# GROUNDANYTHING: RECONCILING PARALLEL DECODING WITH PRECISE VISUAL GROUNDING AT FLASH SPEED

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Autoregressive (AR) grounding models serialize spatial predictions, introducing sequential latency and imposing a causal order on output tokens. We view grounding as visual evidence extraction: objects, locations, and spatial relations are jointly constrained by the image and query, yet their dependencies do not imply an intrinsic left-to-right generation order. This distinction makes bidirectional diffusion a natural fit, allowing spatial hypotheses to emerge in parallel and be jointly refined through iterative denoising. We introduce GroundAnything, a 4B-parameter grounding foundation model that reconciles fast parallel decoding with precise localization through blockwise denoising. Training combines grounding pretraining from public datasets and dedicated data engines, direct AR-to-diffusion conversion with joint AR and diffusion objectives, supervised fine-tuning, and GRPO-based reinforcement post-training. Across 30 grounding benchmarks, our autoregressive variant, GroundAnything-VLM, establishes a new overall state of the art among similarly sized models at 72.42%, remaining competitive with GPT-6 Astra (71.35%). With entropy-guided decoding, GroundAnything also surpasses the prior state of the art at this scale, averaging 61.75% versus 53.32% for the fast MTP-based LocateAnything model. We further explore decoding strategies, showing that an optional self-speculative mode achieves a 4.51× speedup over the AR counterpart with a 0.74 percentage-point drop in COCO F1mIoU. Infrastructure experiments show that progressive inference optimizations translate parallel decoding into practical speedups. These gains support efficient visual grounding in latency-sensitive real-world systems.

## 1 INTRODUCTION

Visual grounding turns visual inputs into structured spatial predictions across object detection, referring expression comprehension, OCR, GUI interaction, and embodied perception. Next-token prediction has become a mainstream paradigm for generative grounding, with models such as Pix2Seq, Florence-2, and Rex-Omni expressing visual predictions as structured token sequences (Jiang et al., 2026; Chen et al., 2022; Xiao et al., 2024). However, autoregressive (AR) decoding serializes labels and coordinates, introducing sequential latency and restricting output interactions to the generated prefix (Chen et al., 2022; Cheng et al., 2024). Dense scenes and complex layouts make these limitations especially costly for real-time grounding.

To reduce AR decoding latency, LocateAnything uses multi-token prediction (MTP) for parallel box decoding (Wang et al., 2026). Diffusion VLMs support parallel generation and bidirectional refinement, with advances in multimodal understanding (You et al., 2026; Wu et al., 2026a), GUI grounding (Kumbhar et al., 2026), and document OCR (Dong et al., 2026). Yet these grounding applications remain domain-specific, leaving diffusion's broader potential for unified structured prediction across diverse visual tasks largely untapped.

To unlock this potential and reconcile fast parallel decoding with broad and precise visual grounding, we start from a familiar observation: we may notice objects or read separate signs in different orders, yet reach the same judgment about what is present and where. We therefore view grounding as structured visual evidence extraction: spatial variables are jointly constrained by the image and



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.

query, without an intrinsic left-to-right generation order. Bidirectional masked diffusion naturally supports this process, allowing hypotheses to emerge in parallel and be refined using shared visual evidence and partially recovered structure (Figure 1).




Figure 1: **GroundAnything**: broad, precise grounding with fast parallel decoding. **Top**: diverse tasks challenge localization. **Bottom**: given a 2D image and prompt, bidirectional diffusion extracts visual evidence into 1D structured grounding tokens in parallel. Self-speculative decoding achieves a 4.51× speedup over GroundAnything-VLM with a 0.74 percentage-point drop in COCO F1mIoU.

We introduce **GroundAnything**, a 4B-parameter grounding foundation model combining broad perceptual capabilities with blockwise parallel decoding. A shared vocabulary with quantized coordinates unifies diverse grounding tasks. Its decoder combines bidirectional attention within blocks with causal dependencies across blocks, enabling parallel prediction and KV-cache reuse (Arriola et al., 2025). Grounding pretraining draws on public datasets and dedicated data engines. We build GroundAnything-VLM and GroundAnything from a shared pretrained checkpoint, applying AR-to-diffusion conversion with joint AR and diffusion objectives to the latter. Both variants then undergo supervised fine-tuning and GRPO-based reinforcement post-training.

Across 30 grounding benchmarks, GroundAnything-VLM establishes a new overall state of the art among similarly sized models at 72.42%, remaining competitive with GPT-6 Astra (71.35%). With entropy-guided decoding, GroundAnything is, to our knowledge, the first diffusion language model to surpass the prior overall AR state of the art at this scale, averaging 61.75% at over 2× the speed of GroundAnything-VLM. We use this mode throughout the main benchmarks to evaluate pure diffusion decoding without AR verification. Additional ablations examine key model and training design choices throughout the study.

We further conduct a systematic speed analysis of decoding strategies and their trade-offs, including comparisons with MTP-based generation (Wang et al., 2026). An optional self-speculative mode combines diffusion drafting with AR verification and improves both speed and accuracy over entropy-guided decoding on COCO, achieving a 4.51× speedup over GroundAnything-VLM with a 0.74 percentage-point drop in F1mIoU. Systematic infrastructure experiments with SGLang, CUDA Graph, and FP8 demonstrate further practical speedups through progressive inference optimizations, supporting efficient visual grounding in real-world applications.

In summary, our major contributions are as follows:

- **Unlocking diffusion for broad visual grounding.** We introduce GroundAnything, a 4B foundation model that unifies diverse grounding tasks with precise localization and parallel decoding, trained through grounding pretraining, AR-to-diffusion conversion with joint objectives, supervised fine-tuning, and GRPO-based reinforcement learning.
- **State-of-the-art grounding at 4B scale.** Across 30 benchmarks, GroundAnything surpasses the prior overall state of the art among similarly sized AR models and outperforms the larger Qwen3.7-Max. GroundAnything-VLM establishes a new overall state of the art at this scale and remains competitive with GPT-6 Astra.
- **Systematic acceleration studies.** We analyze decoding strategies in detail, compare with MTP-based generation, and evaluate progressive infrastructure optimizations for practical deployment.

## 2 Related Work

**Visual Grounding.** YOLO efficiently localizes objects from predefined categories (Redmon et al., 2016). Grounded vision–language pretraining enables language-conditioned, open-set localization (Li et al., 2022; Liu et al., 2024). Generative methods cast detection as autoregressive next-token prediction (Chen et al., 2022) and integrate grounding into multimodal language models (Peng et al., 2023). Structured generation unifies perception tasks (Xiao et al., 2024), while quantized coordinates, large-scale data, and reinforcement post-training improve localization (Jiang et al., 2026). However, serial AR decoding makes dense spatial predictions costly. MTP-based parallel decoding reduces this overhead by treating boxes and points as atomic units (Wang et al., 2026). This motivates exploring alternative parallel mechanisms for broad and precise grounding.

**Diffusion Vision-Language Models.** Masked diffusion enables parallel generation through bidirectional conditioning and iterative token recovery. Diffusion VLMs show promise in visual instruction following and multimodal understanding (You et al., 2026; Li et al., 2026b). AR-to-diffusion conversion adapts pretrained VLMs for efficient blockwise generation (Wu et al., 2026a). Specialized applications include GUI grounding with geometry-aware masking (Kumbhar et al., 2026) and structured document OCR with blockwise denoising (Dong et al., 2026). Yet a unified diffusion grounding foundation for precise localization across heterogeneous tasks remains underexplored. GroundAnything combines grounding-focused training with a shared structured output interface for broad perception and efficient parallel decoding.

**Parallel Decoding.** Prior work explores blockwise generation (Arriola et al., 2025), adaptive parallel sampling (Wu et al., 2026b; Ben-Hamu et al., 2026), diffusion-based speculative decoding (Liu et al., 2026; Wu et al., 2026a), and optimized inference infrastructure (Liu et al., 2025). Building on these advances, we analyze decoding speed–accuracy trade-offs for GroundAnything, benchmark against MTP-based generation, and evaluate progressive infrastructure optimizations.

## 3 Method

### 3.1 Problem Formulation and Model Architecture

**Problem Formulation: Grounding as Structured Visual Evidence Extraction.** Given an image $I$ and query $Q$, grounding extracts semantic identifiers and spatial evidence (boxes, points, or ordered point sequences), serialized as $Y = (y_1, \ldots, y_L)$. We use a shared vocabulary of text, structural markers, and quantized coordinates (Jiang et al., 2026) (I/O conventions: Section E.1). Token recovery need not follow causal order, while semantic order (e.g., trajectory time order) is preserved

(Figure 1). We use blockwise masked diffusion (Arriola et al., 2025; Wu et al., 2026a), modeling $p_\theta(Y \mid I, Q) = \prod_{k=1}^{K} p_\theta(Y^{(k)} \mid I, Q, Y^{(<k)})$ over $K$ consecutive blocks. Bidirectional attention enables parallel recovery within blocks, with completed blocks as a causal prefix.

**Model Architecture.** GroundAnything is a diffusion model directly converted from GroundAnything-VLM, an AR grounding model trained through multimodal and spatial pretraining, supervised fine-tuning, and reinforcement post-training. It combines a MoonViT-V2 (Kimi K3) visual encoder (Team et al., 2026), a projector with $2 \times 2$ spatial aggregation and a two-layer MLP, and a Qwen3-4B language decoder (Yang et al., 2025) (Figure 2). Variable-length outputs use a shared vocabulary of semantic labels, protocol markers, and 1,000 quantized coordinate tokens (`<0>`–`<999>`); absent queried categories receive a `None` payload. Conversion adds a mask token and jointly trains causal prediction and blockwise denoising ($B = 32$) with a shared decoder and vocabulary head, enabling parallel grounding and optional self-speculative decoding (Section 3.2). Training combines public datasets with annotations from our data engine; the complete data description and pipeline are provided in Section A.

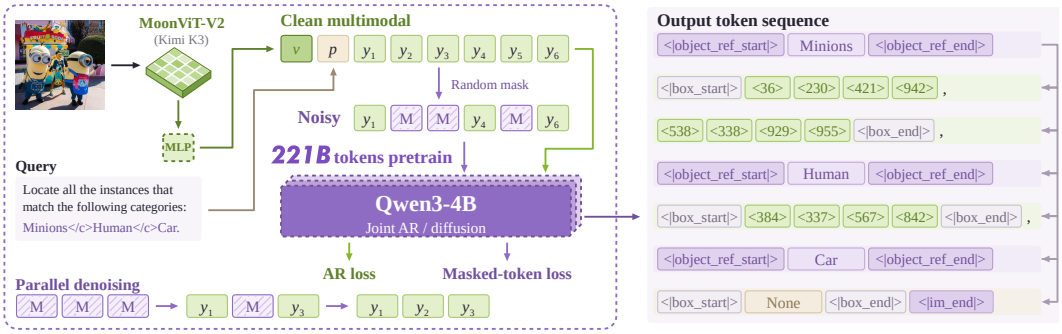


Figure 2: **GroundAnything architecture.** Clean multimodal inputs and response-corrupted text views share one visual encoding and jointly supervise causal prediction and masked denoising. One noisy view is illustrated, with its uncorrupted prompt copy omitted for clarity. The bottom-left panel illustrates parallel denoising, and the grounding example includes two minions and one human and an absent car category. Architecture specifications and parameter counts are provided in Section E.2.

**Efficient parallel inference.** After image/query prefill, GroundAnything denoises response blocks bidirectionally, processing sub-blocks left to right. Our default *Entropy-Guided Decoding* commits masked positions with entropy $H_j = -\sum_v p_j(v) \log p_j(v) \le \tau$, computed from unmodified token distributions. If none qualifies, it commits the lowest-entropy position. Committed tokens remain fixed; a causal pass caches completed blocks without AR verification. Dynamic and Static Decoding use confidence thresholds and fixed quotas, respectively (Section F.1). Shared weights support optional self-speculative decoding with causal verification (Section B).

### 3.2 Training Design

**Base VLM training.** Base pretraining has three stages. *Vision–language connector alignment* trains the projector on image–text pairs with both backbones frozen. *Joint multimodal pretraining* updates all modules using text and multimodal supervision. *General visual and video understanding* develops instruction following and image/video understanding. This autoregressive foundation supports subsequent spatial specialization through supervised fine-tuning (SFT) and reinforcement post-training. GroundAnything additionally undergoes the conversion below; training settings are provided in Sections E.3 and E.5.

**AR-to-diffusion conversion.** Following multimodal AR-to-diffusion conversion (Wu et al., 2026a; Dong et al., 2026), we retain the pretrained backbone and grounding vocabulary and add a mask token. For a clean sequence $x = [V, P, Y]$ of visual embeddings, text prompt, and response, we partition $Y$ into $B = 32$ token blocks, independently of geometric-tuple boundaries. Sampling $t_b \sim \mathcal{U}(0, 1)$ independently per block, we corrupt only response positions to form complementary noisy views $w^{(a)} = [P, \widetilde{Y}^{(a)}]$, $a \in \{1, 2\}$; EOS is masked in both. The joint input $[w^{(1)}; w^{(2)}; x]$ contains visual embeddings only in $x$.

Under the mask in Figure 3, each noisy block attends bidirectionally within itself and reads clean image/prompt context and strictly preceding clean response blocks; clean tokens attend causally without accessing noisy views (Figure 3). The shared decoder minimizes $\mathcal{L} = 0.5(\mathcal{L}_{\text{MDM}} + \mathcal{L}_{\text{AR}})$, combining masked response prediction and clean next-token prediction, both retaining the pretrained token shift. Continued SFT uses the same objective, first updating all modules and then only the language parameters (details: Sections E.4 and E.5).

**Reinforcement post-training.** We apply GRPO (Shao et al., 2024) with causal rollouts and exact causal token likelihoods (Figure 4). Box grounding combines set-completeness $R_{\text{set}}$ and strict localization $R_{\text{strict}}$ with weights $(0.7, 0.3)$; OCR scores annotation-aware text–geometry agreement; pointing scores location, count, label, and format. Advantages are $A_i = Z_G(\sum_c w_c Z_G(R_{c,i}))$, where $Z_G$ standardizes within each prompt group; OCR and pointing each supply one scalar reward. We minimize $\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_i[|\mathcal{T}_i|^{-1} \sum_{t \in \mathcal{T}_i} \min(r_{i,t}A_i, \text{clip}(r_{i,t}, 0.8, 1.2)A_i)]$, where $r_{i,t}$ is the current-to-old causal token-probability ratio at rollout temperature and $\mathcal{T}_i$ indexes valid response tokens including EOS. One update per rollout trains the projector and language parameters with the vision encoder frozen and no reference-policy KL term; the updated weights are reused for diffusion inference. Reward and training details appear in Sections E.6 to E.8.

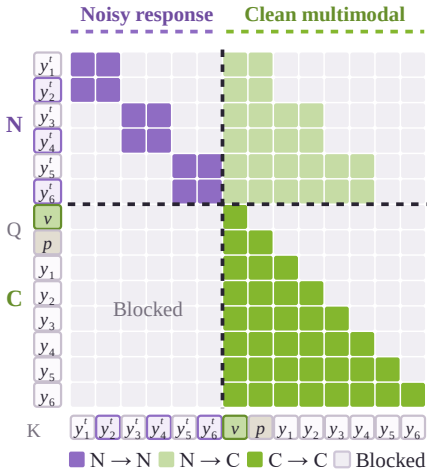


Figure 3: **Conversion attention mask.** Rows/columns are queries/keys. Purple denotes within-block bidirectionality, light green the permitted clean prefix, and green clean-stream causality. Other entries are blocked. The symbols $v$ and $p$ condense visual and prompt tokens. One noisy view with two-token blocks is shown; noisy prompt copies are omitted. Training uses $B = 32$.

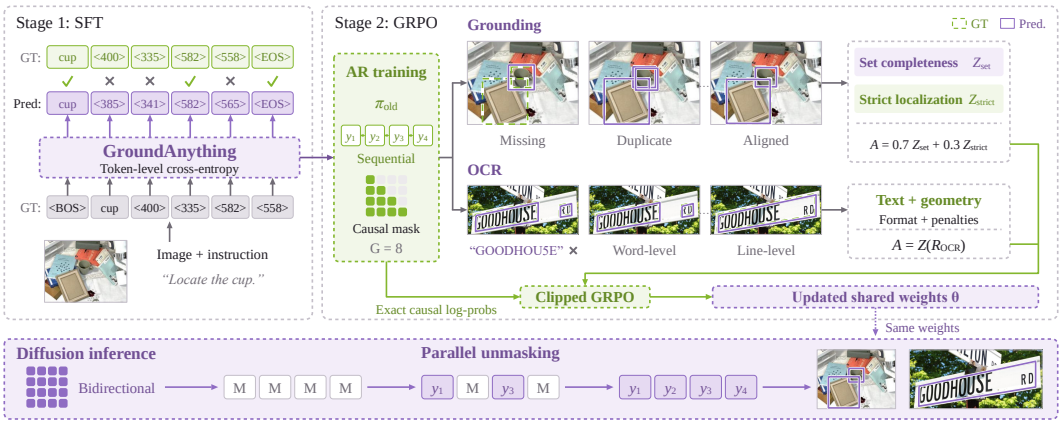


Figure 4: **Training and inference pipeline.** SFT and causal GRPO update shared parameters reused for blockwise diffusion inference. The schematic illustrates token-level supervision and grounding/OCR reward aggregation; the final advantage includes the additional group normalization in Equation (S12). Rollouts and unmasking states are illustrative.

## 4 Experiments

We ask: **Grounding capability:** how far can GroundAnything push the breadth and precision of visual grounding? **Speed:** how far can parallel grounding accelerate inference with minimal loss of precision, and what makes its speedups practical? **Design:** which model, training, and output choices support precise and efficient parallel grounding?

Table 1: **Detection and referring grounding (F1mIoU).** External rows follow Jiang et al. (2026). Full results appear in Sections H.1 to H.3.

| Model | Common | Long-tailed | Dense & Tiny | | Referring grounding | | |
|---|---|---|---|---|---|---|---|
| | COCO | LVIS | Dense200 | VisDrone | RefCOCOg val | RefCOCOg test | RefCOCO avg |
| **Open-set Specialized Detectors** | | | | | | | |
| GroundingDINO (Liu et al., 2024) | 60.56 | 52.61 | 24.92 | 34.47 | 49.77 | 50.43 | 45.15 |
| **Vision-Language Models (<10B)** | | | | | | | |
| Qwen3-VL-4B (Bai et al., 2025a) | 46.53 | 49.86 | 14.02 | 31.14 | 75.27 | 75.88 | 75.24 |
| Rex-Omni (Jiang et al., 2026) | 56.28 | 46.74 | 53.29 | 27.19 | 73.90 | 74.76 | 69.29 |
| LocateAnything Fast (Wang et al., 2026) | 53.06 | 42.90 | 20.70 | 9.82 | 75.30 | 76.50 | 76.62 |
| LocateAnything Hybrid (Wang et al., 2026) | 59.12 | 49.56 | 50.07 | 28.57 | 76.43 | 77.67 | 78.06 |
| SenseNova-Vision (Han et al., 2026) | 57.49 | 56.12 | 68.13 | **42.35** | 78.69 | 79.48 | 77.55 |
| GroundAnything-VLM | **63.70** | **56.63** | **75.55** | 40.74 | 84.37 | 83.56 | 83.43 |
| GroundAnything | 60.72 | 52.59 | 70.04 | 33.76 | **91.61** | **91.10** | **85.33** |
| **Vision-Language Models (>1T)** | | | | | | | |
| Qwen3.7-Max (Qwen Team, 2026c) | 62.79 | 53.28 | 31.20 | 41.59 | 80.54 | 81.71 | 80.20 |
| Kimi-K2.6 | 61.16 | 51.19 | 42.67 | 29.95 | 70.33 | 71.87 | 69.66 |
| GPT-6 Astra | 62.75 | 54.97 | 65.04 | 37.12 | 74.98 | 78.91 | 77.81 |

### 4.1 Main Results.

**Evaluation setup.** We evaluate both variants on 30 grounding benchmarks; the main-table GroundAnything model uses entropy-guided decoding without AR verification. All our model results average ten runs (five random seeds, two runs each); baselines evaluated locally for the main leaderboard average three runs (three random seeds, one run each). GPT-6 Astra uses `High` thinking effort. Training and evaluation details are in Sections E and H.

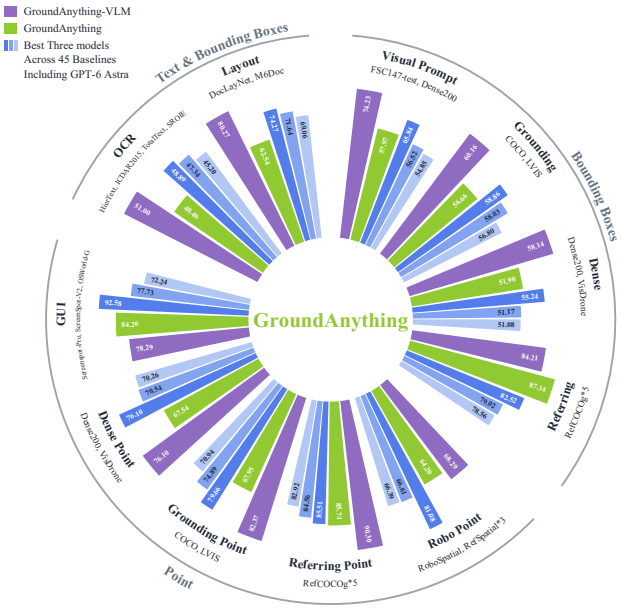


Figure 5: **Grounding across task groups.** GroundAnything-VLM, GroundAnything, and leading baselines across box, point, text, and exemplar-conditioned tasks. Complete results: Section H.

**Overall performance.** GroundAnything-VLM leads similarly sized models overall at 72.42%, remaining competitive with GPT-6 Astra (71.35%). GroundAnything reaches 61.75% with pure diffusion decoding, surpassing the previous overall AR best at comparable scale (Figure 5). **Detection and referring grounding.** GroundAnything reaches 70.04 F1mIoU on Dense200 and 91.61/91.10 on RefCOCOg val/test (Table 1); tiny objects remain challenging. **Robot, spatial, and GUI grounding.** GroundAnything matches its AR counterpart on RefSpatial Unseen (62.34%) and improves ScreenSpot-Pro from 65.34% to 75.96%; Astra remains stronger on RefSpatial and GUI tasks (Table 2). **OCR, layout, and visual prompting.** GroundAnything-VLM scores 85.78 on DocLayNet and 74.76 on M6Doc; GroundAnything achieves 63.52 on Dense200 visual prompting (Table 3). OCR and layout show larger diffusion–AR gaps than referring and GUI grounding. **Object pointing.** GroundAnything-VLM leads five of six selected comparisons.

***Finding 1:*** Across 30 benchmarks, GroundAnything extends broad and precise grounding to pure diffusion, surpassing the prior overall AR state of the art at comparable scale. GroundAnything-VLM sets a new overall best at this scale, remaining competitive with GPT-6 Astra.

Table 2: **Robot, spatial, and GUI grounding (accuracy).** External RefSpatial and JEDI/UI-R1 scores follow Jiang et al. (2026); GUI-Owl scores follow Wang et al. (2026). Full results appear in Sections H.5 and H.7.

| Model | Robot and spatial pointing | | | GUI grounding | | |
|---|---|---|---|---|---|---|
| | RefSpatial (avg) | RefSpatial Unseen | RoboSpatial Context | ScreenSpot-Pro | ScreenSpot-V2 | OSWorld-G |
| **Open-set Specialized Detectors** | | | | | | |
| GroundingDINO (Liu et al., 2024) | 14.25 | 4.33 | 4.92 | N/A* | N/A* | N/A* |
| **Vision-Language Models (<10B)** | | | | | | |
| Qwen3-VL-4B (Bai et al., 2025a) | 49.00 | 27.27 | 64.75 | 56.74 | 92.30 | 56.91 |
| Rex-Omni (Jiang et al., 2026) | 51.75 | 37.01 | 59.02 | 36.75 | 88.29 | 46.10 |
| LocateAnything Fast (Wang et al., 2026) | 36.00 | 16.88 | 15.57 | 56.04 | 88.44 | 59.93 |
| LocateAnything Hybrid (Wang et al., 2026) | 36.17 | 20.78 | 14.75 | 57.05 | 89.94 | 60.46 |
| SenseNova-Vision (Han et al., 2026) | 20.38 | 8.54 | 0.82 | N/A† | N/A† | N/A† |
| GroundAnything-VLM | 69.00 | 62.34 | **72.13** | 65.34 | 94.89 | 74.65 |
| GroundAnything | 63.00 | 62.34 | 69.67 | 75.96 | 95.60 | 81.21 |
| **Vision-Language Models (>1T)** | | | | | | |
| Qwen3.7-Max (Qwen Team, 2026c) | 68.75 | 57.14 | 69.67 | 55.06 | 81.13 | 49.29 |
| Kimi-K2.6 | 41.33 | 41.56 | 30.33 | 6.07† | 52.36† | 10.11† |
| GPT-6 Astra | **85.93** | **81.93** | 65.69 | **93.17** | **97.88** | **86.70** |

## 4.2 Speed Evaluation

### 4.2.1 Decoding Strategy Analysis

***Question 2.1:*** Which decoding choices govern the speed–accuracy trade-off?

We report tokens per forward (TPF), output tokens per second (TPS), and COCO F1mIoU. Decoding protocols and additional analysis appear in Section G.1.




Figure 6: **Self-speculative schedules.** TPF and TPS across block sizes; dashed curves show the GroundAnything-VLM baseline in every panel.

***Finding 2.1:*** In direct DLM decoding, the entropy threshold controls the speed–precision trade-off, providing a decoding-effort knob analogous to thinking effort. We favor $\tau = 0.8$ in the $B = 32$ COCO sweep. Optional self-speculation improves both quality and throughput over this direct-decoding point, reaching $4.51\times$ the AR throughput with a 0.74 percentage-point F1mIoU.

**Forward efficiency versus latency.** At $B = 32$, quadratic self-speculation has higher TPF than the linear schedule (3.17 versus 2.18) but lower throughput (49.9 versus 223.8 TPS; Figure 6). Its $O(B^2)$ query-token cost offsets fewer forward calls, so the linear schedule is preferable here.

**Entropy threshold and block size.** Relaxing the entropy threshold increases parallel commitment, but COCO quality drops beyond $\tau = 0.8$ (Figure 7). At $B = 32$, $\tau = 0.8$ gives the best F1mIoU in this sweep (60.72) at 127.0 TPS, or $2.56\times$ GroundAnything-VLM throughput. Entropy-guided quality peaks at $B = 16$, whereas linear self-speculation peaks at $B = 32$ (Figure 8); larger blocks are not uniformly better.

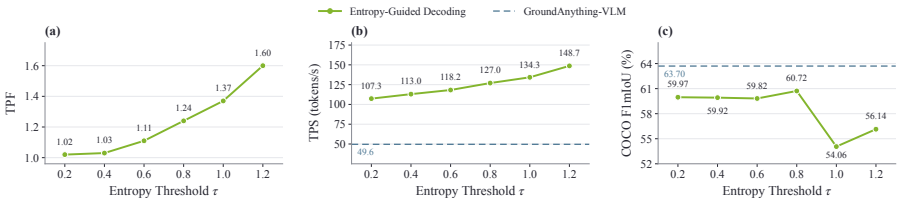


Figure 7: **Entropy-threshold sensitivity.** TPF, TPS, and COCO F1mIoU, at $B = 32$ with a sub-block size 4.

Table 3: **OCR, layout, and visual prompting (F1mIoU).** External rows follow Jiang et al. (2026). SenseNova-Vision uses published HierText/ICDAR2015 scores (Han et al., 2026) and locally evaluated TotalText/SROIE scores. Full results appear in Sections H.6, H.8 and H.9.

| Model | OCR | | | | Layout grounding | | Visual prompting | |
|---|---|---|---|---|---|---|---|---|
| | HierText | ICDAR2015 | TotalText | SROIE | DocLayNet | M6Doc | FSC147 | Dense200 |
| **Vision-Language Models (<10B)** | | | | | | | | |
| Qwen3-VL-4B (Bai et al., 2025a) | 23.48 | 28.41 | 38.35 | 40.41 | 40.81 | 24.73 | 18.90 | 2.49 |
| Rex-Omni (Jiang et al., 2026) | 34.46 | 45.65 | 52.35 | 48.35 | 68.06 | 54.95 | 57.15 | 55.50 |
| LocateAnything Fast (Wang et al., 2026) | 22.59 | 26.73 | 44.40 | 24.89 | 49.37 | 47.03 | N/A* | N/A* |
| LocateAnything Hybrid (Wang et al., 2026) | 26.65 | 27.48 | 45.49 | 30.05 | 77.34 | 65.94 | N/A* | N/A* |
| SenseNova-Vision (Han et al., 2026) | 31.20* | **49.50*** | 11.40 | 36.26 | 85.53 | 35.62 | **62.51** | 62.84 |
| GroundAnything-VLM | 41.33 | 42.50 | 49.03 | **71.15** | **85.78** | **74.76** | 60.29 | **75.26** |
| GroundAnything | 33.31 | 41.81 | 43.08 | 43.64 | 68.38 | 56.69 | 52.43 | 63.52 |
| **Vision-Language Models (>1T)** | | | | | | | | |
| Qwen3.7-Max (Qwen Team, 2026c) | **42.95** | 43.07 | 48.64 | 38.54 | 46.93 | 32.31 | 21.62 | 31.81 |
| Kimi-K2.6 | 25.26† | 32.25† | 40.99† | 46.83† | 15.59† | 13.50† | 32.47 | 32.41 |
| GPT-6 Astra | 39.58 | 48.87 | **53.55** | 53.57 | 77.54 | 60.59 | 61.04 | 68.94 |

Table 4: **Object pointing (F1@Point).** External rows follow Jiang et al. (2026), where Molmo denotes Molmo-7B-D. Full results appear in Section H.4.

| Model | Referring object pointing | | Common / long-tailed | | Dense / tiny | |
|---|---|---|---|---|---|---|
| | RefCOCOg val | RefCOCOg test | COCO | LVIS | Dense200 | VisDrone |
| **Open-set Specialized Detectors** | | | | | | |
| GroundingDINO (Liu et al., 2024) | 49.34 | 49.97 | 70.41 | 55.07 | 32.93 | 39.45 |
| **Vision-Language Models (<10B)** | | | | | | |
| Qwen3-VL-4B (Bai et al., 2025a) | 76.43 | 77.64 | 65.33 | 55.08 | 21.72 | 23.50 |
| Rex-Omni (Jiang et al., 2026) | 84.96 | 85.32 | 79.74 | 70.04 | 76.66 | 51.97 |
| LocateAnything Fast (Wang et al., 2026) | 73.84 | 74.94 | 71.53 | 62.13 | 65.63 | 56.36 |
| LocateAnything Hybrid (Wang et al., 2026) | 75.89 | 76.65 | 73.78 | 64.89 | 78.07 | 57.30 |
| SenseNova-Vision (Han et al., 2026) | 74.63 | 75.42 | 72.96 | 62.66 | 78.71 | 61.81 |
| Molmo-7B* (Deitke et al., 2025) | 83.70 | 83.60 | 77.30 | 40.30 | 33.10 | 29.20 |
| GroundAnything-VLM | **90.44** | **91.03** | **84.92** | **79.81** | 84.16 | **68.03** |
| GroundAnything | 85.75 | 85.73 | 71.32 | 64.57 | 77.14 | 57.94 |
| **Vision-Language Models (>1T)** | | | | | | |
| Qwen3.7-Max (Qwen Team, 2026c) | 71.21 | 72.40 | 72.13 | 66.94 | 66.28 | 59.04 |
| Kimi-K2.6 | 39.59† | 39.60† | 30.65† | 25.51† | 35.18† | 13.63† |
| GPT-6 Astra | 87.80 | 84.90 | 82.17 | 77.14 | **86.57** | 65.62 |

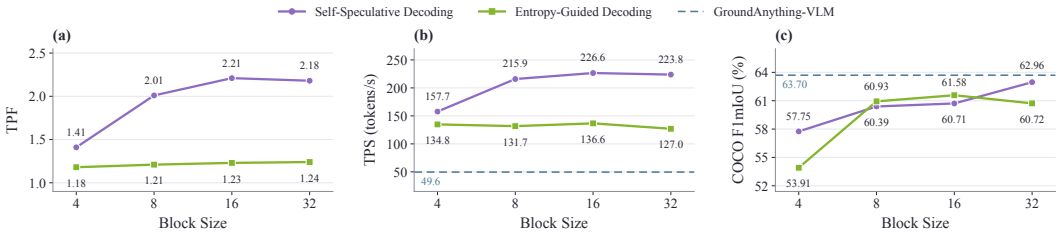


Figure 8: **Decoding strategy and block size.** Linear self-speculation and entropy guidance ($\tau = 0.8$, sub-block size 4) compared on COCO. Dashed lines show the GroundAnything-VLM baseline for visual reference.

### 4.2.2 DLM vs. MTP

With matched visual inputs, adaptation data, tokenization, prefixes, and a frozen shared AR verifier, four questions distinguish diffusion from causal MTP (Section C). **(1) At the same candidate length, which drafter achieves higher acceptance?** At $K = 16$, one-step DLM accepts 3.48 draft tokens per round versus 2.69 for causal MTP. **(2) How much wall-clock time does each accepted draft token cost?** Including drafting, verification, and rejected work, DLM reduces this cost from 7.71 to 5.38 ms. **(3) Can additional DLM refinement repay its extra forward-pass cost?** A second DLM pass reduces the cost to 4.70 ms; further passes increase it. MTP remains preferable when latency budgets are extremely tight. **(4) Which grounding failures benefit from bidirectional at-**

**tention?** On Dense200, bidirectional DLM reduces unmatched valid proposals from 24.89% for causal MTP to 14.79%, versus 21.74% for retrained causal DLM. Bidirectional block-MTP also improves, so the benefit is not unique to diffusion.

These gains concern draft efficiency and reliability; exact verification preserves the shared AR verifier's completed output.

#### 4.2.3 Infrastructure Analysis

***Question 2.2:*** How can parallel decoding translate into practical inference speedups?

Progressive integration with SGLang, CUDA Graph, and FP8 reduces time per output token across all four decoding modes (Figure 9). The complete stack yields 3.45–4.06× speedups over each mode's native PyTorch implementation, with self-speculation reaching 3.73 ms/token. Implementation gains are measured within each decoding mode, not algorithmic speedups (Section G.2).

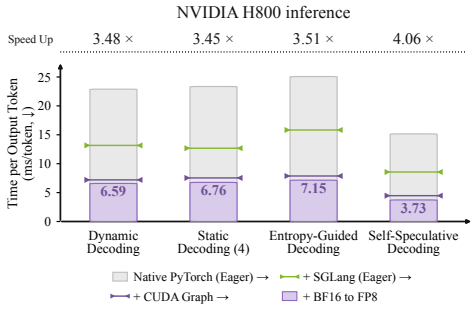


Figure 9: **Progressive inference optimization.** Time per output token (↓) and cumulative speedup relative to each decoding mode's native PyTorch implementation.

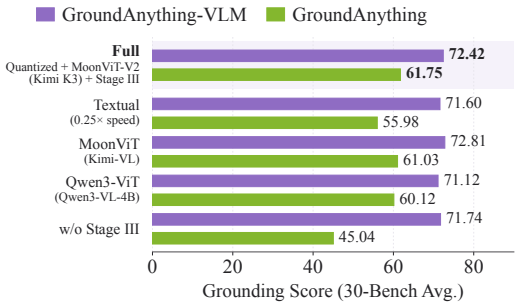


Figure 10: **Architecture and training ablations.** Coordinate representation, visual encoder, and Stage III with ablation protocol (Section G.3).

### 4.3 Ablation Studies

**Coordinate representation.** Quantized coordinates improve the reported ablation average by 1.43 pp for GroundAnything-VLM and 5.68 pp for GroundAnything over textual coordinates (Figure 10). **ViT choice.** MoonViT-V2 gives the strongest AR and DLM scores among the tested encoders. Section G.3.1 analyzes when multi-level visual evidence can help coordinate prediction; the encoder ablation supports our choice under this recipe. **Training.** Removing Stage III causes the largest DLM drop (16.62 pp). Output serialization and generation-cost analysis are provided in Section D.

### 4.4 Discussion

***Takeaway 1:*** Grounding as visual evidence extraction naturally suits bidirectional diffusion: related spatial fields share the same image evidence and can be filled jointly, then selectively refined before commitment. This connects the task's structure to parallel decoding.

***Takeaway 2:*** Diffusion is a generation strategy within a Transformer backbone (Wu et al., 2026a). Its distinctive value lies in suitable workloads, such as parallel filling and structured outputs, where joint prediction and limited refinement can repay their computational cost.

## 5 Conclusion

We introduced GroundAnything, reconciling fast parallel decoding with broad and precise visual grounding. Structured grounding naturally suits DLMs: its predictions are jointly constrained by shared visual evidence. Diffusion complements the Transformer backbone through a generation process with distinctive strengths in parallel filling and structured outputs. Causal conditioning supports long outputs across blocks, while diffusion accelerates prediction within them. GroundAnything thus highlights a broader principle: generation should follow task structure, and spatial evidence need not be inferred in the order it is serialized.

### AI USE STATEMENT

In this work, we used generative AI tools to assist with code development, to polish the writing, and to produce some of the figures. We also used generative AI models as part of the data annotation pipeline to generate pseudo-labels for model training. We did not use generative AI tools to develop the research ideas or methodology, to design or interpret the experiments, or to outline the paper. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, data annotations, or artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This work does not involve human-subject studies. All datasets and evaluation benchmarks used in this work were obtained from publicly available or appropriately licensed sources in accordance with their respective terms and licenses. The experiments focus on visual grounding and structured visual understanding and do not involve real-world robotic control, autonomous driving, or interaction with people. GroundAnything is developed as a general-purpose visual grounding model for research rather than for safety-critical real-world deployment. Any future integration into embodied or autonomous systems should be subject to appropriate safety evaluation and human oversight. The authors declare no conflicts of interest.

### REPRODUCIBILITY STATEMENT

We will publicly release the code, model weights of GroundAnything and GroundAnything-VLM, and evaluation data to support systematic and reproducible evaluation. The release will include training and decoding configurations, inference implementations, and evaluation scripts with standardized task prompts, output parsers, and metric implementations. The shared architecture, tokenizer, visual processing, and input/output protocol are specified in Sections E.1 and E.2; the data engine and validation pipeline in Section A.1; and base VLM training, AR-to-diffusion conversion, and continued supervised fine-tuning in Sections E.3 to E.5. Variant-specific GRPO configurations and task rewards are detailed in Sections E.6 to E.8. Direct diffusion decoding, optional self-speculative decoding, and execution optimizations are described in Sections B.1, F.1 and F.2. Controlled DLM–MTP comparisons are documented in Section C, while speed metrics, decoding sensitivity, and infrastructure comparisons appear in Sections G.1 and G.2. Architecture and training ablations and output-token efficiency analyses are provided in Sections D and G.3. The benchmark suite, evaluation metrics, reporting conventions, and complete task-level results are given in Section H.

## REFERENCES

Marianne Arriola, Aaron Gokaslan, Justin Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. In *International Conference on Learning Representations*, volume 2025, pp. 50726–50753, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Heli Ben-Hamu, Itai Gat, Daniel Severo, Niklas S Nolte, and Brian Karrer. Accelerated sampling from masked diffusion models via entropy bounded unmasking. *Advances in Neural Information Processing Systems*, 38:55981–56007, 2026.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection, 2022. URL https://arxiv.org/abs/2109.10852.

Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. M$^6$doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis, 2023. URL https://arxiv.org/abs/2305.08719.

Zesen Cheng, Kehan Li, Peng Jin, Siheng Li, Xiangyang Ji, Li Yuan, Chang Liu, and Jie Chen. Parallel vertex diffusion for unified visual grounding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 1326–1334, 2024.

Chee Kheng Ch'Ng and Chee Seng Chan. Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pp. 935–942. IEEE, 2017.

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.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. Paddleocr 3.0 technical report, 2025. URL https://arxiv.org/abs/2507.05595.

Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7369–7378. ieee, 2021.

Ronghao Dang, Jiayan Guo, Bohan Hou, Sicong Leng, Kehan Li, Xin Li, Jiangpin Liu, Yunxuan Mao, Zhikai Wang, Yuqian Yuan, Minghao Zhu, Xiao Lin, Yang Bai, Qian Jiang, Yaxi Zhao, Minghua Zeng, Junlong Gao, Yuming Jiang, Jun Cen, Siteng Huang, Liuyi Wang, Wenqiao Zhang, Chengju Liu, Jianfei Yang, Shijian Lu, and Deli Zhao. Rynnbrain: Open embodied foundation models, 2026. URL https://arxiv.org/abs/2602.14979.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 91–104. IEEE, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025. URL https://arxiv.org/abs/2505.14683.

Hejun Dong, Junbo Niu, Bin Wang, Weijun Zeng, Wentao Zhang, and Conghui He. Mineru-diffusion: Rethinking document ocr as inverse rendering via diffusion decoding. In *European Conference on Computer Vision*, pp. 439–456. Springer, 2026.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5351–5359. IEEE, 2019.

Xiaoyang Han, Jianhua Li, Kewang Deng, Zukai Chen, Xuanke Shi, Sihan Wang, Boxuan Li, Linyan Wang, Siyi Xie, Xin You, Jinsheng Quan, Zhongang Cai, Haiwen Diao, Ziwei Liu, Lei Yang, Dahua Lin, and Quan Wang. Vision as unified multimodal generation, 2026. URL https://arxiv.org/abs/2607.06560.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1516–1520. IEEE, 2019.

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.

Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1156–1160, 2015. doi: 10.1109/ICDAR.2015.7333942.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Shrinidhi Kumbhar, Haofu Liao, Srikar Appalaraju, and Kunwar Yashraj Singh. Towards gui agents: Vision-language diffusion models for gui grounding, 2026. URL https://arxiv.org/abs/2603.26211.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. Screenspot-pro: Gui grounding for professional high-resolution computer use. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 8778–8786, 2025.

Kehan Li, Bohan Hou, Minghao Zhu, Tianyi Zhang, Zesen Cheng, Zhikai Wang, Sicong Leng, Xin Li, Xiao Lin, Biying Yao, Minghua Zeng, Jiangpin Liu, Ronghao Dang, Jiayan Guo, Siteng Huang, Haoyu Zhao, Heng Ping, Yaxi Zhao, Tong Zhao, Kexiang Wang, Tong Lu, Shengke Xue, Jiahao Tang, Yulei Wang, Zejing Wang, Jianwei Gao, Shijian Lu, Chengju Liu, Jianfei Yang, Mingxiu Chen, and Deli Zhao. Rynnbrain 1.1: Towards more capable and generalizable embodied foundation model, 2026a. URL https://arxiv.org/abs/2607.17977.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10955–10965. IEEE, 2022.

Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. Lavida: A large diffusion language model for multimodal understanding. *Advances in Neural Information Processing Systems*, 38: 105101–105134, 2026b.

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.

Aiwei Liu, Minghua He, Shaoxun Zeng, Sijun Zhang, Linhao Zhang, Chuhan Wu, Wei Jia, Yuan Liu, Xiao Zhou, and Jie Zhou. Wedlm: Reconciling diffusion language models with standard causal attention for fast inference, 2025. URL https://arxiv.org/abs/2512.22737.

Jingyu Liu, Xin Dong, Zhifan Ye, Rishabh Mehta, Yonggan Fu, Vartika Singh, Ce Zhang, and Pavlo Molchanov. Tidar: Think in diffusion, talk in autoregression. *Proceedings of Machine Learning and Systems*, 8:748–762, 2026.

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, 2024.

Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, and Michalis Raptis. Towards end-to-end unified scene text detection and layout analysis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1039–1049. IEEE, 2022.

Zhengxi Lu, Yuxiang Chai, Yaxuan Guo, Xi Yin, Liang Liu, Hao Wang, Han Xiao, Shuai Ren, Pengxiang Zhao, Guangyi Liu, et al. Ui-r1: Enhancing efficient action prediction of gui agents by reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17608–17616, 2026.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 11–20, 2016.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. *Advances in Neural Information Processing Systems*, 37:23464–23487, 2024.

Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *European conference on computer vision*, pp. 792–807. Springer, 2016.

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, 2023. URL https://arxiv.org/abs/2306.14824.

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter Staar. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 3743–3751, 2022.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, Wanjun Zhong, Kuanye Li, Jiale Yang, Yu Miao, Woyu Lin, Longxiang Liu, Xu Jiang, Qianli Ma, Jingyu Li, Xiaojun Xiao, Kai Cai, Chuang Li, Yaowei Zheng, Chaolin Jin, Chen Li, Xiao Zhou, Minchao Wang, Haoli Chen, Zhaojian Li, Haihua Yang, Haifeng Liu, Feng Lin, Tao Peng, Xin Liu, and Guang Shi. Ui-tars: Pioneering automated gui interaction with native agents, 2025. URL https://arxiv.org/abs/2501.12326.

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

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-27b.

Qwen Team. Qwen3.7: The agent frontier, May 2026c. URL https://qwen.ai/blog?id=qwen3.7.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026d. URL https://qwen.ai/blog?id=qwen3.8.

Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3393–3402. IEEE, 2021.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15768–15780. IEEE, 2025.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Shihao Wang, Shilong Liu, Yuanguo Kuang, Xinyu Wei, Yangzhou Liu, Zhiqi Li, Yunze Man, Guo Chen, Andrew Tao, Guilin Liu, et al. Locateanything: Fast and high-quality vision-language grounding with parallel box decoding. In *European Conference on Computer Vision*, pp. 336–357. Springer, 2026.

Chengyue Wu, Shiyi Lan, Yonggan Fu, Sensen Gao, Jin Wang, Jincheng Yu, Jose M. Alvarez, Pavlo Molchanov, Ping Luo, Song Han, Ligeng Zhu, and Enze Xie. Fast-dvlm: Efficient block-diffusion vlm via direct conversion from autoregressive vlm, 2026a. URL https://arxiv.org/abs/2604.06832.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. In *International Conference on Learning Representations*, volume 2026, pp. 57027–57051, 2026b.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, et al. Os-atlas: Foundation action model for generalist gui agents. In *International Conference on Learning Representations*, volume 2025, pp. 5090–5108, 2025.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4818–4829. IEEE, 2024.

Tianbao Xie, Jiaqi Deng, Xiaochuan Li, Junlin Yang, Haoyuan Wu, Jixuan Chen, Wenjing Hu, Xinyuan Wang, Yuhui Xu, Zekun Wang, et al. Scaling computer-use grounding via user interface decomposition and synthesis. *Advances in Neural Information Processing Systems*, 38, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jiabo Ye, Xi Zhang, Haiyang Xu, Haowei Liu, Junyang Wang, Zhaoqing Zhu, Ziwei Zheng, Feiyu Gao, Junjie Cao, Zhengxi Lu, Jitong Liao, Qi Zheng, Fei Huang, Jingren Zhou, and Ming Yan. Mobile-agent-v3: Fundamental agents for gui automation, 2025. URL https://arxiv.org/abs/2508.15144.

Zebin You, Shen Nie, Xiaolu Zhang, JUN ZHOU, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. Llada-v: Large language diffusion models with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10093–10105, 2026.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pp. 69–85. Springer, 2016.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics, 2024. URL https://arxiv.org/abs/2406.10721.

Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, et al. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 2025.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. URL https://arxiv.org/abs/2203.03605.

Zhiyuan Zhao, Hengrui Kang, Bin Wang, and Conghui He. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception, 2024. URL https://arxiv.org/abs/2410.12628.

Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. *Advances in Neural Information Processing Systems*, 38:28404–28481, 2026.

Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, and Qinghua Hu. Vision meets drones: A challenge, 2018. URL https://arxiv.org/abs/1804.07437.

# APPENDIX

## A GROUNDANYTHING DATA

Training data combines public datasets, summarized in Figure S1, with annotations produced by data engine.

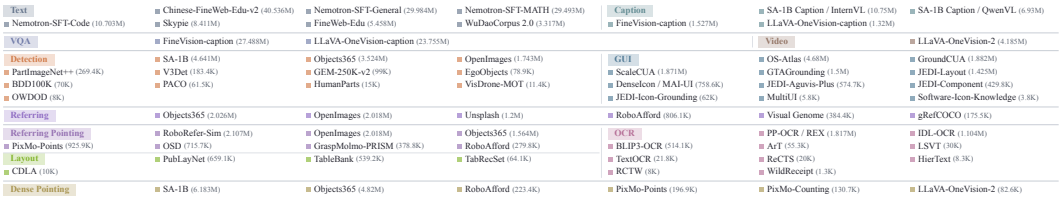


Figure S1: Public datasets grouped by supervision type.

Our data engine (Figure S2) fuses multi-teacher annotations at the field level and applies task-specific validation. Accepted labels train a unified grounding expert for iterative annotation; complementary teachers and local observations resolve uncertain cases. This process expands supervision while refining existing labels. Field validity and query-level coverage are checked separately, so an empty result alone does not establish absence. Fusion, validation, and expert iteration are detailed in Section A.1.

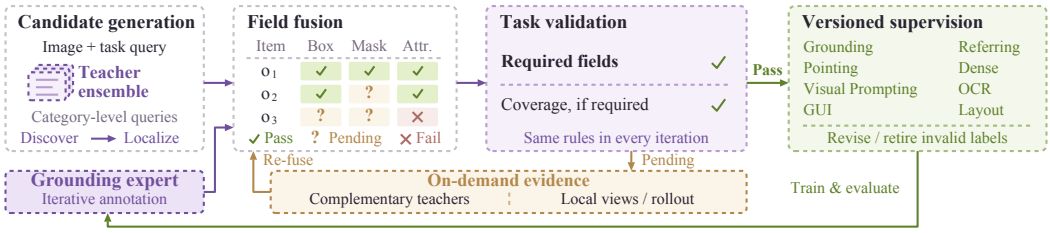


Figure S2: **Data engine.** Field-level teacher fusion and task-specific validation guide expert iteration, with targeted observations resolving uncertain annotations.

### A.1 EXPERT-DRIVEN DATA ENGINE

Field-level fusion. Complementary teachers provide semantic, localization, segmentation, text, interface, and layout evidence. Category discovery precedes category-conditioned localization; category names are normalized while attributes needed for referring expressions are retained. Predictions are mapped back to the original image and aligned only when their query scope and annotation granularity agree. Fusion operates on individual fields, allowing reliable text or geometry to be retained without accepting an entire inconsistent annotation.

Validation and coverage. Task-specific checks assess geometric validity, semantic consistency, text–region agreement, and cross-view support. Each dependency has a validation state $a(f) \in \{\text{pass}, \text{fail}, \text{pending}\}$; query-level coverage $c(I, Q)$ uses the same state space. Coverage checks whether the required query scope has been examined and relevant unresolved proposals and matching conflicts have been addressed. Unresolved dependencies trigger complementary teachers or targeted local observations.

Required dependencies and acceptance. Let $\mathcal{F}(I, Q, Y)$ contain the fields and semantic facts required to validate candidate $Y$, including query–target correspondence. The task protocol and query semantics determine these requirements, which are then instantiated for the candidate. Missing required fields remain in $\mathcal{F}$ with status pending: omitting a field cannot remove its validation requirement. Let $h(Q) \in \{0, 1\}$ indicate whether the query requires exhaustive instance coverage. Acceptance requires

$$\mathcal{A}(I, Q, Y) = \mathbf{1}[\forall f \in \mathcal{F}(I, Q, Y),\ a(f) = \text{pass}]\, \mathbf{1}[h(Q) = 0 \ \vee\ c(I, Q) = \text{pass}]\,, \tag{S1}$$

so every required dependency must pass, and exhaustive queries additionally require validated coverage. For example, validating the returned boxes does not establish that every queried instance has been found; an all-instance query remains unaccepted while coverage is pending.

Empty-target claims. An answer with no instances still requires verification of its query–answer relation. Its dependency set explicitly includes a query-scoped absence fact $f_{\text{abs}} \in \mathcal{F}(I, Q, Y)$, so an empty prediction never yields an empty dependency set. An unsupported absence claim remains pending; contrary evidence makes it fail. It passes only when absence is verified and $c(I, Q) = \text{pass}$, even when $h(Q) = 0$. Missing outputs, failed annotation runs, and unresolved empty predictions do not constitute evidence of absence. These requirements prevent empty predictions from being accepted through a vacuously true check over an empty set.

Derived box, point, referring, and visual-prompt supervision preserves its dependencies on the validated source evidence; crops and coordinate transforms remain consistent with visible content.

Expert iteration. Accepted annotations train a unified grounding expert, which then proposes annotations for further validation. Difficult cases receive targeted crops or complementary teacher evidence. New evidence can correct or retire previous labels, with versioned annotation snapshots keeping each training input fixed. This loop improves both coverage and field consistency rather than merely accumulating predictions.

## B Optional Self-Speculative Decoding

The shared weights also support diffusion drafting with causal verification (Wu et al., 2026a), accepting the longest matching prefix (Figure S3). Linear decoding uses two network function evaluations (NFEs) and $2B$ query tokens per round; quadratic decoding fuses verification and proposal generation in one NFE after initialization, using $B(B+1)$ query tokens. These counts describe queries, not Transformer FLOPs. Acceptance and proposal reuse are detailed in Section B.1; CUDA Graph and selective FP8 optimizations appear in Section F.2.

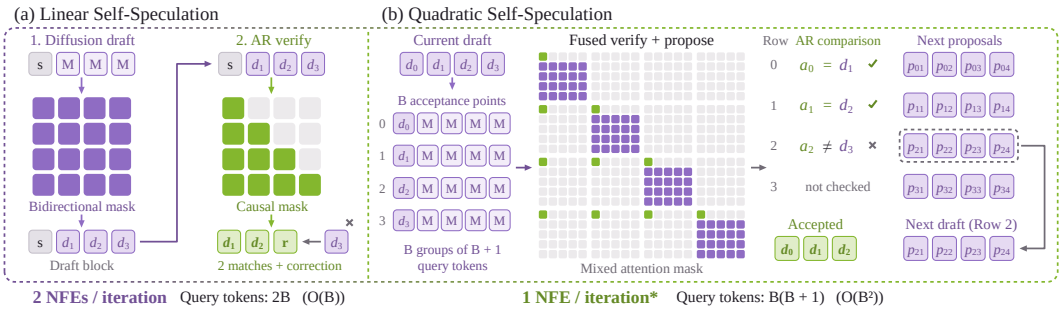


Figure S3: **Self-speculative decoding with shared model weights ($B = 4$).** Linear decoding separates drafting from causal verification: $d_1$ and $d_2$ are accepted, while the mismatched $d_3$ is replaced by the AR correction $r$. Quadratic decoding compares $a_i$ with $d_{i+1}$: $a_0 = d_1$, $a_1 = d_2$, and $a_2 \neq d_3$, thereby committing $[d_0, d_1, d_2]$ and reusing Row 2 as the next still-unverified proposal. Purple denotes bidirectional proposals and green causal verification. Matrix rows and columns denote queries and keys, respectively; the cached prefix is omitted. $M$ denotes `[MASK]`, $s$ the last accepted seed, and $p_{ij}$ proposal token $j$ from row $i$. One NFE per quadratic round applies after initialization.

### B.1 Acceptance and Proposal Reuse

Self-speculative decoding (Wu et al., 2026a) uses GroundAnything's shared weights for bidirectional drafting and causal verification, as illustrated in Figure S3. Verification accepts only the longest consecutive draft prefix agreeing with the causal predictions. The first disagreement ends acceptance; later coincidental matches are discarded. Rejected suffix states are removed from the cache.

Linear schedule. Given a committed seed, a bidirectional pass predicts a candidate block, and a causal pass verifies it in parallel. Accepted draft tokens are committed with the verifier correction at the first mismatch. Each round uses two NFEs and $2B$ query tokens. The seed is counted once and is excluded from draft-acceptance statistics.

Quadratic schedule. The fused query contains $B$ groups of $B+1$ tokens, with group $i$ arranged as $[d_i, M, \ldots, M]$ at overlapping positions $[t{+}i, \ldots, t{+}i{+}B]$. Row leaders support causal verification; each row's mask positions form a bidirectional proposal group. All queries read the cached prefix. Compare row prediction $a_i$ with $d_{i+1}$, starting the accepted count at one for the known leading token $d_0$. At the first mismatch, select that row's proposal as the next draft. This selection reuses computation but does not verify the selected proposal. After initialization, each round uses one NFE and $B(B+1)$ query tokens. The linear/quadratic terminology concerns query-token counts, not total Transformer FLOPs or a fixed number of accepted tokens.

These schedules use greedy causal comparisons. Their verification semantics concern the converted model's causal branch and do not establish equivalence to sampling an arbitrary autoregressive distribution or to an earlier model's outputs.

## C DLM–MTP Comparison: Analysis and Experimental Details

***Question C.1:*** Why use diffusion rather than MTP when both can draft for the same AR verifier?

We match visual inputs, adaptation data, tokenization, and prefixes, and freeze the same converted model's causal branch as verifier. The primary MTP baseline is an in-house causal drafter; bidirectional block-MTP and retrained causal-DLM controls test the role of attention (Section C).

We address four questions:

> ***Question C.1.1:*** **At the same candidate length, which drafter achieves higher acceptance?**
> ***Question C.1.2:*** **How much wall-clock time does each accepted draft token cost?**
> ***Question C.1.3:*** **Can additional DLM refinement repay its extra forward-pass cost?**
> ***Question C.1.4:*** **Which grounding failures benefit from bidirectional attention?**

Draft acceptance. At $K = 16$, one-step DLM accepts 3.48 draft tokens per round versus 2.69 for causal MTP (Figure S4), yielding more useful proposals at the same candidate length.

Cost per accepted draft token. Including drafting, verification, and rejected work, one-step DLM reduces full-round cost per accepted draft token from 7.71 to 5.38 ms (Figure S5). Higher acceptance therefore translates into lower wall-clock cost at this operating point.

Refinement versus latency. At $K = 16$, a second DLM refinement step lowers this cost to 4.70 ms; further refinement increases it. With validation-selected policies on the long-output subset, DLM commits 97.1 versus 84.6 tokens within 400 ms, although MTP leads at 50 ms (Figure S6). Refinement is thus useful when its acceptance gain repays its additional latency.

Bidirectional attention and grounding failures. Bidirectional DLM reduces Dense200's unmatched valid-proposal rate from 24.89% for causal MTP to 14.79%; its retrained causal counterpart scores 21.74% (Figure S7). This supports bidirectional attention for drafting jointly constrained spatial predictions. Bidirectional block-MTP also improves draft quality, so the benefit is not exclusive to diffusion.

These gains concern draft efficiency and reliability; exact shared verification preserves the AR verifier's completed output.

***Finding C.1:*** Bidirectional diffusion with a limited number of refinement steps improves draft-generation efficiency over the tested causal MTP baseline. At $K = 16$, two passes reduce cost per accepted draft token to 4.70 ms; further refinement raises the cost.

### C.1 Controlled Model Comparisons

The drafters share visual preprocessing, the projector, coordinate vocabulary, serialization, and autoregressive backbone initialization. A single frozen causal branch of GroundAnything verifies all proposals. Adaptation matches ordered examples, supervised-token exposure, optimizer updates, effective batch size, and candidate-length sampling. These controls do not imply equal trainable parameter counts or training FLOPs.

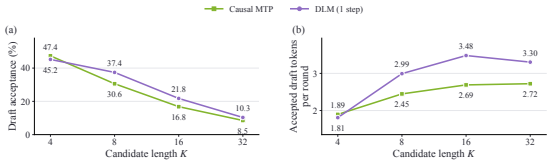


Figure S4: **Acceptance at matched candidate length.** One-step DLM and causal MTP share prefixes and a frozen AR verifier. Accepted tokens exclude verifier corrections and bonuses.

The primary in-house MTP baseline uses a causal recurrent future-token module with weights shared across prediction horizons. It is distinct from LocateAnything, which uses bidirectional attention within box-aligned MTP blocks and autoregressive fallback in Hybrid Mode (Wang et al., 2026). An additional in-house block-MTP control predicts a fully masked candidate block in one bidirectional pass; it shares our data and token format but is not a reproduction of LocateAnything.

The attention ablation compares separately trained causal and bidirectional DLMs with the same architecture, initialization, denoising objective, corruption distribution, training exposure, and refinement schedule. Only candidate-block attention differs during training and inference. All candidate positions can read the committed prefix; the prefix cannot read candidates. This pair isolates attention, whereas comparing block MTP with DLM also changes the objective and refinement procedure.

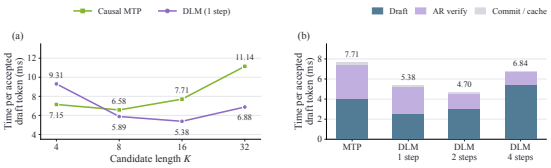


Figure S5: **Cost per accepted draft token.** Full-round latency includes rejected work; the stage breakdown uses $K = 16$.

### C.2 Draft Acceptance and Token Cost

All configurations receive the same bank of verifier-generated prefixes from held-out COCO prompts, with validation and test prefixes separated by image. The deadline comparison uses long-output prefixes and therefore characterizes decode progress on that subset rather than full-dataset request throughput. Candidate length $K$ counts newly proposed tokens, excluding the committed anchor, padding, and verifier correction or bonus. Greedy verification accepts the longest matching prefix, then emits a correction at the first mismatch or a bonus after complete acceptance, subject to response termination.

For round $r$, let $K_r$, $A_r$, and $C_r$ denote proposed, accepted-draft, and committed-output token counts, and let $T_r$ be full-round latency in milliseconds. We report

$$R_{\text{draft}} = \frac{\sum_r A_r}{\sum_r K_r}, \qquad \bar{A} = \frac{1}{N_r}\sum_r A_r, \qquad \bar{C} = \frac{1}{N_r}\sum_r C_r, \tag{S2}$$

$$c_A = \frac{\sum_r T_r}{\sum_r A_r}, \qquad c_C = \frac{\sum_r T_r}{\sum_r C_r}, \qquad \bar{T} = \frac{1}{N_r}\sum_r T_r, \tag{S3}$$

where $N_r > 0$ is the number of rounds; each ratio requires a positive denominator. Ratios of totals retain the cost of zero-acceptance and malformed-proposal rounds. For a full nonterminal round, $C_r = A_r + 1$; terminal rounds use actual emitted counts. Output throughput is $1000/c_C$, while $1000/c_A$ measures accepted draft tokens per second.

Round timing begins after the shared prefix cache is ready and ends after output and cache commitment. It includes drafting, verification, rejected work, candidate selection, scheduling, and cache maintenance; shared visual encoding and prefix prefill are outside this decode-only cost. All methods use the same execution conditions.

Acceptance and latency results. At $K = 16$, one-step DLM accepts 21.76% of candidates (3.48 tokens per round), compared with 16.81% (2.69 tokens) for causal MTP. This reduces $c_A$ from 7.71 to 5.38 ms, a 30.2% reduction at matched candidate length. Across the tested lengths, DLM's lowest cost is 5.38 ms at $K = 16$, while MTP's is 6.58 ms at $K = 8$. However, MTP is preferable at $K = 4$: its cost is 7.15 ms versus 9.31 ms for DLM. At $K = 32$, DLM's accepted prefix decreases to 3.30 tokens despite the longer proposal. These results favor selecting candidate length by accepted work and elapsed time rather than nominal parallelism.

### C.3 Refinement and Fixed-Deadline Decoding

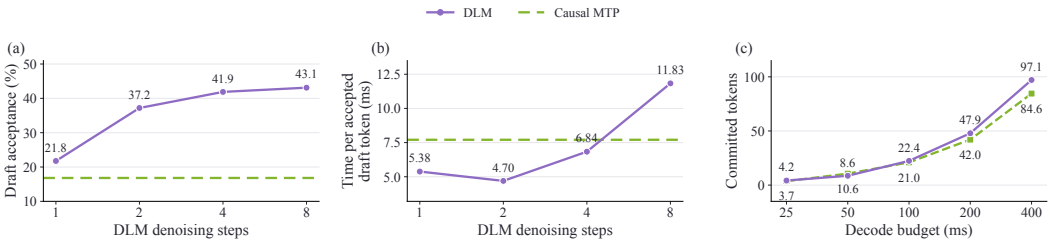


Figure S6: **Refinement and available decode time.** (a,b) Varying refinement depth at $K = 16$. (c) Committed tokens under equal deadlines on long-output prefixes, using validation-selected policies.

The refinement comparison fills all $K$ candidate positions on the first pass. Before pass $j \in \{2, \ldots, S\}$, it remasks the lowest-confidence

$$m_j = \left\lceil \frac{K(S - j + 1)}{S} \right\rceil \tag{S4}$$

positions and predicts them jointly. Both DLM attention variants use this schedule, followed by causal verification. This controlled draft-refinement experiment is distinct from monotone entropy-guided commitment in direct generation.

For two refinement depths $S$ and $S'$ with positive mean latency and accepted-token counts, extra refinement lowers accepted-token cost precisely when

$$\frac{\bar{T}_{S'}}{\bar{T}_S} < \frac{\bar{A}_{S'}}{\bar{A}_S}. \tag{S5}$$

The condition for committed-token cost replaces $\bar{A}$ by $\bar{C}$. Thus, improved draft acceptance alone does not establish a wall-clock benefit.

For a deadline $D$, candidate length and denoising depth are selected on validation prefixes and fixed for testing. Methods may execute repeated rounds, and only rounds completed by the deadline contribute:

$$\bar{N}(D) = \frac{1}{N} \sum_{i=1}^{N} \sum_{r} C_{i,r} \mathbf{1}[t_{i,r}^{\text{finish}} \leq D]. \tag{S6}$$

This measures progress at equal elapsed time, counting verifier corrections and bonuses.

Refinement and deadline results. At $K = 16$, two DLM passes reduce $c_A$ from 5.38 to 4.70 ms, whereas four and eight passes increase it to 6.84 and 11.83 ms. Cost per committed output token decreases by only 3.8%, because the verifier contributes a correction or bonus token in each full round. At 400 ms, DLM commits 97.1 tokens versus 84.6 for MTP; at 50 ms, MTP leads with 10.6 versus 8.6. Thus, refinement is beneficial over part of the latency range, but additional passes and very short budgets can reverse the advantage.

### C.4 Structural and Geometric Diagnostics

Each drafter receives the same diagnostic windows: category transitions from COCO, crowded objects from Dense200, document regions from DocLayNet, and GUI targets from ScreenSpot-Pro. Candidate length is matched, both DLM variants use two refinement steps, and the bidirectional

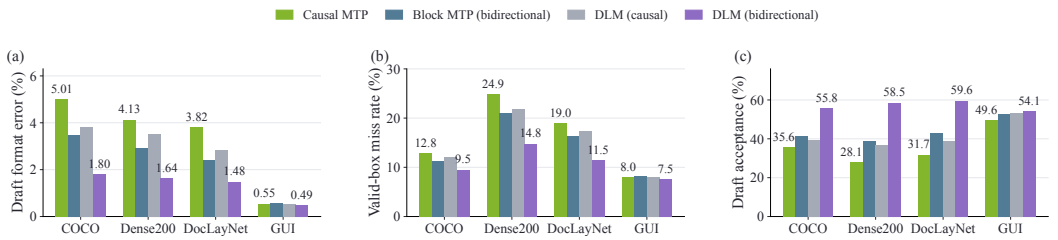


Figure S7: **Draft diagnostics before verification.** Comparisons at $K = 8$; both DLMs use two refinement steps. Valid-box misses count unmatched proposals, not missed ground-truth instances. GUI denotes ScreenSpot-Pro.

block-MTP control uses one pass. Diagnostic windows are selected from the common reference serialization before inspecting proposals. Raw format error detects illegal structural transitions, token types, or coordinate arity, using the parser state inherited from the prefix. Ending partway through an otherwise valid tuple at the candidate boundary is not itself a format error. Diagnostics precede grammar filtering, fallback, and verification.

The valid-box miss rate is

$$M_{\text{box}} = \frac{\#\text{ unmatched complete, syntactically valid proposed boxes}}{\#\text{ complete, syntactically valid proposed boxes}}. \tag{S7}$$

Matching is one-to-one against relevant remaining references, constrained by category or referring target and IoU at least 0.5. Unmatched duplicates, reversed corners, and zero-area boxes count as geometric misses. The numerator counts unmatched complete, syntactically valid proposed boxes; the denominator counts all complete, syntactically valid proposed boxes. The rate is undefined when this denominator is zero. It measures the fraction of eligible proposals that remain unmatched and is interpreted alongside format errors; it is not a ground-truth miss rate.

Verifier agreement, structural validity, and geometric accuracy measure different properties. Exact greedy verification produces the same sequence as the shared causal verifier; a finite deadline can yield different-length prefixes of that sequence. These diagnostics therefore assess draft reliability and rejected computation, rather than improved final localization under a fixed exact verifier.

Where bidirectional refinement helps. Relative to causal MTP, bidirectional DLM reduces COCO format errors from 5.01% to 1.80%. Its valid-box miss rate decreases by 10.10 pp on Dense200 and 7.52 pp on DocLayNet, compared with only 0.46 pp on GUI. These differences are consistent with greater benefits for proposals containing several spatially related elements. The retrained DLM pair provides the attention control: bidirectional attention lowers the Dense200 rate from 21.74% to 14.79% at the same refinement depth. Bidirectional block-MTP also improves over causal MTP, so bidirectionality itself is not a uniquely diffusion-based advantage. Its one-pass comparison with two-step DLM changes objective and refinement cost; it does not isolate diffusion training at equal elapsed time. The separate deadline analysis provides the matched-time comparison.

Relation to the grounding formulation. Together, the acceptance, cost, and deadline results support Finding C.1: bidirectional diffusion with limited refinement makes draft generation more efficient than the tested causal MTP baseline over the reported operating range. The attention control supports joint access to visual evidence, while the refinement sweep identifies when an additional denoising pass repays its cost; neither establishes superiority over every MTP design. The geometric diagnostic measures proposal reliability before verification, separately from ScreenSpot accuracy, F1mIoU, and ground-truth recall. Exact verification preserves the shared causal model's completed output.

# D Output Efficiency and Generation Cost

The shared output format uses 7.6 tokens/box on COCO and 5.1 on Dense200, compared with 148.8 and 74.5 for the SEED1.5-VL reference reported by Jiang et al. (2026) (Table S1). Figure S8 shows how generation time and output length vary with predicted object count: GroundAnything reduces generation time across the displayed ranges, with larger absolute savings for longer outputs (Section D).

Table S1: **Output token efficiency.** Mean boxes and tokens per image, and tokens per box. SEED1.5-VL values are external references from Jiang et al. (2026); output-length ratios are not runtime speedups.

| Model | Boxes/img | Tokens/img | Tokens/box |
|---|---|---|---|
| **COCO** | | | |
| SEED1.5-VL (Guo et al., 2025) | 4.2 | 631.0 | 148.8 |
| GroundAnything-VLM | **6.1** | **46.4** | **7.6** |
| GroundAnything | **6.1** | **46.4** | **7.6** |
| **Dense200** | | | |
| SEED1.5-VL (Guo et al., 2025) | 73.1 | 5446.3 | 74.5 |
| GroundAnything-VLM | **87.6** | **446.8** | **5.1** |
| GroundAnything | **87.6** | **446.8** | **5.1** |

Compact serialization. Table S1 follows the output-length analysis of Jiang et al. (2026). Its SEED1.5-VL row is taken from that work. The GroundAnything variants use the same structured vocabulary and have identical reported output statistics: 7.6 tokens/box on COCO and 5.1 on Dense200. Compared with 148.8 and 74.5 tokens/box for SEED1.5-VL, these are approximately $19.6\times$ and $14.6\times$ lower token counts per box. The ratios reflect serialization length, not matched-hardware latency; differences in detected instance counts and model execution preclude reading them as speedups.

Generation cost across object counts. Figure S8 groups outputs by predicted box count and reports average generation time alongside output-token count. Longer coordinate lists increase decoding work for both variants, while GroundAnything reduces generation time across the displayed ranges, with larger absolute savings for longer outputs. The token-efficiency table characterizes serialization length; this figure characterizes generation cost as the number of predicted instances grows. The speed–quality comparisons in Section G.1 complement this output-length analysis.

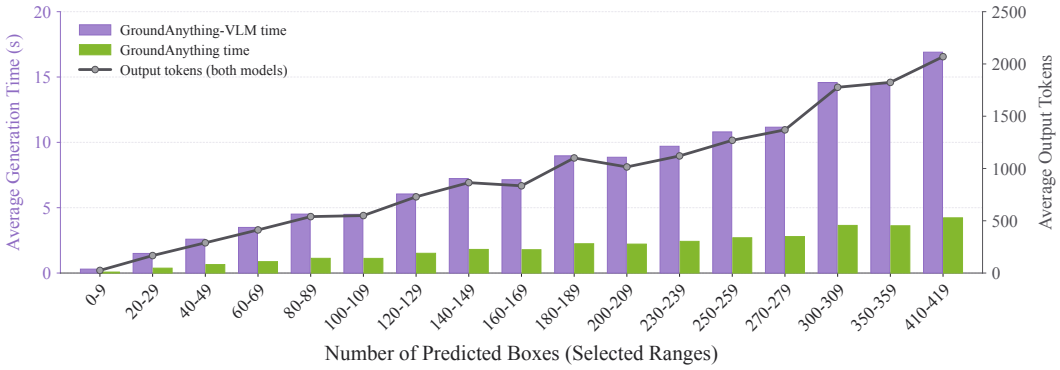


Figure S8: **Generation cost versus predicted object count.** Average generation time for GroundAnything-VLM and GroundAnything, with output-token counts across box-count ranges.

# E Model and Training Details

## E.1 Input/Output Protocol

GroundAnything and GroundAnything-VLM share a structured interface: the query specifies what evidence to extract, and each response entry associates a semantic identifier with spatial coordinates. The geometry is task-dependent, while the enclosing grammar remains unchanged, following the unified grounding formulation of Jiang et al. (2026).

Table S2: Task-specific input instructions. Category names and referring expressions are preserved in the response identifiers.

| Task | Instruction |
|---|---|
| Category grounding | Locate all the instances that match the following categories: [CATEGORIES]. |
| Referring grounding | Locate the target referred to by the following description: [EXPRESSION]. |
| Category pointing | Point to: [CATEGORIES]. |
| Referring pointing | Point to the target referred to by the following description: [EXPRESSION]. |
| GUI grounding | Point to the UI element to click for the following instruction: [INSTRUCTION]. |
| OCR | OCR task detect all the text in box format. |
| Layout grounding | Detect all document layout elements that match the following categories: [CATEGORIES]. |
| Visual prompting | Given reference boxes <\|box_start\|>[BOXES]<\|box_end\|> indicating one or more objects, find all similar objects in the image and output their bounding boxes. |

Query templates. Table S2 lists the task instructions. Bracketed terms are placeholders. A category list joins the requested names with `</c>`, without intervening spaces; for example, `person</c>car`. Visual reference boxes use the same coordinate tokens as outputs. The image and instruction form the user turn; the structured response forms the assistant turn.

Response grammar. The entry template and illustrative payloads are shown below. Displayed line breaks are for readability and are omitted in the serialized response.

**Entry template**

```
<|object_ref_start|>[IDENTIFIER]<|object_ref_end|>
<|box_start|>[PAYLOAD]<|box_end|>
```

**Boxes: two instances of one category**

```
<|object_ref_start|>person<|object_ref_end|>
<|box_start|><10><20><30><40>,<50><60><70><80><|box_end|>
```

**Point: one target location**

```
<|object_ref_start|>button<|object_ref_end|>
<|box_start|><512><384><|box_end|>
```

**OCR: transcription as identifier**

```
<|object_ref_start|>OPEN<|object_ref_end|>
<|box_start|><100><200><300><260><|box_end|>
```

**Explicitly queried but absent category**

```
<|object_ref_start|>car<|object_ref_end|>
<|box_start|>None<|box_end|>
```

Boxes use four atomic coordinate tokens in $(x_1, y_1, x_2, y_2)$ order; points use two in $(x, y)$ order. Instances sharing an identifier occupy one wrapper, with tuples separated by commas without spaces. Distinct entries are separated by a comma and a space. Ordered point sequences retain their temporal order. The ordinary text `None` denotes an absent queried target, not a padding element. OCR transcriptions occupy the identifier field rather than a separate text channel.

Coordinates and canonicalization. Coordinates are normalized by image width and height and quantized to `<0>`–`<999>`. For annotations on a $[0, 1000]$ grid, the conversion is

$$q(v) = \left\lfloor \frac{999v + 500}{1000} \right\rfloor . \tag{S8}$$

Input `xywh` boxes are converted to `xyxy`; polygons used for box supervision become axis-aligned enclosing boxes. Non-finite or out-of-range coordinates and inverted or degenerate boxes fail validation. Box instances are stably sorted by $x_1$, and unordered points by $x$; semantic sequence order is never replaced by spatial sorting.

Table S3: Grounding and multimodal token conventions.

| Token(s) | Function |
|---|---|
| <\|object_ref_start\|>, <\|object_ref_end\|> | Delimit a semantic identifier |
| <\|box_start\|>, <\|box_end\|> | Delimit a geometric payload |
| <0>–<999> | Atomic quantized coordinates |
| </c> | Separate requested categories |
| <\|vision_start\|>, <\|vision_end\|> | Delimit visual input |
| <\|image_pad\|> | Image placeholder in the input template |
| <\|im_start\|>, <\|im_end\|> | Turn boundaries; the latter terminates the response |
| \|<MASK>\| | Internal diffusion mask |

Table S4: Shared backbone architecture. Positional capacity is independent of the training sequence lengths.

| Component | Configuration |
|---|---|
| Vision encoder | 27 layers; width 1024; FFN width 4096; 12 heads; QKV width 1536; patch size 14 |
| Projector | LayerNorm(1024), $2 \times 2$ aggregation, bias-free linear layers $4096 \to 4096 \to 2560$ with GELU, then RMSNorm(2560); normalization $\epsilon = 10^{-5}$ |
| Language backbone | 36 layers; width 2560; FFN width 9728; 32 query heads and 8 KV heads; head dimension 128 |
| Language numerics | SiLU; attention dropout 0; RMSNorm $\epsilon = 10^{-6}$; RoPE base $5{,}000{,}000$ |
| Positional capacity | 262,144 positions |

Structural tokens and termination. Table S3 distinguishes entry boundaries from response termination. Structural and coordinate tokens are retained when decoding text for the parser. The diffusion mask is an internal input token and is excluded from the generatable vocabulary.

### E.2 Architecture, Tokenizer, and Visual Processing

The variants share a MoonViT-V2 (Kimi K3) visual encoder, a multimodal projector, and a Qwen3-4B language architecture. All language layers use full attention and one-dimensional rotary position indices. Dynamic image processing supplies at most 1024 projected visual tokens, corresponding to 4096 patches before $2 \times 2$ aggregation; channel means and standard deviations are both $(0.5, 0.5, 0.5)$. Table S4 gives the module specifications.

Vocabulary and parameter sharing. GroundAnything-VLM extends the 151,669-entry base vocabulary with 1000 coordinate tokens and </c>, giving 152,670 entries. Coordinate IDs are 151669–152668; the category delimiter has ID 152669. GroundAnything adds |<MASK>| at ID 152670, giving 152,671 entries. EOS and padding have distinct IDs, 151645 and 151643; the image, vision-start, and vision-end IDs are 151655, 151652, and 151653.

Input embeddings and output heads are untied. Within GroundAnything, causal and diffusion modes share all model weights. The new mask row is initialized separately in each vocabulary matrix using the FP32 mean of the rows for <|im_start|>, <|im_end|>, <|vision_start|>, and <|vision_end|>, then cast to the model dtype. This adds 5120 parameters. The complete multimodal model has approximately 4.844B parameters (Table S5).

### E.3 Autoregressive Foundation and Spatial Adaptation

Three-stage base training. Connector alignment trains only the projector on image–text supervision. Joint multimodal pretraining unfreezes the visual and language backbones and combines text modeling with image-conditioned learning. General visual and video understanding then develops instruction following, visual question answering, and image/video understanding. Text-only examples supervise valid text tokens; multimodal examples supervise assistant responses, excluding

Table S5: Parameter counts. The 4B designation refers to the language backbone.

| Module | GroundAnything-VLM | GroundAnything |
|---|---|---|
| Vision encoder | 401,214,464 | 401,214,464 |
| Projector | 27,267,584 | 27,267,584 |
| Decoder excluding vocabulary | 3,633,511,936 | 3,633,511,936 |
| Input embedding | 390,835,200 | 390,837,760 |
| Output head | 390,835,200 | 390,837,760 |
| Total | 4,843,664,384 | 4,843,669,504 |

Table S6: Final base-stage and initial supervised spatial-adaptation settings.

| Setting | Final base stage | Spatial adaptation |
|---|---|---|
| GPUs | 256 | 256 |
| Per-device / global batch | 2 / 512 | 6 / 1536 |
| Sequence length | 8192 | 2048 |
| Trainable modules | All | All |
| Language peak LR | $10^{-5}$ | $2 \times 10^{-5}$ |
| Projector / vision peak LR | $10^{-6}$ / $10^{-6}$ | $2 \times 10^{-6}$ / $2 \times 10^{-6}$ |
| Gradient checkpointing | Language and vision | Vision |
| Seed / data seed | 42 / 42 | 42 / 42 |

visual inputs, prompts, and template tokens. The text and multimodal losses are reduced within their respective domains and summed. Aggregate pretraining exposure exceeds 700B tokens.

Spatial specialization uses supervised adaptation followed by SFT and reinforcement post-training. Conversion is initialized from an autoregressive grounding model with initial spatial adaptation under assistant-only causal cross-entropy. Table S6 gives the final base-stage and initial adaptation settings; subsequent variant-specific stages are described below.

Both stages use BF16, ZeRO-1, one gradient-accumulation step, AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$ and $\epsilon = 10^{-8}$, weight decay 0.1, and gradient clipping at 1.0. A cosine schedule uses 3% warmup and a minimum learning-rate ratio of 0.1.

### E.4 Direct Conversion: Corruption, Attention, and Supervision

Direct multimodal conversion (Wu et al., 2026a) retains the autoregressive grounding vocabulary and output grammar. Let $x = [V, P, Y]$ contain visual embeddings $V$, the text prompt $P$ encoding query $Q$, and response $Y$. Supervised response spans are partitioned into consecutive blocks of $B = 32$ tokens. Blocks may cross geometric-tuple boundaries, stop at response-span boundaries, and have a shorter final block; they require no semantic padding.

Complementary corruption. For each response block $b$, sample $t_b \sim \mathcal{U}(0, 1)$. For every supervised non-EOS position $i$ in that block, draw $m_i^{(1)} \sim \text{Bernoulli}(t_b)$ and set $m_i^{(2)} = 1 - m_i^{(1)}$. View $a$ replaces the target by `|<MASK>|` if $m_i^{(a)} = 1$, giving $w^{(a)} = [P, \widetilde{Y}^{(a)}]$. EOS is masked in both views. Images, prompts, headers, unsupervised turns, and empty thinking prefixes are not corrupted.

Attention visibility. The joint input $[w^{(1)}; w^{(2)}; x]$ shares one clean multimodal stream and one visual encoding; each noisy view retains an uncorrupted prompt copy. Within a view, a noisy response block attends bidirectionally to itself and reads the clean visual/prompt context and strictly preceding clean response blocks. It cannot read current or future clean response targets or the other noisy view. Clean tokens use token-level causal attention and cannot attend to either noisy view. Figure 3 illustrates one view with two-token blocks for clarity.

Table S7: GroundAnything conversion configuration.

| Setting | Phase I | Phase II |
|---|---|---|
| GPUs | 256 | 256 |
| Per-device / global batch | 32 / 8192 | 32 / 8192 |
| Clean sequence length | 4096 | 4096 |
| Trainable modules | Projector and language | All |
| Language / projector peak LR | $2 \times 10^{-5}$ / $2 \times 10^{-5}$ | $2 \times 10^{-5}$ / $2 \times 10^{-5}$ |
| Vision peak LR | Frozen | $2 \times 10^{-6}$ |
| Gradient checkpointing | Language | Language and vision |
| Seed / data seed | 32 / 32 | 42 / 42 |

Table S8: Supervised fine-tuning settings. GA denotes GroundAnything; VLM denotes GroundAnything-VLM.

| Setting | SFT1 | SFT2 |
|---|---|---|
| GPUs / accumulation | 256 / 1 | 256 / 1 |
| Sequence length | 8192 | 2048 |
| Trainable modules | All | Language |
| Language peak LR | $7 \times 10^{-6}$ | $7 \times 10^{-6}$ |
| Projector peak LR | $2 \times 10^{-6}$ | Frozen |
| Vision peak LR | $7 \times 10^{-7}$ | Frozen |
| GA per-device / global batch | 32 / 8192 | 32 / 8192 |
| VLM per-device / global batch | 1 / 256 | 2 / 512 |
| GA gradient checkpointing | Language and vision | Language |
| VLM gradient checkpointing | Language and vision | Disabled |
| Seed / data seed | 42 / 42 | 42 / 42 |

Joint objective. For masked target sets $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$, and valid clean response targets $\mathcal{T}$, the two losses are

$$\mathcal{L}_{\mathrm{MDM}} = -\frac{\sum_{a=1}^{2} \sum_{i \in \mathcal{M}^{(a)}} \log p_\theta(y_i \mid x, w^{(a)}; \mathcal{A}_a)}{|\mathcal{M}^{(1)}| + |\mathcal{M}^{(2)}|}, \tag{S9}$$

$$\mathcal{L}_{\mathrm{AR}} = -\frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \log p_\theta(y_i \mid V, P, y_{<i}). \tag{S10}$$

Here $\mathcal{A}_a$ enforces the visibility rules above, so writing $x$ as an input does not expose the clean target. The MDM mean includes both masked EOS targets and has no inverse-noise weighting. Both losses retain the next-token shift: hidden position $i-1$ predicts target $i$. Prompts, visual positions, headers, padding, unsupervised turns, and empty thinking prefixes are excluded. Valid target losses are averaged across the training batch and combined as $\mathcal{L} = 0.5\mathcal{L}_{\mathrm{MDM}} + 0.5\mathcal{L}_{\mathrm{AR}}$.

### E.5 Conversion and Continued Supervised Fine-Tuning

Conversion phases. Phase I freezes the vision encoder and trains the projector and language parameters; Phase II updates all modules. Both retain the fixed block size and joint objective. The optimizer and scheduler restart between phases. Table S7 reports the settings.

Conversion uses BF16, ZeRO-1, one gradient-accumulation step, and AdamW with betas $(0.9, 0.95)$, epsilon $10^{-8}$, weight decay 0.1, and gradient clipping at 1.0. Both phases use cosine decay with 3% warmup.

Continued SFT. Both variants first update all modules (SFT1), then freeze the vision encoder and projector and update only the language parameters (SFT2). Language parameters include the untied input embedding and output head. GroundAnything retains the joint AR/diffusion objective; GroundAnything-VLM uses assistant-only causal cross-entropy. Table S8 separates their batch and checkpointing settings while consolidating shared hyperparameters.

Table S9: Training and inference stages of GroundAnything.

| GroundAnything stage | Objective or generation rule |
|---|---|
| AR-to-diffusion conversion | $0.5\mathcal{L}_{\mathrm{MDM}} + 0.5\mathcal{L}_{\mathrm{AR}}$; shared weights |
| Continued SFT | Same joint objective; all modules, then language only |
| Causal GRPO | Causal rollouts and likelihoods; task rewards with nested group normalization |
| Diffusion inference | Bidirectional block recovery with the updated shared weights |

SFT uses BF16, ZeRO-1, and AdamW with betas $(0.9, 0.95)$, epsilon $10^{-8}$, weight decay 0.1, and gradient clipping at 1.0. Each stage has a separate cosine schedule with 3% warmup and a minimum learning-rate ratio of 0.1. The projected visual-token budget remains 1024.

Table S9 summarizes the stages illustrated in Figure 4.

### E.6 Causal Reinforcement Post-Training

GroundAnything uses causal attention for both rollout generation and likelihood evaluation. For each prompt, $G = 8$ responses are sampled at temperature $T = 0.7$. Before updating, a no-gradient teacher-forcing pass records detached old-policy log-probabilities; a second pass computes current-policy gradients. Both use FP32 log-softmax at the rollout temperature. These are exact likelihoods of the causal policy, not likelihoods of a diffusion denoising trajectory.

Advantage normalization. For a scalar component $u$ within one prompt group, define

$$Z_G(u_i) = \frac{u_i - \bar{u}}{\sqrt{G^{-1}\sum_{j=1}^{G}(u_j - \bar{u})^2 + 10^{-6}}}, \qquad \bar{u} = G^{-1}\sum_{j=1}^{G} u_j. \tag{S11}$$

Each active reward component is normalized before weighting, and their sum is normalized again within the same group:

$$A_i = Z_G\left(\sum_{c\in\mathcal{C}_{\mathrm{task}}} w_c Z_G(R_{c,i})\right). \tag{S12}$$

Box rewards have weights $(0.7, 0.3)$; OCR and pointing each supply one complete scalar reward with unit weight. Internal OCR subterms are not normalized separately. Constant groups produce zero advantage.

Policy objective. Let $\pi_{\theta,T}$ denote the causal policy at temperature $T$, and let $\mathcal{T}_i$ contain response tokens including EOS, excluding prompts and post-EOS padding. Expanding the response-wise averaging described in Section 3.2, the update minimizes

$$r_{i,t}(\theta) = \frac{\pi_{\theta,T}(y_{i,t} \mid I, Q, y_{i,<t})}{\pi_{\mathrm{old},T}(y_{i,t} \mid I, Q, y_{i,<t})}, \tag{S13}$$

$$\mathcal{L}_{\mathrm{GRPO}} = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathcal{T}_i|}\sum_{t\in\mathcal{T}_i}\min(r_{i,t}A_i, \mathrm{clip}(r_{i,t}, 1-\epsilon, 1+\epsilon)A_i), \tag{S14}$$

with $\epsilon = 0.2$. There is one update per rollout and no reference-policy KL term. The vision encoder is frozen; the projector and language parameters are updated. Causal and diffusion attention modes subsequently reuse these parameters without changing the output protocol.

### E.7 GroundAnything-VLM Reinforcement Settings

GroundAnything-VLM uses the same task-reward definitions but a distinct optimization configuration. It freezes both the vision encoder and projector and retains a frozen supervised reference with KL coefficient 0.02. Active reward components are standardized within each prompt group using the sample standard deviation and epsilon $10^{-8}$; after weighted summation, advantages are standardized across the complete response batch. GroundAnything instead uses population standard deviations and a second within-group normalization. Table S10 keeps these recipes separate.

Table S10: Variant-specific GRPO settings. Global batches count generated responses, not distinct prompts.

| Setting | GroundAnything | GroundAnything-VLM |
|---|---|---|
| GPUs | 256 | 256 |
| Per-device response batch / accumulation | 1 / 1 | 1 / 1 |
| Responses per prompt | 8 | 8 |
| Prompt / response batch | 32 / 256 | 32 / 256 |
| Updates per rollout | 1 | 1 |
| Learning rate | $2 \times 10^{-6}$ | $5 \times 10^{-7}$ |
| Schedule / warmup | Constant / none | Constant / none |
| AdamW betas / epsilon | $(0.9, 0.999)$ / $10^{-8}$ | $(0.9, 0.999)$ / $10^{-8}$ |
| Weight decay / gradient clipping | 0 / 1.0 | 0.01 / 1.0 |
| Policy clipping | 0.2 | 0.2 |
| Reference-policy KL | No reference | 0.02 |
| Trainable modules | Projector and language | Language |
| Precision / sharding | BF16 / ZeRO-1 | BF16 / ZeRO-2 |
| Seed | 20260902 | 20260812 |

GroundAnything maintains FP32 optimizer states. Its rollout and gradient passes use the same causal policy; the inference implementations described below do not alter this training factorization.

### E.8 Task Rewards

The rewards below evaluate complete structured responses. We write $n$ and $m$ for the numbers of predictions and reference instances, respectively, and $\mathrm{clip}_{[0,1]}$ for clipping to the unit interval. Content scores below assume nonempty references; missing predictions receive zero content agreement.

#### E.8.1 Box Grounding

Set completeness. For each reference box $g_j$, select the predicted box with maximum IoU before checking its label:

$$i^*(j) = \arg\max_i \mathrm{IoU}(b_i, g_j), \qquad s_j = \mathrm{IoU}(b_{i^*(j)}, g_j)\,\mathbf{1}[\ell_{i^*(j)} = \ell_j]. \tag{S15}$$

With $S = \sum_j s_j$, soft precision and recall are $P = S/n$ and $R = S/m$, and

$$R_{\mathrm{set}} = \frac{2PR}{P + R + 10^{-8}}. \tag{S16}$$

This GT-wise matching permits a prediction to support multiple references. Consequently, $R_{\mathrm{set}}$ is a completeness proxy rather than a bounded one-to-one F1 metric. It is combined with a stricter reward that penalizes duplicate and oversized predictions.

Strict localization. Let $\bar{F}$ average localization F1 at IoU thresholds $0.50$, $0.75$, and $0.95$. The strict reward is

$$\begin{aligned} R_{\mathrm{strict}} = \mathrm{clip}_{[0,1]}\big(&0.10R_{\mathrm{fmt}} + 0.10R_{\mathrm{count}} + 0.50\bar{F} + 0.25R_{\mathrm{IoU}} + 0.05R_{\mathrm{sort}} \\ &- 0.07P_{\mathrm{big}} - 0.03P_{\mathrm{dup}}\big). \end{aligned} \tag{S17}$$

The auxiliary terms score protocol validity, instance-count agreement, matched IoU, and canonical ordering. Penalties discourage unsupported large boxes and duplicate predictions. The set and strict rewards are normalized separately before combination, rather than treating their raw weighted sum as a single reward.

#### E.8.2 OCR

Text–geometry matching. OCR uses one-to-one Hungarian matching. The main text key applies Unicode NFKC normalization and case folding and retains alphanumeric characters; symbol-only

strings retain codepoint-based keys. Let $E_{ij}$ be one minus normalized Levenshtein distance between the resulting keys, and $I_{ij}$ the box IoU. For an affinity matrix $A$, define

$$F(A) = \frac{2}{n+m} \max_{\mathcal{M}} \sum_{(i,j)\in\mathcal{M}} A_{ij}, \tag{S18}$$

where $\mathcal{M}$ is a one-to-one assignment. Hard affinity requires equal text keys and $I_{ij} \geq \tau$. Write its score as $H_\tau$ and its average over $\tau \in \{0.50, 0.55, \ldots, 0.95\}$ as $\bar{H}$. The soft affinity and single-reference score are

$$A_{ij}^{\text{soft}} = \sqrt{I_{ij}}\, E_{ij}\, \mathbf{1}[I_{ij} \geq 0.10,\ E_{ij} \geq 0.20], \tag{S19}$$

$$V(P,T) = \text{clip}_{[0,1]}\left(0.45\bar{H} + 0.35F(A^{\text{soft}}) + 0.10H_{0.50} + 0.10C\right), \tag{S20}$$

with count agreement $C = \min(n,m)/\max(n,m)$.

Word/line consistency. Let $V_{\text{hi}}$ and $V_{\text{lo}}$ be the higher and lower scores against complete word- and line-level references. For a region with alternative granularities, local agreement selects the better *complete* representation, not independently chosen fragments. Singleton consensus uses the maximum text and box similarities among the reference alternatives, followed by one-to-one assignment, with score $0.60\bar{H}^* + 0.40S^*$. Aggregating local agreements gives $\Gamma$, with lower weights for text-conflict regions. The global and local terms are combined as

$$D = \alpha V_{\text{hi}} + (1-\alpha)V_{\text{lo}}, \qquad Q_g = \beta D + (0.90-\beta)\Gamma, \tag{S21}$$

with $(\alpha,\beta) = (0.90, 0.35)$ for conflicting granularities and $(0.75, 0.50)$ otherwise. Define

$$m_g = \max(|T_{\text{word}}|, |T_{\text{line}}|, 1), \qquad C_g = \frac{\min(n, m_g)}{m_g}.$$

The reward is

$$R_g = \text{clip}_{[0,1]}\left((Q_g + 0.10 f_g C_g) f_g^2 - 0.05P_{\text{dup}} - 0.10P_{\text{over}} - 0.05P_{\text{invalid}}\right). \tag{S22}$$

Complementary-reference consistency. For complex text, each reference score additionally preserves surface-form accuracy: $W = 0.85V + 0.15U$, where $U$ is exact-string F1 after NFKC and case folding, retaining punctuation and requiring IoU at least 0.50. Two primary references and a complementary reference yield

$$W_{\text{text}} = 0.82(0.70\max(W_a, W_b) + 0.30\min(W_a, W_b)) + 0.18W_c. \tag{S23}$$

When auxiliary geometry is available, $Q_c = 0.95W_{\text{text}} + 0.05G_{\text{aux}}$; otherwise $Q_c = W_{\text{text}}$. The geometry-only score is $G_{\text{aux}} = 0.60\bar{H}_{\text{geo}} + 0.30S_{\text{geo}} + 0.10C_{\text{geo}}$, excluding transcription agreement. Let $m_c$ be the rounded median reference count and $C_c = \min(n, m_c)/m_c$. Then

$$\begin{aligned} R_c = \text{clip}_{[0,1]}\big((0.90Q_c + 0.10 f_c C_c) f_c^2 - 0.05P_{\text{dup}} - 0.10P_{\text{over}} \\ - 0.05P_{\text{invalid}} - 0.08P_{\text{large}}\big). \end{aligned} \tag{S24}$$

Format gating and penalties. The parse grades in Table S11 gate the full content score quadratically. Penalties measure duplicated, excessive, or invalid predictions; $P_{\text{large}}$ additionally penalizes boxes covering over 80% of the image without geometric reference support. The annotation structure selects $R_g$ or $R_c$, which enters GRPO as one scalar reward.

### E.8.3 POINTING

Point predictions are matched one-to-one within each label. For distance $d$ on the normalized 0–999 coordinate grid, similarity is

$$s(d) = \begin{cases} \exp[-\frac{1}{2}(d/75)^2], & d \leq 200, \\ 0, & d > 200. \end{cases} \tag{S25}$$

The matched similarities define soft precision $P$ and recall $R$. With $F_2 = 5PR/(4P+R)$, defined as zero when the denominator vanishes, the reward is

$$R_{\text{point}} = 0.55F_2 + 0.25R_{\text{count}} + 0.10F_{1,\text{label}} + 0.10R_{\text{fmt}}. \tag{S26}$$

Count agreement is the smaller-to-larger count ratio when the larger count is positive. Format validity is 1 for a valid response, 0.5 when valid points remain recoverable from a malformed response, and 0 otherwise.

Table S11: OCR format grades used for squared reward gating.

| Branch | Parse status | Grade |
|---|---|---|
| Word/line | Complete, valid structure with no invalid instances | 1.00 |
| | Incomplete structure with reliably parsed instances | 0.60 |
| | Only some valid instances can be recovered | 0.30 |
| | No reliable parse | 0.00 |
| Complementary references | Strictly valid structure with no invalid content | 1.00 |
| | Complete main structure with extraneous content | 0.45 |
| | Incomplete structure with parseable instances | 0.35 |
| | Recoverable instances without a compliant overall structure | 0.15 |
| | No reliable parse | 0.00 |

# F Parallel Inference

## F.1 Direct Diffusion Decoding

Block state and token alignment. Image/query prefill establishes a causal prefix cache and the first anchor token. A physical block contains this known anchor followed by $B - 1$ mask positions. Predictions retain the training token shift. Sub-blocks are completed from left to right; every denoising forward still evaluates the full physical block. Completed tokens remain visible to subsequent masked positions, and the prefix cache is reused throughout.

Entropy-guided commitment. Following the use of entropy as a reliability signal in diffusion decoding (Liu et al., 2025), we compute

$$p_j = \mathrm{softmax}(z_j), \qquad H_j = -\sum_{v \in \mathcal{V}_{\mathrm{gen}}} p_j(v) \log p_j(v), \tag{S27}$$

over the generatable vocabulary, excluding the mask token. For the still-masked positions $\mathcal{U}$ of the active sub-block, commit all positions in $\mathcal{S} = \{j \in \mathcal{U} : H_j \leq \tau\}$. If $\mathcal{S}$ is empty, commit only $\arg\min_{j \in \mathcal{U}} H_j$. Candidate sampling and reliability assessment are separate operations. The procedure repeats until the sub-block is complete; committed tokens are neither remasked nor revised.

Causal cache construction. Bidirectional hidden states depend on positions to their right and cannot serve directly as causal history. After completing a block, one causal forward recomputes its authoritative KV entries and predicts the next anchor. This pass does not verify, reject, or replace the completed block. Thus, a block requiring $D$ denoising passes uses $D + 1$ NFEs, excluding initial prefill. The number of passes depends on commitment decisions rather than a fixed per-response denoising count.

Other commitment policies. Dynamic Decoding uses a confidence threshold with a forced best-position fallback. Static Decoding commits the highest-confidence remaining positions under a prescribed per-pass quota, distributing the remaining positions across the remaining passes. These policies change token commitment, while preserving blockwise generation. They are distinct from AR verification; their confidence scores should not be identified with the unmodified-logit entropy used above.

## F.2 Execution Optimizations

The implementation progresses from Native PyTorch (Eager) to SGLang (Eager), CUDA Graph replay, and selective FP8 execution. Denoising and causal passes retain their respective attention and cache semantics. FP8 applies to eligible language-model linear operations, with the remaining components kept in BF16. These changes reduce execution overhead or arithmetic cost; quantization can still alter logits and decoding decisions. The corresponding latency results are analyzed in Section G.2.

# G Additional Experimental Analysis

## G.1 Speed Reporting and Decoding Sensitivity

Metrics and scope. TPF denotes committed output tokens per model forward; TPS denotes output tokens per second. Forward counts must include the causal cache-construction or verification passes required by the relevant schedule. The physical block size $B$ in the decoding sweeps includes the known anchor, whereas candidate length $K$ in the DLM–MTP comparison counts newly proposed tokens. The two quantities should not be identified. All quality values in the decoding sweeps are COCO F1mIoU, and their speed–accuracy trade-offs should not be extrapolated to a 30-benchmark quality average.

Reference operating points. The AR reference is 49.6 TPS with 63.70 F1mIoU. At $B = 32$, throughput is 223.8 TPS for linear self-speculation, 49.9 TPS for quadratic self-speculation, and 127.0 TPS for entropy guidance at $\tau = 0.8$. The corresponding COCO quality is 62.96 for linear self-speculation and 60.72 for entropy guidance. Thus, the two main speedups are $223.8/49.6 = 4.51\times$ and $127.0/49.6 = 2.56\times$, with quality gaps of 0.74 and 2.98 pp.

### Interpreting the Decoding Sweeps

Linear versus quadratic self-speculation. Linear TPF rises from 1.41 at $B = 4$ to 2.21 at $B = 16$, then changes little at $B = 32$ (2.18). Quadratic TPF reaches 3.17 at both $B = 16$ and $B = 32$. At $B = 32$, its 45.4% higher TPF nevertheless accompanies much lower TPS than the linear schedule. Fusing verification and proposal reduces forward-call count but increases the number of query tokens, illustrating why TPF alone is not a latency metric. From $B = 16$ to $B = 32$, linear throughput changes from 226.6 to 223.8 TPS, whereas quadratic throughput falls from 88.0 to 49.9 TPS.

Entropy threshold. As $\tau$ increases from 0.2 to 1.2, TPF increases from 1.02 to 1.60 and TPS from 107.3 to 148.7. COCO F1mIoU remains near 60 for $\tau \leq 0.8$, peaks at 60.72 at $\tau = 0.8$, falls to 54.06 at $\tau = 1.0$, and partly recovers to 56.14 at $\tau = 1.2$. The non-monotonic quality curve supports a speed–accuracy trade-off rather than a universal monotonic accuracy law. Increasing $\tau$ permits more uncertain tokens to be committed together; the observed quality drop is consistent with the risk of premature commitment.

Block size and decoding mode. Entropy-guided TPF changes only from 1.18 to 1.24 over the tested block sizes, while its best COCO score is 61.58 at $B = 16$. Linear self-speculation attains its best tested COCO score at $B = 32$. Consequently, the $B = 32$ operating point used for the headline comparison is not the maximum-quality choice for every decoding strategy. Block-size selection therefore depends on task quality and output length; the COCO sweep alone does not establish a universal optimum.

What exact verification preserves. Self-speculation verifies against the converted model's causal branch. Exact greedy verification preserves that branch's outputs, not necessarily the outputs of the separately trained GroundAnything-VLM. The 0.74 pp gap to GroundAnything-VLM therefore does not imply that the verifier accepts mismatching tokens. Direct entropy-guided decoding has no AR acceptance test and remains the appropriate main-table setting for assessing diffusion generation itself.

## G.2 Infrastructure Analysis

Figure 9 adds SGLang (Eager), CUDA Graph, and FP8 sequentially to each decoding mode. The final time per output token is 6.59 ms for Dynamic Decoding, 6.76 ms for Static Decoding, 7.15 ms for Entropy-Guided Decoding, and 3.73 ms for Self-Speculative Decoding. Relative to each mode's Native PyTorch (Eager) implementation, the cumulative speedups are $3.48\times$, $3.45\times$, $3.51\times$, and $4.06\times$, respectively. The consistent reductions support the claim that parallel decoding requires corresponding execution support to realize practical acceleration.

These ratios compare implementations within a decoding mode. They are not multiplied by the separate $4.51\times$ algorithmic comparison, whose denominator is GroundAnything-VLM at a different reported operating point. The figure reports latency rather than a matched quality ablation for every

infrastructure stage. In particular, it does not establish that FP8 preserves accuracy; selective quantization can change logits and thus commitment or verification decisions. The implementation-level distinction is described in Section F.2.

### G.3 Architecture and Training Ablations

Coordinate representation. In Figure 10, replacing quantized coordinates with textual coordinates reduces the AR score from 73.03 to 71.60 and the DLM score from 61.66 to 55.98. The respective 1.43 and 5.68 pp drops support the compact coordinate vocabulary as a useful part of both variants, with a larger effect in the diffusion setting. The figure's $0.25\times$ annotation concerns the textual-representation comparison; it should not be identified with the external SEED1.5-VL token-count ratios.

Visual encoder. With MoonViT (Kimi-VL), the AR and DLM scores are 72.81 and 61.03; with Qwen3-ViT, they are 71.12 and 60.12. MoonViT-V2 is consistently strongest among these tested choices. These comparisons support the selected encoder within this training recipe; they do not establish a universal ranking across vision architectures. Section G.3.1 analyzes the task-dependent value of multi-level visual injection at the coordinate-token readout; this is a separate design question from the encoder comparison.

Stage III. Removing Stage III gives 71.74 for GroundAnything-VLM and 45.04 for GroundAnything, drops of 1.29 and 16.62 pp. The larger DLM sensitivity indicates that the final training stage is particularly important for recovering useful grounding behavior after conversion. The experiment changes a complete stage and does not isolate individual reward components or optimization choices within it.

Comparison scope. The full ablation scores (73.03/61.66) differ from the final 30-benchmark summary (72.42/61.75). We retain the ablation values and interpret only within-comparison changes, without replacing either set of scores or assuming identical aggregation.

#### G.3.1 Multi-level Visual Injection and Coordinate Readout

Connection to the encoder choice. GroundAnything uses MoonViT-V2 with an input-level projector and no intermediate visual injection. DeepStack supplies additional visual evidence at intermediate language layers (Meng et al., 2024); Qwen3-VL adds three projected ViT feature levels to its first three language layers and reports fine-grained understanding gains (Bai et al., 2025a). The relevant question here is whether grounding adaptation makes this evidence useful to the coordinate-token readout under both causal and masked contexts. Our encoder ablation supports MoonViT-V2 under the current recipe but does not isolate the injection scheme.

Readout sensitivity. Fix model parameters, image, query, and one decoding context: an AR prefix or a DLM block with fixed masks and visible tokens. For vectorized hidden states, write $h_{\ell+1} = F_\ell(h_\ell) + S_\ell u_\ell$ and $q = g(h_L)$, where $u_\ell$ contains projected visual features, $S_\ell$ inserts them at visual-token positions, and $q \in \mathbb{R}^V$ contains logits over a fixed vocabulary at one coordinate position. The readout $g$ includes final normalization and the output head; $\mathcal{I}$ indexes injection layers. With $h_0$ fixed, changing injected features by $\delta u_\ell$ gives

$$\begin{aligned} d &= \sum_{\ell \in \mathcal{I}} B_\ell \delta u_\ell + O(\|\delta u\|_2^2), \\ B_\ell &= Dg(h_L) J_{L-1} \cdots J_{\ell+1} S_\ell, \qquad J_j = DF_j(h_j). \end{aligned} \tag{S28}$$

Here $d$ is the exact logit change, $\delta u$ concatenates the feature changes, and empty products are identities; all derivatives are evaluated at the unchanged interface. The chain rule gives each path contribution; locally bounded second derivatives give the remainder. This is sensitivity relative to the specified interface, not prediction error relative to an ideal alignment.

Task alignment. Let $y$ denote the reference coordinate token and $\mathcal{L}_y(q) = -\log p_y$ its cross-entropy loss. Write $p = \mathrm{softmax}(q)$ and $p' = \mathrm{softmax}(q + d)$ for the original and updated token distributions. For the exact change $d$,

$$\begin{aligned} \mathcal{L}_y(q+d) - \mathcal{L}_y(q) &= \log\left(\sum_{v=1}^{V} p_v e^{d_v}\right) - d_y \\ &= (p - e_y)^\top d + D_{\mathrm{KL}}(p\|p'), \end{aligned} \tag{S29}$$

where $e_y$ is the one-hot target and $D_{\mathrm{KL}}(p\|p') = \sum_v p_v \log(p_v/p'_v)$. The first equality follows from log-softmax; expanding the KL divergence gives the second. Additional evidence lowers coordinate loss precisely when its contribution opposing the current loss gradient exceeds the nonnegative KL remainder. Averaging over matched reference coordinate positions and decoding contexts gives the corresponding token-loss comparison. The loss identity is exact; substituting the linearized change from Equation (S28) requires controlling its remainder.

If additional branches can be zeroed with all other components unchanged, the multi-injection model retains the single-interface model as a special case; path count alone cannot raise optimal task loss. The testable issue is whether finite grounding data and optimization learn useful contributions across causal and masked contexts. A controlled comparison should fix encoder and language-backbone initialization, coordinate vocabulary, data, and training budget, then compare input-only and multi-level interfaces with comparable capacity. Coordinate loss across adaptation budgets and dense/tiny-object localization would test this hypothesis; entropy-based commitment and final box metrics require separate evaluation beyond the fixed-context analysis. Disabling branches after training measures sensitivity only. Thus, the current ablation supports our encoder choice, while attributing its advantage to the absence of DeepStack requires an injection-specific experiment.

# H Comprehensive Grounding Benchmark Results

We expand the selected main-text comparisons with complete baseline tables, threshold-specific localization metrics, and task-level analysis. GroundAnything uses Entropy-Guided Decoding throughout these tables; GroundAnything-VLM uses AR decoding. External results and unresolved evaluations retain their original qualifications.

## Benchmark Suite and Metrics

Evaluation setup. We evaluate GroundAnything-VLM and GroundAnything against 44 baselines on 30 benchmarks spanning 11 perceptual capabilities, covering specialized detectors, general-purpose VLMs, grounding specialists, and embodied foundations. GPT-6 Astra uses thinking effort `High`. We report F1mIoU for box grounding and OCR, F1@Point for object pointing, and task-specific accuracy for spatial and GUI grounding.

For a fair comparison of diffusion language model (DLM) decoding, all GroundAnything results in the main benchmark tables (Tables 1–4) use entropy-guided diffusion decoding, without self-speculative decoding. The self-speculative setting provides higher throughput and quality at the reported operating point, but combines diffusion drafting with causal autoregressive (AR) verification and correction; it is therefore not pure DLM decoding. We analyze this setting separately in Section 4.2 and highlight it in the teaser (Figure 1).

Main-table reporting conventions. Scores are percentages; bold marks column bests, including ties (lower is better only for parse error). Model-name stars denote externally reported rows; entry-level stars denote source or task-interface exceptions. `–`, `N/A`, and `UNK` indicate unreported values, unsupported or unresolved evaluations, and unspecified zero-shot status, respectively. Daggers flag uncertain prompt/protocol alignment, so affected scores are descriptive. Under our protocols, GroundingDINO lacks GUI/OCR/layout interfaces, and Kimi-K3 lacks compatible pointing/OCR/layout outputs; SenseNova-Vision's GUI evaluation remains unresolved. LocateAnything lacks a supported visual-prompt interface. Starred SenseNova-Vision HierText/ICDAR2015 scores follow Han et al. (2026); its other OCR scores are local. Complete sources and exceptions appear in Section H.

Tasks and datasets. The suite covers common and long-tailed detection on COCO (Lin et al., 2014) and LVIS (Gupta et al., 2019); dense and tiny-object detection on Dense200 (Jiang et al.,

2026) and VisDrone (Zhu et al., 2018); referring grounding on RefCOCO, RefCOCO+, and RefCOCOg (Yu et al., 2016; Mao et al., 2016; Nagaraja et al., 2016); and object pointing on COCO, LVIS, Dense200, VisDrone, and RefCOCOg val/test. Spatial pointing uses RefSpatial (Zhou et al., 2026) and RoboSpatial (Song et al., 2025), while GUI grounding uses ScreenSpot-V2 (Wu et al., 2025), ScreenSpot-Pro (Li et al., 2025), and OSWorld-G (Xie et al., 2026). OCR covers HierText (Long et al., 2022), ICDAR2015 (Karatzas et al., 2015), TotalText (Ch'Ng & Chan, 2017), and SROIE (Huang et al., 2019); layout grounding uses DocLayNet (Pfitzmann et al., 2022) and M6Doc (Cheng et al., 2023); visual prompting uses FSC147 (Ranjan et al., 2021) and Dense200.

Metrics and aggregation. Following Jiang et al. (2026), SAM-derived masks (Kirillov et al., 2023) determine point correctness, with F1@Point balancing misses and false positives. Following the grounding evaluation convention of Jiang et al. (2026), box tasks report recall, precision, and F1 at IoU 0.50 and 0.95, together with their reported aggregates over IoU thresholds from 0.50 to 0.95. F1mIoU denotes the threshold-aggregated F1 score, not mean matched-box IoU. OCR additionally requires transcription agreement under the loose-match evaluation protocol and reports parse-error rates separately. Object pointing uses point-in-mask F1; spatial pointing uses point-in-mask accuracy, ScreenSpot uses action accuracy, and OSWorld-G uses exact accuracy. RefCOCO avg is the unweighted mean of the three family entries; RefSpatial avg averages Location and Placement. Reported benchmark entries are preserved, and missing metrics are not reconstructed from other scores.

Reporting conventions and external scores. The reporting conventions above apply throughout. R and P denote recall and precision; lower is better only for parse error. Model-name stars identify external-only rows, while entry-level stars identify individual substitutions or task-interface exceptions. The relevant captions specify their sources. In particular, selected detector, SEED1.5-VL, Molmo, spatial, and GUI references are taken from Jiang et al. (2026); GUI-Owl references follow Wang et al. (2026). Size groups refer to language-backbone configurations; the 7B understanding branch is used for MoT models and total language parameters for MoE models. Qwen3.7-Max, Kimi-K2.6, Kimi-K3, and GPT-6 Astra are grouped separately in the $>$ 1T category.

Interpreting unsupported or uncertain evaluations. N/A means that the required task interface or a reliable prompt/parser combination was unavailable; it does not establish that a model intrinsically lacks the capability. The $^{\dagger}$ marker identifies uncertain prompt/protocol alignment, including the affected Kimi, DeepSeek, and SenseNova-Vision evaluations. Repeated DeepSeek GUI attempts did not yield a suitable prompt/parser combination. Starred BAGEL and MiMo observations under suspected protocol incompatibility are retained descriptively. Where external scores replace unreproduced local results, only the matching model and available metrics are used: DeepSeek-VL2-Small scores are never substituted for the full model, and missing parse-error rates remain unreported. The following comparisons use the reported compatible entries.

### H.1 Common and Long-tailed Object Detection

COCO. GroundAnything-VLM achieves 63.70 F1mIoU, while direct diffusion retains 60.72, exceeding Rex-Omni by 4.44 pp. The threshold breakdown qualifies this result: GroundAnything's F1 at IoU 0.95 is 22.39, below LocateAnything Hybrid's 27.61, despite its higher aggregate. Thus, the overall gain does not imply uniformly tighter boxes at the strictest threshold.

Table S12: **COCO.** Complete box-grounding metrics. Model-name stars denote external scores from Table 2 of Jiang et al. (2026).

| Model | Zero-shot | IoU 0.50 | | | IoU 0.95 | | | mIoU | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | R | P | F1 | R | P | F1 | R | P | F1 |
| **Closed-set Specialized Detectors** | | | | | | | | | | |
| DINO-R50* (Zhang et al., 2022) | NO | 62.60 | 76.50 | 68.80 | 17.80 | 25.80 | 21.10 | 50.00 | 62.40 | 55.60 |
| DETR-R50* (Carion et al., 2020) | NO | 59.60 | 73.90 | 65.90 | 10.60 | 19.00 | 13.60 | 42.90 | 55.30 | 48.30 |
| DyHead-R50* (Dai et al., 2021) | NO | 58.10 | 76.60 | 66.10 | 11.90 | 20.60 | 15.00 | 44.80 | 60.10 | 51.30 |
| **Open-set Specialized Detectors** | | | | | | | | | | |
| GroundingDINO (Liu et al., 2024) | YES | 79.60 | 83.23 | 81.37 | 24.66 | 28.07 | 26.25 | 61.13 | 63.65 | 60.56 |
| **Vision-Language Models (<10B)** | | | | | | | | | | |
| Rex-Omni (Jiang et al., 2026) | YES | 69.90 | 80.99 | 75.04 | 17.93 | 21.73 | 19.65 | 53.92 | 59.36 | 56.28 |
| LocateAnything Fast (Wang et al., 2026) | YES | 66.15 | 73.34 | 69.56 | 24.73 | 26.90 | 25.77 | 50.89 | 59.26 | 53.06 |
| LocateAnything Hybrid (Wang et al., 2026) | YES | 74.67 | 80.80 | 77.61 | 26.64 | 28.66 | 27.61 | 52.40 | 65.16 | 59.12 |
| LocateAnything Slow NTP (Wang et al., 2026) | YES | 75.02 | 81.87 | 78.30 | 24.97 | 27.19 | 26.03 | 54.70 | 65.09 | 59.37 |
| Qwen2.5-VL-7B (Bai et al., 2025b) | UNK | 57.18 | 62.87 | 59.89 | 12.16 | 13.23 | 12.67 | 44.67 | 49.05 | 46.27 |
| Qwen3-VL-2B (Bai et al., 2025a) | UNK | 63.50 | 67.14 | 65.27 | 21.77 | 22.57 | 22.16 | 42.64 | 44.86 | 42.48 |
| Qwen3-VL-4B (Bai et al., 2025a) | UNK | 66.05 | 70.64 | 68.27 | 23.69 | 24.87 | 24.27 | 44.87 | 47.76 | 46.53 |
| Qwen3-VL-8B (Bai et al., 2025a) | UNK | 65.70 | 69.48 | 67.54 | 23.91 | 25.11 | 24.49 | 44.80 | 47.29 | 46.30 |
| Qwen3.5-4B (Qwen Team, 2026a) | UNK | 65.86 | 67.75 | 66.79 | 22.94 | 23.67 | 23.30 | 48.40 | 45.71 | 47.58 |
| Qwen3.5-9B (Qwen Team, 2026a) | UNK | 73.17 | 68.13 | 70.56 | 23.83 | 24.48 | 24.15 | 58.50 | 46.30 | 51.99 |
| RynnBrain1.1 (Li et al., 2026a) | UNK | 36.49 | 60.51 | 45.53 | 11.56 | 16.53 | 13.60 | 28.72 | 45.31 | 35.15 |
| SenseNova-Vision (Han et al., 2026) | UNK | 79.07 | 80.54 | 79.80 | **27.19** | **30.19** | **28.61** | 63.13 | 55.37 | 57.49 |
| MiMo-VL-7B-SFT (Yue et al., 2025) | UNK | 65.48 | 72.18 | 68.67 | 10.02 | 11.11 | 10.54 | 45.72 | 50.47 | 47.98 |
| MiMo-VL-7B-RL (Yue et al., 2025) | UNK | 65.64 | 74.52 | 69.79 | 9.17 | 10.29 | 9.69 | 44.48 | 50.24 | 47.18 |
| BAGEL (Deng et al., 2025) | UNK | 65.57 | 75.20 | 70.06 | 11.04 | 13.01 | 11.94 | 46.30 | 47.15 | 45.98 |
| RynnBrain (Dang et al., 2026) | UNK | 21.27 | 49.89 | 29.82 | 4.89 | 10.87 | 6.74 | 15.46 | 35.37 | 21.51 |
| GroundAnything-VLM | YES | 81.53 | 81.43 | **81.48** | 23.91 | 23.83 | 23.87 | 63.75 | 63.64 | **63.70** |
| GroundAnything | YES | 81.02 | 70.38 | 75.33 | 26.38 | 19.45 | 22.39 | 63.70 | 54.91 | 60.72 |
| **Vision-Language Models (10B–1T)** | | | | | | | | | | |
| Qwen3.6-27B (Qwen Team, 2026b) | UNK | 78.19 | 79.00 | 78.59 | 25.02 | 25.68 | 25.35 | 61.26 | 62.19 | 61.72 |
| Qwen3-VL-32B (Bai et al., 2025a) | UNK | 76.47 | 77.79 | 77.12 | 23.05 | 23.84 | 23.44 | 59.50 | 60.84 | 60.16 |
| Qwen3.8-27B (Qwen Team, 2026d) | UNK | 77.48 | 79.51 | 78.48 | 23.60 | 24.44 | 24.01 | 59.97 | 61.72 | 60.84 |
| Qwen3.5-35B-A3B (Qwen Team, 2026a) | UNK | 78.37 | 79.65 | 79.00 | 25.20 | 25.92 | 25.56 | 61.45 | 62.71 | 62.07 |
| DeepSeek-VL2-Small-16B (Wu et al., 2024) | UNK | 20.86 | 45.22 | 28.55 | 0.85 | 1.73 | 1.14 | 10.12 | 21.19 | 13.70 |
| DeepSeek-VL2-27B (Wu et al., 2024) | UNK | 34.58 | **84.65** | 49.10 | 12.89 | 28.12 | 17.67 | 28.60 | **67.97** | 40.25 |
| SEED1.5-VL* (Guo et al., 2025) | YES | 65.30 | 78.60 | 71.30 | 12.70 | 16.40 | 14.30 | 46.80 | 56.90 | 51.40 |
| **Vision-Language Models (>1T)** | | | | | | | | | | |
| Qwen3.7-Max (Qwen Team, 2026c) | UNK | 78.98 | 82.62 | 80.76 | 24.19 | 25.15 | 24.66 | 60.97 | 64.72 | 62.79 |
| Kimi-K2.6 | UNK | 72.27 | 80.10 | 75.99 | 25.04 | 27.19 | 26.07 | 58.28 | 64.34 | 61.16 |
| Kimi-K3 (Team et al., 2026) | UNK | 72.69 | 84.33 | 78.08 | 23.40 | 26.13 | 24.69 | 56.99 | 65.37 | 60.89 |
| GPT-6 Astra | UNK | **82.65** | 77.22 | 79.81 | 26.50 | 25.74 | 26.11 | **64.61** | 61.03 | 62.75 |

LVIS. GroundAnything reaches 52.59 F1mIoU, above Rex-Omni (46.74) and LocateAnything Fast (42.90), while GroundAnything-VLM reaches 56.63. This extends competitive grounding beyond common categories, although SenseNova-Vision remains stronger than the DLM variant at 56.12. The AR variant's 75.46 F1 at IoU 0.50 versus 23.23 at IoU 0.95 also shows that rare-category coverage and very tight localization remain distinct challenges.

Table S13: **LVIS.** Complete box-grounding metrics. The starred SEED1.5-VL scores are from Table 3 of Jiang et al. (2026).

| Model | Zero-shot | IoU 0.50 | | | IoU 0.95 | | | mIoU | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | R | P | F1 | R | P | F1 | R | P | F1 |
| **Open-set Specialized Detectors** | | | | | | | | | | |
| GroundingDINO (Liu et al., 2024) | YES | 52.64 | **82.61** | 64.30 | 23.34 | 26.55 | 24.84 | 44.30 | 59.58 | 52.61 |
| **Vision-Language Models (<10B)** | | | | | | | | | | |
| Rex-Omni (Jiang et al., 2026) | YES | 58.50 | 73.11 | 64.99 | 17.76 | 20.04 | 18.83 | 44.02 | 46.58 | 46.74 |
| LocateAnything Fast (Wang et al., 2026) | YES | 48.06 | 61.58 | 53.98 | 20.37 | 25.12 | 22.50 | 38.43 | 48.55 | 42.90 |
| LocateAnything Hybrid (Wang et al., 2026) | YES | 55.98 | 71.98 | 62.98 | 22.43 | 27.72 | 24.79 | 44.30 | 56.25 | 49.56 |
| LocateAnything Slow NTP (Wang et al., 2026) | YES | 57.98 | 75.03 | 65.41 | 20.96 | 26.03 | 23.22 | 44.95 | 57.46 | 50.44 |
| Qwen2.5-VL-7B (Bai et al., 2025b) | UNK | 49.52 | 66.52 | 56.78 | 8.16 | 9.87 | 8.93 | 33.37 | 43.60 | 37.80 |
| Qwen3-VL-2B (Bai et al., 2025a) | UNK | 55.20 | 68.93 | 61.30 | 16.50 | 19.12 | 17.71 | 40.66 | 49.71 | 44.73 |
| Qwen3-VL-4B (Bai et al., 2025a) | UNK | 59.60 | 76.16 | 66.87 | 18.71 | 22.25 | 20.32 | 44.84 | 56.15 | 49.86 |
| Qwen3-VL-8B (Bai et al., 2025a) | UNK | 59.47 | 74.39 | 66.10 | 18.00 | 21.19 | 19.46 | 44.25 | 54.49 | 48.84 |
| Qwen3.5-4B (Qwen Team, 2026a) | UNK | 58.19 | 70.41 | 63.72 | 17.87 | 20.52 | 19.10 | 42.65 | 50.95 | 46.43 |
| Qwen3.5-9B (Qwen Team, 2026a) | UNK | 60.86 | 72.03 | 65.98 | 19.33 | 22.00 | 20.58 | 45.28 | 53.07 | 48.87 |
| RynnBrain1.1 (Li et al., 2026a) | UNK | 26.86 | 52.27 | 35.48 | 8.76 | 13.56 | 10.64 | 20.18 | 36.64 | 26.01 |
| SenseNova-Vision (Han et al., 2026) | UNK | 64.50 | 75.97 | 69.76 | **27.09** | **30.94** | **28.88** | 52.08 | **60.84** | 56.12 |
| MiMo-VL-7B-SFT (Yue et al., 2025) | UNK | 42.88 | 60.78 | 50.28 | 6.26 | 8.50 | 7.21 | 28.39 | 39.89 | 33.17 |
| MiMo-VL-7B-RL (Yue et al., 2025) | UNK | 43.50 | 61.87 | 51.09 | 5.55 | 7.52 | 6.38 | 27.53 | 38.74 | 32.18 |
| BAGEL (Deng et al., 2025) | UNK | 34.68 | 51.47 | 41.44 | 8.74 | 11.58 | 9.96 | 24.88 | 35.76 | 29.34 |
| RynnBrain (Dang et al., 2026) | UNK | 15.06 | 43.29 | 22.35 | 3.24 | 8.90 | 4.75 | 10.34 | 28.98 | 15.25 |
| GroundAnything-VLM | YES | **72.91** | 78.18 | **75.46** | 22.75 | 23.72 | 23.23 | **54.94** | 58.42 | **56.63** |
| GroundAnything | YES | 69.98 | 67.68 | 68.81 | 27.01 | 19.78 | 22.84 | 53.50 | 54.73 | 52.59 |
| **Vision-Language Models (10B–1T)** | | | | | | | | | | |
| Qwen3.6-27B (Qwen Team, 2026b) | UNK | 62.89 | 75.25 | 68.52 | 19.87 | 22.54 | 21.12 | 47.15 | 55.67 | 51.05 |
| Qwen3-VL-32B (Bai et al., 2025a) | UNK | 61.21 | 72.44 | 66.36 | 18.01 | 20.44 | 19.15 | 45.41 | 53.20 | 48.99 |
| Qwen3.8-27B (Qwen Team, 2026d) | UNK | 62.86 | 74.17 | 68.05 | 18.79 | 21.14 | 19.90 | 46.43 | 54.15 | 49.99 |
| Qwen3.5-35B-A3B (Qwen Team, 2026a) | UNK | 62.87 | 74.82 | 68.33 | 20.04 | 22.75 | 21.31 | 47.14 | 55.41 | 50.94 |
| DeepSeek-VL2-Small-16B (Wu et al., 2024) | UNK | 12.80 | 28.41 | 17.65 | 0.50 | 1.20 | 0.71 | 6.10 | 13.05 | 8.31 |
| DeepSeek-VL2-27B (Wu et al., 2024) | UNK | 26.13 | 70.96 | 38.19 | 10.40 | 24.05 | 14.52 | 21.06 | 54.62 | 30.39 |
| SEED1.5-VL* (Guo et al., 2025) | YES | 54.70 | 82.00 | 65.60 | 15.00 | 28.10 | 19.50 | 38.50 | 59.30 | 46.70 |
| **Vision-Language Models (>1T)** | | | | | | | | | | |
| Qwen3.7-Max (Qwen Team, 2026c) | UNK | 63.86 | 79.06 | 70.65 | 20.27 | 23.49 | 21.76 | 48.54 | 59.04 | 53.28 |
| Kimi-K2.6 | UNK | 57.79 | 76.44 | 65.82 | 21.58 | 26.91 | 23.95 | 45.23 | 58.95 | 51.19 |
| Kimi-K3 (Team et al., 2026) | UNK | 57.17 | 76.66 | 65.50 | 17.46 | 21.54 | 19.29 | 42.08 | 55.14 | 47.73 |
| GPT-6 Astra | UNK | 71.35 | 73.85 | 72.58 | 23.47 | 24.60 | 24.02 | 54.23 | 55.73 | 54.97 |

### H.2 Dense and Tiny Object Detection

Dense200. GroundAnything-VLM and GroundAnything reach 75.55 and 70.04 F1mIoU, respectively, both above GPT-6 Astra (65.04), Rex-Omni (53.29), and LocateAnything Hybrid (50.07). GroundAnything also scores 26.86 at IoU 0.95 versus 12.64 for GPT-6 Astra, so its advantage is not restricted to coarse overlap. These results directly support precise visual evidence extraction when many neighboring instances must be represented in one response.

Table S14: **Dense200.** Complete box-grounding metrics. Local BAGEL and DeepSeek runs did not reproduce the reference results, so only available external scores are reported for these models. The starred BAGEL scores are taken from Table 1 of Han et al. (2026). No matching external result is available for full DeepSeek-VL2; Small and Tiny checkpoint results are not substituted. The starred DeepSeek-VL2-Small and SEED1.5-VL scores are from Table 4 of Jiang et al. (2026), also reported in Table 2 of Wang et al. (2026).

| Model | IoU 0.50 | | | IoU 0.95 | | | mIoU | | |
|---|---|---|---|---|---|---|---|---|---|
| | R | P | F1 | R | P | F1 | R | P | F1 |
| **Open-set Specialized Detectors** | | | | | | | | | |
| GroundingDINO (Liu et al., 2024) | 22.43 | 36.30 | 27.73 | 10.92 | 18.38 | 13.70 | 20.16 | 32.62 | 24.92 |
| **Vision-Language Models (<10B)** | | | | | | | | | |
| Rex-Omni (Jiang et al., 2026) | 70.61 | 75.13 | 72.80 | 8.66 | 9.17 | 8.91 | 51.79 | 54.88 | 53.29 |
| LocateAnything Fast (Wang et al., 2026) | 22.45 | 26.12 | 24.15 | 9.16 | 10.50 | 9.78 | 19.30 | 22.31 | 20.70 |
| LocateAnything Hybrid (Wang et al., 2026) | 58.06 | 63.32 | 60.57 | 21.52 | 22.90 | 22.19 | 48.13 | 52.18 | 50.07 |
| LocateAnything Slow NTP (Wang et al., 2026) | 74.89 | 81.46 | 78.03 | 20.64 | 22.11 | 21.35 | 59.65 | 64.62 | 62.04 |
| Qwen2.5-VL-7B (Bai et al., 2025b) | 0.50 | 0.57 | 0.53 | 0.00 | 0.00 | 0.00 | 0.22 | 0.24 | 0.23 |
| Qwen3-VL-2B (Bai et al., 2025a) | 9.93 | 13.28 | 11.36 | 1.27 | 1.93 | 1.53 | 7.18 | 9.65 | 8.23 |
| Qwen3-VL-4B (Bai et al., 2025a) | 17.61 | 22.94 | 19.92 | 2.68 | 3.76 | 3.13 | 12.33 | 16.23 | 14.02 |
| Qwen3-VL-8B (Bai et al., 2025a) | 15.42 | 16.40 | 15.90 | 2.26 | 2.58 | 2.41 | 11.13 | 11.77 | 11.44 |
| Qwen3.5-4B (Qwen Team, 2026a) | 43.44 | 48.47 | 45.82 | 8.20 | 9.01 | 8.58 | 32.20 | 36.12 | 34.04 |
| Qwen3.5-9B (Qwen Team, 2026a) | 39.02 | 42.00 | 40.46 | 7.37 | 8.19 | 7.76 | 28.82 | 31.33 | 30.02 |
| RynnBrain1.1 (Li et al., 2026a) | 0.10 | 4.00 | 0.19 | 0.02 | 0.50 | 0.04 | 0.06 | 2.40 | 0.12 |
| SenseNova-Vision (Han et al., 2026) | 79.18 | 86.84 | 82.83 | 24.52 | 25.72 | 25.10 | 65.31 | 71.20 | 68.13 |
| MiMo-VL-7B-SFT (Yue et al., 2025) | 11.75 | 12.38 | 12.06 | 0.14 | 0.14 | 0.14 | 5.62 | 5.95 | 5.78 |
| MiMo-VL-7B-RL (Yue et al., 2025) | 13.10 | 14.14 | 13.60 | 0.08 | 0.07 | 0.07 | 6.31 | 6.88 | 6.58 |
| BAGEL* (Deng et al., 2025) | – | – | – | – | – | – | – | – | 42.40 |
| RynnBrain (Dang et al., 2026) | 0.03 | 1.50 | 0.06 | 0.00 | 0.00 | 0.00 | 0.02 | 0.80 | 0.03 |
| GroundAnything-VLM | **90.38** | **92.93** | **91.64** | **27.75** | **28.45** | **28.09** | **74.58** | **76.54** | **75.55** |
| GroundAnything | 83.16 | 89.25 | 86.10 | 26.30 | 27.45 | 26.86 | 69.73 | 70.35 | 70.04 |
| **Vision-Language Models (10B–1T)** | | | | | | | | | |
| Qwen3.6-27B (Qwen Team, 2026b) | 51.74 | 56.09 | 53.83 | 9.91 | 10.76 | 10.32 | 39.12 | 42.56 | 40.77 |
| Qwen3-VL-32B (Bai et al., 2025a) | 24.97 | 28.35 | 26.56 | 2.89 | 4.04 | 3.37 | 17.78 | 20.69 | 19.12 |
| Qwen3.8-27B (Qwen Team, 2026d) | 44.31 | 47.44 | 45.82 | 8.92 | 9.56 | 9.23 | 33.28 | 35.40 | 34.30 |
| Qwen3.5-35B-A3B (Qwen Team, 2026a) | 46.43 | 49.58 | 47.95 | 9.22 | 10.00 | 9.60 | 35.18 | 37.36 | 36.24 |
| DeepSeek-VL2-Small-16B* (Wu et al., 2024) | – | – | 16.00 | – | – | 3.90 | – | – | 12.70 |
| DeepSeek-VL2-27B* (Wu et al., 2024) | – | – | – | – | – | – | – | – | – |
| SEED1.5-VL* (Guo et al., 2025) | – | – | 76.90 | – | – | 5.30 | – | – | 53.20 |
| **Vision-Language Models (>1T)** | | | | | | | | | |
| Qwen3.7-Max (Qwen Team, 2026c) | 37.56 | 43.69 | 40.40 | 7.20 | 8.65 | 7.86 | 28.94 | 33.84 | 31.20 |
| Kimi-K2.6 | 55.64 | 59.00 | 57.27 | 9.21 | 9.91 | 9.55 | 41.38 | 44.04 | 42.67 |
| Kimi-K3 (Team et al., 2026) | 65.95 | 78.99 | 71.88 | 10.40 | 11.88 | 11.09 | 47.61 | 56.40 | 51.64 |
| GPT-6 Astra | 89.01 | 84.63 | 86.76 | 12.90 | 12.39 | 12.64 | 66.54 | 63.61 | 65.04 |

VisDrone. GroundAnything improves over Rex-Omni from 27.19 to 33.76 F1mIoU, but remains below GroundingDINO (34.47) and SenseNova-Vision (42.35). GroundAnything-VLM scores 40.74. Both variants have low F1 at IoU 0.95 (3.42 and 2.24), identifying precise tiny-object boundaries as a remaining limitation despite the stronger Dense200 results.

Table S15: **VisDrone.** Complete box-grounding metrics. Local BAGEL and DeepSeek runs did not reproduce the reference results, so only available external scores are reported for these models. The starred BAGEL scores are taken from Table 1 of Han et al. (2026). No matching external result is available for full DeepSeek-VL2; Small and Tiny checkpoint results are not substituted. The starred DeepSeek-VL2-Small and SEED1.5-VL scores are from Table 4 of Jiang et al. (2026), also reported in Table 2 of Wang et al. (2026).

| Model | IoU 0.50 | | | IoU 0.95 | | | mIoU | | |
|---|---|---|---|---|---|---|---|---|---|
| | R | P | F1 | R | P | F1 | R | P | F1 |
| **Open-set Specialized Detectors** | | | | | | | | | |
| GroundingDINO (Liu et al., 2024) | 32.38 | **87.09** | 47.21 | 2.81 | **7.42** | **4.08** | 23.65 | **63.54** | 34.47 |
| **Vision-Language Models (<10B)** | | | | | | | | | |
| Rex-Omni (Jiang et al., 2026) | 42.47 | 52.33 | 46.89 | 1.07 | 1.27 | 1.16 | 24.78 | 30.12 | 27.19 |
| LocateAnything Fast (Wang et al., 2026) | 13.58 | 15.44 | 14.45 | 0.85 | 0.97 | 0.90 | 9.27 | 10.44 | 9.82 |
| LocateAnything Hybrid (Wang et al., 2026) | 40.66 | 46.99 | 43.59 | 2.34 | 2.64 | 2.48 | 26.78 | 30.62 | 28.57 |
| LocateAnything Slow NTP (Wang et al., 2026) | 59.28 | 70.06 | 64.22 | 2.73 | 3.13 | 2.91 | 37.32 | 43.81 | 40.31 |
| Qwen2.5-VL-7B (Bai et al., 2025b) | 27.63 | 42.83 | 33.59 | 0.44 | 0.63 | 0.52 | 15.48 | 23.55 | 18.67 |
| Qwen3-VL-2B (Bai et al., 2025a) | 37.16 | 46.15 | 41.17 | 1.16 | 1.46 | 1.29 | 23.01 | 28.23 | 25.36 |
| Qwen3-VL-4B (Bai et al., 2025a) | 47.85 | 52.36 | 50.00 | 1.78 | 1.86 | 1.82 | 29.92 | 32.47 | 31.14 |
| Qwen3-VL-8B (Bai et al., 2025a) | 46.61 | 44.84 | 45.71 | 1.74 | 1.70 | 1.72 | 28.80 | 27.81 | 28.29 |
| Qwen3.5-4B (Qwen Team, 2026a) | 54.93 | 50.78 | 52.77 | 2.60 | 2.53 | 2.57 | 34.91 | 32.71 | 33.77 |
| Qwen3.5-9B (Qwen Team, 2026a) | 57.52 | 45.04 | 50.52 | 2.87 | 2.48 | 2.66 | 36.75 | 29.69 | 32.84 |
| RynnBrain1.1 (Li et al., 2026a) | 7.42 | 46.41 | 12.80 | 0.38 | 2.05 | 0.64 | 4.83 | 29.23 | 8.29 |
| SenseNova-Vision (Han et al., 2026) | **65.10** | 69.99 | **67.45** | 2.86 | 3.07 | 2.96 | **40.91** | 43.88 | **42.35** |
| MiMo-VL-7B-SFT (Yue et al., 2025) | 26.42 | 28.73 | 27.53 | 0.14 | 0.15 | 0.15 | 12.60 | 13.88 | 13.21 |
| MiMo-VL-7B-RL (Yue et al., 2025) | 27.94 | 32.86 | 30.20 | 0.28 | 0.34 | 0.31 | 13.79 | 16.62 | 15.07 |
| BAGEL* (Deng et al., 2025) | – | – | – | – | – | – | – | – | 23.00 |
| RynnBrain (Dang et al., 2026) | 8.78 | 32.85 | 13.86 | 0.20 | 0.75 | 0.32 | 4.97 | 18.52 | 7.84 |
| GroundAnything-VLM | 57.81 | 69.16 | 62.98 | **3.17** | 3.71 | 3.42 | 37.55 | 44.52 | 40.74 |
| GroundAnything | 46.94 | 54.60 | 50.48 | 2.09 | 2.42 | 2.24 | 32.52 | 35.10 | 33.76 |
| **Vision-Language Models (10B–1T)** | | | | | | | | | |
| Qwen3.6-27B (Qwen Team, 2026b) | 57.36 | 56.32 | 56.83 | 2.61 | 2.73 | 2.67 | 36.44 | 36.43 | 36.44 |
| Qwen3-VL-32B (Bai et al., 2025a) | 48.94 | 46.14 | 47.50 | 1.48 | 1.46 | 1.47 | 29.51 | 28.24 | 28.86 |
| Qwen3.8-27B (Qwen Team, 2026d) | 56.44 | 49.01 | 52.46 | 2.76 | 2.62 | 2.69 | 35.48 | 31.42 | 33.33 |
| Qwen3.5-35B-A3B (Qwen Team, 2026a) | 59.78 | 50.81 | 54.93 | 2.92 | 2.76 | 2.84 | 37.97 | 32.86 | 35.23 |
| DeepSeek-VL2-Small-16B* (Wu et al., 2024) | – | – | 35.80 | – | – | 1.70 | – | – | 23.30 |
| DeepSeek-VL2-27B* (Wu et al., 2024) | – | – | – | – | – | – | – | – | – |
| SEED1.5-VL* (Guo et al., 2025) | – | – | 55.90 | – | – | 0.60 | – | – | 27.40 |
| **Vision-Language Models (>1T)** | | | | | | | | | |
| Qwen3.7-Max (Qwen Team, 2026c) | 60.27 | 71.40 | 65.36 | 2.84 | 3.45 | 3.11 | 38.28 | 45.52 | 41.59 |
| Kimi-K2.6 | 41.36 | 65.79 | 50.79 | 1.20 | 1.73 | 1.42 | 24.42 | 38.73 | 29.95 |
| Kimi-K3 (Team et al., 2026) | 41.51 | 76.49 | 53.81 | 0.98 | 1.59 | 1.21 | 23.57 | 42.47 | 30.31 |
| GPT-6 Astra | 64.60 | 65.83 | 65.18 | 2.69 | 2.80 | 2.74 | 36.70 | 37.58 | 37.12 |

Why strict overlap is difficult for small objects. For equal axis-aligned boxes of width $w > 0$ and height $h > 0$ separated only horizontally by $\delta$, with $|\delta| < w$,

$$\mathrm{IoU}(\delta) = \frac{w - |\delta|}{w + |\delta|}, \qquad \mathrm{IoU} \geq \eta \iff |\delta| \leq w\frac{1-\eta}{1+\eta}, \quad 0 < \eta < 1. \tag{S30}$$

The intersection and union areas are $(w-|\delta|)h$ and $(w+|\delta|)h$; IoU is zero for $|\delta| \geq w$. At $\eta = 0.75$, the permissible shift is only $w/7$. Under this translation-only model, the failure rate at a given width equals the probability of exceeding this displacement threshold, a prediction testable from measured offsets. The calculation explains size sensitivity without identifying its architectural cause; size errors and missed instances require separate analysis.

### H.3 Referring Object Detection

RefCOCOg val/test. On RefCOCOg, GroundAnything reaches 91.61 F1mIoU on validation and 91.10 on test, versus 84.37 and 83.56 for its AR counterpart. LocateAnything Hybrid scores 76.43 and 77.67. The reported gains persist at IoU 0.95, where GroundAnything scores 88.63/87.15. Within this evaluation, the improvement therefore reflects precise language-conditioned localization rather than only successful target identification.

Table S16: **RefCOCOg validation and test.** Complete recorded F1 metrics at IoU 0.50, 0.95, and mIoU. The starred BAGEL scores are taken from Table 1 of Han et al. (2026), because local runs did not reproduce the reported performance. The starred SEED1.5-VL scores are from Table 5 of Jiang et al. (2026).

| Model | RefCOCOg val | | | RefCOCOg test | | |
|---|---|---|---|---|---|---|
| | F1@.50 | F1@.95 | F1mIoU | F1@.50 | F1@.95 | F1mIoU |
| **Open-set Specialized Detectors** | | | | | | |
| GroundingDINO (Liu et al., 2024) | 58.37 | 22.47 | 49.77 | 58.52 | 24.38 | 50.43 |
| **Vision-Language Models (<10B)** | | | | | | |
| Rex-Omni (Jiang et al., 2026) | 87.01 | 35.23 | 73.90 | 87.36 | 36.51 | 74.76 |
| LocateAnything Fast (Wang et al., 2026) | 87.99 | 39.39 | 75.30 | 88.34 | 42.08 | 76.50 |
| LocateAnything Hybrid (Wang et al., 2026) | 88.50 | 40.40 | 76.43 | 88.91 | 42.63 | 77.67 |
| LocateAnything Slow NTP (Wang et al., 2026) | 88.18 | 34.81 | 74.93 | 88.64 | 36.99 | 76.46 |
| Qwen2.5-VL-7B (Bai et al., 2025b) | 78.19 | 15.10 | 61.86 | 78.49 | 16.31 | 62.93 |
| Qwen3-VL-2B (Bai et al., 2025a) | 85.25 | 32.13 | 71.80 | 85.87 | 33.88 | 72.64 |
| Qwen3-VL-4B (Bai et al., 2025a) | 88.53 | 36.61 | 75.27 | 88.69 | 36.61 | 75.88 |
| Qwen3-VL-8B (Bai et al., 2025a) | 88.92 | 35.61 | 75.79 | 89.26 | 37.03 | 76.26 |
| Qwen3.5-4B (Qwen Team, 2026a) | 89.24 | 34.89 | 75.03 | 89.00 | 35.14 | 75.54 |
| Qwen3.5-9B (Qwen Team, 2026a) | 89.89 | 36.63 | 76.20 | 89.42 | 36.89 | 76.28 |
| RynnBrain1.1 (Li et al., 2026a) | 83.21 | 24.36 | 67.66 | 83.75 | 25.14 | 68.24 |
| SenseNova-Vision (Han et al., 2026) | 89.94 | 43.58 | 78.69 | 89.85 | 44.91 | 79.48 |
| MiMo-VL-7B-SFT (Yue et al., 2025) | 86.51 | 14.72 | 66.44 | 86.59 | 15.26 | 66.78 |
| MiMo-VL-7B-RL (Yue et al., 2025) | 88.28 | 12.56 | 64.89 | 87.69 | 12.92 | 64.19 |
| BAGEL* (Deng et al., 2025) | – | – | 76.40 | – | – | 77.80 |
| RynnBrain (Dang et al., 2026) | 73.72 | 17.32 | 57.99 | 73.77 | 18.29 | 58.38 |
| GroundAnything-VLM | **94.11** | 51.46 | 84.37 | **93.75** | 48.32 | 83.56 |
| GroundAnything | 92.66 | **88.63** | **91.61** | 92.78 | **87.15** | **91.10** |
| **Vision-Language Models (10B–1T)** | | | | | | |
| Qwen3.6-27B (Qwen Team, 2026b) | 91.66 | 38.19 | 77.76 | 91.36 | 38.80 | 78.43 |
| Qwen3-VL-32B (Bai et al., 2025a) | 86.27 | 34.17 | 73.79 | 86.03 | 34.83 | 74.02 |
| Qwen3.8-27B (Qwen Team, 2026d) | 90.14 | 35.69 | 76.03 | 90.86 | 36.98 | 77.37 |
| Qwen3.5-35B-A3B (Qwen Team, 2026a) | 90.92 | 38.46 | 77.32 | 90.88 | 39.09 | 78.38 |
| DeepSeek-VL2-Small-16B (Wu et al., 2024) | 59.86 | 0.90 | 26.18 | 61.06 | 0.89 | 27.49 |
| DeepSeek-VL2-27B (Wu et al., 2024) | 92.46 | 55.36 | 82.73 | 92.19 | 58.33 | 84.07 |
| SEED1.5-VL* (Guo et al., 2025) | 84.70 | 30.90 | 71.90 | 85.20 | 32.10 | 73.20 |
| **Vision-Language Models (>1T)** | | | | | | |
| Qwen3.7-Max (Qwen Team, 2026c) | 91.56 | 48.45 | 80.54 | 92.36 | 48.75 | 81.71 |
| Kimi-K2.6 | 80.15 | 41.17 | 70.33 | 81.19 | 42.36 | 71.87 |
| Kimi-K3 (Team et al., 2026) | 88.02 | 34.39 | 73.39 | 87.92 | 35.06 | 73.99 |
| GPT-6 Astra | 92.13 | 35.89 | 74.98 | 89.93 | 40.25 | 78.91 |

RefCOCO family. GroundAnything scores 88.49, 83.06, and 84.43 on RefCOCO, RefCOCOg, and RefCOCO+, giving an 85.33 mean versus 83.43 for GroundAnything-VLM. All three exceed LocateAnything Fast, although the larger DeepSeek-VL2-27B remains stronger on these family

entries. These results support broad referring competence without claiming a universal lead over every model size or evaluation split.

Table S17: **RefCOCO family.** The three datasets are reported separately; their F1mIoU arithmetic mean is RefCOCO avg in the main text.

| Model | RefCOCO | | | RefCOCOg | | | RefCOCOplus | | |
|---|---|---|---|---|---|---|---|---|---|
| | F1@.50 | F1@.95 | F1mIoU | F1@.50 | F1@.95 | F1mIoU | F1@.50 | F1@.95 | F1mIoU |
| **Open-set Specialized Detectors** | | | | | | | | | |
| GroundingDINO (Liu et al., 2024) | 51.19 | 23.28 | 44.33 | 57.99 | 23.17 | 49.69 | 49.20 | 20.89 | 41.42 |
| **Vision-Language Models (<10B)** | | | | | | | | | |
| Rex-Omni (Jiang et al., 2026) | 84.37 | 32.85 | 71.43 | 84.74 | 34.74 | 72.34 | 76.92 | 29.53 | 64.10 |
| LocateAnything Fast (Wang et al., 2026) | 92.48 | 43.95 | 80.61 | 88.47 | 40.69 | 76.17 | 84.43 | 39.53 | 73.08 |
| LocateAnything Hybrid (Wang et al., 2026) | 92.73 | 44.32 | 81.37 | 89.42 | 41.46 | 77.73 | 85.80 | 40.57 | 75.08 |
| LocateAnything Slow NTP (Wang et al., 2026) | 92.43 | 38.43 | 80.21 | 88.65 | 35.63 | 76.08 | 85.22 | 35.69 | 73.90 |
| Qwen2.5-VL-7B (Bai et al., 2025b) | 82.42 | 17.71 | 66.93 | 75.28 | 15.56 | 60.28 | 73.04 | 16.39 | 59.27 |
| Qwen3-VL-2B (Bai et al., 2025a) | 88.33 | 26.65 | 72.39 | 85.89 | 25.73 | 69.81 | 81.21 | 24.34 | 66.47 |
| Qwen3-VL-4B (Bai et al., 2025a) | 92.09 | 32.75 | 77.63 | 88.79 | 32.25 | 74.73 | 86.76 | 30.94 | 73.36 |
| Qwen3-VL-8B (Bai et al., 2025a) | 91.11 | 29.87 | 75.83 | 88.74 | 28.82 | 73.33 | 86.43 | 29.02 | 71.98 |
| Qwen3.5-4B (Qwen Team, 2026a) | 90.11 | 29.07 | 74.46 | 88.42 | 26.98 | 71.61 | 84.13 | 27.79 | 69.48 |
| Qwen3.5-9B (Qwen Team, 2026a) | 91.97 | 35.09 | 78.31 | 89.53 | 33.47 | 75.27 | 87.45 | 33.90 | 74.64 |
| RynnBrain1.1 (Li et al., 2026a) | 82.61 | 20.88 | 65.80 | 84.28 | 21.48 | 67.57 | 72.45 | 19.18 | 57.52 |
| SenseNova-Vision (Han et al., 2026) | 90.12 | 45.07 | 79.66 | 89.18 | 44.73 | 78.69 | 84.40 | 41.68 | 74.28 |
| MiMo-VL-7B-SFT (Yue et al., 2025) | 86.05 | 14.45 | 66.26 | 84.87 | 14.33 | 64.76 | 76.79 | 13.13 | 58.86 |
| MiMo-VL-7B-RL (Yue et al., 2025) | 90.44 | 14.15 | 68.45 | 87.69 | 13.11 | 64.80 | 84.83 | 13.55 | 64.42 |
| BAGEL (Deng et al., 2025) | 79.39 | 25.09 | 65.23 | 77.69 | 22.20 | 62.30 | 68.34 | 20.45 | 54.45 |
| RynnBrain (Dang et al., 2026) | 73.99 | 13.60 | 55.59 | 76.49 | 14.29 | 57.91 | 65.51 | 11.72 | 48.88 |
| GroundAnything-VLM | **95.54** | 48.01 | 85.53 | 92.66 | 43.78 | 81.62 | **92.65** | 48.01 | 83.15 |
| GroundAnything | 91.52 | **78.64** | 88.49 | 88.00 | **68.78** | 83.06 | 87.73 | **73.94** | 84.43 |
| **Vision-Language Models (10B–1T)** | | | | | | | | | |
| Qwen3.6-27B (Qwen Team, 2026b) | 93.46 | 39.83 | 80.78 | 91.33 | 38.69 | 78.37 | 89.57 | 38.36 | 77.44 |
| Qwen3-VL-32B (Bai et al., 2025a) | 90.33 | 34.50 | 77.21 | 84.49 | 31.54 | 71.75 | 85.34 | 33.08 | 73.17 |
| Qwen3.8-27B (Qwen Team, 2026d) | 92.93 | 38.44 | 79.83 | 90.37 | 37.15 | 76.91 | 89.07 | 37.76 | 76.91 |
| Qwen3.5-35B-A3B (Qwen Team, 2026a) | 92.57 | 36.44 | 79.30 | 90.62 | 34.77 | 76.49 | 87.61 | 34.38 | 75.06 |
| DeepSeek-VL2-Small-16B (Wu et al., 2024) | 66.95 | 1.10 | 30.95 | 64.47 | 1.29 | 29.60 | 64.93 | 1.25 | 30.58 |
| DeepSeek-VL2-27B (Wu et al., 2024) | 94.73 | 74.35 | **90.43** | **94.02** | 65.74 | **87.32** | 91.33 | 71.62 | **87.19** |
| **Vision-Language Models (>1T)** | | | | | | | | | |
| Qwen3.7-Max (Qwen Team, 2026c) | 95.11 | 45.55 | 82.99 | 92.78 | 44.87 | 80.38 | 91.78 | 32.41 | 77.23 |
| Kimi-K2.6 | 82.55 | 39.91 | 72.27 | 82.06 | 40.18 | 71.83 | 76.00 | 33.87 | 64.89 |
| Kimi-K3 (Team et al., 2026) | 88.46 | 35.78 | 75.12 | 88.57 | 34.51 | 74.14 | 81.86 | 33.77 | 69.31 |
| GPT-6 Astra | 93.61 | 37.14 | 79.45 | 87.21 | 44.27 | 76.15 | 90.45 | 36.30 | 77.84 |

### H.4 Object Pointing

Following Jiang et al. (2026), SAM (Kirillov et al., 2023) converts ground-truth boxes to object masks. A predicted point is correct when it lies inside the corresponding mask; F1@Point balances point precision and recall.

Referring pointing. On RefCOCOg val/test, GroundAnything-VLM reaches 90.44 and 91.03 F1@Point. GroundAnything scores 85.75/85.73, modestly exceeding Rex-Omni (84.96/85.32) while remaining below its AR counterpart. Unlike referring boxes, referring points do not improve after conversion in these results, demonstrating that the diffusion–AR trade-off depends on the output primitive.

Table S18: **Referring object pointing.** Starred BAGEL scores are retained despite unreliable support for the unified pointing protocol. Kimi-K3, both MiMo variants, and both DeepSeek variants are N/A under that protocol. These outcomes do not establish intrinsic pointing capability. The starred Molmo and SEED1.5-VL scores are from Table 7 of Jiang et al. (2026); its Molmo checkpoint is Molmo-7B-D.

| Model | RefCOCOg val | RefCOCOg test |
|---|---|---|
| | F1@Point | F1@Point |
| **Open-set Specialized Detectors** | | |
| GroundingDINO (Liu et al., 2024) | 49.34 | 49.97 |
| **Vision-Language Models (<10B)** | | |
| Rex-Omni (Jiang et al., 2026) | 84.96 | 85.32 |
| LocateAnything Fast (Wang et al., 2026) | 73.84 | 74.94 |
| LocateAnything Hybrid (Wang et al., 2026) | 75.89 | 76.65 |
| LocateAnything Slow NTP (Wang et al., 2026) | 77.17 | 77.59 |
| Qwen2.5-VL-7B (Bai et al., 2025b) | 81.65 | 82.48 |
| Qwen3-VL-2B (Bai et al., 2025a) | 76.17 | 76.01 |
| Qwen3-VL-4B (Bai et al., 2025a) | 76.43 | 77.64 |
| Qwen3-VL-8B (Bai et al., 2025a) | 81.97 | 82.09 |
| Qwen3.5-4B (Qwen Team, 2026a) | 79.35 | 79.31 |
| Qwen3.5-9B (Qwen Team, 2026a) | 77.59 | 77.85 |
| RynnBrain1.1 (Li et al., 2026a) | 74.42 | 74.17 |
| SenseNova-Vision (Han et al., 2026) | 74.63 | 75.42 |
| MiMo-VL-7B-SFT (Yue et al., 2025) | N/A* | N/A* |
| MiMo-VL-7B-RL (Yue et al., 2025) | N/A* | N/A* |
| BAGEL (Deng et al., 2025) | 55.92* | 54.57* |
| RynnBrain (Dang et al., 2026) | 73.34 | 73.67 |
| Molmo-7B* (Deitke et al., 2025) | 83.70 | 83.60 |
| GroundAnything-VLM | **90.44** | **91.03** |
| GroundAnything | 85.75 | 85.73 |
| **Vision-Language Models (10B–1T)** | | |
| Qwen3.6-27B (Qwen Team, 2026b) | 82.73 | 82.88 |
| Qwen3-VL-32B (Bai et al., 2025a) | 76.52 | 76.23 |
| Qwen3.8-27B (Qwen Team, 2026d) | 75.84 | 75.86 |
| Qwen3.5-35B-A3B (Qwen Team, 2026a) | 75.35 | 76.09 |
| DeepSeek-VL2-Small-16B (Wu et al., 2024) | N/A* | N/A* |
| DeepSeek-VL2-27B (Wu et al., 2024) | N/A* | N/A* |
| SEED1.5-VL* (Guo et al., 2025) | 83.60 | 84.20 |
| **Vision-Language Models (>1T)** | | |
| Qwen3.7-Max (Qwen Team, 2026c) | 71.21 | 72.40 |
| Kimi-K2.6 | 39.59† | 39.60† |
| Kimi-K3 (Team et al., 2026) | N/A* | N/A* |
| GPT-6 Astra | 87.80 | 84.90 |

Common and long-tailed pointing. GroundAnything-VLM attains 84.92 on COCO and 79.81 on LVIS; GroundAnything decreases to 71.32 and 64.57. For the DLM variant, precision is lower than recall on both datasets: 66.42 versus 77.00 on COCO and 58.67 versus 71.80 on LVIS. This imbalance points to excess or incorrectly localized point predictions as an important source of the gap, rather than missed instances alone.

Table S19: **Object pointing on COCO and LVIS.** Starred BAGEL scores are retained despite unreliable support for the unified pointing protocol. Kimi-K3, both MiMo variants, and both DeepSeek variants are N/A under that protocol. These outcomes do not establish intrinsic pointing capability. The starred Molmo and SEED1.5-VL scores are from Table 7 of Jiang et al. (2026); its Molmo checkpoint is Molmo-7B-D.

| Model | COCO | | | LVIS | | |
|---|---|---|---|---|---|---|
| | R@Point | P@Point | F1@Point | R@Point | P@Point | F1@Point |
| **Open-set Specialized Detectors** | | | | | | |
| GroundingDINO (Liu et al., 2024) | 68.92 | 71.97 | 70.41 | 44.91 | 71.15 | 55.07 |
| **Vision-Language Models (<10B)** | | | | | | |
| Rex-Omni (Jiang et al., 2026) | 77.81 | 81.77 | 79.74 | 63.46 | 78.14 | 70.04 |
| LocateAnything Fast (Wang et al., 2026) | 68.99 | 74.26 | 71.53 | 55.81 | 70.05 | 62.13 |
| LocateAnything Hybrid (Wang et al., 2026) | 73.80 | 73.77 | 73.78 | 60.96 | 69.36 | 64.89 |
| LocateAnything Slow NTP (Wang et al., 2026) | 74.68 | 75.05 | 74.86 | 62.36 | 72.42 | 67.01 |
| Qwen2.5-VL-7B (Bai et al., 2025b) | 61.24 | 65.76 | 63.42 | 46.46 | 56.82 | 51.12 |
| Qwen3-VL-2B (Bai et al., 2025a) | 55.65 | 54.98 | 55.31 | 43.70 | 50.11 | 46.69 |
| Qwen3-VL-4B (Bai et al., 2025a) | 63.13 | 67.69 | 65.33 | 49.45 | 62.17 | 55.08 |
| Qwen3-VL-8B (Bai et al., 2025a) | 64.92 | 66.76 | 65.83 | 52.25 | 61.41 | 56.46 |
| Qwen3.5-4B (Qwen Team, 2026a) | 70.15 | 68.85 | 69.50 | 55.61 | 64.72 | 59.82 |
| Qwen3.5-9B (Qwen Team, 2026a) | 71.87 | 72.56 | 72.21 | 58.90 | 70.07 | 64.00 |
| RynnBrain1.1 (Li et al., 2026a) | 21.06 | 32.96 | 25.70 | 13.35 | 25.63 | 17.56 |
| SenseNova-Vision (Han et al., 2026) | 70.85 | 75.21 | 72.96 | 57.21 | 69.26 | 62.66 |
| MiMo-VL-7B-SFT (Yue et al., 2025) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| MiMo-VL-7B-RL (Yue et al., 2025) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| BAGEL (Deng et al., 2025) | 33.57* | 39.16* | 36.15* | 22.73* | 31.65* | 26.46* |
| RynnBrain (Dang et al., 2026) | 9.46 | 17.73 | 12.33 | 5.38 | 13.73 | 7.73 |
| Molmo-7B* (Deitke et al., 2025) | – | – | 77.30 | – | – | 40.30 |
| GroundAnything-VLM | 84.77 | **85.08** | **84.92** | **76.56** | **83.35** | **79.81** |
| GroundAnything | 77.00 | 66.42 | 71.32 | 71.80 | 58.67 | 64.57 |
| **Vision-Language Models (10B–1T)** | | | | | | |
| Qwen3.6-27B (Qwen Team, 2026b) | 74.01 | 71.98 | 72.98 | 62.98 | 72.61 | 67.45 |
| Qwen3-VL-32B (Bai et al., 2025a) | 72.70 | 70.24 | 71.45 | 60.44 | 66.42 | 63.29 |
| Qwen3.8-27B (Qwen Team, 2026d) | 72.69 | 75.38 | 74.01 | 61.98 | 74.47 | 67.65 |
| Qwen3.5-35B-A3B (Qwen Team, 2026a) | 72.21 | 71.42 | 71.81 | 61.05 | 70.99 | 65.65 |
| DeepSeek-VL2-Small-16B (Wu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| DeepSeek-VL2-27B (Wu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| SEED1.5-VL* (Guo et al., 2025) | – | – | 78.20 | – | – | 70.70 |
| **Vision-Language Models (>1T)** | | | | | | |
| Qwen3.7-Max (Qwen Team, 2026c) | 71.61 | 72.65 | 72.13 | 60.97 | 74.22 | 66.94 |
| Kimi-K2.6 | 29.83† | 31.52† | 30.65† | 23.20† | 28.32† | 25.51† |
| Kimi-K3 (Team et al., 2026) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| GPT-6 Astra | **86.56** | 78.14 | 82.17 | 74.92 | 79.50 | 77.14 |

Dense and tiny-object pointing. GroundAnything scores 77.14 on Dense200 and 57.94 on VisDrone, exceeding Rex-Omni on both. GroundAnything-VLM reaches 84.16 and 68.03, while GPT-6 Astra remains stronger on Dense200 at 86.57. The remaining gap to the AR variant confirms that strong dense box grounding does not automatically guarantee equally strong point-set prediction.

Table S20: **Object pointing on Dense200 and VisDrone.** Starred BAGEL scores are retained despite unreliable support for the unified pointing protocol. Kimi-K3, both MiMo variants, and both DeepSeek variants are N/A under that protocol. These outcomes do not establish intrinsic pointing capability. The starred Molmo and SEED1.5-VL scores are from Table 7 of Jiang et al. (2026); its Molmo checkpoint is Molmo-7B-D.

| Model | Dense200 | | | VisDrone | | |
|---|---|---|---|---|---|---|
| | R@Point | P@Point | F1@Point | R@Point | P@Point | F1@Point |
| **Open-set Specialized Detectors** | | | | | | |
| GroundingDINO (Liu et al., 2024) | 22.00 | 65.45 | 32.93 | 27.20 | 71.76 | 39.45 |
| **Vision-Language Models (<10B)** | | | | | | |
| Rex-Omni (Jiang et al., 2026) | 75.59 | 77.76 | 76.66 | 47.81 | 56.92 | 51.97 |
| LocateAnything Fast (Wang et al., 2026) | 64.63 | 66.65 | 65.63 | 55.22 | 57.55 | 56.36 |
| LocateAnything Hybrid (Wang et al., 2026) | 77.43 | 78.73 | 78.07 | 59.18 | 55.54 | 57.30 |
| LocateAnything Slow NTP (Wang et al., 2026) | 78.87 | 81.39 | 80.11 | 61.17 | 60.77 | 60.97 |
| Qwen2.5-VL-7B (Bai et al., 2025b) | 12.12 | 36.42 | 18.19 | 12.81 | 18.21 | 15.04 |
| Qwen3-VL-2B (Bai et al., 2025a) | 14.06 | 15.95 | 14.95 | 10.64 | 11.98 | 11.27 |
| Qwen3-VL-4B (Bai et al., 2025a) | 14.13 | 46.93 | 21.72 | 21.27 | 26.25 | 23.50 |
| Qwen3-VL-8B (Bai et al., 2025a) | 20.61 | 32.96 | 25.36 | 17.68 | 18.54 | 18.10 |
| Qwen3.5-4B (Qwen Team, 2026a) | 56.82 | 60.30 | 58.51 | 32.72 | 31.57 | 32.13 |
| Qwen3.5-9B (Qwen Team, 2026a) | 61.02 | 70.33 | 65.35 | 44.42 | 45.04 | 44.73 |
| RynnBrain1.1 (Li et al., 2026a) | 2.10 | 34.50 | 3.95 | 7.68 | 46.55 | 13.18 |
| SenseNova-Vision (Han et al., 2026) | 74.86 | 82.98 | 78.71 | 59.36 | 64.48 | 61.81 |
| MiMo-VL-7B-SFT (Yue et al., 2025) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| MiMo-VL-7B-RL (Yue et al., 2025) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| BAGEL (Deng et al., 2025) | 8.48* | 16.84* | 11.28* | 5.09* | 7.33* | 6.01* |
| RynnBrain (Dang et al., 2026) | 0.51 | 1.00 | 0.68 | 5.64 | 34.46 | 9.69 |
| Molmo-7B* (Deitke et al., 2025) | – | – | 33.10 | – | – | 29.20 |
| GroundAnything-VLM | 81.93 | **86.51** | 84.16 | 61.77 | **75.70** | **68.03** |
| GroundAnything | 76.50 | 77.79 | 77.14 | 56.27 | 59.72 | 57.94 |
| **Vision-Language Models (10B–1T)** | | | | | | |
| Qwen3.6-27B (Qwen Team, 2026b) | 70.52 | 75.69 | 73.01 | 46.68 | 43.94 | 45.27 |
| Qwen3-VL-32B (Bai et al., 2025a) | 44.04 | 46.11 | 45.05 | 27.88 | 25.71 | 26.75 |
| Qwen3.8-27B (Qwen Team, 2026d) | 73.97 | 75.15 | 74.55 | 51.07 | 51.25 | 51.16 |
| Qwen3.5-35B-A3B (Qwen Team, 2026a) | 68.54 | 72.89 | 70.65 | 45.64 | 43.92 | 44.76 |
| DeepSeek-VL2-Small-16B (Wu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| DeepSeek-VL2-27B (Wu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| SEED1.5-VL* (Guo et al., 2025) | – | – | 72.10 | – | – | 56.70 |
| **Vision-Language Models (>1T)** | | | | | | |
| Qwen3.7-Max (Qwen Team, 2026c) | 57.56 | 78.12 | 66.28 | 55.04 | 63.65 | 59.04 |
| Kimi-K2.6 | 36.14† | 34.27† | 35.18† | 11.22† | 17.37† | 13.63† |
| Kimi-K3 (Team et al., 2026) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| GPT-6 Astra | **89.52** | 83.74 | **86.57** | **64.39** | 66.99 | 65.62 |

### H.5 Robot and Spatial Pointing

Relational localization and placement. GroundAnything retains the AR variant's 62.34% RefSpatial Unseen accuracy and reaches 69.67% on RoboSpatial Context, above GPT-6 Astra's 65.69%. Its RefSpatial Location/Placement scores are 59.00/67.00, below the AR variant's 67.00/71.00 and GPT-6 Astra's 86.00/85.86. Parallel grounding therefore transfers to relational and free-space targets, while more difficult spatial reasoning remains a clear source of headroom. The external RoboRefer comparison uses the setting without a depth prior.

Table S21: **Robot and spatial pointing.** Point-in-mask accuracy is reported for each dataset. The starred RefSpatial baselines follow Table 11 of Jiang et al. (2026); RoboRefer uses the setting without a depth prior. Values retain the precision recorded in our evaluation tables.

| Model | Point-in-mask accuracy | | | |
|---|---|---|---|---|
| | RefSpatial Location | RefSpatial Placement | RefSpatial Unseen | RoboSpatial Context |
| **Open-set Specialized Detectors** | | | | |
| GroundingDINO (Liu et al., 2024) | 26.50 | 2.00 | 4.33 | 4.92 |
| **Vision-Language Models (<10B)** | | | | |
| Rex-Omni (Jiang et al., 2026) | 51.00 | 52.50 | 37.01 | 59.02 |
| LocateAnything Fast (Wang et al., 2026) | 53.00 | 19.00 | 16.88 | 15.57 |
| LocateAnything Hybrid (Wang et al., 2026) | 55.00 | 17.33 | 20.78 | 14.75 |
| LocateAnything Slow NTP (Wang et al., 2026) | 54.00 | 27.20 | 16.88 | 16.39 |
| Qwen2.5-VL-7B (Bai et al., 2025b) | 43.00 | 15.50 | 18.18 | 26.23 |
| Qwen3-VL-2B (Bai et al., 2025a) | 42.00 | 24.00 | 11.69 | 32.79 |
| Qwen3-VL-4B (Bai et al., 2025a) | 48.00 | 50.00 | 27.27 | 64.75 |
| Qwen3-VL-8B (Bai et al., 2025a) | 51.00 | 45.00 | 28.57 | 59.02 |
| Qwen3.5-4B (Qwen Team, 2026a) | 65.00 | 38.00 | 38.96 | 50.82 |
| Qwen3.5-9B (Qwen Team, 2026a) | 65.00 | 46.83 | 37.01 | 60.66 |
| RynnBrain1.1 (Li et al., 2026a) | 49.70 | 51.50 | 36.90 | 54.10 |
| SenseNova-Vision (Han et al., 2026) | 34.51 | 6.25 | 8.54 | 0.82 |
| MiMo-VL-7B-SFT (Yue et al., 2025) | 1.00† | 8.03† | 4.64† | 4.10† |
| MiMo-VL-7B-RL (Yue et al., 2025) | 1.00† | 2.20† | 2.61† | 4.92† |
| BAGEL (Deng et al., 2025) | 51.79 | 12.01 | 23.38 | 13.93 |
| RynnBrain (Dang et al., 2026) | 42.00 | 37.00 | 22.08 | 28.69 |
| Molmo-7B* (Deitke et al., 2025) | 21.90 | 12.80 | 12.20 | – |
| RoboRefer* (Zhou et al., 2026) | 51.00 | 49.00 | 39.00 | – |
| GroundAnything-VLM | 67.00 | 71.00 | 62.34 | **72.13** |
| GroundAnything | 59.00 | 67.00 | 62.34 | 69.67 |
| **Vision-Language Models (10B–1T)** | | | | |
| Qwen3.6-27B (Qwen Team, 2026b) | 72.00 | 66.00 | 61.04 | 63.93 |
| Qwen3-VL-32B (Bai et al., 2025a) | 62.00 | 52.00 | 41.56 | 63.93 |
| Qwen3.8-27B (Qwen Team, 2026d) | 64.00 | 56.00 | 46.75 | 63.93 |
| Qwen3.5-35B-A3B (Qwen Team, 2026a) | 70.00 | 58.00 | 53.25 | 71.31 |
| DeepSeek-VL2-Small-16B (Wu et al., 2024) | 2.11† | 1.33† | 0.32† | 5.74† |
| DeepSeek-VL2-27B (Wu et al., 2024) | 3.75† | 0.66† | 0.72† | 5.74† |
| SpaceLLaVA* | 5.80 | 4.30 | 4.00 | – |
| RoboPoint* (Yuan et al., 2024) | 22.90 | 9.30 | 8.40 | – |
| Molmo-72B* (Deitke et al., 2025) | 45.80 | 14.70 | 21.20 | – |
| Gemini-2.5-Pro* (Comanici et al., 2025) | 47.00 | 24.20 | 27.10 | – |
| **Vision-Language Models (>1T)** | | | | |
| Qwen3.7-Max (Qwen Team, 2026c) | 71.50 | 66.00 | 57.14 | 69.67 |
| Kimi-K2.6 | 54.09 | 28.57 | 41.56 | 30.33 |
| Kimi-K3 (Team et al., 2026) | 67.85 | 50.00 | 54.98 | 54.92 |
| GPT-6 Astra | **86.00** | **85.86** | **81.93** | 65.69 |

### H.6 OCR

HierText and ICDAR2015. GroundAnything obtains 33.31/41.81 F1mIoU, compared with 41.33/42.50 for GroundAnything-VLM. Both variants have zero reported parse errors on these two datasets, so their quality gap cannot be explained by malformed responses alone. The DLM variant exceeds LocateAnything Fast and Hybrid on both datasets but trails Rex-Omni; the external SenseNova-Vision entries provide F1mIoU only and do not support conclusions about its parsing reliability.

Table S22: **OCR on HierText and ICDAR2015.** Each dataset reports four loose-match F1 measures and parse-error rate. GroundingDINO is N/A because OCR is unsupported. Kimi-K3 and both DeepSeek variants are N/A because their outputs do not satisfy the evaluation protocol; this does not establish a lack of OCR capability. SenseNova-Vision uses the HierText and ICDAR2015 F1mIoU scores of 31.20 and 49.50 reported in Table 1 of Han et al. (2026), because its local evaluation prompts could not be aligned. Other metrics for these datasets are unavailable; TotalText and SROIE use local results. The starred PaddleOCRv5 and SEED1.5-VL scores use the BBOX results in Table 10 of Jiang et al. (2026).

| Model | HierText | | | | | ICDAR2015 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | F1@.50 | F1@.75 | F1@.95 | F1mIoU | Parse err. | F1@.50 | F1@.75 | F1@.95 | F1mIoU | Parse err. |
| **Closed-set Specialized Detectors** | | | | | | | | | | |
| PaddleOCRv5* (Cui et al., 2025) | 45.20 | – | 3.40 | 30.50 | – | 38.20 | – | 1.20 | 25.60 | – |
| **Open-set Specialized Detectors** | | | | | | | | | | |
| GroundingDINO (Liu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| **Vision-Language Models (<10B)** | | | | | | | | | | |
| Rex-Omni (Jiang et al., 2026) | 54.16 | 35.67 | 2.10 | 34.46 | 1.92 | 73.39 | 50.07 | 0.96 | 45.65 | **0.00** |
| LocateAnything Fast (Wang et al., 2026) | 29.91 | 25.83 | 3.13 | 22.59 | 31.46 | 44.51 | 28.32 | 0.40 | 26.73 | 3.43 |
| LocateAnything Hybrid (Wang et al., 2026) | 35.50 | 30.43 | 3.61 | 26.65 | 19.04 | 45.61 | 29.41 | 0.40 | 27.48 | 1.21 |
| LocateAnything Slow NTP (Wang et al., 2026) | 58.34 | **48.65** | **5.28** | 42.94 | 1.10 | 53.07 | 33.29 | 0.65 | 31.81 | **0.00** |
| Qwen2.5-VL-7B (Bai et al., 2025b) | 29.61 | 14.28 | 0.51 | 15.49 | **0.00** | 55.50 | 23.37 | 1.19 | 27.72 | **0.00** |
| Qwen3-VL-2B (Bai et al., 2025a) | 25.08 | 11.38 | 0.33 | 12.65 | 1.74 | 49.47 | 24.52 | 1.93 | 27.43 | 0.40 |
| Qwen3-VL-4B (Bai et al., 2025a) | 41.06 | 23.42 | 0.97 | 23.48 | 0.93 | 51.52 | 26.37 | 1.67 | 28.41 | 0.20 |
| Qwen3-VL-8B (Bai et al., 2025a) | 42.32 | 23.47 | 0.99 | 23.89 | 0.29 | 54.20 | 29.71 | 1.40 | 30.82 | **0.00** |
| Qwen3.5-4B (Qwen Team, 2026a) | 30.47 | 16.50 | 0.76 | 17.00 | 12.65 | 41.16 | 16.20 | 0.37 | 19.61 | **0.00** |
| Qwen3.5-9B (Qwen Team, 2026a) | 48.60 | 30.10 | 1.90 | 29.63 | 13.93 | 56.02 | 26.65 | 0.87 | 29.90 | **0.00** |
| RynnBrain1.1 (Li et al., 2026a) | 3.55 | 2.04 | 0.11 | 2.07 | 30.24 | 36.78 | 15.63 | 0.20 | 17.81 | 6.05 |
| SenseNova-Vision* (Han et al., 2026) | – | – | – | 31.20 | – | – | – | – | **49.50** | – |
| MiMo-VL-7B-SFT (Yue et al., 2025) | 24.77 | 12.82 | 0.36 | 13.44 | 1.33 | 64.31 | 30.58 | 0.46 | 34.33 | 1.01 |
| MiMo-VL-7B-RL (Yue et al., 2025) | 26.13 | 13.01 | 0.32 | 13.88 | 0.29 | 56.84 | 24.32 | 0.37 | 28.79 | 7.86 |
| BAGEL (Deng et al., 2025) | 11.91 | 3.30 | 0.05 | 4.87 | 0.75 | 36.29 | 10.41 | 0.17 | 15.48 | 0.20 |
| RynnBrain (Dang et al., 2026) | 4.08 | 1.92 | 0.06 | 2.13 | 0.64 | 36.05 | 15.15 | 0.47 | 17.81 | 0.60 |
| GroundAnything-VLM | 60.18 | 45.45 | 3.82 | 41.33 | **0.00** | 64.52 | 46.35 | 2.02 | 42.50 | **0.00** |
| GroundAnything | 46.97 | 39.47 | 4.97 | 33.31 | **0.00** | 58.53 | 50.80 | **2.09** | 41.81 | **0.00** |
| **Vision-Language Models (10B–1T)** | | | | | | | | | | |
| Qwen3.6-27B (Qwen Team, 2026b) | 36.27 | 25.16 | 1.94 | 23.49 | 4.12 | 53.83 | 32.09 | 1.75 | 31.64 | 0.20 |
| Qwen3-VL-32B (Bai et al., 2025a) | 15.55 | 9.08 | 0.49 | 9.00 | 54.32 | 42.26 | 26.11 | 1.69 | 25.42 | 25.40 |
| Qwen3.8-27B (Qwen Team, 2026d) | 51.50 | 35.06 | 2.46 | 32.95 | 0.06 | 68.67 | 39.98 | 0.95 | 39.72 | **0.00** |
| Qwen3.5-35B-A3B (Qwen Team, 2026a) | 49.83 | 33.22 | 2.16 | 31.38 | 3.71 | 51.85 | 25.41 | 1.22 | 28.04 | 3.02 |
| DeepSeek-VL2-Small-16B (Wu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| DeepSeek-VL2-27B (Wu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| SEED1.5-VL* (Guo et al., 2025) | 27.10 | – | 0.20 | 12.00 | – | 38.60 | – | 0.00 | 18.70 | – |
| **Vision-Language Models (>1T)** | | | | | | | | | | |
| Qwen3.7-Max (Qwen Team, 2026c) | **62.98** | 47.45 | 3.93 | **42.95** | **0.00** | 69.09 | 45.15 | 1.97 | 43.07 | **0.00** |
| Kimi-K2.6 | 45.89† | 24.25† | 1.25† | 25.26† | 0.12† | 56.81† | 31.14† | 1.95† | 32.25† | **0.00**† |
| Kimi-K3 (Team et al., 2026) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| GPT-6 Astra | 61.35 | 41.89 | 3.38 | 39.58 | 0.23 | **79.02** | **55.72** | 0.87 | 48.87 | **0.00** |

TotalText and SROIE. GroundAnything scores 43.08 on TotalText and 43.64 on SROIE, versus 49.03 and 71.15 for its AR counterpart. The 27.51 pp SROIE gap is substantially larger than the 5.95 pp TotalText gap, showing that diffusion's quality cost varies markedly across OCR settings. Although GroundAnything exceeds LocateAnything Hybrid on SROIE, it remains below Rex-Omni and the

dedicated PaddleOCRv5 reference. Its parse-error rates are unreported on these datasets and must not be treated as zero.

Table S23: **OCR on TotalText and SROIE.** Each dataset reports four loose-match F1 measures and parse-error rate. GroundingDINO is N/A because OCR is unsupported. Kimi-K3 and both DeepSeek variants are N/A because their outputs do not satisfy the evaluation protocol; this does not establish a lack of OCR capability. The starred PaddleOCRv5 and SEED1.5-VL scores use the BBOX results in Table 10 of Jiang et al. (2026).

| Model | TotalText | | | | | SROIE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | F1@.50 | F1@.75 | F1@.95 | F1mIoU | Parse err. | F1@.50 | F1@.75 | F1@.95 | F1mIoU | Parse err. |
| **Closed-set Specialized Detectors** | | | | | | | | | | |
| PaddleOCRv5* (Cui et al., 2025) | 40.20 | – | 0.70 | 25.70 | – | 77.70 | – | 5.60 | 58.60 | – |
| **Open-set Specialized Detectors** | | | | | | | | | | |
| GroundingDINO (Liu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| **Vision-Language Models (<10B)** | | | | | | | | | | |
| Rex-Omni (Jiang et al., 2026) | 74.04 | 59.21 | 4.47 | 52.35 | **0.00** | 72.28 | 47.65 | 2.49 | 48.35 | 3.61 |
| LocateAnything Fast (Wang et al., 2026) | 60.96 | 49.64 | 5.48 | 44.40 | 3.00 | 33.18 | 29.73 | 1.86 | 24.89 | 45.83 |
| LocateAnything Hybrid (Wang et al., 2026) | 62.40 | 51.07 | 5.66 | 45.49 | 1.33 | 40.53 | 35.84 | 2.12 | 30.05 | 34.17 |
| LocateAnything Slow NTP (Wang et al., 2026) | 70.79 | 53.40 | 4.88 | 49.13 | 0.33 | 88.23 | 77.85 | 4.49 | 65.49 | 1.11 |
| Qwen2.5-VL-7B (Bai et al., 2025b) | 56.62 | 28.47 | 2.54 | 31.13 | **0.00** | 30.76 | 14.04 | 0.35 | 15.58 | **0.00** |
| Qwen3-VL-2B (Bai et al., 2025a) | 60.76 | 29.25 | 6.45 | 33.55 | **0.00** | 34.86 | 12.51 | 0.15 | 15.93 | **0.00** |
| Qwen3-VL-4B (Bai et al., 2025a) | 65.02 | 36.49 | 7.75 | 38.35 | **0.00** | 68.73 | 41.96 | 0.98 | 40.41 | **0.00** |
| Qwen3-VL-8B (Bai et al., 2025a) | 61.40 | 35.44 | 7.40 | 36.70 | **0.00** | 49.41 | 24.33 | 0.46 | 25.88 | **0.00** |
| Qwen3.5-4B (Qwen Team, 2026a) | 54.40 | 28.83 | 4.75 | 31.56 | **0.00** | 45.96 | 20.70 | 0.30 | 23.21 | 1.11 |
| Qwen3.5-9B (Qwen Team, 2026a) | 63.78 | 35.22 | 4.51 | 37.26 | 1.00 | 51.52 | 20.59 | 0.81 | 26.74 | 1.94 |
| RynnBrain1.1 (Li et al., 2026a) | 30.38 | 13.04 | 4.29 | 16.90 | 10.67 | 7.23 | 3.24 | 0.06 | 3.59 | 0.83 |
| SenseNova-Vision (Han et al., 2026) | 18.30 | 11.21 | 1.02 | 11.40 | **0.00** | 46.97 | 41.90 | 4.23 | 36.26 | **0.00** |
| MiMo-VL-7B-SFT (Yue et al., 2025) | 64.24 | 42.31 | 1.79 | 39.76 | **0.00** | 30.13 | 13.02 | 0.36 | 15.01 | 5.28 |
| MiMo-VL-7B-RL (Yue et al., 2025) | 65.24 | 40.22 | 2.19 | 39.55 | 0.67 | 39.77 | 15.71 | 0.47 | 19.00 | 0.28 |
| BAGEL (Deng et al., 2025) | 51.19 | 19.02 | 1.02 | 24.25 | **0.00** | 18.63 | 4.94 | 0.07 | 7.74 | 1.11 |
| RynnBrain (Dang et al., 2026) | 35.58 | 17.41 | 1.57 | 18.90 | **0.00** | 7.08 | 2.37 | 0.02 | 3.19 | 4.72 |
| GroundAnything-VLM | 72.66 | 53.35 | 4.95 | 49.03 | **0.00** | **91.85** | **83.60** | **7.06** | **71.15** | **0.00** |
| GroundAnything | 60.46 | 46.01 | 5.41 | 43.08 | – | 49.31 | 44.85 | 3.02 | 43.64 | – |
| **Vision-Language Models (10B–1T)** | | | | | | | | | | |
| Qwen3.6-27B (Qwen Team, 2026b) | 63.52 | 42.49 | 5.38 | 41.21 | 2.33 | 52.05 | 29.37 | 0.63 | 29.37 | **0.00** |
| Qwen3-VL-32B (Bai et al., 2025a) | 44.51 | 28.93 | **8.27** | 28.84 | 18.67 | 10.12 | 6.31 | 0.20 | 6.02 | 73.89 |
| Qwen3.8-27B (Qwen Team, 2026d) | 65.73 | 42.57 | 3.99 | 41.57 | **0.00** | 61.27 | 38.97 | 1.70 | 37.24 | 0.28 |
| Qwen3.5-35B-A3B (Qwen Team, 2026a) | 59.86 | 34.72 | 5.31 | 35.95 | 1.67 | 65.83 | 36.31 | 0.92 | 36.64 | 0.28 |
| DeepSeek-VL2-Small-16B (Wu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| DeepSeek-VL2-27B (Wu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| SEED1.5-VL* (Guo et al., 2025) | 35.00 | – | 0.30 | 19.50 | – | 51.90 | – | 0.80 | 28.10 | – |
| **Vision-Language Models (>1T)** | | | | | | | | | | |
| Qwen3.7-Max (Qwen Team, 2026c) | 72.58 | 52.18 | 7.92 | 48.64 | **0.00** | 57.72 | 44.25 | 1.50 | 38.54 | **0.00** |
| Kimi-K2.6 | 67.66† | 40.90† | 5.52† | 40.99† | **0.00**† | 76.94† | 49.43† | 1.92† | 46.83† | **0.00**† |
| Kimi-K3 (Team et al., 2026) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| GPT-6 Astra | **74.61** | **62.13** | 4.96 | **53.55** | **0.00** | 70.81 | 62.54 | 5.27 | 53.57 | **0.00** |

## H.7 GUI GROUNDING

ScreenSpot-Pro. GroundAnything improves overall action accuracy from 65.34% for GroundAnything-VLM to 75.96%, with gains in every reported text/icon domain subset. For example, CAD icon accuracy rises from 48.44% to 71.88%, and scientific-interface icon accuracy from 61.82% to 81.82%. These gains support precise parallel localization in complex interfaces, although GPT-6 Astra remains stronger overall at 93.17%. Uncertain prompt/protocol results are not used to infer intrinsic GUI capability.

Table S24: **ScreenSpot-Pro.** Action accuracy is broken down by domain and target type, followed by overall action accuracy and parse-error rate. GroundingDINO is N/A because GUI grounding is unsupported. BAGEL is N/A because its outputs do not satisfy the unified protocol. The starred JEDI, UI-R1, and UI-TARS scores are from Table 8 of Jiang et al. (2026); GUI-Owl-32B scores are from Table 3 of Wang et al. (2026).

| Model | Dev | | Creative | | CAD | | Sci | | Office | | OS | | Overall | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Text | Icon | Text | Icon | Text | Icon | Text | Icon | Text | Icon | Text | Icon | Action acc. | Parse err. |
| **Open-set Specialized Detectors** | | | | | | | | | | | | | | |
| GroundingDINO (Liu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| **Vision-Language Models (<10B)** | | | | | | | | | | | | | | |
| Rex-Omni (Jiang et al., 2026) | 61.04 | 9.66 | 53.03 | 12.59 | 23.35 | 9.38 | 57.64 | 26.36 | 65.54 | 24.53 | 42.06 | 13.48 | 36.75 | – |
| LocateAnything Fast (Wang et al., 2026) | 68.18 | 44.83 | 59.60 | 32.87 | 58.38 | 35.94 | 71.53 | 53.64 | 75.14 | 56.60 | 51.40 | 37.08 | 56.04 | – |
| LocateAnything Hybrid (Wang et al., 2026) | 70.13 | 44.83 | 59.60 | 36.36 | 58.88 | 37.50 | 71.53 | 51.82 | 72.88 | 58.49 | 58.88 | 40.45 | 57.05 | – |
| LocateAnything Slow NTP (Wang et al., 2026) | 70.78 | 48.28 | 61.11 | 39.86 | 60.41 | 39.06 | 75.69 | 51.82 | 74.58 | 54.72 | 57.94 | 40.45 | 58.57 | – |
| Qwen2.5-VL-7B (Bai et al., 2025b) | 40.91 | 3.45 | 36.36 | 9.09 | 18.27 | 3.12 | 47.22 | 6.36 | 56.50 | 13.21 | 34.58 | 11.24 | 26.57 | – |
| Qwen3-VL-2B (Bai et al., 2025a) | 47.40 | 7.59 | 29.29 | 8.39 | 23.86 | 7.81 | 38.89 | 18.18 | 49.15 | 22.64 | 37.38 | 20.22 | 27.77 | – |
| Qwen3-VL-4B (Bai et al., 2025a) | 72.73 | 31.72 | 66.67 | 25.17 | 57.87 | 26.56 | 77.08 | 35.45 | 84.75 | 47.17 | 78.50 | 34.83 | 56.74 | – |
| Qwen3-VL-8B (Bai et al., 2025a) | 75.32 | 30.34 | 71.21 | 20.98 | 60.91 | 26.56 | 76.39 | 40.00 | 83.62 | 37.74 | 73.83 | 33.71 | 56.86 | – |
| Qwen3.5-4B (Qwen Team, 2026a) | 79.22 | 42.07 | 67.68 | 25.17 | 65.99 | 35.94 | 74.31 | 34.55 | 82.49 | 50.94 | 69.16 | 43.82 | 59.27 | – |
| Qwen3.5-9B (Qwen Team, 2026a) | 70.78 | 37.93 | 64.14 | 30.77 | 40.10 | 28.12 | 72.92 | 33.64 | 74.58 | 49.06 | 69.16 | 38.20 | 53.13 | – |
| RynnBrain1.1 (Li et al., 2026a) | 52.60 | 15.86 | 45.96 | 12.59 | 21.32 | 10.94 | 52.08 | 21.82 | 60.45 | 20.75 | 47.66 | 20.22 | 34.66 | 0.38 |
| SenseNova-Vision (Han et al., 2026) | N/A† | N/A† | N/A† | N/A† | N/A† | N/A† | N/A† | N/A† | N/A† | N/A† | N/A† | N/A† | N/A† | N/A† |
| MiMo-VL-7B-SFT (Yue et al., 2025) | 28.57 | 4.83 | 31.82 | 4.20 | 19.29 | 7.81 | 52.78 | 12.73 | 40.11 | 20.75 | 24.30 | 6.74 | 23.21 | 6.83 |
| MiMo-VL-7B-RL (Yue et al., 2025) | 27.92 | 2.07 | 38.38 | 4.20 | 27.41 | 9.38 | 57.64 | 17.27 | 53.11 | 24.53 | 29.91 | 7.87 | 27.58 | 9.30 |
| BAGEL (Deng et al., 2025) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| RynnBrain (Dang et al., 2026) | 44.81 | 7.59 | 35.35 | 5.59 | 15.23 | 7.81 | 40.28 | 13.64 | 54.80 | 18.87 | 42.06 | 11.24 | 27.07 | 2.47 |
| JEDI* (Xie et al., 2026) | 61.00 | 13.80 | 53.50 | 8.40 | 27.40 | 9.40 | 54.20 | 18.20 | 64.40 | 32.10 | 38.30 | 9.00 | 36.10 | – |
| UI-R1* (Lu et al., 2026) | 22.70 | 4.10 | 27.30 | 3.50 | 11.20 | 6.30 | 42.40 | 11.80 | 32.20 | 11.30 | 13.10 | 4.50 | 17.80 | – |
| UI-TARS* (Qin et al., 2025) | 47.40 | 4.10 | 42.90 | 6.30 | 17.80 | 4.70 | 56.90 | 17.30 | 50.30 | 17.00 | 21.50 | 5.60 | 27.70 | – |
| GroundAnything-VLM | 75.97 | 55.86 | 72.73 | 58.04 | 55.84 | 48.44 | 86.11 | 61.82 | 74.58 | 62.26 | 62.62 | 48.31 | 65.34 | 0.06 |
| GroundAnything | 77.92 | 70.34 | 85.86 | 66.43 | 71.57 | 71.88 | 90.97 | 81.82 | 83.62 | 66.04 | 64.49 | 60.67 | 75.96 | – |
| **Vision-Language Models (10B–1T)** | | | | | | | | | | | | | | |
| Qwen3.6-27B (Qwen Team, 2026b) | 87.01 | 53.79 | 76.77 | 44.76 | 75.13 | 46.88 | 81.25 | 48.18 | 87.57 | 64.15 | 79.44 | 57.30 | 69.64 | – |
| Qwen3-VL-32B (Bai et al., 2025a) | 73.38 | 20.00 | 76.26 | 24.48 | 59.39 | 34.38 | 86.11 | 30.91 | 83.62 | 41.51 | 60.75 | 22.47 | 55.66 | 0.13 |
| Qwen3.8-27B (Qwen Team, 2026d) | 85.71 | 64.83 | 45.45 | 41.96 | 50.76 | 14.06 | 67.36 | 36.36 | 87.57 | 56.60 | 74.77 | 47.19 | 58.76 | 0.76 |
| Qwen3.5-35B-A3B (Qwen Team, 2026a) | 82.47 | 44.83 | 56.57 | 27.97 | 54.82 | 14.06 | 53.47 | 29.09 | 61.58 | 41.51 | 53.27 | 20.22 | 49.08 | 0.38 |
| DeepSeek-VL2-Small-16B (Wu et al., 2024) | 0.00† | 0.00† | 0.00† | 0.00† | 0.00† | 0.00† | 0.00† | 0.00† | 0.00† | 1.89† | 0.93† | 0.00† | 0.13† | 25.62† |
| DeepSeek-VL2-27B (Wu et al., 2024) | 0.00† | 0.00† | 0.00† | 0.00† | 0.51† | 0.00† | 0.00† | 0.00† | 0.00† | 0.00† | 0.00† | 0.00† | 0.06† | 34.28† |
| GUI-Owl* (Ye et al., 2025) | 84.40 | 39.30 | 65.20 | 18.20 | 62.40 | 28.10 | 82.60 | 39.10 | 81.40 | 39.60 | 70.10 | 36.00 | 58.00 | – |
| **Vision-Language Models (>1T)** | | | | | | | | | | | | | | |
| Qwen3.7-Max (Qwen Team, 2026c) | 79.22 | 37.50 | 62.63 | 34.27 | 46.19 | 31.25 | 77.08 | 31.82 | 78.53 | 47.17 | 61.68 | 38.20 | 55.06 | 10.57 |
| Kimi-K2.6 | – | – | – | – | – | – | – | – | – | – | – | – | 6.07† | 8.22† |
| Kimi-K3 (Team et al., 2026) | 29.87† | 20.00† | 43.94† | 25.17† | 25.38† | 3.12† | 29.86† | 12.73† | 27.12† | 13.21† | 28.97† | 8.99† | 25.36† | 3.73† |
| GPT-6 Astra | **96.10** | **84.83** | **95.45** | **89.51** | **95.43** | **85.94** | **94.44** | **87.27** | **98.87** | **96.23** | **97.20** | **89.89** | **93.17** | **0.00** |

ScreenSpot-V2 and OSWorld-G. GroundAnything reaches 95.60% ScreenSpot-V2 accuracy and 81.21% OSWorld-G exact accuracy, improving over GroundAnything-VLM by 0.71 and 6.56 pp. The smaller gain on ScreenSpot-V2 reflects a setting where the AR baseline already scores 94.89%; OSWorld-G leaves more room for improvement. GroundAnything does not improve every individual text/icon subset, and GPT-6 Astra still leads both overall metrics. Unreported DLM parse-error rates remain unavailable.

Table S25: **ScreenSpot-V2 and OSWorld-G.** ScreenSpot-V2 includes text/icon results for mobile, desktop, and web environments, overall action accuracy, and parse-error rate; OSWorld-G reports exact accuracy and parse-error rate. GroundingDINO is N/A because GUI grounding is unsupported. BAGEL is N/A because its outputs do not satisfy the unified protocol. The starred ScreenSpot-V2 scores are from Table 8 of Jiang et al. (2026).

| Model | ScreenSpot-V2 | | | | | | | | OSWorld-G | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mobile text | Mobile icon | Desktop text | Desktop icon | Web text | Web icon | Action acc. | Parse err. | Exact acc. | Parse err. |
| **Open-set Specialized Detectors** | | | | | | | | | | |
| GroundingDINO (Liu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| **Vision-Language Models (<10B)** | | | | | | | | | | |
| Rex-Omni (Jiang et al., 2026) | 96.90 | 82.46 | 97.94 | 80.71 | 89.74 | 76.35 | 88.29 | – | 46.10 | – |
| LocateAnything Fast (Wang et al., 2026) | 94.48 | 81.04 | 93.81 | 86.43 | 88.89 | 83.25 | 88.44 | – | 59.93 | – |
| LocateAnything Hybrid (Wang et al., 2026) | 96.21 | 83.89 | 92.78 | 89.29 | 88.89 | 86.21 | 89.94 | – | 60.46 | – |
| LocateAnything Slow NTP (Wang et al., 2026) | 95.86 | 83.41 | 93.30 | 89.29 | 90.60 | 85.22 | 90.02 | – | 61.17 | – |
| Qwen2.5-VL-7B (Bai et al., 2025b) | 98.28 | 82.94 | 91.75 | 67.86 | 92.74 | 79.31 | 87.34 | – | 34.22 | – |
| Qwen3-VL-2B (Bai et al., 2025a) | 93.10 | 70.62 | 79.90 | 61.43 | 79.91 | 62.07 | 76.49 | – | 33.51 | – |
| Qwen3-VL-4B (Bai et al., 2025a) | 97.93 | 87.20 | 96.91 | 87.14 | 94.44 | 86.21 | 92.30 | – | 56.91 | – |
| Qwen3-VL-8B (Bai et al., 2025a) | 98.62 | 89.57 | 98.45 | 87.86 | 94.87 | 89.16 | 93.71 | – | 56.91 | – |
| Qwen3.5-4B (Qwen Team, 2026a) | **98.97** | 88.15 | 97.94 | 89.29 | 95.73 | 90.15 | 93.95 | – | 54.61 | – |
| Qwen3.5-9B (Qwen Team, 2026a) | 96.55 | 88.15 | 97.94 | 93.57 | 87.61 | 78.82 | 90.57 | – | 60.99 | – |
| RynnBrain1.1 (Li et al., 2026a) | 82.07 | 71.09 | 75.26 | 52.86 | 73.08 | 57.64 | 70.44 | 4.25 | 33.33 | 4.96 |
| SenseNova-Vision (Han et al., 2026) | N/A† | N/A† | N/A† | N/A† | N/A† | N/A† | N/A† | N/A† | N/A† | N/A† |
| MiMo-VL-7B-SFT (Yue et al., 2025) | 91.03 | 68.25 | 86.60 | 50.00 | 63.68 | 56.65 | 71.54 | 8.18 | 34.04 | 6.38 |
| MiMo-VL-7B-RL (Yue et al., 2025) | 90.69 | 67.30 | 85.57 | 49.29 | 78.21 | 62.56 | 74.69 | 5.66 | 37.06 | 2.30 |
| BAGEL (Deng et al., 2025) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| RynnBrain (Dang et al., 2026) | 89.31 | 69.19 | 80.93 | 56.43 | 79.91 | 60.10 | 74.69 | 5.27 | 35.99 | 1.77 |
| JEDI* (Xie et al., 2026) | 96.60 | 81.50 | 96.90 | 78.60 | 88.50 | 83.70 | 88.60 | – | – | – |
| UI-R1* (Lu et al., 2026) | 84.30 | **96.20** | 75.40 | 89.20 | 63.60 | 92.30 | 85.40 | – | – | – |
| UI-TARS* (Qin et al., 2025) | 95.20 | 79.10 | 90.70 | 68.60 | 87.20 | 78.30 | 84.70 | – | – | – |
| GroundAnything-VLM | 97.24 | 89.10 | 98.45 | 94.29 | 95.73 | 93.60 | 94.89 | **0.00** | 74.65 | **0.00** |
| GroundAnything | 98.62 | 91.00 | 97.94 | 95.00 | 94.87 | 95.07 | 95.60 | – | 81.21 | – |
| **Vision-Language Models (10B–1T)** | | | | | | | | | | |
| Qwen3.6-27B (Qwen Team, 2026b) | **98.97** | 92.42 | 98.45 | 92.86 | 96.58 | 89.16 | 95.13 | – | 68.44 | – |
| Qwen3-VL-32B (Bai et al., 2025a) | 98.62 | 90.05 | 98.45 | 86.43 | 96.58 | 91.63 | 94.34 | 0.08 | 60.28 | 0.71 |
| Qwen3.8-27B (Qwen Team, 2026d) | 97.93 | 90.52 | 98.45 | 94.29 | 94.87 | 89.66 | 94.50 | 0.24 | 63.48 | **0.00** |
| Qwen3.5-35B-A3B (Qwen Team, 2026a) | 97.59 | 89.10 | 96.39 | 92.14 | 94.02 | 86.70 | 93.00 | 1.42 | 62.23 | 0.18 |
| DeepSeek-VL2-Small-16B (Wu et al., 2024) | 2.07† | 0.00† | 4.64† | 1.43† | 0.85† | 0.49† | 1.57† | 41.82† | 0.89† | 11.52† |
| DeepSeek-VL2-27B (Wu et al., 2024) | 6.90† | 1.42† | 2.58† | 0.00† | 2.56† | 1.48† | 2.91† | 25.63† | 0.00† | 32.62† |
| **Vision-Language Models (>1T)** | | | | | | | | | | |
| Qwen3.7-Max (Qwen Team, 2026c) | 89.31 | 82.94 | 71.13 | 76.43 | 81.62 | 79.80 | 81.13 | 13.21 | 49.29 | 24.47 |
| Kimi-K2.6 | – | – | – | – | – | – | 52.36† | 7.15† | 10.11† | 7.98† |
| Kimi-K3 (Team et al., 2026) | 95.86† | 88.63† | 97.42† | 89.29† | 63.68† | 61.08† | 82.70† | 1.57† | 68.26† | 4.96† |
| GPT-6 Astra | **98.97** | 95.26 | **98.97** | **98.57** | **98.29** | **97.04** | **97.88** | 0.24 | **86.70** | 0.18 |

## H.8 Layout Grounding

Layout grounding predicts labeled document regions and uses the box evaluation convention of Jiang et al. (2026).

DocLayNet. GroundAnything-VLM reaches 85.78 F1mIoU, slightly above SenseNova-Vision (85.53) and above the external DocLayout-YOLO reference (81.10). GroundAnything scores 68.38, close to Rex-Omni (68.06) but below LocateAnything Hybrid (77.34). Its 37.78 F1 at IoU 0.95 nevertheless exceeds GPT-6 Astra's 34.93, illustrating that strict boundary accuracy and overall region detection can rank models differently.

Table S26: **DocLayNet.** Complete box-grounding metrics. GroundingDINO lacks a compatible document-region interface; it, Kimi-K3, and both DeepSeek variants are N/A. Starred MiMo and BAGEL scores retain observations under suspected output-protocol incompatibility and are not formal capability measurements. The starred DocLayout-YOLO and SEED1.5-VL scores are from Table 9 of Jiang et al. (2026).

| Model | IoU 0.50 | | | IoU 0.95 | | | mIoU | | |
|---|---|---|---|---|---|---|---|---|---|
| | R | P | F1 | R | P | F1 | R | P | F1 |
| **Closed-set Specialized Detectors** | | | | | | | | | |
| DocLayout-YOLO* (Zhao et al., 2024) | – | – | 91.20 | – | – | **52.10** | – | – | 81.10 |
| **Open-set Specialized Detectors** | | | | | | | | | |
| GroundingDINO (Liu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| **Vision-Language Models (<10B)** | | | | | | | | | |
| Rex-Omni (Jiang et al., 2026) | 83.48 | 88.85 | 86.08 | 27.13 | 27.95 | 27.53 | 66.25 | 69.98 | 68.06 |
| LocateAnything Fast (Wang et al., 2026) | 55.10 | 59.26 | 57.11 | 25.03 | 26.55 | 25.77 | 47.73 | 51.12 | 49.37 |
| LocateAnything Hybrid (Wang et al., 2026) | 87.95 | 90.90 | 89.40 | 38.72 | 39.82 | 39.26 | 76.14 | 78.57 | 77.34 |
| LocateAnything Slow NTP (Wang et al., 2026) | 91.55 | 94.58 | 93.04 | 39.56 | 40.64 | 40.09 | 78.98 | 81.45 | 80.19 |
| Qwen2.5-VL-7B (Bai et al., 2025b) | 30.23 | 29.07 | 29.64 | 2.52 | 2.92 | 2.70 | 16.41 | 16.72 | 16.55 |
| Qwen3-VL-2B (Bai et al., 2025a) | 40.38 | 35.70 | 37.89 | 2.80 | 2.83 | 2.81 | 21.10 | 19.19 | 20.09 |
| Qwen3-VL-4B (Bai et al., 2025a) | 64.86 | 69.50 | 67.10 | 8.40 | 8.68 | 8.54 | 39.90 | 41.76 | 40.81 |
| Qwen3-VL-8B (Bai et al., 2025a) | 60.09 | 57.80 | 58.92 | 7.13 | 7.18 | 7.15 | 37.76 | 36.58 | 37.16 |
| Qwen3.5-4B (Qwen Team, 2026a) | 59.37 | 47.02 | 52.48 | 5.04 | 4.47 | 4.74 | 35.03 | 28.51 | 31.43 |
| Qwen3.5-9B (Qwen Team, 2026a) | 63.44 | 49.12 | 55.37 | 6.96 | 6.09 | 6.50 | 39.12 | 31.10 | 34.65 |
| RynnBrain1.1 (Li et al., 2026a) | 5.50 | 24.01 | 8.96 | 1.52 | 6.68 | 2.48 | 3.70 | 16.21 | 6.02 |
| SenseNova-Vision (Han et al., 2026) | 95.45 | 95.78 | 95.61 | **49.23** | **49.43** | 49.33 | 85.38 | 85.67 | 85.53 |
| MiMo-VL-7B-SFT (Yue et al., 2025) | 3.37* | 5.01* | 4.03* | 0.17* | 0.24* | 0.20* | 1.21* | 1.99* | 1.50* |
| MiMo-VL-7B-RL (Yue et al., 2025) | 5.43* | 8.01* | 6.47* | 0.11* | 0.19* | 0.14* | 1.98* | 3.15* | 2.43* |
| BAGEL (Deng et al., 2025) | 10.83* | 10.93* | 10.88* | 0.60* | 0.69* | 0.64* | 4.34* | 4.64* | 4.48* |
| RynnBrain (Dang et al., 2026) | 7.16 | 17.74 | 10.20 | 0.98 | 1.88 | 1.29 | 4.01 | 9.40 | 5.62 |
| GroundAnything-VLM | 96.29 | **96.22** | 96.26 | 44.80 | 44.75 | 44.77 | **85.83** | **85.73** | **85.78** |
| GroundAnything | 78.77 | 72.76 | 75.65 | 38.07 | 37.50 | 37.78 | 71.25 | 65.74 | 68.38 |
| **Vision-Language Models (10B–1T)** | | | | | | | | | |
| Qwen3.6-27B (Qwen Team, 2026b) | 73.85 | 72.14 | 72.98 | 10.44 | 10.10 | 10.27 | 47.15 | 46.10 | 46.62 |
| Qwen3-VL-32B (Bai et al., 2025a) | 40.07 | 36.22 | 38.05 | 4.08 | 4.22 | 4.15 | 21.49 | 20.37 | 20.91 |
| Qwen3.8-27B (Qwen Team, 2026d) | 69.25 | 70.21 | 69.73 | 8.62 | 8.74 | 8.68 | 41.97 | 42.46 | 42.21 |
| Qwen3.5-35B-A3B (Qwen Team, 2026a) | 63.83 | 56.73 | 60.07 | 7.40 | 6.99 | 7.19 | 40.02 | 36.03 | 37.92 |
| DeepSeek-VL2-Small-16B (Wu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| DeepSeek-VL2-27B (Wu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| SEED1.5-VL* (Guo et al., 2025) | – | – | 54.90 | – | – | 4.30 | – | – | 28.70 |
| **Vision-Language Models (>1T)** | | | | | | | | | |
| Qwen3.7-Max (Qwen Team, 2026c) | 75.85 | 65.60 | 70.35 | 12.76 | 12.08 | 12.41 | 49.90 | 44.31 | 46.93 |
| Kimi-K2.6 | 29.23† | 28.21† | 28.71† | 2.29† | 2.43† | 2.36† | 15.64† | 15.55† | 15.59† |
| Kimi-K3 (Team et al., 2026) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| GPT-6 Astra | **97.45** | 95.72 | **96.58** | 35.04 | 34.83 | 34.93 | 78.24 | 76.86 | 77.54 |

M6Doc. GroundAnything-VLM obtains 74.76 F1mIoU, exceeding LocateAnything Slow NTP (68.35). GroundAnything reaches 56.69, above Rex-Omni (54.95) but below LocateAnything Hybrid (65.94). Together with DocLayNet, these results establish coverage of document-region grounding while identifying a substantial diffusion–AR gap in structured layout extraction.

Table S27: **M6Doc.** Complete box-grounding metrics. GroundingDINO lacks a compatible document-region interface; it, Kimi-K3, and both DeepSeek variants are N/A. Starred MiMo and BAGEL scores retain observations under suspected output-protocol incompatibility and are not formal capability measurements. The starred SEED1.5-VL scores are from Table 9 of Jiang et al. (2026).

| Model | IoU 0.50 | | | IoU 0.95 | | | mIoU | | |
|---|---|---|---|---|---|---|---|---|---|
| | R | P | F1 | R | P | F1 | R | P | F1 |
| **Open-set Specialized Detectors** | | | | | | | | | |
| GroundingDINO (Liu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| **Vision-Language Models (<10B)** | | | | | | | | | |
| Rex-Omni (Jiang et al., 2026) | 73.09 | 78.27 | 75.59 | 18.16 | 18.78 | 18.46 | 53.36 | 56.64 | 54.95 |
| LocateAnything Fast (Wang et al., 2026) | 59.82 | 61.54 | 60.67 | 18.06 | 18.50 | 18.28 | 46.39 | 47.69 | 47.03 |
| LocateAnything Hybrid (Wang et al., 2026) | 83.54 | 84.94 | 84.23 | 26.18 | 26.59 | 26.39 | 65.38 | 66.50 | 65.94 |
| LocateAnything Slow NTP (Wang et al., 2026) | 88.16 | 89.64 | 88.89 | 25.62 | 26.08 | 25.85 | 67.78 | 68.94 | 68.35 |
| Qwen2.5-VL-7B (Bai et al., 2025b) | 19.20 | 22.91 | 20.89 | 2.24 | 2.65 | 2.43 | 11.86 | 14.26 | 12.95 |
| Qwen3-VL-2B (Bai et al., 2025a) | 22.31 | 24.68 | 23.44 | 2.24 | 2.58 | 2.40 | 12.61 | 14.08 | 13.31 |
| Qwen3-VL-4B (Bai et al., 2025a) | 37.94 | 44.13 | 40.80 | 4.88 | 5.62 | 5.22 | 23.03 | 26.69 | 24.73 |
| Qwen3-VL-8B (Bai et al., 2025a) | 35.84 | 36.64 | 36.23 | 4.73 | 5.01 | 4.87 | 21.56 | 22.36 | 21.95 |
| Qwen3.5-4B (Qwen Team, 2026a) | 17.32 | 15.67 | 16.45 | 1.47 | 1.54 | 1.50 | 8.44 | 8.13 | 8.28 |
| Qwen3.5-9B (Qwen Team, 2026a) | 30.92 | 28.08 | 29.43 | 4.08 | 4.17 | 4.12 | 17.80 | 16.94 | 17.35 |
| RynnBrain1.1 (Li et al., 2026a) | 4.11 | 26.77 | 7.13 | 0.89 | 6.20 | 1.56 | 2.60 | 17.51 | 4.53 |
| SenseNova-Vision (Han et al., 2026) | 50.01 | 63.15 | 55.82 | 8.87 | 10.60 | 9.66 | 32.12 | 39.99 | 35.62 |
| MiMo-VL-7B-SFT (Yue et al., 2025) | 2.88* | 5.77* | 3.84* | 0.03* | 0.10* | 0.05* | 0.91* | 2.14* | 1.28* |
| MiMo-VL-7B-RL (Yue et al., 2025) | 4.63* | 9.36* | 6.19* | 0.07* | 0.12* | 0.09* | 1.53* | 3.24* | 2.07* |
| BAGEL (Deng et al., 2025) | 10.62* | 15.75* | 12.68* | 0.40* | 0.62* | 0.48* | 4.94* | 7.61* | 5.99* |
| RynnBrain (Dang et al., 2026) | 3.98 | 21.54 | 6.72 | 0.32 | 2.06 | 0.55 | 2.00 | 10.88 | 3.38 |
| GroundAnything-VLM | **91.91** | **93.15** | **92.53** | **31.57** | **31.90** | **31.74** | **74.29** | **75.24** | **74.76** |
| GroundAnything | 69.65 | 73.29 | 71.42 | 25.96 | 24.09 | 24.99 | 55.91 | 57.49 | 56.69 |
| **Vision-Language Models (10B–1T)** | | | | | | | | | |
| Qwen3.6-27B (Qwen Team, 2026b) | 45.19 | 47.32 | 46.23 | 6.69 | 6.91 | 6.80 | 27.92 | 29.17 | 28.53 |
| Qwen3-VL-32B (Bai et al., 2025a) | 46.89 | 48.85 | 47.85 | 7.02 | 7.37 | 7.19 | 29.66 | 31.02 | 30.32 |
| Qwen3.8-27B (Qwen Team, 2026d) | 34.42 | 35.88 | 35.13 | 4.19 | 4.47 | 4.33 | 19.43 | 20.43 | 19.92 |
| Qwen3.5-35B-A3B (Qwen Team, 2026a) | 44.90 | 46.66 | 45.76 | 6.63 | 6.99 | 6.81 | 28.07 | 29.32 | 28.68 |
| DeepSeek-VL2-Small-16B (Wu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| DeepSeek-VL2-27B (Wu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| SEED1.5-VL* (Guo et al., 2025) | – | – | 48.00 | – | – | 3.40 | – | – | 28.00 |
| **Vision-Language Models (>1T)** | | | | | | | | | |
| Qwen3.7-Max (Qwen Team, 2026c) | 48.23 | 50.84 | 49.50 | 9.36 | 9.87 | 9.61 | 31.50 | 33.16 | 32.31 |
| Kimi-K2.6 | 21.96† | 23.87† | 22.87† | 2.80† | 3.16† | 2.97† | 12.89† | 14.16† | 13.50† |
| Kimi-K3 (Team et al., 2026) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| GPT-6 Astra | 78.30 | 83.51 | 80.84 | 19.68 | 21.54 | 20.58 | 58.50 | 62.78 | 60.59 |

### H.9 Visual Prompting

FSC147. GroundAnything-VLM and GroundAnything achieve 60.29 and 52.43 F1mIoU, respectively, below SenseNova-Vision's 62.51. The DLM variant has higher F1 at IoU 0.95 than its AR counterpart (9.37 versus 4.48), despite the lower aggregate. This threshold dependence again cautions against equating a single strict-overlap score with overall grounding quality. LocateAnything's N/A entries indicate an unavailable visual-prompt interface, not zero accuracy.

Table S28: **FSC147 visual prompting.** Complete box-grounding metrics. LocateAnything variants lack a supported visual-prompt interface; both DeepSeek variants have incompatible output protocols. Their entries are N/A. Starred GroundingDINO scores come from an unsupported visual-prompt task interface; starred MiMo and BAGEL scores are retained observations under suspected output-protocol incompatibility.

| Model | IoU 0.50 | | | IoU 0.95 | | | mIoU | | |
|---|---|---|---|---|---|---|---|---|---|
| | R | P | F1 | R | P | F1 | R | P | F1 |
| **Open-set Specialized Detectors** | | | | | | | | | |
| GroundingDINO (Liu et al., 2024) | 7.94* | 20.16* | 11.39* | 1.56* | 4.94* | 2.37* | 6.34* | 16.01* | 9.08* |
| **Vision-Language Models (<10B)** | | | | | | | | | |
| Rex-Omni (Jiang et al., 2026) | 78.97 | 77.92 | 78.44 | 9.13 | 9.01 | 9.07 | 57.56 | 56.75 | 57.15 |
| LocateAnything Fast (Wang et al., 2026) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| LocateAnything Hybrid (Wang et al., 2026) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| LocateAnything Slow NTP (Wang et al., 2026) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| Qwen2.5-VL-7B (Bai et al., 2025b) | 14.70 | 34.29 | 20.57 | 0.09 | 1.20 | 0.16 | 7.54 | 17.39 | 10.49 |
| Qwen3-VL-2B (Bai et al., 2025a) | 22.55 | 39.85 | 28.80 | 0.30 | 0.73 | 0.42 | 11.21 | 19.07 | 14.11 |
| Qwen3-VL-4B (Bai et al., 2025a) | 18.74 | 83.98 | 30.65 | 0.59 | 1.60 | 0.86 | 11.60 | 51.16 | 18.90 |
| Qwen3-VL-8B (Bai et al., 2025a) | 19.80 | 74.93 | 31.33 | 0.39 | 0.85 | 0.53 | 11.09 | 40.67 | 17.41 |
| Qwen3.5-4B (Qwen Team, 2026a) | 27.06 | 41.98 | 32.91 | 0.14 | 0.21 | 0.17 | 11.70 | 18.62 | 14.37 |
| Qwen3.5-9B (Qwen Team, 2026a) | 29.37 | 60.29 | 39.49 | 0.64 | 1.21 | 0.84 | 15.31 | 31.30 | 20.56 |
| RynnBrain1.1 (Li et al., 2026a) | 5.25 | 47.70 | 9.47 | 0.02 | 0.14 | 0.03 | 2.68 | 22.33 | 4.78 |
| SenseNova-Vision (Han et al., 2026) | 78.56 | 88.88 | 83.40 | **10.25** | **10.78** | **10.51** | 59.46 | **65.89** | **62.51** |
| MiMo-VL-7B-SFT (Yue et al., 2025) | 13.59* | 16.68* | 14.98* | 0.04* | 0.04* | 0.04* | 4.65* | 5.79* | 5.15* |
| MiMo-VL-7B-RL (Yue et al., 2025) | 13.58* | 21.88* | 16.76* | 0.01* | 0.02* | 0.01* | 4.51* | 7.23* | 5.56* |
| BAGEL (Deng et al., 2025) | 10.94* | 52.39* | 18.09* | 0.31* | 1.06* | 0.48* | 6.62* | 31.67* | 10.95* |
| RynnBrain (Dang et al., 2026) | 2.04 | 37.23 | 3.87 | 0.02 | 0.17 | 0.03 | 1.00 | 17.12 | 1.88 |
| GroundAnything-VLM | 87.40 | 88.52 | 87.95 | 4.50 | 4.46 | 4.48 | 60.09 | 60.49 | 60.29 |
| GroundAnything | 66.73 | 72.09 | 69.31 | 9.24 | 9.51 | 9.37 | 51.03 | 53.90 | 52.43 |
| **Vision-Language Models (10B–1T)** | | | | | | | | | |
| Qwen3.6-27B (Qwen Team, 2026b) | 28.56 | 53.78 | 37.31 | 1.60 | 3.45 | 2.19 | 19.05 | 35.48 | 24.78 |
| Qwen3-VL-32B (Bai et al., 2025a) | 29.33 | 79.24 | 42.82 | 1.25 | 1.90 | 1.51 | 18.24 | 47.65 | 26.36 |
| Qwen3.8-27B (Qwen Team, 2026d) | 14.80 | 86.51 | 25.27 | 0.60 | 2.42 | 0.97 | 9.84 | 57.57 | 16.80 |
| Qwen3.5-35B-A3B (Qwen Team, 2026a) | 37.48 | 80.34 | 51.12 | 1.44 | 2.31 | 1.77 | 23.36 | 49.13 | 31.65 |
| DeepSeek-VL2-Small-16B (Wu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| DeepSeek-VL2-27B (Wu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| **Vision-Language Models (>1T)** | | | | | | | | | |
| Qwen3.7-Max (Qwen Team, 2026c) | 19.91 | **90.30** | 32.62 | 0.82 | 1.92 | 1.15 | 13.31 | 57.98 | 21.62 |
| Kimi-K2.6 | 52.17 | 55.60 | 53.83 | 2.75 | 3.06 | 2.89 | 31.33 | 33.69 | 32.47 |
| Kimi-K3 (Team et al., 2026) | 67.29 | 76.03 | 71.39 | 3.98 | 4.16 | 4.07 | 40.56 | 44.83 | 42.59 |
| GPT-6 Astra | **89.83** | 87.12 | **88.48** | 8.34 | 8.03 | 8.18 | **61.83** | 60.25 | 61.04 |

Dense200 with exemplars. GroundAnything-VLM scores 75.26 F1mIoU, while GroundAnything reaches 63.52, above Rex-Omni (55.50) and SenseNova-Vision (62.84). GroundAnything's F1 at IoU 0.95 is 25.97 versus 12.29 for GPT-6 Astra, but its IoU-0.50 recall is lower than the AR variant's (67.13 versus 88.68). Thus, the DLM model can localize exemplar-matched instances precisely, while recovering the full instance set remains more difficult.

Table S29: **Dense200 visual prompting.** Complete box-grounding metrics. LocateAnything variants lack a supported visual-prompt interface; both DeepSeek variants have incompatible output protocols. Their entries are N/A. Kimi-K3 is N/A because the required prompt format is unsupported. Starred GroundingDINO scores come from an unsupported visual-prompt task interface; starred MiMo and BAGEL scores are retained observations under suspected output-protocol incompatibility.

| Model | IoU 0.50 | | | IoU 0.95 | | | mIoU | | |
|---|---|---|---|---|---|---|---|---|---|
| | R | P | F1 | R | P | F1 | R | P | F1 |
| **Open-set Specialized Detectors** | | | | | | | | | |
| GroundingDINO (Liu et al., 2024) | 2.03* | 9.98* | 3.38* | 1.35* | 5.64* | 2.18* | 1.90* | 9.00* | 3.14* |
| **Vision-Language Models (<10B)** | | | | | | | | | |
| Rex-Omni (Jiang et al., 2026) | 73.05 | 74.81 | 73.92 | 10.89 | 11.12 | 11.00 | 54.93 | 56.08 | 55.50 |
| LocateAnything Fast (Wang et al., 2026) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| LocateAnything Hybrid (Wang et al., 2026) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| LocateAnything Slow NTP (Wang et al., 2026) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| Qwen2.5-VL-7B (Bai et al., 2025b) | 0.75 | 23.66 | 1.46 | 0.00 | 0.00 | 0.00 | 0.44 | 13.33 | 0.86 |
| Qwen3-VL-2B (Bai et al., 2025a) | 7.76 | 60.82 | 13.76 | 0.69 | 1.78 | 1.00 | 5.08 | 33.85 | 8.79 |
| Qwen3-VL-4B (Bai et al., 2025a) | 1.76 | **93.00** | 3.45 | 0.21 | 8.00 | 0.41 | 1.27 | 62.90 | 2.49 |
| Qwen3-VL-8B (Bai et al., 2025a) | 8.26 | 87.71 | 15.10 | 0.99 | 3.91 | 1.58 | 5.79 | 55.05 | 10.45 |
| Qwen3.5-4B (Qwen Team, 2026a) | 2.82 | 78.69 | 5.44 | 0.34 | 7.71 | 0.65 | 2.01 | 54.25 | 3.87 |
| Qwen3.5-9B (Qwen Team, 2026a) | 26.66 | 65.61 | 37.91 | 4.19 | 10.33 | 5.96 | 19.00 | 46.27 | 26.93 |
| RynnBrain1.1 (Li et al., 2026a) | 1.16 | 62.50 | 2.28 | 0.07 | 4.00 | 0.14 | 0.74 | 38.25 | 1.45 |
| SenseNova-Vision (Han et al., 2026) | 69.56 | 84.68 | 76.38 | 19.15 | 22.06 | 20.50 | 57.21 | 69.70 | 62.84 |
| MiMo-VL-7B-SFT (Yue et al., 2025) | 2.87* | 20.40* | 5.03* | 0.01* | 0.00* | 0.00* | 1.08* | 6.38* | 1.83* |
| MiMo-VL-7B-RL (Yue et al., 2025) | 2.58* | 14.41* | 4.38* | 0.01* | 0.01* | 0.01* | 0.89* | 5.40* | 1.51* |
| BAGEL (Deng et al., 2025) | 1.28* | 56.29* | 2.50* | 0.02* | 1.25* | 0.05* | 0.74* | 29.87* | 1.45* |
| RynnBrain (Dang et al., 2026) | 1.26 | 64.50 | 2.47 | 0.02 | 1.00 | 0.03 | 0.74 | 36.35 | 1.44 |
| GroundAnything-VLM | 88.68 | 92.85 | 90.72 | **28.51** | **30.01** | **29.24** | **73.59** | **77.00** | **75.26** |
| GroundAnything | 67.13 | 79.38 | 72.74 | 24.97 | 27.06 | 25.97 | 58.16 | 69.96 | 63.52 |
| **Vision-Language Models (10B–1T)** | | | | | | | | | |
| Qwen3.6-27B (Qwen Team, 2026b) | 31.92 | 59.87 | 41.64 | 5.79 | 7.69 | 6.61 | 23.81 | 45.89 | 31.33 |
| Qwen3-VL-32B (Bai et al., 2025a) | 18.96 | 55.73 | 28.30 | 2.51 | 6.05 | 3.55 | 13.84 | 41.11 | 20.69 |
| Qwen3.8-27B (Qwen Team, 2026d) | 25.08 | 68.06 | 36.65 | 4.61 | 14.62 | 7.01 | 19.15 | 53.40 | 28.19 |
| Qwen3.5-35B-A3B (Qwen Team, 2026a) | 41.09 | 71.35 | 52.15 | 5.85 | 7.14 | 6.43 | 29.95 | 51.04 | 37.73 |
| DeepSeek-VL2-Small-16B (Wu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| DeepSeek-VL2-27B (Wu et al., 2024) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| **Vision-Language Models (>1T)** | | | | | | | | | |
| Qwen3.7-Max (Qwen Team, 2026c) | 28.05 | 72.46 | 40.44 | 5.40 | 9.51 | 6.89 | 22.37 | 55.16 | 31.81 |
| Kimi-K2.6 | 43.14 | 51.19 | 46.82 | 5.48 | 6.84 | 6.09 | 29.90 | 35.39 | 32.41 |
| Kimi-K3 (Team et al., 2026) | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* | N/A* |
| GPT-6 Astra | **95.44** | 88.71 | **92.04** | 12.84 | 11.75 | 12.29 | 71.48 | 66.44 | 68.94 |

### H.10 Qualitative Analysis

We examine selected visualizations across grounding tasks, emphasizing the spatial and semantic demands visible in each example. Source images and their displayed annotations are preserved. Panels marked "User-curated candidate" are curated illustrations; panels marked "GT-completed display" include ground-truth completion and are not presented as raw model predictions. These examples provide qualitative context, not additional estimates of accuracy or recall. A panel marked N/A denotes an unavailable comparison.

#### H.10.1 General Object Grounding

Grounding in a cluttered indoor scene. The indoor scene in Figure S9 combines people, furniture, books, containers, and small accessories. Large objects provide scene context, while partially visible

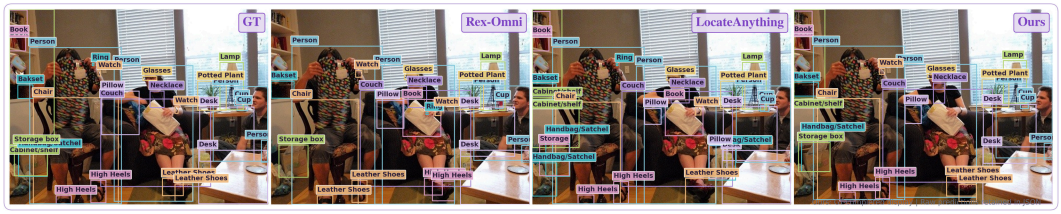


Figure S9: Multi-scale object grounding in a cluttered indoor scene.

items and accessories require finer spatial discrimination. The displayed annotations illustrate the need to retain instance identity across scale and occlusion, rather than replacing a collection of nearby objects with one broad region. In particular, distinguishing an accessory from its wearer requires semantic association as well as localization. The supplied Ours panel is explicitly marked as a GT-completed display and is treated as an illustrative rendering, not as an unedited prediction set.

### H.10.2 Dense Object Grounding

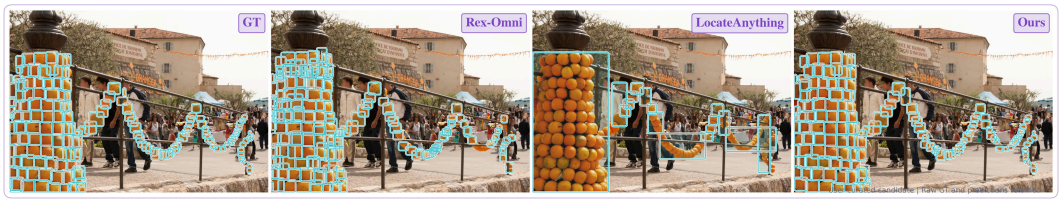


Figure S10: Dense instance grounding of fruit decorations.

Dense instances under occlusion. The repeated fruit decorations in Figure S10 vary in apparent size and overlap along a central support and surrounding garlands. The comparison illustrates a distinction between localizing individual fruits and enclosing an entire decorated structure. Several broad boxes in the comparison panels merge multiple instances, whereas the displayed Ours panel retains finer instance granularity. Adjacent fruits with similar color make both duplicate suppression and boundary separation difficult. The curated visualization illustrates these error modes without assigning additional detection scores.

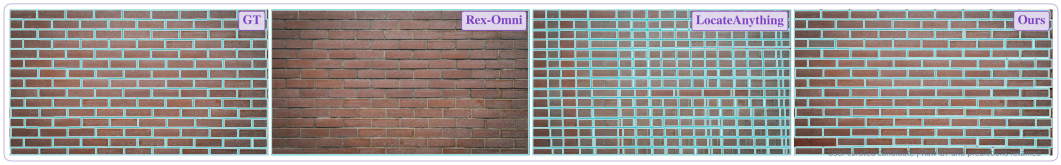


Figure S11: Dense grounding of repeated bricks.

Repeated boundaries and fine spatial structure. Figure S11 uses a brick wall to probe repeated structures with small appearance differences. The displayed Ours boxes follow individual bricks and preserve the staggered rows, whereas crossing or oversized boxes in the comparison panels mix neighboring units. Success here requires aligning the queried unit with local mortar boundaries while maintaining consistency across a large set of nearly interchangeable instances. The figure is a qualitative illustration of spatial granularity; an unannotated comparison panel alone does not identify the underlying cause of a missing display.

### H.10.3 Referring Grounding and Complex Visual Configurations

Complex case: visually deceptive overlap. Figure S12 presents a visually deceptive configuration: two birds overlap so closely that their similar feather textures and aligned body contours can resemble a single two-headed bird. A third bird is only partly visible at the lower image boundary.

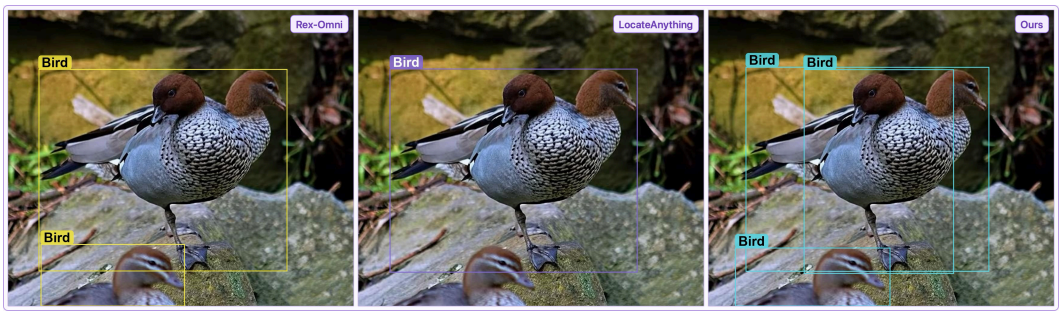


Figure S12: Separating overlapping birds with deceptive silhouettes.

The Rex-Omni panel encloses the overlapping pair in one large box and marks the foreground bird separately; LocateAnything shows one broad box; the Ours panel distinguishes the two overlapping instances and the truncated foreground instance. The relevant evidence includes two distinct heads and beaks, differently oriented necks, and the relationship between each head and its body contour. A purely salient-region interpretation can merge those cues into one object. Resolving the ambiguity calls for instance-level semantic understanding, occlusion reasoning, and spatial part association, motivating a strong vision–language foundation rather than localization based on isolated texture or outline alone. This example illustrates a demanding capability; a single selected case does not establish a particular internal reasoning mechanism or a universal advantage over all competing models.

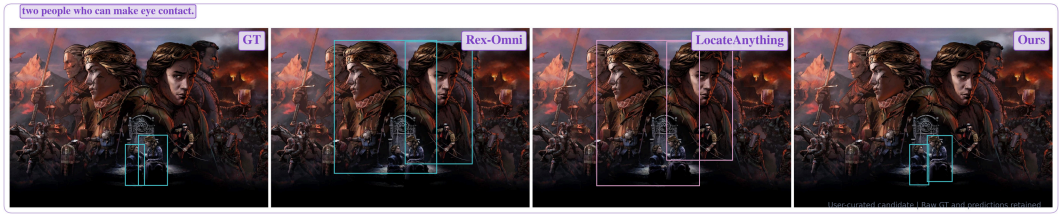


Figure S13: Grounding a pair specified by a visual relationship.

Relational referring beyond visual salience. The query in Figure S13 asks for two people who can make eye contact. The relevant pair occupies a small area near the bottom of a poster dominated by much larger portraits. The displayed Ours regions select that pair, while other panels emphasize the larger faces. The challenge is to interpret a relation between two instances and their orientation, rather than rank people by size or salience. This example concerns the visible relational configuration, not an inference about the depicted people's identity or mental state.




Figure S14: Grounding a spatially specified, partially visible person.

Referring at the image boundary. In Figure S14, the phrase "the person on the far right" refers to a partially truncated spectator at the image boundary, rather than the prominent foreground participant. The scene contains many people with similar hats and clothing, making category recognition alone insufficient. The displayed Ours box follows the boundary instance. The example illustrates the need to resolve a relative spatial expression over all visible candidates, including small and partly occluded instances, before predicting the target extent.

### H.10.4 Referring Point-in-Mask Grounding




Figure S15: Point grounding within an occluded person's visible mask.

Referring points inside visible target regions. Figure S15 separates semantic target selection from the geometric requirement that a point lie inside the target's visible mask. The child is partially hidden by a float, so the enclosing box center can fall on an occluder. The displayed Ours point lies on the visible face, while the comparison points fall on the float or another person. A valid point therefore requires both identifying the intended instance and selecting visible evidence belonging to it, rather than using a generic scene center or box-center heuristic.

### H.10.5 Dense Point Grounding




Figure S16: Dense point grounding of overlapping balloons.

Dense pointing with repeated appearances. Figure S16 contains many balloons with repeated colors, touching outlines, partial occlusions, and instances cut by the image boundary. The displayed points span the arch rather than collapsing the group to a few salient locations. The task combines set coverage with one-point-per-instance consistency: missing small balloons and placing repeated points on the same balloon are distinct errors. Marker size is part of the visualization and should not be interpreted as localization uncertainty or predicted object size.

### H.10.6 Dense Point-in-Mask Grounding




Figure S17: Dense point grounding of vehicles in a night scene.

Dense point placement in low light. Figure S17 shows vehicles from above under low illumination, including moving vehicles on the road and a compact parked group. Ground-truth masks identify the visible vehicle regions, while point predictions must remain within those regions despite shadows and bright headlights. The displayed Ours points cover both isolated and clustered vehicles. The example illustrates why coverage and interior-point placement must be evaluated together: a point on a light streak or adjacent road surface can be close to a vehicle while missing its visible mask.

### H.10.7 GUI Grounding

GUI grounding among overlapping windows. The desktop in Figure S18 contains an editor, a terminal, a plot window, and a property inspector. The marked reference region is the small plot-title

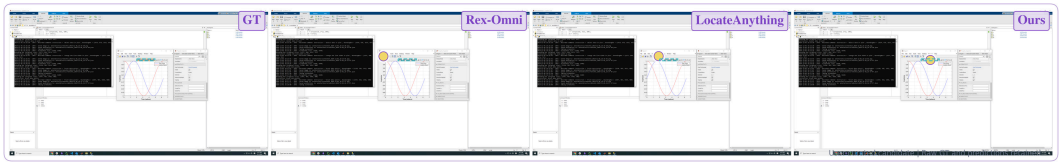


Figure S18: GUI target grounding in a multi-window desktop.

area, surrounded by nearby menu and toolbar controls. The displayed Ours point lies in that region, whereas the comparison points are displaced toward neighboring controls. This case requires assigning a target to the correct window and interpreting local interface structure. It illustrates visual target grounding, not successful execution of an unshown action sequence.

### H.10.8 OCR AND ARTISTIC TEXT UNDERSTANDING




Figure S19: OCR of artistic lettering embedded in illustration.

Complex case: artistic letters and decorative texture. The word "DRAW" in Figure S19 is distributed across staggered, tilted letters embedded in brush strokes and dense line art. Decorative contours resemble character strokes, while the letters do not share a conventional horizontal baseline. The Ours panel reads the two spatial groups as "DR" and "AW"; LocateAnything shows "DR" and "AN", and Rex-Omni uses a single "DRAW" region. This distinction matters: a complete word box and two correct fragments can both be reasonable annotation granularities, so the split alone is not evidence of superior recognition. The informative feature is maintaining the correct letter identity and its spatial support despite stylization, especially the final W. Such reading requires integrating local strokes with word-level context and distinguishing typography from illustration. It motivates strong visual–linguistic representations while also showing why OCR comparisons must specify both transcription and grouping conventions.

Complex case: shape-based typography. In Figure S20, the poster combines small conventional words with the oversized, heavily deformed word "FAN" and an illustrated electric fan. Rounded, interlocking glyphs and decorative stars blur the boundary between text and drawing. The Ours panel preserves the large "FAN" region while also identifying smaller words such as "BE", "YOUR", and "OWN". Reading the slogan requires tracking character order across sizes and styles rather than treating all blue shapes as one object. The fan illustration supplies useful semantic context, but a grounded transcription must still be supported by visible letter strokes. This is a demanding vision–language reading example; it does not imply that contextual plausibility can substitute for character evidence or that every small printed line is recovered perfectly.




Figure S20: OCR of highly stylized poster lettering.




Figure S21: OCR of outdoor signs with perspective distortion.

Scene text with perspective and scale variation. Figure S21 includes large lettering along a curved entrance, a hanging business sign, and smaller roadside text. Perspective changes character orientation and spacing, while foliage and uneven contrast complicate localization. The displayed Ours regions retain separate text groups across these scales; comparison panels illustrate merged phrases and transcription changes. The case highlights joint text–geometry consistency: a plausible phrase in a region that spans multiple signs is different from a transcription aligned to the intended word or line.

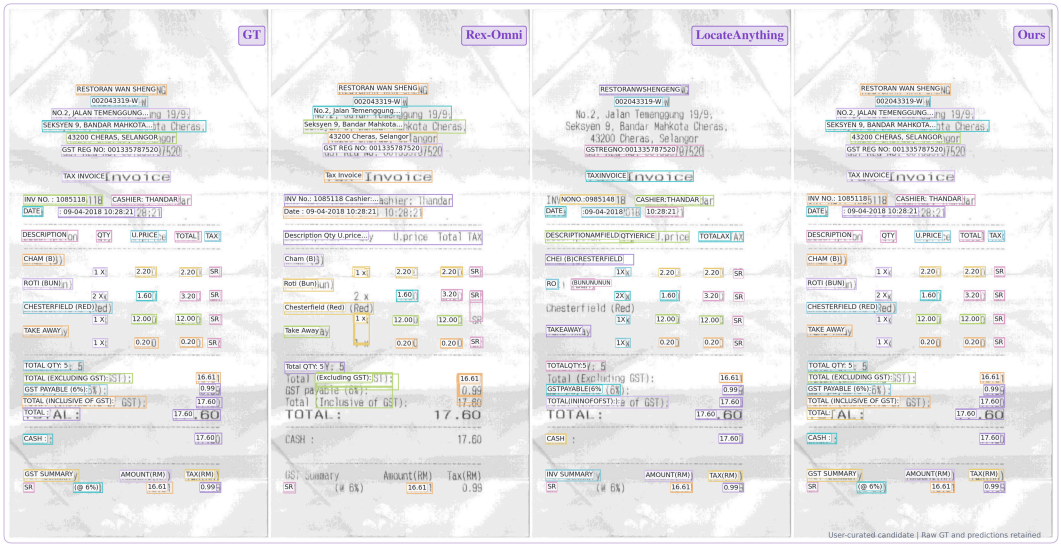


Figure S22: OCR of a low-contrast receipt with aligned fields.

Document text and tabular alignment. The faded receipt in Figure S22 contains headings, item descriptions, quantities, unit prices, totals, and tax fields. The displayed Ours regions preserve sepa-

rate text cells and repeated numerical entries across the document, while comparison panels illustrate merged headers and missed or fragmented lines. Useful OCR must retain both character identity and the layout that associates each value with its row and column. The visualization concerns recognition and localization; it does not establish arithmetic verification or downstream financial correctness.

### H.10.9 VISUAL-PROMPT GROUNDING

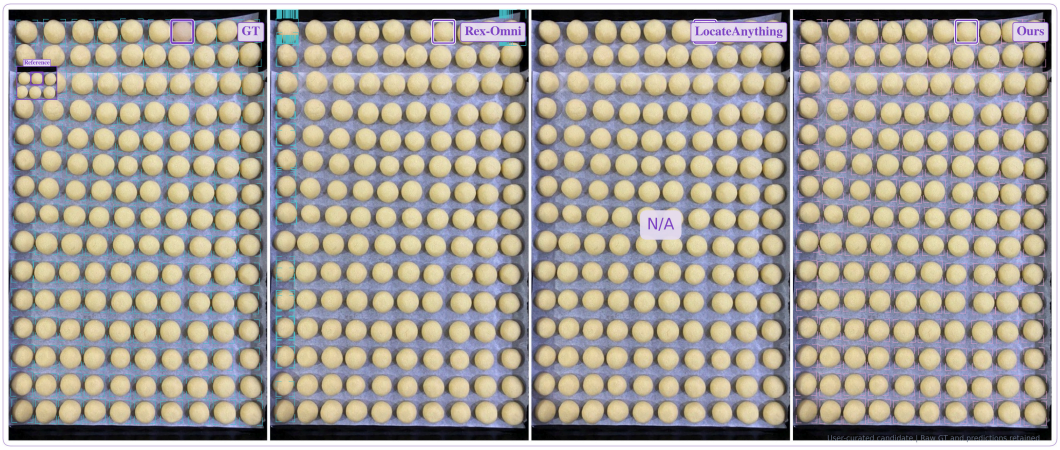


Figure S23: Exemplar-guided grounding of repeated rounded objects.

Visual exemplars in highly repetitive scenes. In Figure S23, a boxed visual exemplar specifies the repeated rounded objects without relying on a category name. The target instances form dense rows with mild shape and size variation. The displayed Ours boxes cover instances across the tray, while the Rex-Omni panel contains sparse and repeated edge-aligned proposals. The LocateAnything panel is explicitly marked N/A and is not treated as a scored failure. The example illustrates transferring an exemplar's visual identity to many instances while avoiding duplicate localization.

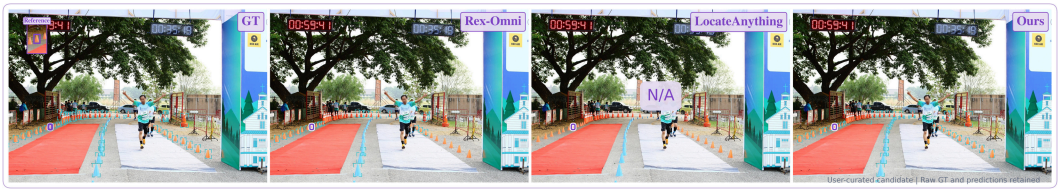


Figure S24: Exemplar-guided grounding of traffic cones.

Visual prompting with perspective variation. Figure S24 uses a local visual reference to specify cones distributed around a running course. Their apparent size changes strongly with depth, and the scene includes both orange and blue cones. The displayed Ours boxes follow instances from the foreground toward the distant gate, illustrating visual category transfer beyond the exact reference crop. The comparison highlights partial coverage and repeated localization. The panel marked N/A denotes an unavailable comparison, not an observed zero score.

## H.11 POTENTIAL APPLICATION SCENARIOS

The following prospective workflows combine changing queries with substantial spatial output; they are opportunities for evaluation rather than demonstrated deployments.

High-throughput industrial inspection. Language or exemplar queries can specify defects, components, assembly checks, and labels across changing product lines. Parallel output is particularly relevant when each image requires many localized results.

Embodied and driving data annotation. Batch processing of robot views, manipulation keyframes, and road images can produce candidate object, part, text, and relation-conditioned spatial annota-

tions. Lower decoding cost could expand annotation and review capacity as images, queries, and instances multiply.

Dense remote sensing. Image tiles of ports, car parks, airports, and urban areas require locating many queried instances. Open queries and long coordinate lists make this a useful setting for evaluating parallel spatial extraction.

Batch OCR and document processing. Receipts, forms, archives, and complex pages require text together with region locations. Long transcriptions and multi-region outputs make this a practical test of decoding savings at matched extraction quality.

Medical and microscopy annotation. Candidate boxes or points for cells, nuclei, and repeated structures could assist research counting and expert review. Domain-specific reliability requires separate evaluation before using these annotations in specialist workflows.

Sports and video analysis. Sampled frames can be queried for players, balls, officials, and jersey numbers, including appearance or spatial conditions. Cheaper frame-level extraction could support denser sampling; temporal association remains a downstream task.

Retail inventory and agricultural counting. Crowded shelves, logistics bins, orchards, and nurseries combine repeated instances with changing appearance or region queries. Exemplar conditioning provides a useful interface when category names alone are insufficient.

Interactive annotation and GUI grounding. Users can revise descriptions and confirm candidate boxes, points, or click locations through repeated queries. Short responses should be assessed by end-to-end interaction latency, including visual processing and prefill.