Spatial Latent Reasoning for Embodied Reference Understanding

Anonymous authors
Paper under double-blind review

ABSTRACT

Pointing-gesture visual grounding requires connecting hand geometry with the visual identity and extent of a referred object. A central challenge for continuous latent reasoning is how to organize these complementary cues into useful intermediate supervision. We propose *Spatial Latent Reasoning (SLR)*, a framework that structures this supervision around an ordered sequence of geometric and visual states. A spatial ray state is supervised by fingertip position and pointing direction, followed by four states aligned with target-region features. To construct the visual targets, we introduce parity pooling, which applies polyphase grouping to average region tokens on four interleaved spatial supports. All states are generated recurrently during training and inference; auxiliary annotations are required only during training. On EgoPoint-Ground, the framework improves mIoU over same-backbone supervised fine-tuning by **2.8, 17.5, and 21.1 percentage points** on Qwen3.5-4B, Qwen2.5-VL-7B, and Qwen3-VL-8B, respectively, with improvements on both hard evaluation sets. On YouRefIt, it achieves **77.6%** precision at IoU 0.5, a numerical margin of **5.2 percentage points** over the reported state of the art under differing evaluation protocols. Ablations support joint geometric and visual supervision on the standard and similar-object sets, and favor parity over three alternative pooling operators on the standard set. These results support task-structured supervision for continuous pointing grounding. We will release the code and supporting materials.

1 INTRODUCTION

Embodied reference understanding identifies the object a person refers to through language and gesture in a shared environment (Chen et al., 2021). We focus on its image-based pointing setting: given an image and a language query, the model predicts the referred object's bounding box (Li et al., 2026c). Consider several similar cups on a table and the query "this cup." Recognizing the cups does not resolve which instance is intended; interpreting the hand's direction does not determine the object's full extent. Resolving the reference requires connecting gesture geometry with visual evidence of the target's identity and spatial extent.

Intermediate reasoning offers a way to organize these cues before predicting the intended referent. Explicit approaches generate textual steps or invoke visual tools (Hu et al., 2024; Zheng et al., 2025); continuous approaches feed generated hidden states back as input embeddings (Hao et al., 2025). LVR (Li et al., 2026a) supervises visual-feature reconstruction with teacher-forced region tokens, while RIS (Cui et al., 2026) combines box and semantic supervision with staged internalization. These methods establish region-grounded supervision as a foundation for latent reasoning. Embodied reference understanding with pointing gestures requires modeling a specific relation: how the observed hand indicates the intended referent. Region reconstruction and box prediction do not directly supervise that hand geometry. A fixed latent budget also requires a choice of how to distribute target-region evidence across states. This motivates our central question: *how should gesture geometry and target appearance supervise latent reasoning for embodied reference understanding?*

In this paper, we propose *Spatial Latent Reasoning (SLR)*, organizing gesture geometry and target appearance into ordered, model-generated continuous states. A *spatial ray state* is trained through a lightweight auxiliary head to predict the fingertip as the ray origin and the finger-base-to-fingertip direction as its orientation. Direction is defined on the visual-token grid to account for nonsquare layouts. These labels depend only on the observed hand, providing supervision distinct from target-region reconstruction.

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.

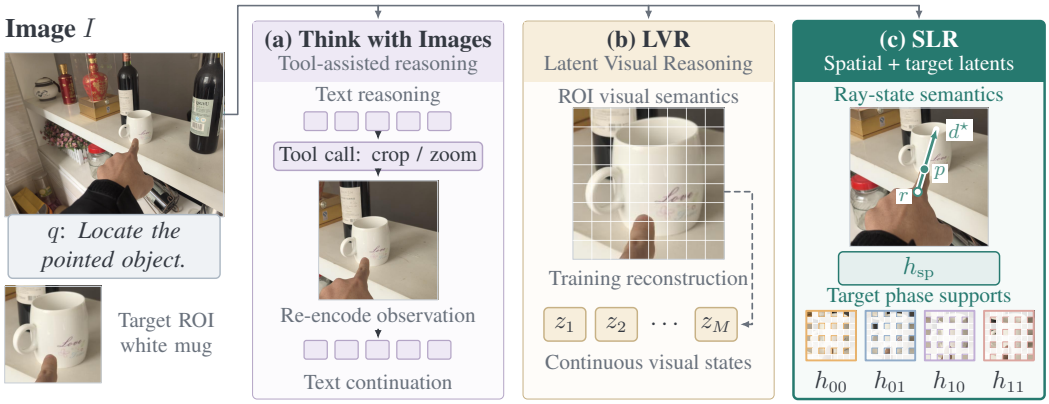


Figure 1: Comparison of reasoning approaches on an EgoPoint-Ground example (Li et al., 2026c). (a) Tool-assisted reasoning crops and re-encodes image regions during text generation (Zheng et al., 2025). (b) LVR (Li et al., 2026a) supervises continuous states by reconstructing annotated region features. (c) SLR combines one spatial ray state with four states supervised by parity-pooled target features. Image crops and ray overlays illustrate the reasoning operations and supervision targets; generic latent boundary markers are omitted for clarity.

The recurrence carries the full-dimensional spatial state; the head's predictions are not fed back. Geometric supervision, subsequent visual alignment, and answer generation jointly train this state to retain task-relevant context alongside pointing information. Four subsequent *target visual states* align with referred-region features, conditioning their generation on the geometry-supervised state. Training and inference share this model-generated recurrence; auxiliary labels enter only the losses.

We introduce *parity pooling* to distribute target-region evidence across four visual states. Drawing on region aggregation (He et al., 2014) and polyphase decomposition (Smith, 2011), it groups tokens by global row and column parity and averages each group. The resulting targets summarize interleaved positions across a variable-size ROI, providing distributed regional evidence without crop re-encoding or textual descriptions. Comparisons with quadrant grouping, random grouping, and max aggregation evaluate this construction. Figure 1 contrasts the reasoning paths; Appendix A provides further comparisons.

Our contributions are threefold: (1) We propose **SLR**, an ordered supervision scheme that couples a full-dimensional, geometry-supervised ray state with subsequent target visual states for embodied reference understanding. (2) We introduce **parity pooling** for latent supervision, converting variable-size target regions into four phase-specific feature targets with interleaved spatial support. (3) SLR improves EgoPoint-Ground mIoU over same-backbone SFT by 2.8, 17.5, and 21.1 percentage points on Qwen3.5-4B (Qwen Team, 2026), Qwen2.5-VL-7B (Bai et al., 2025b), and Qwen3-VL-8B (Bai et al., 2025a), respectively, with gains on both hard evaluation sets. On YouRefIt, it reaches 77.6% precision at IoU 0.5.

2 Method

Building on Latent Visual Reasoning (LVR) (Li et al., 2026a), SLR generates one spatial ray state and four target visual states before decoding the bounding box:

$$(I, q) \longrightarrow h_{\text{sp}} \longrightarrow h_{00} \longrightarrow h_{01} \longrightarrow h_{10} \longrightarrow h_{11} \longrightarrow y_{\text{bbox}}. \tag{1}$$

Figure 2 summarizes the geometric and visual supervision of this sequence.

2.1 Problem Formulation

Let $I \in \mathbb{R}^{H_0 \times W_0 \times 3}$ be an image with a pointing gesture and q a grounding query. We learn

$$f_\theta : (I, q) \mapsto \widehat{b}, \qquad b = (x_1, y_1, x_2, y_2) \in \mathcal{B}, \tag{2}$$

where $\mathcal{B} = \{b \in [0,1]^4 : x_1 < x_2,\ y_1 < y_2\}$ is the space of normalized, axis-aligned boxes. Training examples form $\mathcal{D} = \{(I_n, q_n, b_n^\star)\}_{n=1}^N$, where $b_n^\star$ is the annotated target box. We serialize each box as an answer sequence $y^\star = \tau(b^\star)$ using integer coordinates on a 1000×1000 grid.

Training examples may also provide finger-base and fingertip annotations (r_n, p_n). Geometry and target-region features supply auxiliary losses only and are never inputs to the generated states.

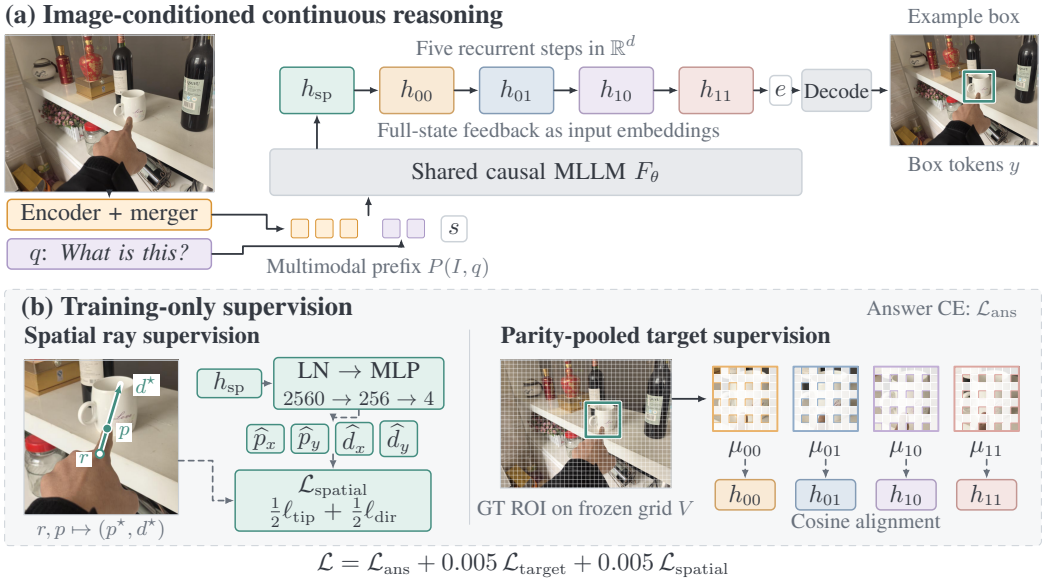


Figure 2: Overview of SLR. (a) A frozen visual encoder and merger provide image features. The causal MLLM generates five recurrent states in its hidden space, then decodes the target box; s and e denote latent boundary markers. (b) A spatial readout predicts fingertip position and pointing direction, while visual states align with parity-pooled region features with stopped gradients. Auxiliary readouts and annotations are used only for training. The ray overlays and grid are schematic; readout dimensions correspond to Qwen3.5-4B.

Model representations. A frozen visual encoder and projector E_ϕ map the image to a post-merge grid $V = E_\phi(I) \in \mathbb{R}^{H_g \times W_g \times d}$. The multimodal prompt $P(I, q)$ packs these visual tokens with query embeddings and modality/position markers. A causal language backbone F_θ generates continuous states $Z_\theta(I, q) = (z_1, \ldots, z_K)$, where $z_k \in \mathbb{R}^d$, before the box answer. Appendix A.6 gives the autoregressive formulation and the original LVR supervision objective.

2.2 Recurrent Continuous Latent States

Given the multimodal prompt $P(I, q)$, a latent-start marker s, and a latent-end marker e, each continuous slot receives the final hidden vector from the preceding position:

$$\begin{aligned} z_1 &= F_\theta([P(I,q), s])_{\text{last}}\,, \\ z_k &= F_\theta([P(I,q), s, z_1, \ldots, z_{k-1}])_{\text{last}}\,, \quad 2 \le k \le K. \end{aligned} \tag{3}$$

Here $F_\theta(\cdot)_{\text{last}}$ is the final-layer hidden vector before vocabulary projection. We use $K = 5$, with

$$(z_1, z_2, z_3, z_4, z_5) = (h_{\text{sp}}, h_{00}, h_{01}, h_{10}, h_{11}), \tag{4}$$

and decode the answer after $[P, s, z_1, z_2, z_3, z_4, z_5, e]$. All states retain dimension d (2560 for Qwen3.5-4B). Slot roles follow their order and supervision; we retain Coconut's generic boundary markers without role-specific markers or embeddings.

Auxiliary losses supervise the producer vectors z_k injected into the slots, rather than the outputs at those slots after recomputation. Gradients propagate through the recurrence; causal masking prevents later placeholders and teacher-forced answer tokens from influencing earlier states.

2.3 Spatial Ray-State Latent

An auxiliary head supervises the origin and orientation of a pointing ray from h_{sp}. Position anchors the gesture in the image, while direction specifies its orientation; together they describe a pointing cue without selecting a candidate object. The full-dimensional state also receives gradients from the subsequent visual and answer objectives, allowing the recurrence to combine this geometric cue with evidence needed for referent selection.

Geometric supervision in two coordinate systems. Let $r = (x_r, y_r)$ and $p = (x_p, y_p)$ denote the annotated finger base and fingertip in normalized-1000 image coordinates, with the horizontal axis

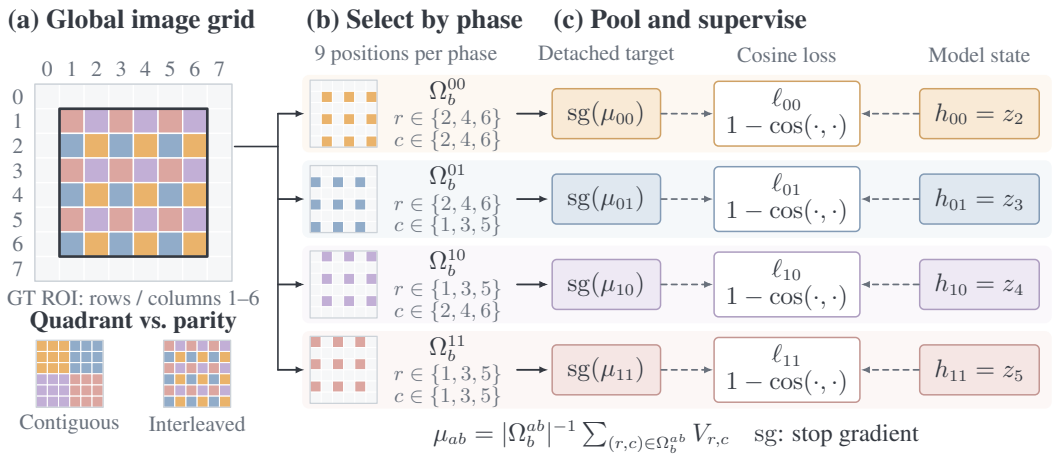


Figure 3: Global-grid parity pooling. The target ROI spans rows and columns 1–6 of an illustrative 8×8 grid. Each branch averages nine feature vectors selected by global row and column parity and aligns the detached mean with its generated state $h_{ab} = z_{2+2a+b}$. Dashed arrows connect both inputs to the cosine loss; pooled features serve only as training targets. The inset compares contiguous quadrants with interleaved parity groups.

pointing right and the vertical axis pointing down. The fingertip regression target lies in $[-1, 1]^2$:

$$p^\star = \left(2\frac{x_p}{1000} - 1,\ 2\frac{y_p}{1000} - 1\right). \tag{5}$$

We scale the finger-base-to-fingertip displacement to the post-merge grid before normalizing:

$$\Delta_g = \left(W_g\frac{x_p - x_r}{1000},\ H_g\frac{y_p - y_r}{1000}\right), \qquad d^\star = \frac{\Delta_g}{\max(\|\Delta_g\|_2, \epsilon)}, \qquad \epsilon = 10^{-12}. \tag{6}$$

This scaling accounts for nonsquare grids; normalizing afterward removes displacement magnitude while preserving orientation in the visual grid. The position and direction targets use different scales; expressing both in grid coordinates defines the pointing ray

$$p_g = \left(W_g\frac{x_p}{1000},\ H_g\frac{y_p}{1000}\right), \qquad \mathcal{R}(t) = p_g + t\,d^\star, \quad t \geq 0. \tag{7}$$

The finger base determines direction only; the spatial labels do not use target boxes or candidate objects. Answer and target-alignment gradients can nevertheless make the state target-aware.

Auxiliary readout. A lightweight training head reads h_{sp} using LayerNorm and two linear maps:

$$\begin{aligned} o &= W_2\ \mathrm{GELU}(W_1\ \mathrm{LN}(h_{\mathrm{sp}}) + b_1) + b_2, \\ \widehat{p} &= o_{1:2}, \qquad \widehat{d} = \frac{o_{3:4}}{\max(\|o_{3:4}\|_2, \epsilon)}. \end{aligned} \tag{8}$$

The head maps $d \to 256 \to 4$, with unconstrained fingertip outputs and a normalized direction. Its predictions are not injected into the recurrent path.

Spatial loss and validity masking. For each valid example, we average SmoothL1 over the two fingertip coordinates and combine it with directional cosine distance:

$$\ell_{\mathrm{spatial}} = \tfrac{1}{2}\,\mathrm{SmoothL1}_{\beta=0.1}(\widehat{p}, p^\star) + \tfrac{1}{2}(1 - \widehat{d}^\top d^\star). \tag{9}$$

Let $\mathcal{B}_{\mathrm{sp}}$ contain minibatch examples with finite finger-base and fingertip annotations and $\|p - r\|_2 > 10^{-6}$ in normalized-1000 coordinates. Averaging ℓ_{spatial} over this set gives $\mathcal{L}_{\mathrm{spatial}}$; an empty set gives zero loss. Examples with invalid finger geometry skip only the spatial auxiliary term; their answer loss and any valid target-region supervision remain active.

2.4 Parity-Pooled Visual Targets

Target support. On the post-merge visual grid V, token (r, c) has normalized center $u_{r,c} = ((c + \frac{1}{2})/W_g, (r + \frac{1}{2})/H_g)$. We select tokens whose centers lie inside the target box:

$$\Omega_b = \{(r, c) : u_{r,c} \in [x_1, x_2] \times [y_1, y_2]\}, \qquad M_b(r, c) = \mathbf{1}[(r, c) \in \Omega_b]. \tag{10}$$

The support is defined directly by the annotated bounding box.

Polyphase decomposition. In multirate signal processing, a discrete signal is decomposed into subsequences indexed by sampling phase (Smith, 2011). For a two-dimensional signal X, stride-two sampling along each axis gives four polyphase components:

$$(\mathcal{P}_{\alpha\beta}X)_{i,j} = X_{2i+\alpha,\,2j+\beta}, \qquad (\alpha,\beta) \in \{0,1\}^2. \tag{11}$$

The components partition the lattice, as in polyphase representations for image sampling (Chaman & Dokmanić, 2021); we partition the ROI mask and visual features to define supervision targets.

Region-conditioned phase averages. For phase (α,β), define

$$\Omega_b^{\alpha\beta} = \{(r,c) \in \Omega_b : r \bmod 2 = \alpha,\ c \bmod 2 = \beta\}, \qquad n_{\alpha\beta} = |\Omega_b^{\alpha\beta}|. \tag{12}$$

For each nonempty support ($n_{\alpha\beta} > 0$), parity pooling computes a mean feature

$$\begin{aligned} \mu_{\alpha\beta}(V,b) &= \frac{1}{n_{\alpha\beta}} \sum_{(r,c)\in\Omega_b^{\alpha\beta}} V_{r,c} \\ &= \frac{\sum_{i,j}(\mathcal{P}_{\alpha\beta}M_b)_{i,j}(\mathcal{P}_{\alpha\beta}V)_{i,j}}{\sum_{i,j}(\mathcal{P}_{\alpha\beta}M_b)_{i,j}}. \end{aligned} \tag{13}$$

The descriptors supervise (z_2, z_3, z_4, z_5) in order $(00, 01, 10, 11)$.

Figure 3 contrasts these interleaved supports with contiguous quadrants. For example, phase $(0,1)$ pools tokens at even rows and odd columns throughout the ROI, rather than selecting one side of the object. The fixed phase-to-state correspondence assigns each recurrent state a consistent sampling rule across differently sized regions.

Each ROI token contributes to exactly one phase. Empty phases are omitted from the auxiliary loss without duplicating tokens, while all four recurrent slots remain. Appendix C gives the parity-pooling properties.

2.5 Visual Alignment and Joint Training

Let $\mu_{n,k}$ denote the phase mean assigned to state $z_{n,k+1}$, and let $\mathcal{A}$ collect minibatch example–slot pairs with nonempty phase supports. Using stop-gradient targets, we minimize

$$\mathcal{L}_{\text{target}} = \frac{1}{|\mathcal{A}|} \sum_{(n,k)\in\mathcal{A}} \left(1 - \frac{z_{n,k+1}^\top \operatorname{sg}(\mu_{n,k})}{\|z_{n,k+1}\|_2 \,\|\operatorname{sg}(\mu_{n,k})\|_2}\right), \tag{14}$$

with stabilized norms and zero loss if $\mathcal{A}$ is empty. The average is over active pairs, and cosine alignment constrains direction rather than magnitude.

Let $\mathcal{T}$ be the set of supervised answer-token positions in the minibatch. The complete objective is

$$\begin{aligned} \mathcal{L}_{\text{ans}} &= -\frac{1}{|\mathcal{T}|} \sum_{(n,t)\in\mathcal{T}} \log p_\theta(y^\star_{n,t} \mid I_n, q_n, Z_\theta(I_n,q_n), y^\star_{n,<t}), \\ \mathcal{L} &= \mathcal{L}_{\text{ans}} + \lambda_{\text{target}}\mathcal{L}_{\text{target}} + \lambda_{\text{spatial}}\mathcal{L}_{\text{spatial}}, \qquad \lambda_{\text{target}} = \lambda_{\text{spatial}} = 0.005. \end{aligned} \tag{15}$$

Only the serialized box and end-of-sequence token contribute teacher-forced cross-entropy; all prompt, marker, latent, and padding positions are masked. We freeze the visual encoder and merger/projector, and train the language-side parameters, including text embeddings, and spatial head. All five recurrent states are model-generated. Answer and target-alignment gradients propagate through earlier states, coupling intermediate supervision to the final grounding objective.

2.6 Inference

Given (I, q), the model generates five states, appends the latent-end marker, and greedily decodes the box until the end-of-sequence token or output limit. The spatial readout is unused, and no ray–object intersection is performed. Compute depends on prompt and answer length, caching, and the fixed latent budget. Appendices B.2 and B.3 detail parsing, constrained decoding, and timing.

3 Experiments

3.1 Datasets and Evaluation

We evaluate on **EgoPoint-Ground** (Li et al., 2026c), which provides egocentric hand–target annotations, and **YouRefIt** (Chen et al., 2021), an embodied reference benchmark with language and gesture. In addition to the standard EgoPoint-Ground test set, we evaluate on two separate hard

Table 1: Results on **EgoPoint-Ground**. SLR uses one spatial and four parity-supervised visual states. Shading marks SLR; bold denotes the best grounding scores per backbone, including ties. Human scores provide a reference; token counts and times appear in Table 9 in Appendix E.

Model	Method	P@0.3 ↑	P@0.5 ↑	P@0.7 ↑	mIoU ↑
Human Study	Human Participants	0.922	0.909	0.830	0.805
Qwen2.5-VL-7B	Zero-shot	0.650	0.589	0.500	0.546
	Standard SFT	0.644	0.589	0.489	0.538
	Text CoT (Wei et al., 2022) (adapted)	0.744	0.661	0.494	0.575
	PointVG-R (Li et al., 2026d)	0.833	0.761	0.639	0.670
	SLR	**0.839**	**0.806**	**0.706**	**0.713**
Qwen3-VL-8B	Zero-shot	0.556	0.528	0.506	0.516
	Standard SFT	0.594	0.550	0.517	0.533
	Text CoT (Wei et al., 2022) (adapted)	0.794	0.778	0.717	0.701
	SLR	**0.828**	**0.800**	**0.761**	**0.744**
Qwen3.5-4B	Zero-shot (Non-thinking)	0.356	0.322	0.289	0.332
	Zero-shot (Thinking)	0.578	0.544	0.456	0.484
	Standard SFT	0.861	0.839	0.761	0.750
	LVR (Li et al., 2026a)	0.600	0.561	0.511	0.525
	SLR	**0.906**	**0.878**	**0.811**	**0.778**

evaluation sets: *Hard-Similar*, which focuses on visually similar objects, and *Hard-Complex*, which focuses on complex scenes. Neither hard evaluation set is a subset of the standard test set.

We report mIoU and localization accuracy P@τ on EgoPoint-Ground at $\tau \in \{0.3, 0.5, 0.7\}$. YouRefIt uses thresholds $\{0.25, 0.5, 0.75\}$; each row identifies its evaluation metric. We organize the main comparisons by backbone and use mIoU to summarize box overlap across EgoPoint-Ground examples. Thresholded accuracy additionally measures localization at increasing precision requirements. Table 9 reports token counts and times; Appendix B.3 details their measurement.

3.2 Comparison Methods

The comparison covers direct prediction (zero-shot and standard SFT), explicit reasoning (Text CoT (Wei et al., 2022) and PointVG-R (Li et al., 2026d)), and continuous visual reasoning (LVR (Li et al., 2026a)). Tables 1 and 2 show the available comparisons on Qwen2.5-VL-7B, Qwen3-VL-8B, and Qwen3.5-4B. Rows marked "adapted" denote backbone-specific adaptations. Tables group methods by backbone to distinguish improvements from the proposed framework from differences in the underlying model; human performance is reported separately as a reference.

3.3 Implementation and Checkpoint Selection

Our Qwen3.5-4B implementation freezes the visual encoder and merger and trains the language model and spatial readout. It uses AdamW with learning rate 10^{-5}, weight decay 0.01, and an effective batch size of 12 across four devices. Table 6 summarizes the reference settings; Appendix B.1 details the architecture, training, and preprocessing.

We select checkpoints by validation mIoU using free greedy generation with a 96-token output limit and native post-latent caching. The highest-scoring checkpoint is selected, with the earlier epoch breaking an exact tie, and is fixed across the standard test set and both hard evaluation sets. This protocol separates checkpoint selection from final evaluation and avoids selecting separate checkpoints for individual evaluation sets.

3.4 Main Results

Improvements across models and localization criteria. Tables 1 and 2 show higher mIoU than same-backbone SFT in all nine backbone–evaluation combinations. Standard-set gains range from 2.8 to 21.1 percentage points, supporting the framework across different backbones. The benefit also extends beyond the direct-prediction baseline: on Qwen2.5-VL-7B and Qwen3-VL-8B, SLR exceeds the strongest reported reasoning baseline in standard-set mIoU by 4.3 points in each case. SLR leads all three localization thresholds within each backbone on the standard set. For Qwen3.5-4B, P@0.7 rises from 0.761 to 0.811, showing that the gain includes boxes meeting a stricter overlap criterion. This supports improved localization quality.

Table 2: Key comparisons on the two **EgoPoint-Ground** hard evaluation sets. Shading identifies SLR; bold scores mark the best reported result within each backbone and evaluation set, including ties. Table 8 in Appendix D reports all methods, localization metrics, token counts, and times.

Model	Method	Hard-Similar		Hard-Complex	
		P@0.5 ↑	mIoU ↑	P@0.5 ↑	mIoU ↑
Qwen2.5-VL-7B	Standard SFT	0.229	0.202	0.458	0.389
	Text CoT (Wei et al., 2022) (adapted)	0.188	0.198	0.448	0.411
	PointVG-R (Li et al., 2026d)	0.313	0.276	0.573	0.473
	SLR	**0.323**	**0.327**	**0.583**	**0.511**
Qwen3-VL-8B	Standard SFT	0.281	0.290	0.385	0.349
	Text CoT (Wei et al., 2022) (adapted)	**0.333**	**0.353**	0.500	0.452
	SLR	0.302	0.315	**0.542**	**0.495**
Qwen3.5-4B	Standard SFT	0.396	0.361	0.594	0.536
	LVR (Li et al., 2026a)	0.167	0.158	0.396	0.396
	SLR	**0.458**	**0.421**	**0.656**	**0.551**

Performance depends on the grounding condition. Ours achieves higher mIoU than SFT on both hard evaluation sets across all three backbones. For Qwen3.5-4B, the gain is larger on Hard-Similar than on Hard-Complex (6.0 versus 1.5 percentage points), motivating a closer examination of visually similar distractors in the component analysis below. The relative ranking also depends on the evaluation metric. On Hard-Complex with Qwen2.5-VL-7B, PointVG-R outperforms Ours at P@0.3 (0.656 versus 0.615), whereas Ours performs better at P@0.7 (0.469 versus 0.365) and in mIoU (0.511 versus 0.473). The reversal between P@0.3 and P@0.7 reflects differences in performance under stricter overlap requirements. On Hard-Similar with Qwen3-VL-8B, Text CoT achieves higher mIoU than Ours (0.353 versus 0.315), showing that the comparison with explicit reasoning is setting-dependent. Appendix D provides the full results. Figures 4 and 5 present qualitative predictions and attention diagnostics, respectively.

Evidence on a second benchmark. On YouRefIt, the gains over the same-backbone zero-shot model are similar across IoU thresholds 0.25/0.5/0.75: 8.4/8.0/8.1 points (Table 3). The persistence of the gain at the strictest threshold supports an improvement in precise grounding on this benchmark.

Table 3: Results on **YouRefIt**. Published baselines use their original metrics and evaluation protocols. Shading identifies SLR; the zero-shot model uses the same backbone.

Method	Reported Metric	@0.25 ↑	@0.50 ↑	@0.75 ↑
Human Participants (Chen et al., 2021)	Accuracy	0.942	0.858	0.533
YouRefIt Full (Chen et al., 2021)	Accuracy	0.547	0.405	0.140
REP (Shi & Yang, 2022)	Precision	0.588	0.457	0.188
Touch-Line (VTL) (Li et al., 2023)	Precision	0.711	0.635	0.390
AD-DINO (ADTL) (Guo et al., 2024)	Accuracy	0.763	0.724	0.554
CAPE (Eyiokur et al., 2026a)	mAP	0.750	0.654	0.357
DA-ERU (Eyiokur et al., 2026b)	mAP	0.787	0.676	0.381
Zero-shot (Qwen3.5-4B)	Precision	0.766	0.696	0.460
SLR (Qwen3.5-4B)	Precision	0.850	0.776	0.541

Qualitative grounding. Figure 4 illustrates referent selection among similar products and localization in a cluttered scene. Our predictions overlap the annotated targets in both examples, complementing the aggregate results with concrete instances of spatial grounding.

3.5 Ablation Studies

Using Qwen3.5-4B, we evaluate components across all three EgoPoint-Ground sets and pooling on the standard set. Figure 6 and Table 7 in Appendix B.5 report the target-state count sweep.

Geometry is most effective with visual supervision on Hard-Similar. Table 4 shows that the full model achieves the best mIoU on the standard and Hard-Similar sets. Adding the spatial state

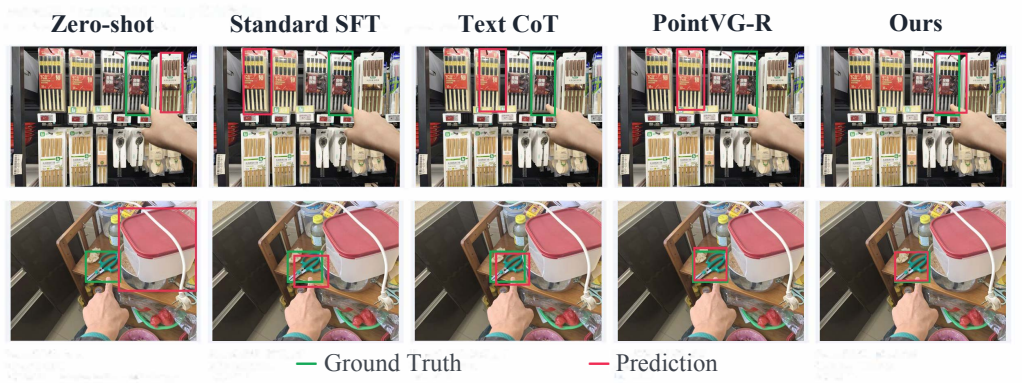


Figure 4: Grounding comparisons for two pointing examples. Columns show Zero-shot, Standard SFT, Text CoT, PointVG-R, and SLR (labeled "Ours" in the supplied visualization). Green and red boxes indicate the annotated targets and predicted regions, respectively, for each of the five compared methods.

Table 4: Component ablations on **Qwen3.5-4B**. Shading marks the full model (one spatial and four target states); bold marks the best grounding scores per evaluation set, including ties. Spatial-only retains the ray state; Target-only retains the four visual states; SFT uses no latent segment.

Backbone	Variant	P@0.3 ↑	P@0.5 ↑	P@0.7 ↑	mIoU ↑
EgoPoint-Ground					
Qwen3.5-4B	**Spatial + Parity (SLR)**	**0.906**	**0.878**	**0.811**	**0.778**
	Spatial-only	0.867	0.844	0.767	0.747
	Target-only	0.883	0.856	0.783	0.764
	SFT (no latent states)	0.861	0.839	0.761	0.750
Hard-Similar: multiple visually similar objects					
Qwen3.5-4B	**Spatial + Parity (SLR)**	**0.500**	**0.458**	**0.396**	**0.421**
	Spatial-only	0.323	0.302	0.250	0.286
	Target-only	0.427	0.396	0.344	0.369
	SFT (no latent states)	0.417	0.396	0.344	0.361
Hard-Complex: complex scenes					
Qwen3.5-4B	**Spatial + Parity (SLR)**	0.719	0.656	0.479	0.551
	Spatial-only	0.688	0.625	0.479	0.558
	Target-only	**0.760**	**0.677**	**0.521**	**0.597**
	SFT (no latent states)	0.667	0.594	0.458	0.536

to Target-only yields a larger gain on Hard-Similar than on the standard set (5.2 versus 1.4 points). Crucially, Spatial-only falls below SFT on Hard-Similar (0.286 versus 0.361), whereas the joint model reaches 0.421. The strongest result therefore comes from supervising both cues together, consistent with the motivation that pointing direction and target appearance provide complementary information when similar objects compete as referents. The Hard-Similar gain also holds at the strictest overlap threshold: joint supervision raises P@0.7 from 0.344 for Target-only to 0.396. Thus, the improvement extends to more precisely localized predictions. Spatial-only does not show this gain, suggesting that a directional cue is most useful when paired with supervision of the target's visual extent. On Hard-Complex, Target-only leads the full model (0.597 versus 0.551), showing that this benefit depends on scene conditions.

Target construction matters at a fixed state budget. Table 5 compares four target states without a spatial state. Parity exceeds Quadrant, Random, and Max in mIoU by 3.6, 2.9, and 1.8 points and leads at all three thresholds. Since all variants use four states, the advantage is not explained by a larger latent-step budget. The controls probe different aspects of the target construction: Quadrant changes interleaved support to contiguous regions, Random changes group membership, and Max changes aggregation on the parity supports (Appendix B.4). Their ordering favors the combination of spatial interleaving and mean aggregation on the standard set. Each parity target aggregates

Table 5: Pooling-operator ablations on **EgoPoint-Ground** with Qwen3.5-4B, four target latents, and no spatial state. Shading identifies parity pooling; bold scores mark the best grounding result, including ties. Appendix B.4 defines the operators.

Backbone	Pooling	P@0.3 ↑	P@0.5 ↑	P@0.7 ↑	mIoU ↑
Qwen3.5-4B	**Parity**	**0.883**	**0.856**	**0.783**	**0.764**
	Quadrant	0.839	0.817	0.744	0.728
	Random	0.844	0.822	0.756	0.735
	Max	0.850	0.833	0.756	0.746

interleaved locations throughout the ROI, preserving phase-specific differences while providing distributed regional evidence. Mean aggregation incorporates all sampled features, whereas Max retains their componentwise extrema. The budget sweep (Appendix B.5) gives a complementary result: four states lead in mIoU, but one state is close (0.764 versus 0.761), and larger budgets bring no improvement. Together, these findings support a compact supervision budget with deliberately constructed visual targets and identify target organization as an important design choice for recurrent visual grounding.

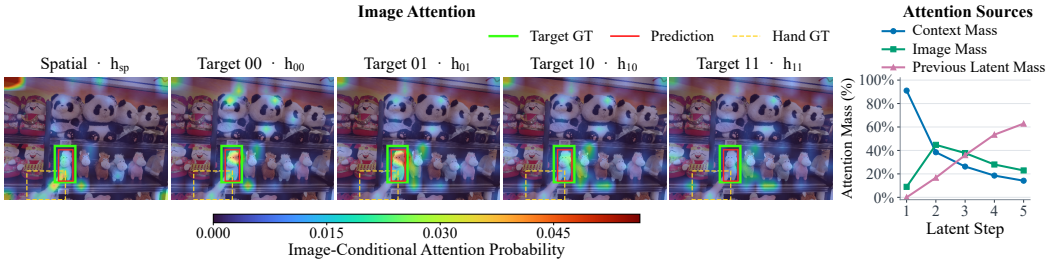


Figure 5: Attention across five recurrent latent steps. Left: image-token attention for one pointing example for the spatial state h_{sp} and target states $h_{00}, h_{01}, h_{10}, h_{11}$, averaged over all 16 heads in layers 20, 24, 28, and 32 and normalized over image tokens using a shared color scale. Green and red boxes mark the ground-truth and predicted target, respectively; the dashed yellow box marks the ground-truth hand region. Right: mean attention mass assigned to context, image tokens, and previous latent states over all evaluation samples.

Attention across latent states. Figure 5 shows a change in attention allocation across the five latent states. The first spatial state attends predominantly to context. The four subsequent visual states combine attention to context, image tokens, and previous latent states. Image attention peaks at the first visual state, while attention to earlier latent states increases through the sequence. This pattern is consistent with a progression from contextual processing toward integrating visual scene evidence with recurrent context. The heatmaps show one example normalized over image tokens, whereas the curves average source allocation over all evaluation samples; concentrated image attention therefore need not dominate total attention.

4 Conclusion

SLR organizes pointing geometry and target appearance into recurrent continuous states. Improvements over same-backbone SFT across three backbones and both EgoPoint-Ground hard evaluation sets, alongside positive YouRefIt results, support task-structured intermediate supervision for embodied reference understanding.

5 Scope and Future Work

Beyond the current single-image, two-dimensional setting, future work includes depth and temporal supervision, weaker annotation requirements, and scene-adaptive states. Controlled state interventions could clarify the reasoning mechanism, while repeated runs and broader evaluations would assess robustness.

AI USE STATEMENT

We used large language models to assist with manuscript preparation, including drafting and revising passages, improving grammar and clarity, organizing the presentation of technical material, and refining figure and table captions. The authors are responsible for reviewing and verifying all AI-assisted content and take full responsibility for the accuracy and integrity of the final manuscript.

ETHICS STATEMENT

Egocentric imagery can expose people and private environments beyond the intended target. Image use and redistribution must respect dataset access conditions and participant privacy. The photographs and overlays in Figures 1–3 illustrate the method; Figures 4 and 5 present prediction and attention examples. A predicted region should not be interpreted as a definitive account of a person's intention. Human evaluation requires documented participant and consent procedures.

REPRODUCIBILITY STATEMENT

Section 2 specifies the recurrence, supervision targets, and objectives. Appendix B.1 documents the reference training configuration; Appendices B.2 and B.3 specify decoding, metric computation, and timing scope. The available source package contains configuration templates, an evaluation-record merger, source hashes, and contract tests for masking, target construction, and gradient boundaries. It does not include model weights, dataset records, or final per-example results. Reproducing these comparisons also requires split manifests, checkpoints, and run configurations.

REFERENCES

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. *arXiv preprint arXiv:2511.21631*, 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. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

Anadi Chaman and Ivan Dokmanić. Truly shift-invariant convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3773–3783, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chaman_Truly_Shift-Invariant_Convolutional_Neural_Networks_CVPR_2021_paper.html.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_SpatialVLM_Endowing_Vision-Language_Models_with_Spatial_Reasoning_Capabilities_CVPR_2024_paper.html.

Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, Song-Chun Zhu, Tao Gao, Yixin Zhu, and Siyuan Huang. YouRefIt: Embodied reference understanding with language and gesture. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1385–1395,

2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Chen_YouRefIt_Embodied_Reference_Understanding_With_Language_and_Gesture_ICCV_2021_paper.html.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f38cb4cf9a5eaa92b3cfa481832719c6-Abstract-Conference.html.

Jin Cui, Xinyue Long, Xunyong Zhang, Yadong Zhang, Chuanchang Su, Jingye Gan, Boran Zhao, and Pengju Ren. Retrieve, integrate, and synthesize: Spatial-semantic grounded latent visual reasoning. *arXiv preprint arXiv:2605.07106*, 2026. URL https://arxiv.org/abs/2605.07106.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. TransVG: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1769–1779, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Deng_TransVG_End-to-End_Visual_Grounding_With_Transformers_ICCV_2021_paper.html.

Shuai Dong, Siyuan Wang, Xingyu Liu, Chenglin Li, Haowen Hou, and Zhongyu Wei. Interleaved latent visual reasoning with selective perceptual modeling. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29316–29335, 2026. URL https://aclanthology.org/2026.acl-long.1351/.

Yifan Du, Kun Zhou, Yingqian Min, Yue Ling, Wayne Xin Zhao, Youbin Wu, and Ji-Rong Wen. Revisiting the necessity of lengthy chain-of-thought in vision-centric reasoning generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12019–12029, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Du_Revisiting_the_Necessity_of_Lengthy_Chain-of-Thought_in_Vision-centric_Reasoning_Generalization_CVPR_2026_paper.html.

Fevziye Irem Eyiokur, Dogucan Yaman, Hazım Kemal Ekenel, and Alexander Waibel. CAPE: A CLIP-aware pointing ensemble of complementary heatmap cues for embodied reference understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3939–3950, 2026a. URL https://openaccess.thecvf.com/content/WACV2026/html/Eyiokur_CAPE_A_CLIP-Aware_Pointing_Ensemble_of_Complementary_Heatmap_Cues_for_WACV_2026_paper.html.

Fevziye Irem Eyiokur, Dogucan Yaman, Hazım Kemal Ekenel, and Alexander Waibel. A multimodal depth-aware method for embodied reference understanding. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2026b. URL https://arxiv.org/abs/2510.08278.

Hao Guo, Wei Fan, Baichun Wei, Jianfei Zhu, Jin Tian, Chunzhi Yi, and Feng Jiang. AD-DINO: Attention-dynamic DINO for distance-aware embodied reference understanding. *arXiv preprint arXiv:2411.08451*, 2024. URL https://arxiv.org/abs/2411.08451.

Ziyu Guo, Rain Liu, Xinyan Chen, and Pheng-Ann Heng. ATLAS: Agentic or latent visual reasoning? one word is enough for both. *arXiv preprint arXiv:2605.15198*, 2026. URL https://arxiv.org/abs/2605.15198.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. In *Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2412.06769.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision*, pp. 346–361, 2014. URL https://arxiv.org/abs/1406.4729.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/He_Mask_R-CNN_ICCV_2017_paper.html.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/fb82011040977c7712409fbdb5456647-Abstract-Conference.html.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/efa0939caa8a906fd0a61a9c60e46e8b-Paper-Conference.pdf.

Kelvin Li, Chuyi Shang, Leonid Karlinsky, Rogerio Feris, Trevor Darrell, and Roei Herzig. Latent implicit visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33457–33466, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_Latent_Implicit_Visual_Reasoning_CVPR_2026_paper.html.

Ling Li, Bowen Liu, Zinuo Zhan, Peng Jie, Jianhui Zhong, Kenglun Chang, and Zhidong Deng. Beyond language: Grounding referring expressions with hand pointing in egocentric vision. In *European Conference on Computer Vision*, 2026c.

Ling Li, Bowen Liu, Zinuo Zhan, Jianhui Zhong, Ziyu Zhu, Bingcai Wei, Kenglun Chang, and Zhidong Deng. PointVG-R: Internalizing geometric reasoning in MLLMs for precise pointing localization via visual chain of thought. *arXiv preprint arXiv:2606.24539*, 2026d. URL https://arxiv.org/abs/2606.24539.

Yang Li, Xiaoxue Chen, Hao Zhao, Jiangtao Gong, Guyue Zhou, Federico Rossano, and Yixin Zhu. Understanding embodied reference with touch-line transformer. In *International Conference on Learning Representations*, 2023. URL https://yzhu.io/publication/pointing2023iclr/.

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. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=lLmqxkfSIw.

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

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0ff38d72a2e0aa6dbe42de83a17b2223-Abstract-Datasets_and_Benchmarks_Track.html.

Cheng Shi and Sibei Yang. Spatial and visual perspective-taking via view rotation and relation reasoning for embodied reference understanding. In *European Conference on Computer Vision*, 2022. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136960199.pdf.

Julius O. Smith. *Spectral Audio Signal Processing*. W3K Publishing, 2011. URL https://www.dsprelated.com/freebooks/sasp/Polyphase_Decomposition.html.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_V_Guided_Visual_Search_as_a_Core_Mechanism_in_Multimodal_CVPR_2024_paper.html.

Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33577–33587, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wu_Grounded_Chain-of-Thought_for_Multimodal_Large_Language_Models_CVPR_2026_paper.html.

Jiaer Xia, Bingkui Tong, Yuhang Zang, Rui Shao, and Kaiyang Zhou. Bootstrapping grounded chain-of-thought in multimodal LLMs for data-efficient model adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 208–217, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Xia_Bootstrapping_Grounded_Chain-of-Thought_in_Multimodal_LLMs_for_Data-Efficient_Model_Adaptation_ICCV_2025_paper.html.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. SoftCoT: Soft chain-of-thought for efficient reasoning with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23336–23351, 2025. URL https://aclanthology.org/2025.acl-long.1137/.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.07704.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. MAttNet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1307–1315, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Yu_MAttNet_Modular_Attention_CVPR_2018_paper.html.

Richard Zhang. Making convolutional networks shift-invariant again. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7324–7334, 2019. URL https://proceedings.mlr.press/v97/zhang19a.html.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing "thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025. URL https://arxiv.org/abs/2505.14362.

APPENDIX

A RELATED WORK

A.1 REFERRING EXPRESSIONS AND EMBODIED POINTING

Referring expression comprehension localizes the object specified by a language expression. MAttNet (Yu et al., 2018) separates subject appearance, location, and inter-object relationships into modules whose contributions depend on the expression. TransVG (Deng et al., 2021) instead uses Transformer-based multimodal fusion and directly regresses bounding-box coordinates. These approaches establish that localization depends on both object content and its relation to the surrounding scene. Pointing introduces a further source of reference: the gesture can disambiguate objects whose appearance and linguistic descriptions are similar.

YouRefIt (Chen et al., 2021) explicitly studies embodied reference understanding with language and gesture, including the perspective differences between the person making a reference and the observer. EgoPoint (Li et al., 2026c) develops egocentric hand-pointing data and a structured visual chain-of-thought baseline. PointVG-R (Li et al., 2026d) pursues precise pointing localization through geometric reasoning traces, supervised fine-tuning, and reinforcement learning. Together, these works motivate treating pointing as a relation between a gesture and a candidate referent, rather than identifying the hand or recognizing the target category alone. Our setting follows this grounding task, but studies how to supervise the model's recurrent continuous states. The fingertip and root-to-tip direction supervise a two-dimensional ray representation. The box is predicted from the image, query, and latent sequence, without a prescribed ray–object intersection rule.

A.2 GROUNDED MULTIMODAL MODELS AND SPATIAL KNOWLEDGE

Grounded multimodal language models connect linguistic representations to image regions. Kosmos-2 (Peng et al., 2024) associates text spans with bounding boxes and represents location information in its output sequence. Ferret (You et al., 2024) combines discrete spatial coordinates with continuous visual features to represent regions at different granularities. These models demonstrate the value of spatially indexed features. Prompted regions and output coordinates serve different roles from recurrent hidden states generated before the answer.

SpatialVLM (Chen et al., 2024) improves qualitative and quantitative spatial reasoning through automatically generated spatial question–answer data. SpatialRGPT (Cheng et al., 2024) develops grounded spatial reasoning using region information, a spatial data pipeline, and optional depth features. Their emphasis is learning spatial knowledge and answering spatial questions, including relations in three dimensions. Our geometry supervision is narrower: a fingertip position and an image-grid-scaled direction specify a two-dimensional pointing state. This auxiliary objective does not reconstruct the three-dimensional scene or uniquely determine the referred object. It is combined with target-region visual supervision because gesture geometry and candidate appearance supply different information for the grounding decision.

A.3 EXPLICIT VISUAL REASONING AND THINKING WITH IMAGES

Chain-of-thought prompting makes intermediate reasoning explicit in text (Wei et al., 2022). Multimodal extensions can also expose visual evidence or manipulate images during reasoning. V* (Wu & Xie, 2024) uses guided visual search and visual working memory to acquire fine-grained evidence. Visual Sketchpad (Hu et al., 2024) enables models to draw and use visual tools, while DeepEyes (Zheng et al., 2025) trains image-tool use with reinforcement learning. We use *thinking with images* as a family-level description for such explicit visual interactions; Figure 1 depicts a representative crop-and-inspect path, not a shared implementation of all these methods.

A complementary line of work grounds intermediate rationales in image regions. Visual CoT (Shao et al., 2024) provides bounding-box annotations for intermediate visual reasoning and supports iterative attention to relevant regions. Grounded Chain-of-Thought (Wu et al., 2026) connects reasoning steps to localized visual cues and evaluates answer–grounding consistency. Xia et al. (2025) bootstrap grounded reasoning traces for data-efficient model adaptation. PointVG-R (Li et al., 2026d)

specializes explicit geometric reasoning to pointing localization. These approaches expose intermediate evidence and incorporate geometry into explicit reasoning.

The relevant design question is which intermediate representation is useful for the task. In controlled vision-centric generalization experiments, Du et al. (2026) find that concise grounded reasoning can be more effective than lengthy textual chains. This supports examining reasoning format and budget, but does not establish that text is universally inadequate or that continuous states are necessary. Our model uses recurrent hidden-state feedback without intermediate textual rationales or visual-tool calls. Any accuracy or latency advantage over explicit reasoning must nevertheless be measured under matched evaluation conditions; shorter output omits recurrent computation costs.

A.4 Continuous and Latent Visual Reasoning

Coconut (Hao et al., 2025) provides the central continuous transition used here: a hidden state is fed back as the next input embedding without first selecting a vocabulary token. SoftCoT (Xu et al., 2025) offers a different route to soft reasoning representations, using an assistant model and a learned projection to supply soft thoughts to a language model. These methods motivate embedding-space computation without specifying the gesture geometry or target appearance to encode.

LVR (Li et al., 2026a) directly addresses visual grounding of latent computation by aligning predicted continuous states with visual embeddings from annotated regions. It is the closest algorithmic foundation for our target supervision, and already supports reasoning with a bounded number of visual latents. ILVR (Dong et al., 2026) interleaves textual and latent reasoning and selectively models perceptual features from intermediate images. LIVR (Li et al., 2026b) uses learned latent tokens and an attention-bottleneck training stage to encourage implicit visual processing without intermediate visual annotations. Its learned prompt tokens differ from the freely generated hidden vectors recursively fed back in Coconut and our model. ATLAS (Guo et al., 2026) also internalizes visual operations, but uses functional vocabulary tokens rather than the same full-dimensional continuous recurrence. Thus, "latent visual reasoning" encompasses distinct mechanisms and supervision.

Especially relevant is RIS (Cui et al., 2026), which grounds latent reasoning in both spatial and semantic information. RIS uses auxiliary box and semantic prediction heads, a curriculum from explicit grounded steps to latent reasoning, and a later attention-masking stage. Its reported setup also uses five latent tokens. RIS therefore provides a direct precedent for spatial-semantic latent supervision and auxiliary readouts under a compact reasoning budget. Our design assigns task-specific supervision to an ordered recurrent sequence: one state predicts fingertip position and pointing direction, followed by four states aligned with parity-pooled target features. It does not require intermediate textual reasoning traces or region-description targets. Our focus is this task-specific factorization of geometry and visual evidence within freely generated recurrent states.

Across these comparisons, our auxiliary labels constrain what information is recoverable from the recurrent states. They are not substituted for those states, and are not available at inference. Successful decoding of geometry is evidence of representation content; demonstrating that the box prediction depends on that geometry requires additional interventions or controlled ablations.

A.5 Region Pooling and Polyphase Representations

Region pooling converts a variable spatial support into a fixed-size representation. Spatial pyramid pooling (He et al., 2014) aggregates features in spatial bins at multiple resolutions. RoIAlign, introduced in Mask R-CNN (He et al., 2017), avoids quantizing region boundaries and uses bilinear interpolation to improve feature alignment. These methods motivate a general distinction between the region being represented and the operator used to summarize its features. Parity pooling uses post-merge image tokens within the annotated box, without cropping or re-encoding.

The four groups arise from a classical signal-processing construction. Polyphase decomposition partitions a discrete signal into interleaved subsequences indexed by sampling phase (Smith, 2011). In two dimensions, the row and column residues modulo two define four interleaved lattices. Work on convolutional downsampling further studies how sampling phase affects shift sensitivity: Zhang (2019) applies low-pass filtering before subsampling, and Chaman & Dokmanić (2021) develops adaptive polyphase sampling for shift-invariant networks. These are precedents for reasoning about sampling phase, without implying that all uses of parity groups inherit these invariance properties.

We transfer the four-phase lattice partition to the construction of visual supervision targets. A mean is computed over each parity class within the target support, with phase indices anchored to the full image grid. Unlike four quadrants, the classes are spatially interleaved instead of representing contiguous parts of the box. Unlike one region mean, they retain four separate phase-conditioned summaries. The polyphase partition itself is established, and averaging each class is lossy: our construction does not imply invertibility, anti-aliasing, or translation invariance. Our contribution concerns the use of these phase-conditioned visual targets alongside a supervised ray state in recurrent pointing-grounding computation. Comparisons with quadrant means, count-matched random-group means, and maxima on parity supports test the grouping and aggregation choices. Their relative performance is an empirical question, rather than a consequence of the signal-processing analogy.

A.6 Autoregressive Formulation and LVR Supervision

For a multimodal prompt and model-generated states $Z_\theta(I, q)$, answer generation factorizes as

$$p_\theta(y \mid I, q) = \prod_{t=1}^{|y|} p_\theta(y_t \mid I, q, Z_\theta(I, q), y_{<t}) \,. \tag{16}$$

Visual supervision in LVR. LVR (Li et al., 2026a) gives continuous states a visual training target. An annotated ROI selects projected image tokens $(v_1, \ldots, v_M)$. Its supervised stage places these tokens in the latent segment as teacher-forced inputs and trains the preceding hidden states $\widetilde{h}_j$ to reconstruct them, jointly with next-token prediction:

$$\mathcal{L}_{\mathrm{LVR}}^{\mathrm{SFT}} = \mathcal{L}_{\mathrm{NTP}} + \frac{\lambda_{\mathrm{LVR}}}{M} \sum_{j=1}^{M} \|\widetilde{h}_j - \mathrm{sg}(v_j)\|_2^2. \tag{17}$$

At inference, LVR feeds predicted hidden vectors back between latent boundary markers.

B Additional Experimental Details

B.1 Reference Implementation and Training

Table 6 summarizes the configuration for Qwen3.5-4B. It initializes from the base model, freezes the visual encoder and merger, and trains language-side parameters, including token embeddings, together with the spatial head. The existing generic latent markers are retained without role-specific additions. No intermediate textual reasoning steps are used as training targets. The reference configuration disables selector injection and auxiliary bounding-box heads.

Table 6: Reference training configuration for Qwen3.5-4B with a four-epoch training budget. These settings describe the reference implementation rather than all baseline runs.

Setting	Reference value
Backbone / hidden dimension	Qwen3.5-4B / 2560
Spatial / target states	1 / 4
Image pixel limits	50,176 to 262,144
Optimizer / learning rate	AdamW / 10^{-5}
Weight decay / seed	0.01 / 0
AdamW (β_1, β_2) / ϵ	$(0.9, 0.999)$ / 10^{-8}
Devices / batch per device / accumulation	4 / 1 / 3
Training budget	4 epochs / 800 optimizer updates
Spatial / target loss weight	0.005 / 0.005
Fingertip / direction coefficient	0.5 / 0.5
SmoothL1 parameter	$\beta = 0.1$
Generation limit / cache	96 tokens / native post-latent cache enabled

The archived environment is documented as Linux with Python 3.10, PyTorch 2.10.0 (CUDA 12.8), Transformers 5.10.2, and four H100 GPUs. Training uses fully sharded data parallelism with activation checkpointing. Model parameters are converted directly to BF16; the export preserves its

BF16 Adam-state strategy without FP32 master weights. AdamW is reinitialized at each epoch, and the path uses a constant learning rate without a scheduler. Precision, optimizer-state continuity, and scheduling are integral to this recipe; changing any of them alters the training configuration.

The dataset and distributed sampler are shuffled, with the sampler epoch updated each epoch. The nominal effective batch is $4 \times 1 \times 3 = 12$. Checkpoints are saved each epoch as model weights, without a complete optimizer and random-state resume snapshot. Selection uses the validation-generation protocol in Section 3.3. The training configuration's separate 48-token field is inactive for online generation evaluation; the exported evaluation worker explicitly uses 96 tokens. Actual post-merge grid dimensions are read from the image processor rather than a fixed input resolution.

The source package includes configuration and source-identity checks and CPU contract tests for the spatial/parity objectives, valid masks, detached targets, and direct gradient paths. These checks establish implementation contracts, not reproduced GPU training or benchmark performance.

B.2 Bounding-Box Decoding Protocols

Free greedy decoding evaluates localization together with structured answer generation. The reference parser accepts an explicit `bbox_2d` array or either supported box-tag form, and uses the last structured box in output order. Four unmarked numbers are not accepted as a box. The scorer sorts each pair of horizontal and vertical endpoints, computes continuous-area IoU, and does not clip coordinates to the image range. A missing parse or a nonpositive union yields zero IoU. Parse success identifies a coordinate structure but does not guarantee a valid in-image box.

The tokenizer-aware constrained protocol operates during autoregressive answer generation. Fixed syntax tokens are supplied by the decoder; coordinate values are selected from the model logits using a prefix trie over tokenizations of admissible integers. At each coordinate step, the decoder selects from the allowed next-token set $\mathcal{C}_t(\widehat{y}_{<t}) \subseteq \mathcal{V}$ according to

$$\widehat{y}_t = \arg \max_{v \in \mathcal{C}_t(\widehat{y}_{<t})} p_\theta(v \mid I, q, Z_\theta(I, q), \widehat{y}_{<t}). \tag{18}$$

The trie handles multi-token integers using the tokenizer, grammar, and decoded coordinates, without ground-truth boxes. It is a separate control available in the shared implementation; the worker uses free greedy decoding. A matched free/constrained comparison must keep the checkpoint and latent recurrence fixed and report each protocol and its caching settings separately.

B.3 Metrics, Record Integrity, and Timing

For N evaluated examples, let $s_i = \mathrm{IoU}(\widehat{b}_i, b_i^\star)$, with unparseable predictions assigned zero. We report the following mean overlap and thresholded accuracy

$$\mathrm{mIoU} = \frac{1}{N} \sum_{i=1}^{N} s_i, \qquad \mathrm{P}\,@\tau = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[s_i \geq \tau], \quad \tau \in \{0.3, 0.5, 0.7\}. \tag{19}$$

All evaluated examples remain in the denominator. The worker requires an annotated target box for every example. Parse failures, end-of-sequence termination, and output-limit hits are logged separately so that changes in answer formatting can be distinguished from localization quality.

The launcher assigns four disjoint shards without padding. Its merger checks complete example coverage, common checkpoint identity, data and record hashes, and unchanged model parameters. Metrics are aggregated over examples rather than obtained by an unweighted average of shard means. The package retains generated token IDs and per-example elapsed time in JSONL records.

Generated-token counts can be computed from the stored output IDs, including the terminal token when emitted. They exclude the image/query prefix and its latent carrier and boundary markers. The five recurrent states must therefore be reported separately from output length. The recorded elapsed time starts before sample collation and includes device transfer, provenance hashing, instrumented generation, text decoding, and scoring; it excludes model loading and the initial full-model hash. These instrumented times are not pure decoding latency. A dedicated latency comparison requires a common hardware, batch size, caching, warm-up, synchronization, and timing boundary, with diagnostic overhead removed or reported consistently.

B.4 Reference Pooling Controls

All four reference operators use the same ROI token set and produce four detached targets, with empty groups masked out of the alignment loss. Parity averages each global row/column parity group. Quadrant partitions the ROI tokens at the box midpoint, assigning tokens on the vertical or horizontal dividing line to the left or top half, respectively, and averages each group in top-left, top-right, bottom-left, bottom-right order. Max retains the parity groups and takes an elementwise maximum within each group, isolating the aggregation rule from the spatial partition.

The reference Random operator deterministically permutes ROI token indices by a SHA-256 ordering keyed by the partition version, partition seed, image-relative path, grid dimensions, and ROI-index fingerprint. It splits this ordering into four disjoint groups whose sizes match the corresponding parity groups, and averages each group. For a fixed key and ROI, the partition is repeatable and does not resample during training. This controls the number of tokens per target while changing group membership.

B.5 Target-Latent Budget Analysis

Figure 6 and Table 7 compare four target-state budgets without a spatial state. Four states attain the highest reported mIoU (0.764), closely followed by one state (0.761); one state has higher P@0.7 (0.794 versus 0.783). Increasing the budget to nine or sixteen does not improve mIoU. This sweep supports a compact supervision budget and motivates the four-state reference used in the pooling comparison. Changes in target grouping accompany changes in state count, so the comparison concerns the complete supervision configurations.

For $K = m^2$ target states, ROI tokens are partitioned by global (row mod m, column mod m) indices in row-major order. Each nonempty group supplies a detached mean target. Empty groups are excluded from the alignment loss while all K recurrent states are retained.

Table 7: Ablation of target-state count on **EgoPoint-Ground** with Qwen3.5-4B and no spatial state. Shading marks the four-state pooling reference; the full model uses five states. Bold values denote the best grounding scores, including ties. See Appendix B.3 for token-count (†) conventions.

Backbone	Pooling latents	P@0.3 ↑	P@0.5 ↑	P@0.7 ↑	mIoU ↑	Tokens†	Time (s) ↓
EgoPoint-Ground							
Qwen3.5-4B	**4 (reference)**	**0.883**	**0.856**	0.783	**0.764**	6.000	1.766
	1	0.867	0.844	**0.794**	0.761	3.000	1.258
	9	0.817	0.789	0.744	0.712	11.000	2.517
	16	0.867	0.844	0.789	0.757	18.000	3.578

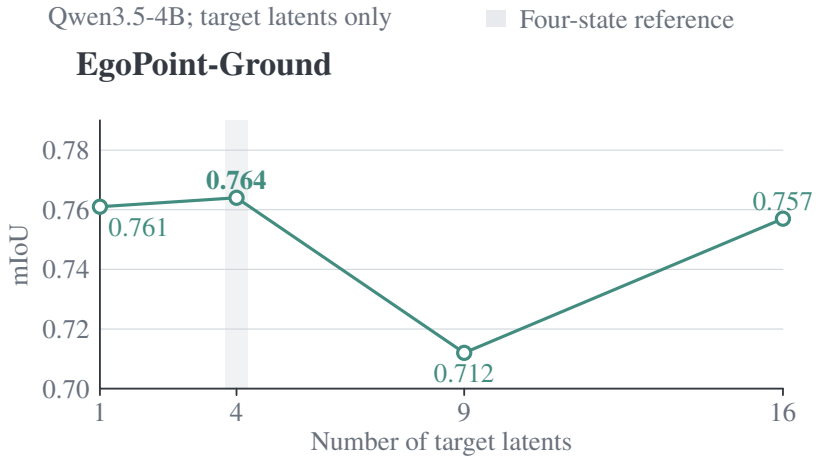


Figure 6: Effect of target-state count on EgoPoint-Ground mIoU with Qwen3.5-4B and no spatial state. Points correspond to Table 7; connecting lines aid visual comparison. Shading marks the four-state reference, and the bold value denotes the highest mIoU.

C PARITY-POOLING PROPERTIES

Structural properties. The supports satisfy $\Omega_b = \bigsqcup_{\alpha,\beta\in\{0,1\}} \Omega_b^{\alpha\beta}$. In particular, each selected visual token contributes to exactly one descriptor. For a nonempty rectangular support spanning h consecutive rows and w consecutive columns, let h_α and w_β count rows and columns of each parity:

$$n_{\alpha\beta} = h_\alpha w_\beta, \qquad h_\alpha \in \{\lfloor h/2 \rfloor, \lceil h/2 \rceil\}, \qquad w_\beta \in \{\lfloor w/2 \rfloor, \lceil w/2 \rceil\}. \tag{20}$$

Hence all four phases are populated when $h, w \geq 2$. Thin or unresolved ROIs can leave phases empty; we omit those supervision terms without duplicating tokens, while retaining all four states.

Moreover, the ordinary ROI mean is recovered from the phase means by

$$\overline{V}_b = \frac{1}{|\Omega_b|} \sum_{(r,c)\in\Omega_b} V_{r,c} = \sum_{\alpha,\beta:n_{\alpha\beta}>0} \frac{n_{\alpha\beta}}{|\Omega_b|} \mu_{\alpha\beta}(V, b). \tag{21}$$

Thus, phase means retain the information needed for the single pooled mean, given their counts, while exposing differences between interleaved subsets. Phase-wise averaging is nevertheless lossy: it does not reconstruct the feature grid or guarantee distinct descriptors. Our use of fixed phase order also does not imply translation invariance of the encoder or of the recurrent predictor.

D FULL RESULTS ON THE HARD EVALUATION SETS

Table 8 expands Table 2 with all reported methods, the human reference, all localization thresholds, token counts, and times. Hard-Similar and Hard-Complex are reported separately. † denotes the reported token counts. Appendix B.3 discusses timing scope and token-count conventions for interpreting the reported token counts and elapsed times.

Table 8: Full results on the **EgoPoint-Ground** Hard-Similar and Hard-Complex evaluation sets. Shading identifies SLR; bold values denote the best grounding scores within each backbone and evaluation set, including ties. Dashes indicate unreported values; N/A indicates inapplicable entries.

Model	Method	P@0.3 ↑	P@0.5 ↑	P@0.7 ↑	mIoU ↑	Tokens†	Time (s) ↓
Hard-Similar: multiple visually similar objects							
Human Study	Human Participants	0.965	0.948	0.917	0.856	N/A	N/A
Qwen2.5-VL-7B	Zero-shot	0.281	0.240	0.167	0.214	0.000	0.864
	Standard SFT	0.281	0.229	0.146	0.202	0.000	0.576
	Text CoT (adapted)	0.292	0.188	0.063	0.198	166.854	3.312
	PointVG-R	**0.354**	0.313	0.177	0.276	175.521	3.511
	SLR	**0.354**	**0.323**	**0.313**	**0.327**	7.000	0.755
Qwen3-VL-8B	Zero-shot	0.271	0.271	0.271	0.271	0.000	1.111
	Standard SFT	0.292	0.281	0.281	0.290	0.000	1.074
	Text CoT (adapted)	**0.375**	**0.333**	**0.323**	**0.353**	179.302	4.473
	SLR	0.313	0.302	0.292	0.315	7.000	0.884
Qwen3.5-4B	Zero-shot (Non-thinking)	0.198	0.167	0.156	0.161	0.000	1.238
	Zero-shot (Thinking)	0.271	0.240	0.188	0.228	130.615	5.149
	Standard SFT	0.417	0.396	0.344	0.361	0.000	1.101
	LVR	0.177	0.167	0.135	0.158	6.000	1.329
	SLR	**0.500**	**0.458**	**0.396**	**0.421**	7.000	1.899
Hard-Complex: complex scenes							
Human Study	Human Participants	0.958	0.920	0.809	0.796	N/A	N/A
Qwen2.5-VL-7B	Zero-shot	0.531	0.490	0.396	0.416	0.000	0.869
	Standard SFT	0.510	0.458	0.323	0.389	0.000	0.564
	Text CoT (adapted)	0.583	0.448	0.208	0.411	168.760	3.368
	PointVG-R	**0.656**	0.573	0.365	0.473	178.938	3.535
	SLR	0.615	**0.583**	**0.469**	**0.511**	7.000	0.748

Continued on next page.

Table 8 (continued)

Model	Method	P@0.3 ↑	P@0.5 ↑	P@0.7 ↑	mIoU ↑	Tokens†	Time (s) ↓
Qwen3-VL-8B	Zero-shot	0.427	0.396	0.302	0.354	0.000	1.116
	Standard SFT	0.427	0.385	0.313	0.349	0.000	1.095
	Text CoT (adapted)	0.531	0.500	0.406	0.452	180.333	4.676
	SLR	**0.583**	**0.542**	**0.448**	**0.495**	7.000	0.930
Qwen3.5-4B	Zero-shot (Non-thinking)	0.323	0.302	0.260	0.274	0.000	1.352
	Zero-shot (Thinking)	0.375	0.333	0.240	0.293	150.188	5.629
	Standard SFT	0.667	0.594	0.458	0.536	0.000	1.145
	LVR	0.500	0.396	0.333	0.396	6.000	1.308
	SLR	**0.719**	**0.656**	**0.479**	**0.551**	7.000	1.889

E SUPPLEMENTARY COMPUTATION RECORDS

Table 9 retains the token counts and times accompanying the main comparisons and ablations. The values are reproduced without converting their units. In particular, the reported zero counts for direct prediction and the fixed counts for latent methods are not total answer lengths. Continuous states, boundary markers, and generated text require separate accounting; these aggregate records do not establish a common counting convention across method families. Timing scope is discussed in Appendix B.3. Localization results in the main text do not rely on an efficiency ranking.

Table 9: Reported token counts and elapsed times for Tables 1, 4, and 5. Appendix B.3 details token-count (†) and timing conventions. Standard denotes EgoPoint-Ground.

Backbone	Method / variant	Evaluation set	Tokens†	Time (s)
Main comparisons				
Qwen2.5-VL-7B	Zero-shot	Standard	0.000	1.181
	Standard SFT	Standard	0.000	0.885
	Text CoT (adapted)	Standard	173.533	4.402
	PointVG-R	Standard	183.622	4.708
	SLR	Standard	7.000	1.362
Qwen3-VL-8B	Zero-shot	Standard	0.000	1.444
	Standard SFT	Standard	0.000	1.421
	Text CoT (adapted)	Standard	176.006	5.602
	SLR	Standard	7.000	1.448
Qwen3.5-4B	Zero-shot (Non-thinking)	Standard	0.000	1.424
	Zero-shot (Thinking)	Standard	150.267	5.685
	Standard SFT	Standard	0.000	1.118
	LVR	Standard	6.000	1.341
	SLR	Standard	7.000	1.859
Component configurations				
Qwen3.5-4B	Spatial + Parity (SLR)	Standard	7.000	1.859
	Spatial-only	Standard	3.000	1.268
	Target-only	Standard	6.000	1.766
	SFT	Standard	0.000	1.118
	Spatial + Parity (SLR)	Hard-Similar	7.000	1.899
	Spatial-only	Hard-Similar	3.000	1.276
	Target-only	Hard-Similar	6.000	1.753
	SFT	Hard-Similar	0.000	1.101
	Spatial + Parity (SLR)	Hard-Complex	7.000	1.889
	Spatial-only	Hard-Complex	3.000	1.243
	Target-only	Hard-Complex	6.000	1.743
	SFT	Hard-Complex	0.000	1.145
Pooling operators				
Qwen3.5-4B	Parity	Standard	6.000	1.766
	Quadrant	Standard	6.000	1.754
	Random	Standard	6.000	1.741
	Max	Standard	6.000	1.740