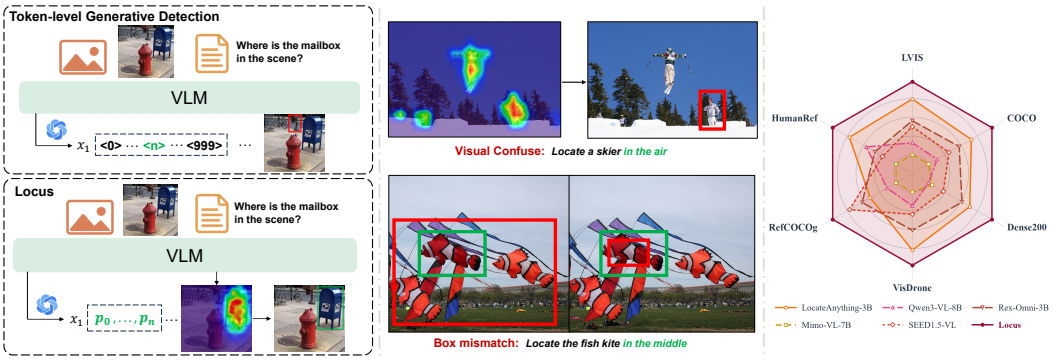


Figure 1: Comparison of conventional token-level generative detection and Locus. Locus combines explicit visual evidence supervision with continuous coordinate estimation from discrete token distributions.

ABSTRACT

Vision-language models increasingly formulate visual grounding as autoregressive generation of discrete coordinate tokens, unifying language understanding and spatial prediction within a single framework. However, standard token-level supervision treats coordinates as categorical labels without explicitly modeling their spatial ordering or geometric dependencies. Language-modeling objectives also supervise object-relevant visual evidence indirectly, leaving the correspondence among referring expressions, visual referents, and spatial predictions insufficiently constrained. In this work, we propose **Locus**, a generative framework for object detection and visual grounding that combines explicit visual evidence supervision with continuous coordinate prediction. Locus supervises the spatial distribution of phrase-conditioned visual attention, encouraging the alignment between linguistic representations and corresponding visual referents. It also interprets coordinate-token probabilities as distributions over an ordered spatial domain, enabling continuous coordinate estimation, distance-aware supervision, and box-level geometric optimization. Notably, Locus preserves the autoregressive generation paradigm and requires no additional detection modules or architectural changes. Experiments across detection and referring expression benchmarks demonstrate consistent gains in localization accuracy and state-of-the-art performance, supporting the effectiveness of combining visual evidence supervision with continuous geometric modeling.

1 INTRODUCTION

Perceiving and understanding the physical environment are fundamental capabilities for embodied intelligence (Intelligence et al., 2025; Agarwal et al., 2026). Visual grounding (Jiang et al., 2026) aims to identify and localize the visual target described by a natural-language expression, connecting language understanding, visual perception, and spatial reasoning. Conventional task-specific approaches (Cheng et al., 2024b; Zhu et al., 2021; Zhang et al., 2022; Ren et al., 2016) typically align linguistic representations with visual regions and rely on detection components, such as object detectors, region proposals, or dedicated regression modules, to predict target locations. More recently, vision-language models (VLMs) (Qwen Team, 2026; Bai et al., 2025a; Team et al., 2025b;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Comanici et al., 2025; Liu et al., 2024a; Agarwal et al., 2026; Singh et al., 2025) have recast visual grounding as a generative task within a unified autoregressive framework. Building on large-scale pretrained models, VLM-based methods (Jiang et al., 2026; Wang et al., 2026; Bai et al., 2025a;b) have achieved strong localization and detection performance across tasks such as referring expression grounding (Jiang et al., 2026; Zhang et al., 2024; Cheng et al., 2024a) and open-vocabulary object detection (Jiang et al., 2024a; Li et al., 2022; Minderer et al., 2022).

Many vision-language models (VLMs) (Bai et al., 2025a; Jiang et al., 2026; Bai et al., 2025b; Wang et al., 2025; Yu et al., 2025; Wang et al., 2026) formulate spatial localization as autoregressive coordinate prediction, representing bounding boxes as discrete token sequences. This formulation integrates perceptual capabilities into the generative framework of general-purpose VLMs (Bai et al., 2025b;a; Liu et al., 2023b) and enables them to benefit from large-scale training. However, this unified formulation raises a fundamental question: ***to what extent is supervision on discrete coordinate outputs sufficient for reliable visual grounding?*** Grounding a referring expression requires both identifying the visual evidence that distinguishes the intended referent and translating that evidence into a geometrically accurate prediction. Standard token-level supervision (Achiam et al., 2023) encourages the correct coordinate sequence, but does not explicitly constrain either the correspondence between the expression and its supporting visual evidence or the geometric quality of prediction. As illustrated in Fig. 1, these requirements reveal two limitations of learning visual grounding solely through next-coordinate-token prediction.

The first limitation concerns the visual evidence underlying localization. Coordinate-token supervision specifies the target location but does not explicitly constrain which visual evidence supports its identification. This may hinder disambiguation when similar candidates differ only in fine-grained attributes or relations. Errors may also propagate during autoregressive decoding if later predictions rely increasingly on generated tokens rather than visual context (Zhao et al., 2025). Another arises from the mismatch between discrete coordinate generation and continuous visual geometry, existing VLM-based methods (Jiang et al., 2026; Wang et al., 2026; You et al., 2024) quantize continuous coordinates into discrete tokens and optimize them as independent categories, ignoring spatial ordering, distance, and geometry quality. For instances, nearby coordinates such as `<500>` and `<501>` are not explicitly distinguished from distant ones such as `<500>` and `<900>`, while coordinate generation provides no direct supervision to evaluate the predicted box.

In this work, we introduce **Locus**, a generative visual detection framework that integrates visual evidence supervision with continuous spatial prediction. While preserving the discrete coordinate interface and autoregressive generation paradigm (Bai et al., 2025a; Chen et al., 2024), Locus provides explicit supervision for both visual evidence acquisition and continuous coordinate estimation, as shown in Fig. 1. Specifically, **Visual Evidence Supervision** leverages grounding annotations to constrain the spatial distribution of visual attention, encouraging the model to acquire localization-relevant evidence from the corresponding target regions while preserving the overall allocation of attention between visual and linguistic inputs. **Continuous Coordinate Prediction** reformulates discrete coordinate classification as probabilistic regression over an ordered coordinate space, recovering continuous spatial estimates from predicted coordinate-token distributions. These continuous estimates further enable distance-aware coordinate supervision and box-level geometric optimization, bringing the learning objective into closer alignment with the underlying visual geometry. Importantly, Locus requires no additional detection modules or architectural modifications, enabling fully end-to-end generative visual detection.

Comprehensive experiments demonstrate that Locus achieves state-of-the-art zero-shot performance in visual grounding and detection across diverse benchmarks, *e.g.*, 53.8% on LVIS (Gupta et al., 2019) and 59.5% on Dense200 (Jiang et al., 2026). The benchmark comparisons and ablations support the complementary roles of visual evidence supervision and continuous geometric modeling in detection and referring expression grounding.

2 Related Work

VLM-based Visual Grounding. Following sequence-based formulations such as Pix2Seq (Chen et al., 2022), recent VLMs (Bai et al., 2025a; Chen et al., 2024) increasingly integrate visual grounding and detection into the language-modeling interface, enabling spatial locations to be predicted or represented together with textual outputs. These methods can be broadly categorized into three

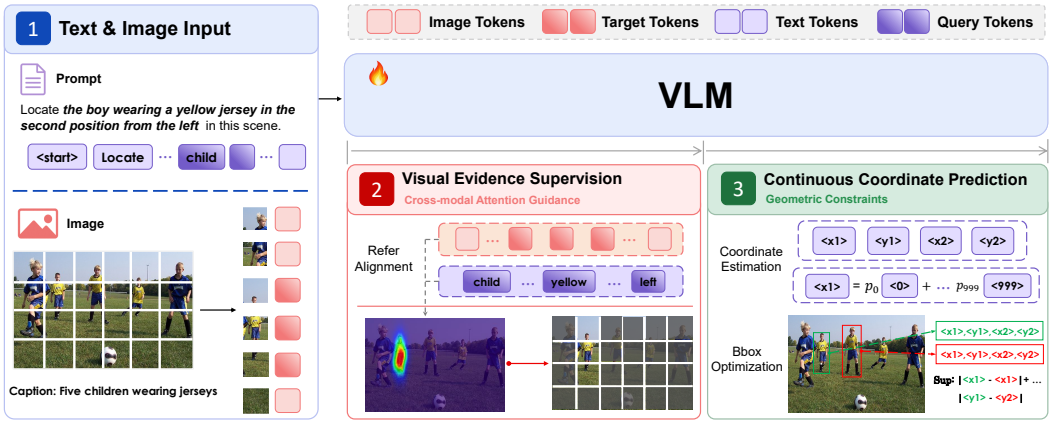


Figure 2: Overview of Locus framework.

paradigms: direct coordinate generation, region-based retrieval, and localization through external decoders. Direct-generation approaches (Chen et al., 2023; Peng et al., 2023; You et al., 2024; Zhan et al., 2024; Jiang et al., 2026; Bai et al., 2025b; Wang et al., 2026) express spatial locations through textual or dedicated coordinate tokens and perform localization within the generative interface. Region-retrieval methods (Ma et al., 2024; Jiang et al., 2024b; Yin et al., 2025; Kang et al., 2026) instead expose localized visual candidates to the language model for selection or reasoning, while decoder-based approaches (Zhang et al., 2024; Wang et al., 2023b; Wu et al., 2024a; Gao et al., 2026) connect language representations or special tokens to dedicated localization modules. Within direct coordinate generation, locations can be represented either as numerical sequences or as normalized coordinates quantized into a finite vocabulary (*e.g.*, <0> - <999>), which reduces prediction complexity and shortens spatial output sequences. Locus follows this autoregressive coordinate-token formulation to preserve a unified end-to-end generative interface, while explicitly supervising the visual evidence underlying localization.

Visual Guidance and Continuous Geometry. Early sequence-based grounding methods formulate localization as discrete coordinate generation. SeqTR (Zhu et al., 2022) quantizes geometric annotations into coordinate sequences trained with token-level cross-entropy, while PEVL (Yao et al., 2022) introduces ordering-aware position supervision to preserve spatial proximity after discretization. Regression-based approaches such as TransVG (Deng et al., 2021) and DETR (Carion et al., 2020) optimize continuous box geometry, often using objectives such as GIoU (Rezatofighi et al., 2019). Closely related to our coordinate formulation, Generalized Focal Loss (Li et al., 2020) learns distributions over discrete box locations, supervises bins adjacent to a continuous target, and recovers coordinate estimates from the predicted distributions. Recent generative grounding methods introduce spatial structure in other ways. Rex-Omni (Jiang et al., 2026) uses geometry-aware reinforcement learning; LocateAnything (Wang et al., 2026) models intra-box dependencies through Parallel Box Decoding; and Motto (Gao et al., 2026) employs spatially grounded latent tokens with a localization decoder. Other methods make grounding explicit through intermediate reasoning or pointing: v1 (Chung et al., 2025) and GRIT (Fan et al., 2026) interleave grounding with reasoning, while MolmoPoint (Clark et al., 2026) hierarchically selects image patches and subpatches. Locus applies distributional coordinate supervision to the native coordinate-token predictions of an autoregressive vision-language model and combines it with regularization of phrase-conditioned visual-token affinity. This combination enables continuous box optimization and target-region supervision within the generative interface, without a dedicated localization decoder.

3 METHODOLOGY

Accurate and reliable visual grounding requires a coherent mapping from referential understanding and visual evidence acquisition to spatial localization. In this section, we present **Locus**, a generative visual grounding framework that integrates visual evidence supervision with continuous coordinate prediction. Section 3.1 introduces the problem formulation and overall framework. Section 3.2 presents **Visual Evidence Supervision**, which constrains visual attention to encourage the acquisi-

tion of localization-relevant evidence. Section 3.3 introduces **Continuous Coordinate Prediction**, which reformulates discrete coordinate-token classification as probabilistic regression, enabling geometric constraints and improving localization quality. Section 3.4 summarizes the training objective and inference procedure.

3.1 PRELIMINARY

As illustrated in Fig. 2, Locus builds upon Qwen3-VL (Bai et al., 2025a), comprising a ViT-based vision encoder, a Qwen3 language decoder Yang et al. (2025), and a MLP projector that maps visual representations into the language-model embedding space. Given an input image $\mathcal{I}$, the vision encoder converts image patches into visual features, which are then transformed by the MLP projector. Task-specific behavior is specified through textual prompts: `<Image> Detect <ref> in this image. Return the coordinates in <box> format`, where `<ref>` denotes the target category or referring expression. Locus generates a structured response that jointly encodes semantics and geometry:

`<|object_ref_start|>PHRASE<|object_ref_end|><box>COORDS</box>`. (1)

Here, `PHRASE` specifies the target semantics, while `COORDS` contains the corresponding spatial coordinates. Bounding boxes are represented as `<x1><y1><x2><y2>`. All outputs are generated autoregressively within the original language-modeling framework.

3.2 VISUAL EVIDENCE SUPERVISION

Existing VLM-based visual grounding methods (Jiang et al., 2026; Bai et al., 2025a; Wang et al., 2026) commonly rely on token-level supervision, leaving the visual evidence underlying localization only implicitly constrained. To address this limitation, we introduce **Visual Evidence Supervision (VES)**. Rather than increasing the overall attention assigned to the visual modality, VES constrains how visual attention is spatially distributed. It consists of two complementary objectives: **Visual Attention Concentration (VAC)**, which directs reference expressions toward their corresponding target regions, and **Target Coverage Regularization**, which prevents the resulting attention from collapsing onto only a few patches within the target.

Visual Attention Concentration. As illustrated in Fig. 2, VAC is applied to the deeper layers of the language decoder and directly operates on the native query and key projections of self-attention, without introducing additional modules. For each reference r in the ground-truth response, we use its semantic phrase P_r as the query source, which has n_r text tokens. Given a supervised layer l and attention head h, the phrase-level query token $\bar{q}_r^{l,h}$ and all image tokens $k_j^{l,h}$ are computed as

$$\bar{q}_r^{l,h} = \frac{1}{n_r} \sum_{t \in P_r} q_t^{l,h}, \qquad s_{r,j}^{l,h} = \frac{\bar{q}_r^{l,h} (k_j^{l,h})^\top}{\sqrt{d_h}}. \tag{2}$$

Let $\mathcal{J}$ be the set of visual tokens and $\mathcal{S}$ the supervised layer–head pairs. At each pair (l, h), we normalize Eq. (2) over visual tokens and measure the attention mass assigned to the ground-truth region of instance i:

$$a_{r,j}^{l,h} = \frac{\exp(s_{r,j}^{l,h})}{\sum_{v \in \mathcal{J}} \exp(s_{r,v}^{l,h})}, \qquad \rho_{r,i}^{l,h} = \sum_{j \in \mathcal{J}} M_{r,i}(j) a_{r,j}^{l,h}. \tag{3}$$

The binary mask $M_{r,i}$ selects visual patches with positive-area overlap with the ground-truth box. The attention concentration loss is

$$\mathcal{L}_{\text{VAC}} = -\frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \frac{1}{N_r} \sum_{i=1}^{N_r} \log \left[\max \left(\frac{1}{|\mathcal{S}|} \sum_{(l,h) \in \mathcal{S}} \rho_{r,i}^{l,h}, \epsilon \right) \right]. \tag{4}$$

Here N_r is the number of target instances for reference r. The objective averages target-region mass across supervised heads and layers before applying the logarithm; instance losses are then averaged within each reference and across references.

Target Coverage Regularization. VAC encourages attention to fall inside the correct target region, but it may still be satisfied by concentrating on only a few highly discriminative patches. This can provide insufficient spatial evidence for accurate localization. We therefore introduce a coverage regularizer that encourages sufficiently broad support within each bounding box.

At each supervised layer l, normalize and every head distributions within the binary target mask:

$$\tilde{a}^{l,h}_{r,i,j} = \frac{M_{r,i}(j)a^{l,h}_{r,j}}{\max(\rho^{l,h}_{r,i}, \epsilon)}, \qquad \bar{a}^{l}_{r,i,j} = \frac{1}{H}\sum_{h=1}^{H} \tilde{a}^{l,h}_{r,i,j}, \tag{5}$$

where ϵ is set to $1e^{-6}$. The effective patch count at this layer is

$$N^{l}_{\text{eff},r,i} = \frac{1}{\sum_{j:M_{r,i}(j)=1}(\bar{a}^{l}_{r,i,j})^2}. \tag{6}$$

Let $N_{r,i} = \sum_j M_{r,i}(j)$ and let m_{cover} be the minimum patch count. The coverage regularizer is

$$\mathcal{R}_{\text{cover}} = \frac{1}{|\mathcal{R}|}\sum_{r\in\mathcal{R}}\frac{1}{N_r}\sum_{i=1}^{N_r}\frac{1}{|\mathcal{L}|}\sum_{l\in\mathcal{L}}\mathbf{1}[N_{r,i}\geq m_{\text{cover}}]\left[\text{ReLU}\left(\frac{r_{\text{cov}}N_{r,i}}{N^{l}_{\text{eff},r,i}}-1\right)\right]^2. \tag{7}$$

Boxes below the minimum patch count contribute zero coverage loss and m_{cover} is set to 4.

The overall visual evidence objective is

$$\mathcal{L}_{\text{VES}} = \lambda_{\text{VAC}}\mathcal{L}_{\text{VAC}} + \lambda_{\text{cover}}\mathcal{R}_{\text{cover}}. \tag{8}$$

Consequently, VAC encourages reference-conditioned attention toward the target, while coverage regularization controls *how sufficiently* that evidence spans the target region. This provides direct supervision over the visual evidence underlying coordinate generation while preserving the original autoregressive generation paradigm.

3.3 Continuous Coordinate Prediction

Autoregressive token-level supervision overlooks the continuous, ordered, and geometric structure of spatial coordinates. To bridge this gap, we introduce **Continuous Coordinate Prediction (CCP)**. As decipt in Fig. 2, CCP enables continuous coordinate supervision, distance-aware distribution optimization, and box-level geometric constraints without modifying the model architecture.

Continuous Coordinate Supervision. Let the coordinate space contain K discrete bins, with $K = 1000$ in our implementation. Given a continuous target coordinate $s \in [0, K-1]$ and $j = \min(\lfloor s \rfloor, K-2)$, we construct a soft target distribution by linear interpolation:

$$q(j) = j + 1 - s, \qquad q(j+1) = s - j. \tag{9}$$

At the upper boundary $s = K-1$, the distribution assigns unit mass to the final bin and never indexes a nonexistent token. For example, $s = 527.3$ yields $q(527) = 0.7$ and $q(528) = 0.3$, preserving the sub-bin information discarded by hard quantization. Rather than normalizing only over coordinate tokens, we supervise coordinate positions over the full language-model vocabulary $\mathcal{V}$:

$$\mathcal{L}_{\text{coord}} = D_{\text{KL}}\left(q \,\|\, P_{\mathcal{V}}\right) = -\sum_i q(i)\log P_{\mathcal{V}}(i) - H(q). \tag{10}$$

This formulation transforms the prediction of discrete coordinate categories into a regression over continuous probability distributions, thereby enabling the modeling of continuous coordinate space and encouraging probability mass to concentrate around continuous target values. Coordinate positions are removed from the standard hard-label cross-entropy loss $\mathcal{L}_{\text{txt}}$, while other valid assistant tokens remain supervised by conventional language modeling.

Ordered Coordinate Distribution. For distance-aware geometric optimization, we further isolate the coordinate subspace and normalize the logits over the K coordinate tokens:

$$p(i \mid \mathcal{C}) = \frac{P_{\mathcal{V}}(i)}{\sum_{m=0}^{K-1} P_{\mathcal{V}}(m)}, \qquad i = 0, \ldots, K-1, \tag{11}$$

where $\mathcal{C}$ denotes the coordinate token set. Particularly, $p(i \mid \mathcal{C})$ represents the model's uncertainty specifically within the coordinate token space, making the relative proximity between coordinate bins explicitly accessible for geometric optimization.

To further encode spatial displacement, we minimize the Wasserstein-1 distance between the predicted and target coordinate distributions. Let their discrete cumulative distribution functions be

$$F_p(k) = \sum_{i=0}^{k} p(i \mid \mathcal{C}), \qquad F_q(k) = \sum_{i=0}^{k} q(i). \tag{12}$$

Let C_{valid} denote the set of valid coordinates in the current sample. The normalized distance loss is

$$\mathcal{L}_{\text{W}} = \frac{1}{|\mathcal{C}_{\text{valid}}|} \sum_{c \in \mathcal{C}_{\text{valid}}} \frac{1}{K-1} \sum_{k=0}^{K-2} \left| F_p^{(c)}(k) - F_q^{(c)}(k) \right|. \tag{13}$$

Compared with categorical supervision, this objective is sensitive to how far the predicted probability mass deviates from the target, rather than only whether it falls on the correct token. As a result, nearby coordinate errors incur smaller penalties than distant ones, allowing the training objective to better reflect the continuous and ordinal structure of spatial localization.

Box-Level Geometric Optimization. The coordinate distributions further enable direct geometric supervision on bounding boxes. We therefore introduce box-level constraints to encourage both valid coordinate ordering and globally consistent localization. For bounding boxes, valid geometry requires the left and top boundaries to precede the right and bottom boundaries, respectively. Let $p_t(k)$ be the coordinate-only distribution at the teacher-forced position of boundary t, and $F_t(k) = \sum_{j=0}^{k} p_t(j)$. We use a product-of-marginals surrogate for each ordering event:

$$P_{\text{ind}}(x_1 \le x_2) = \sum_{k=0}^{K-1} p_{x_2}(k) F_{x_1}(k), \qquad P_{\text{ind}}(y_1 \le y_2) = \sum_{k=0}^{K-1} p_{y_2}(k) F_{y_1}(k).$$

$$\mathcal{L}_{\text{valid}} = -\log\big[\text{clip}(P_{\text{ind}}(x_1 \le x_2), 1e^{-6}, 1)\big] - \log\big[\text{clip}(P_{\text{ind}}(y_1 \le y_2), 1e^{-6}, 1)\big]. \tag{14}$$

Beyond validity, accurate localization also depends on the joint spatial configuration of all four boundaries. We therefore recover each coordinate $\hat{c}_{norm}$ from its predictive distribution by expectation and optimize the resulting continuous box directly:

$$\hat{c}_{norm} = \sum_{i=0}^{K-1} \frac{i}{K-1} p(i \mid \mathcal{C}), \qquad \mathcal{L}_{\text{GIoU}} = 1 - \text{GIoU}(\hat{B}, B^*), \tag{15}$$

where $\hat{B} = (\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2)$ denotes the expected continuous box and B^* is the ground-truth box. When computing GIoU, we order both pairs of expected boundaries using $\min$ and $\max$. Unlike coordinate-wise supervision, GIoU jointly considers the relative arrangement, scale, and overlap of the predicted box, thereby directly optimizing continuous box-level geometry.

The overall continuous coordinate objective is

$$\mathcal{L}_{\text{CCP}} = \mathcal{L}_{\text{coord}} + \lambda_{\text{W}} \mathcal{L}_{\text{W}} + \lambda_{\text{valid}} \mathcal{L}_{\text{valid}} + \lambda_{\text{GIoU}} \mathcal{L}_{\text{GIoU}}. \tag{16}$$

3.4 Training Objective and Inference

Training Objective. The final objective combines language modeling, visual evidence supervision, and continuous coordinate prediction:

$$\mathcal{L} = \mathcal{L}_{\text{txt}} + \mathcal{L}_{\text{VES}} + \mathcal{L}_{\text{CCP}}. \tag{17}$$

Inference. At each decoding step, Locus selects the next token from its native vocabulary and feeds it back into the autoregressive decoder. For each position in a complete generated coordinate group, we compute a continuous coordinate from the recorded logits:

$$\hat{c} = \sum_{i=0}^{K-1} i p(i \mid \mathcal{C}), \qquad t_{\text{AR}} = \arg\max_{v \in \mathcal{V}} P_{\mathcal{V}}(v). \tag{18}$$

After generation, these expectations replace the reported coordinate values. Subsequent tokens remain conditioned on the discrete token t_{AR}, not on the continuous estimates.

Method	Params.	Zero-Shot (LVIS)	LVIS (F1@IoU)			Zero-Shot (COCO)	COCO (F1@IoU)		
			0.5	0.95	Mean		0.5	0.95	Mean
Task-Specialized Detectors									
Grounding DINO-Swin-T (Liu et al., 2024b)	-	Yes	47.7	22.7	38.8	Yes	69.8	23.0	56.6
Faster RCNN-R50 (Ren et al., 2016)	-	-	-	-	-	No	60.6	7.1	48.1
DETR-R50 (Carion et al., 2020)	-	-	-	-	-	No	65.9	13.6	48.3
Deformable-DETR-R50 (Zhu et al., 2021)	-	-	-	-	-	No	69.7	17.7	54.7
DINO-R50 (Zhang et al., 2022)	-	-	-	-	-	No	68.8	21.1	55.6
DINO-Swin-L (Zhang et al., 2022)	-	-	-	-	-	No	**75.6**	**25.4**	**62.1**
Vision-Language Models									
Qwen3-VL (Bai et al., 2025a)	4B	-	59.8	20.0	43.5	-	63.0	14.2	46.1
Qwen3-VL (Bai et al., 2025a)	8B	-	61.5	20.2	44.8	-	62.8	14.0	45.7
OVIS2.5 (Lu et al., 2025)	9B	-	53.1	14.4	35.8	-	51.6	8.2	34.6
DeepSeek-VL2-Small (Wu et al., 2024b)	-	-	56.2	21.0	41.8	-	60.9	14.9	45.9
SEED1.5-VL (Guo et al., 2025)	-	Yes	<u>65.6</u>	19.5	46.7	Yes	71.3	14.3	51.4
MiMo-VL (Team et al., 2025a)	7B	-	49.5	8.8	31.4	-	56.5	6.7	35.9
Cosmos-Reason2 (Cosmos Team, 2025)	8B	-	56.4	9.8	40.2	-	56.4	9.8	39.3
Rex-Omni (Jiang et al., 2026)	3B	Yes	64.3	20.7	46.9	Yes	72.0	15.9	52.9
LocateAnything (Wang et al., 2026)	3B	Yes	62.3	<u>31.1</u>	<u>50.7</u>	Yes	70.1	19.3	54.7
Locus	4B	Yes	**68.5**	**35.4**	**53.8**	Yes	**74.3**	**22.5**	**57.5**

Table 1: Quantitative results on LVIS and COCO. "Params." indicates the model size and "-" denotes that unreported information. **Bold** and <u>underline</u> highlight the best and second-best.

4 EXPERIMENT

4.1 EXPERIMENT SETUP

Implementation Details. We initialize Locus from Qwen3-VL-4B-Instruct (Bai et al., 2025a) and perform supervised fine-tuning with learning rates of 4×10^{-5} for the language decoder and 4×10^{-6} for the SigLIP-2-based vision encoder and MLP projector. For VES, we apply VAC starting from the sixth-to-last intermediate layer to regularize visual referring attention. The minimum coverage ratio r_{cov} is set to 0.4, and the coefficient λ_{VAC} and λ_{cover} are set to 0.5 and 0.06, respectively. For CCP, the coefficients λ_{w}, λ_{GIoU}, and λ_{valid} are set to 0.5, 1.0, and 0.5. Additional training configurations and dataset details are provided in the Supplementary Materials.

Evaluation. Following Rex-omni evaluation protocol (Jiang et al., 2026), we compare Locus with other detector- or VLM-based methods on comprehensive benchmarks across multiple visual grounding and detection tasks. We evaluate object detection on COCO (Lin et al., 2014) for common objects, LVIS (Gupta et al., 2019) for long-tailed distributions, and VisDrone (Du et al., 2019) and Dense200 (Jiang et al., 2026) for dense and tiny object scenarios. Referring expression Grounding tasks include RefCOCOg (Kazemzadeh et al., 2014) and HumanRef (Jiang et al., 2025). We report F1 scores at multiple intersection-over-union (IoU) thresholds. At threshold t, a prediction is counted as a true positive if its IoU with a matched ground-truth box is at least t. We report F1 at IoU thresholds of 0.50 and 0.95, together with mean F1 across the evaluated thresholds.

4.2 RESULT ANALYSIS

Results on General Detection. As shown in Table 1, Locus consistently outperforms existing VLM-based methods on both LVIS and COCO under the zero-shot setting. With only 4B parameters, Locus achieves an F1@mIoU of **53.8%** on LVIS and **57.5%** on COCO, improving over the Qwen3-VL-4B baseline by **10.3%** and **11.4%** points, respectively. Compared with the previous strong zero-shot method LocateAnything, Locus further improves the mean F1 by **3.1** points on LVIS and **2.8** points on COCO. These results indicate that Locus not only improves coarse object localization but also produces more geometrically accurate predictions, supporting the effectiveness of jointly supervising visual evidence acquisition and continuous spatial geometry. Notably, Locus also surpasses several larger VLMs while retaining a compact model scale, demonstrating that the improvements arise from more effective localization learning rather than increased model capacity.

Results on Dense Detection. Table 2 further evaluates Locus on the challenging Dense200 and VisDrone benchmarks, which contain densely distributed and small-scale objects. Locus achieves the best overall performance on both datasets, reaching **59.5%** on Dense200 and **41.0%** on VisDrone. Although LocateAnything retains a slight advantage at the strict IoU@0.95 threshold, Locus

Method	Params.	Dense200 (F1@IoU)			VisDrone (F1@IoU)		
		0.5	0.95	Mean	0.5	0.95	Mean
Task-Specialized Detectors							
Grounding DINO-Swin-T (Liu et al., 2024b)	-	36.9	19.7	33.1	55.2	3.9	38.5
Vision-Language Models							
Qwen3-VL (Bai et al., 2025a)	4B	17.5	2.4	12.5	42.3	1.4	26.0
Qwen3-VL (Bai et al., 2025a)	8B	13.5	1.7	9.6	42.8	1.4	25.8
OVIS2.5 (Lu et al., 2025)	9B	14.0	0.0	5.1	15.8	0.1	6.5
DeepSeek-VL2-Small (Wu et al., 2024b)	-	16.0	3.9	12.7	35.8	1.7	23.3
SEED1.5-VL (Guo et al., 2025)	-	76.9	5.3	53.2	55.9	0.6	27.4
MiMo-VL (Team et al., 2025a)	7B	29.7	0.4	15.9	27.7	0.3	14.3
Cosmos-Reason2 (Cosmos Team, 2025)	8B	25.1	1.1	15.1	40.2	1.3	22.3
Qwen2.5-VL (Bai et al., 2025b)	7B	1.1	0.1	0.6	34.5	1.6	21.7
Rex-Omni (Jiang et al., 2026)	3B	78.4	10.3	58.3	61.6	1.5	35.8
LocateAnything (Wang et al., 2026)	3B	74.0	**18.5**	58.7	<u>63.0</u>	**3.2**	<u>39.9</u>
Locus	4B	**80.0**	<u>16.9</u>	**59.5**	**66.7**	<u>2.8</u>	**41.0**

Table 2: Quantitative results on dense detection benchmarks.

Method	Params.	HumanRef			RefCOCOg val			RefCOCOg test		
		0.5	0.95	Mean	0.5	0.95	Mean	0.5	0.95	Mean
Task-Specialized Detectors										
Grounding DINO-Swin-T (Liu et al., 2024b)	-	28.0	16.5	25.2	52.9	20.9	45.9	53.8	22.9	46.8
Vision-Language Models										
Qwen3-VL (Bai et al., 2025a)	4B	77.7	54.9	71.1	88.0	34.0	74.7	87.6	33.9	74.6
Qwen3-VL (Bai et al., 2025a)	8B	78.6	55.7	72.0	88.6	33.4	74.9	88.6	33.8	75.2
OVIS2.5 (Lu et al., 2025)	9B	73.1	12.4	52.8	88.8	23.5	72.1	88.7	24.2	72.6
DeepSeek-VL2-Small (Wu et al., 2024b)	-	72.0	46.5	64.7	92.4	**45.6**	81.4	91.8	**47.0**	81.6
SEED1.5-VL (Guo et al., 2025)	-	88.2	60.0	81.6	84.7	30.9	71.9	85.2	32.1	73.2
MiMo-VL (Team et al., 2025a)	7B	77.6	26.4	63.4	84.9	14.4	65.3	84.6	14.9	65.5
Qwen2.5-VL (Bai et al., 2025b)	7B	72.9	42.9	64.1	86.2	27.2	70.0	85.7	28.4	70.4
Rex-Omni (Jiang et al., 2026)	3B	85.4	65.4	79.9	86.6	35.3	73.6	86.8	36.6	74.3
LocateAnything (Wang et al., 2026)	3B	82.9	<u>68.8</u>	78.7	88.6	41.5	76.7	88.8	43.4	77.6
Locus	4B	**90.5**	**70.5**	**83.5**	**93.2**	<u>42.5</u>	**82.3**	**92.8**	<u>43.5</u>	**82.4**

Table 3: Quantitative results (F1@IoU) on referring expression comprehension benchmarks.

delivers stronger overall performance across IoU levels. These results demonstrate the robustness of Locus in dense and cluttered scenes, suggesting that explicit visual evidence supervision and continuous geometric modeling improve localization under challenging object distributions.

Results on Referring Comprehension. As shown in Table 3, Locus achieves the best overall performance across HumanRef and both RefCOCOg splits, demonstrating strong referring expression grounding capability. Although DeepSeek-VL2-Small retains the highest F1 at the strict IoU@0.95 threshold on RefCOCOg, Locus consistently achieves the strongest overall performance across the two splits. These gains highlight the effectiveness of visual evidence supervision in explicitly aligning referring expressions with their corresponding visual regions, enabling more reliable referent discrimination and localization.

4.3 Ablation Study

VES components. As shown in Table 4a, VAC improves F1@mean from 52.1% to 55.3% by explicitly aligning attention with the target region. Coverage regularization further improves F1 score when combined with VAC. This indicates that VAC mainly establishes correct visual correspondence, while coverage regularization promotes more complete evidence acquisition.

CCP components. Table 4b shows that $\mathcal{L}_{\text{W}}$ improves F1@0.5 from 67.8% to 69.9% by introducing distance-aware coordinate supervision. Adding $\mathcal{L}_{\text{valid}}$ provides further improvement, while the box-level $\mathcal{L}_{\text{GIoU}}$ yields the largest gain, reaching 74.3%. These results confirm the importance of jointly modeling coordinate-level continuity and box-level geometric consistency.

Main components. Table 4c evaluates the two core components of Locus. Under the same training setting, BASELINE trained only with the autoregressive loss achieves a mean F1 of 46.5%; adding

$\mathcal{L}_{\text{VAC}}$	$\mathcal{R}_{\text{cover}}$	F1@0.5	F1@0.95	F1@mean
		70.3	18.7	52.1
✓		72.8	20.6	55.3
	✓	70.5	19.4	52.5
✓	✓	**74.3**	**22.5**	**57.5**

(a) VES components.

$\mathcal{L}_{\text{W}}$	$\mathcal{L}_{\text{valid}}$	$\mathcal{L}_{\text{GIoU}}$	F1@0.5	F1@0.95
			67.8	15.6
✓			69.9	17.8
✓	✓		70.5	18.6
✓	✓	✓	**74.3**	**22.5**

(b) CCP components.

VES	CCP	F1@0.5	F1@0.95	F1@mean
		65.3	14.8	46.5
✓		67.8	15.6	50.1
	✓	70.3	18.7	52.1
✓	✓	**74.3**	**22.5**	**57.5**

(c) Overall structure.

Table 4: Ablation results on the COCO dataset.



Figure 3: Qualitative results of Locus on object detection, referring expression grounding, and dense detection.

VES increases it to 50.1%, and CCP further provides stronger gains in localization accuracy. Combining VES and CCP achieves the best performance of 57.5%, demonstrating that reliable visual evidence acquisition and continuous geometric modeling are complementary.

4.4 Qualitative Results

As illustrated in Fig. 3, Locus exhibits robust localization across diverse visual scenarios. For object detection, the model accurately localizes multiple instances with varying scales and appearances, while maintaining tight spatial alignment with object boundaries. In referring grounding, Locus successfully resolves fine-grained attributes, spatial relations, and contextual descriptions to distinguish the intended referent from nearby distractors, reflecting the benefit of explicitly supervising localization-relevant visual evidence. Notably, the model also remains effective in dense scenes containing numerous small and closely distributed objects, producing well-separated predictions with limited missed detections or spatial ambiguity. Overall, the visualizations highlight the complementary benefits of VES for referent disambiguation and continuous coordinate modeling for precise localization, particularly in crowded and fine-grained scenarios.

5 Conclusion

In this work, we present Locus to address a central limitation of generative visual grounding: supervision on coordinate tokens alone does not explicitly constrain the visual evidence supporting a prediction. Locus guides attention toward target regions and optimizes continuous coordinates recovered from token distributions, while retaining autoregressive decoding and requiring no additional detection modules. Its improvements across detection and referring expression benchmarks, together with the ablation results, show that visual evidence supervision and geometric optimization make complementary contributions. Since target boxes may omit contextual evidence needed for relational expressions, future work should therefore model evidence beyond the target region and extend this evidence-to-geometry principle to temporal and 3D grounding.

AI USE STATEMENT

In this work, we used generative AI tools solely for language editing and polishing, including improving grammar, clarity, and academic style. We did not use generative AI tools for research ideation, experimental design, data analysis, code generation, or the production of scientific claims or results. All AI-assisted revisions were carefully reviewed and verified by the authors to ensure that the technical meaning and scientific content remained unchanged. We take full responsibility for the final content of this work.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Holger Caesar, Varun Bankiti, Alex H. Lang, et al. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, pp. 213–229. Springer, 2020.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2022.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pp. 24185–24198, 2024.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, et al. SpatialRGPT: Grounded spatial reasoning in vision-language models. *NeurIPS*, 37:135062–135093, 2024a.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16901–16911, 2024b.

Jiwan Chung, Junhyeok Kim, Siyeol Kim, Jaeyoung Lee, Min Soo Kim, and Youngjae Yu. v1: Learning to point visual tokens for multimodal grounded reasoning. *arXiv preprint arXiv:2505.18842*, 2025.

Christopher Clark, Yue Yang, Jae Sung Park, Zixian Ma, Jieyu Zhang, Rohun Tripathi, Mohammadreza Salehi, Sangho Lee, Taira Anderson, Winson Han, et al. Molmopoint: Better pointing for vlms with grounding tokens. *arXiv preprint arXiv:2603.28069*, 2026.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Cosmos Team. Nvidia cosmos-reason2. https://huggingface.co/nvidia/Cosmos-Reason2-8B, 2025.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1769–1779, 2021. doi: 10.1109/ICCV48922.2021.00179.

Jian Ding, Nan Xue, Gui-Song Xia, et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. doi: 10.1109/TPAMI.2021.3117983.

Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Lin, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, et al. Visdrone-det2019: The vision meets drone object detection in image challenge results. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 213–226. IEEE, 2019.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Xinze Guan, and Xin Wang. Grit: Teaching mllms to think with images. *Advances in Neural Information Processing Systems*, 38:116522–116543, 2026.

Tianyi Gao, Han Fang, Tianyi Ding, Hao Li, Xin Wei, Hongbo Sun, Xiaodong Dong, Ye Yuan, Jinglin Xu, Kongming Liang, Hao Sun, and Jingmin Xin. Mixture-of-thought-tokens: Unifying perception and reasoning for free-form multimodal grounding. *arXiv preprint arXiv:2607.24407*, 2026.

Eran Goldman, Roei Herzig, Aviv Eisenschtat, Jacob Goldberger, and Tal Hassner. Precise detection in densely packed scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, et al. Seed1.5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, pp. 5356–5364, 2019.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-rex2: Towards generic object detection via text-visual prompt synergy. In *European Conference on Computer Vision*, pp. 38–57. Springer, 2024a.

Qing Jiang, Gen Luo, Yuqin Yang, Yuda Xiong, Yihao Chen, Zhaoyang Zeng, Tianhe Ren, and Lei Zhang. Chatrex: Taming multimodal llm for joint perception and understanding. *arXiv preprint arXiv:2411.18363*, 2024b.

Qing Jiang, Lin Wu, Zhaoyang Zeng, Tianhe Ren, Yuda Xiong, et al. Referring to any person. In *ICCV*, pp. 21667–21678, 2025.

Qing Jiang, Junan Huo, Xingyu Chen, Yuda Xiong, Zhaoyang Zeng, Yihao Chen, Tianhe Ren, Junzhi Yu, and Lei Zhang. Detect anything via next point prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25472–25483, 2026.

Weitai Kang, Jason Kuen, Mengwei Ren, Zijun Wei, Yan Yan, and Kangning Liu. Vgent: Visual grounding via modular design for disentangling reasoning and prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41160–41170, 2026.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 787–798, 2014.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Marcus Klasson, Cheng Zhang, and Hedvig Kjellstr"om. A hierarchical grocery store image dataset with visual and semantic labels. In *IEEE Winter Conference on Applications of Computer Vision*, 2019.

Ranjay Krishna, Yuke Zhu, Oliver Groth, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 2017.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10965–10975, 2022.

Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21002–21012, 2020.

You Li, Heyu Huang, Chi Chen, et al. Migician: Revealing the magic of free-form multi-image grounding in multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. URL https://aclanthology.org/2025.findings-acl.512/.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024b.

Shiyin Lu, Yang Li, Yu Xia, et al. Ovis2.5 technical report. *arXiv preprint arXiv:2508.11737*, 2025.

Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *Computer Vision – ECCV 2024*, pp. 417–435. Springer, 2024. doi: 10.1007/978-3-031-72658-3_24.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European conference on computer vision*, pp. 728–755. Springer, 2022.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123(1):74–93, 2017.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *PAMI*, 39(6):1137–1149, 2016.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 658–666, 2019.

Shuai Shao, Zijian Zhao, Boxun Li, et al. Crowdhuman: A benchmark for detecting human in a crowd, 2018.

Shuai Shao, Zeming Li, Tianyuan Zhang, et al. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Core Team, Zihao Yue, Zhenru Lin, Yifan Song, and other. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 2025a.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025b.

Jiaqi Wang et al. V3det: Vast vocabulary visual detection dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023a.

Shihao Wang, Shilong Liu, Yuanguo Kuang, Xinyu Wei, Yangzhou Liu, Zhiqi Li, Yunze Man, Guo Chen, Andrew Tao, Guilin Liu, Jan Kautz, Lei Zhang, and Zhiding Yu. Locateanything: Fast and high-quality vision-language grounding with parallel box decoding. *arXiv:2605.27365*, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, et al. InternVL 3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, and Jifeng Dai. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. In *Advances in Neural Information Processing Systems*, volume 36, 2023b.

Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, Ping Luo, Yu Qiao, and Jifeng Dai. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. In *Advances in Neural Information Processing Systems*, volume 37, 2024a.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yuan Yao, Qianyu Chen, Ao Zhang, Wei Ji, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Pevl: Position-enhanced pre-training and prompt tuning for vision-language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11104–11117, 2022.

Heng Yin, Yuqiang Ren, Ke Yan, Shouhong Ding, and Yongtao Hao. Rod-mllm: Towards more reliable object detection in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14358–14368, 2025.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, et al. Ferret: Refer and ground anything anywhere at any granularity. In *ICLR*, 2024.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, et al. Perception-R1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025.

Fisher Yu, Haofeng Chen, Xin Wang, et al. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, 2016.

Yufei Zhan, Yousong Zhu, Zhiyang Chen, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon: Spelling out all object locations at any granularity with large language models. In *Computer Vision – ECCV 2024*, pp. 405–422. Springer, 2024. doi: 10.1007/978-3-031-72946-1_23.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, et al. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, et al. LLaVA-Grounding: Grounded visual chat with large multimodal models. In *ECCV*, pp. 19–35. Springer, 2024.

Haozhe Zhao, Shuzheng Si, Liang Chen, Yichi Zhang, Maosong Sun, Baobao Chang, and Minjia Zhang. Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19666–19690, November 2025.

Qiancheng Zheng, Yunhang Shen, Gen Luo, et al. Omni-referring image segmentation, 2025.

Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *Computer Vision – ECCV 2022*, pp. 598–615. Springer, 2022.

Chenchen Zhu, Fanyi Xiao, Andres Alvarado, et al. Egoobjects: A large-scale egocentric dataset for fine-grained object understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, et al. Deformable DETR: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2021.

A Appendix

A.1 Training Configurations

We initialize Locus from Qwen3-VL-4B-Instruct and fine-tune all model parameters for two epochs using the AdamW optimizer with a global batch size of 512. We use learning rates of 4×10^{-5} for the language decoder and 4×10^{-6} for the vision encoder and visual projector. We use a cosine learning-rate schedule with a warmup ratio of 0.05 and a weight decay of 0.01.

Coordinates are represented with $K = 1000$ discrete tokens. For complete supervised coordinate-token groups, the soft-target KL objective over the full vocabulary replaces ordinary token cross-entropy; other valid assistant-response tokens retain cross-entropy supervision. The token losses share a denominator equal to the number of valid response positions. All geometric losses use teacher-forced prefixes. We set $(\lambda_W, \lambda_{\text{valid}}, \lambda_{\text{GIoU}}) = (0.5, 0.5, 1.0)$. For VES, assistant-response reference phrases serve as queries, and visual patches are assigned to target boxes when they have positive-area overlap. Visual attention concentration is applied to the final six decoder layers. Coverage regularization uses $r_{\text{cov}} = 0.4$ and is enabled only for boxes intersecting at least four merged

Category	Dataset	Scale
Object detection	SA-1B	∼11.0M images
Object detection	Objects365	∼2.0M images
Object detection	V3Det	∼243K images
Object detection	EgoObjects	114K annotated frames
Object detection	BDD100K	100K detection images
Object detection	DOTA-v2.0	11,268 source images
Object detection	nuImages	∼93K annotated images
Object detection	GroceryStore	5,125 images
Referring comprehension	RefCOCO	142,209 expressions (19,994 images)
Referring comprehension	RefCOCO+	141,564 expressions (19,992 images)
Referring comprehension	RefCOCOg	104,560 expressions (26,711 images)
Referring comprehension	gRefCOCO	278,232 expressions (19,994 images)
Referring comprehension	Flickr30K Entities	31,783 images (158,915 captions)
Referring comprehension	HumanRef	103,028 training expressions (34,806 images)
Referring comprehension	Visual Genome	108,077 images (5.4M region descriptions)
Referring comprehension	MGrounding-630K	∼630K grounding examples
Referring comprehension	OmniRef	186,939 prompts (30,956 images)
Dense detection	SKU-110K	11,762 images
Dense detection	CrowdHuman	24,370 images (15,000 training)
Dense detection	FSC-147	6,135 images

Table 5: Overview of the Locus training data. The table summarizes the datasets used for object detection, referring comprehension, and dense detection, and reports the number of source samples from each dataset.

VES	CCP	F1@0.5	F1@0.95	F1@mean
		87.6	33.9	74.6
✓		89.2	37.5	78.1
	✓	90.5	39.7	79.8
✓	✓	**92.8**	**43.5**	**82.4**

(a) Evaluation on RefCOCOg test dataset.

VES	CCP	F1@0.5	F1@0.95	F1@mean
		72.5	12.8	49.2
✓		75.7	14.5	52.1
	✓	78.5	15.2	56.1
✓	✓	**80.0**	**16.9**	**59.5**

(b) Evaluation on Dense200 dataset.

Table 6: Ablation results on the referring expression and dense detection benchmarks.

visual patches. We set $(\lambda_{\text{VAC}}, \lambda_{\text{cover}}) = (0.5, 0.06)$. The target-region attention mass is averaged over supervised heads and layers before the negative logarithm is applied. VES and the CCP training losses are omitted at inference; continuous coordinates are read from the coordinate-token distributions after autoregressive generation.

A.2 Dataset Details

Dataset details. As illustrated in Table 5, we construct a bounding-box training corpus spanning object detection, referring comprehension, and dense detection. For object detection, we draw on SA-1B (Kirillov et al., 2023), Objects365 (Shao et al., 2019), V3Det (Wang et al., 2023a), EgoObjects (Zhu et al., 2023), BDD100K (Yu et al., 2020), DOTA-v2.0 (Ding et al., 2021), nuImages (Caesar et al., 2020), GroceryStore (Klasson et al., 2019). These sources broaden the range of object categories, viewpoints, and visual domains encountered during training. For referring comprehension, we use RefCOCO and RefCOCO+ (Yu et al., 2016), RefCOCOg (Kazemzadeh et al., 2014), gRefCOCO (Liu et al., 2023a), Flickr30K Entities (Plummer et al., 2017), HumanRef (Jiang et al., 2025), Visual Genome (Krishna et al., 2017), MGrounding-630k (Li et al., 2025), and OmniRef (Zheng et al., 2025). These datasets introduce linguistic references that require selecting the intended visual target among competing regions. We further incorporate SKU-110K (Goldman et al., 2019), CrowdHuman (Shao et al., 2018), and FSC-147 (Ranjan et al., 2021) to strengthen localization in crowded scenes and among repeated objects. Across all sources, training targets are represented as bounding boxes in the same autoregressive output format.

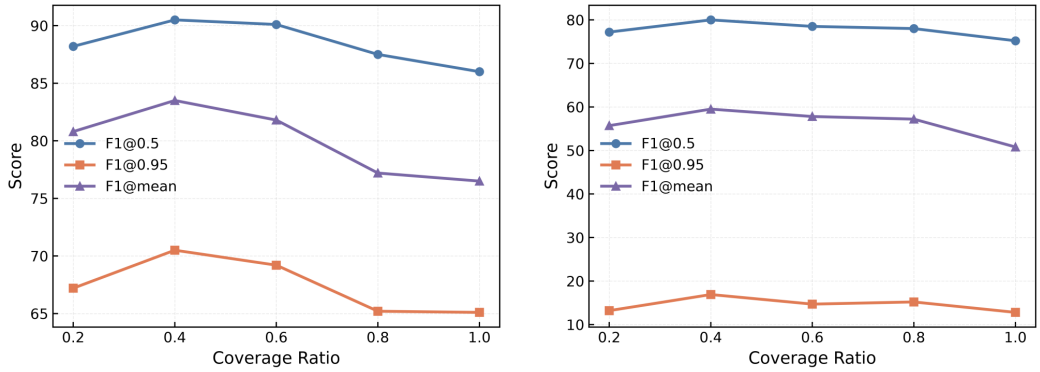


Figure 4: Effect of the target coverage ratio on HumanRef (left) and Dense200 (right). Curves show F1@0.5, F1@0.95, and mean F1; both benchmarks peak near $r_cov = 0.4$.

Benchmark	Model	x (%)	y (%)
COCO	AR only	98.0	98.3
	validity loss	100	99.9
Dense200	AR only	96.2	97.1
	validity loss	100	100

Table 7: Validity rate of detection box coordinates. x and y represent the proportions of valid x and y coordinates, respectively.

A.3 More Experiment Results

Ablation across referring and dense-detection tasks. The additional ablations, as shown in Table 6, extend the analysis beyond COCO to RefCOCOg test and Dense200. On RefCOCOg test, the autoregressive baseline attains F1@0.5, F1@0.95, and mean F1 scores of 87.6%, 33.9%, and 74.6%, respectively. Adding visual evidence supervision (VES) raises these scores to 89.2%, 37.5%, and 78.1%; adding continuous coordinate prediction (CCP) further increases them to 92.8%, 43.5%, and 82.4%. On Dense200, the corresponding baseline scores are 72.5%, 12.8%, and 49.2%. VES improves them to 75.7%, 14.5%, and 52.1%, while the complete model reaches 80.0%, 16.9%, and 59.5%. Thus, the full model improves mean F1 over the baseline by 7.8 points on RefCOCOg test and 10.3 points on Dense200. These results are consistent with the COCO ablation: explicit supervision of target-relevant visual evidence and continuous geometric optimization provide complementary gains across both expression-based and dense localization settings.

Sensitivity to the coverage target and visual evidence. Fig 4 varies the minimum coverage ratio used by the coverage regularizer. The reported F1 curves on both HumanRef and Dense200 peak near a ratio of $r_{\text{cov}} = 0.4$ and decline at larger ratios, supporting the value selected for the main experiments. This trend is consistent with a trade-off between encouraging attention to span the target and imposing an overly broad spatial distribution; the curves alone do not establish which mechanism accounts for each change. Fig. 5 provides qualitative attention maps from the final six decoder layers for two example queries. Relative to the autoregressive BASELINE without vision evidence supervision, VES places visibly more attention on the depicted target regions in these examples. These visualizations illustrate the intended effect of VES, while the quantitative ablations above provide the evidence for its impact on localization performance.

Generated-box validity. The coordinate-validity analysis evaluates whether generated boxes satisfy the ordering constraints $x_1 \leq x_2$ and $y_1 \leq y_2$, as reported in Table 7. We measure the validity rates of the horizontal and vertical boundary pairs separately to assess the structural consistency of the predicted coordinates. With autoregressive decoding alone, these rates are 98.0% and 98.3% for x and y on COCO, respectively, and 96.2% and 97.1% on Dense200. These results indicate

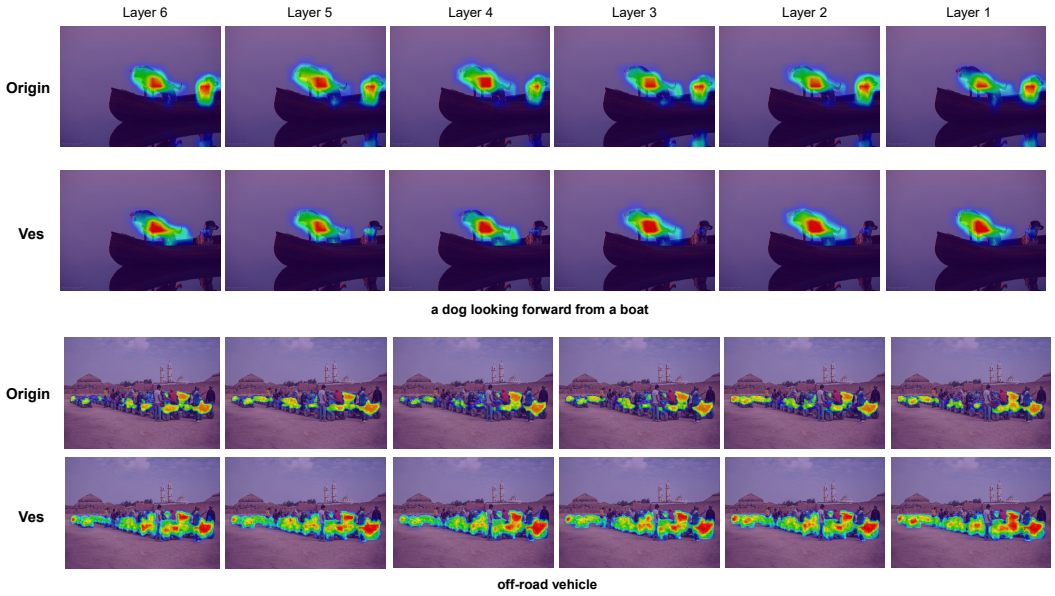


Figure 5: Reference-conditioned attention maps from the final six decoder layers for two example queries, comparing the autoregressive baseline (labeled "Origin") with the VES-trained model.

that autoregressive supervision generally produces correctly ordered coordinates but still permits a small proportion of boundary inversions, with higher violation rates on Dense200. Incorporating the explicit validity loss raises the respective rates to 100.0% and 99.9% on COCO and to 100.0% for both coordinate pairs on Dense200. The improvements across both benchmarks support the effectiveness of explicitly supervising coordinate ordering alongside autoregressive prediction. This analysis complements the IoU-based evaluation by examining a basic structural property of the generated boxes: whether their boundaries are consistently ordered. Moreover, the results show that the validity objective substantially reduces ordering violations in the evaluated outputs.

A.4 Additional Visualization Results

We present further Locus predictions spanning expression-conditioned grounding, general object detection, and dense detection. The referring expression examples illustrate target selection when nearby objects provide plausible alternatives to the language-specified referent. The general-detection examples include multiple categories and instances in a single scene, whereas the dense scenes place greater demands on separating boxes for adjacent, visually similar objects. The examples are organized as follows:

- Referring expression grounding (Fig.6);
- General object detection (Fig.7);
- Dense object detection (Fig.8).



Figure 6: Additional qualitative results for expression-conditioned visual grounding.



Figure 7: Additional qualitative results for multi-instance and multi-category detection.



Figure 8: Additional qualitative results in dense object scenes.