UNDERSTANDING WHAT YOU POINT AT: SPATIALLY GUIDED REFERENCE DISAMBIGUATION IN EGOCENTRIC VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

Pointing gestures link linguistic references to physical objects, providing essential cues for egocentric AI assistants to interpret user intent and interact naturally. However, current multimodal large language models (MLLMs) struggle to distinguish intended targets from distractors in spatially complex and dynamic scenes. Existing benchmarks provide limited systematic coverage of these ambiguities. We introduce PointConfusion, to our knowledge the first egocentric video understanding benchmark designed around a taxonomy of spatial and dynamic pointing ambiguities. Its 5,000 synthetic videos enable systematic evaluation of five ambiguity categories through a unified target-selection task. Our method, PointMod, learns spatial reference heatmaps from hand geometry and uses them to modulate visual features, explicitly incorporating pointing information into the model's visual representation. Extensive evaluations on PointConfusion demonstrate that PointMod outperforms a range of advanced MLLMs and achieves state-of-the-art performance. PointMod-8B reaches 88.33% target-selection accuracy, improving upon task-fine-tuned InternVL3.5-8B by 8.13 percentage points. The dataset and code will be publicly released.

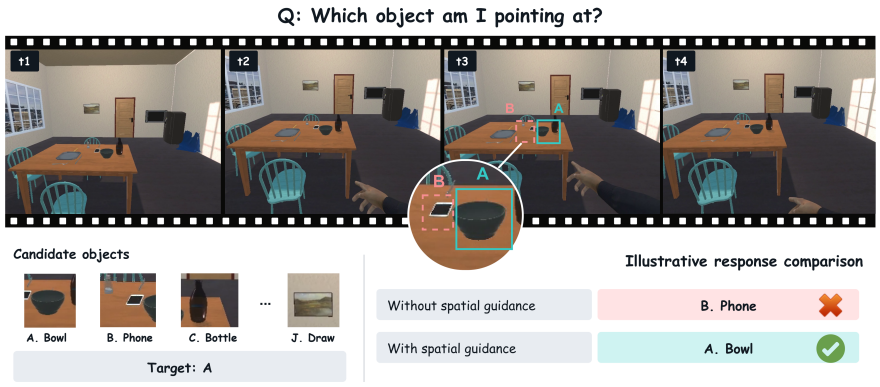

Figure 1: Pointing-based reference disambiguation in egocentric video. Similar pointing directions can correspond to different objects, making the intended target difficult to identify from direction alone. PointConfusion captures these challenging situations, while PointMod translates hand geometry into spatial guidance for target selection.

1 INTRODUCTION

Egocentric AI agents need to interpret user requests by integrating language, actions, and the surrounding scene (Yang et al., 2026). When a user points to an object on a table and asks "What is that?", the gesture identifies a referent left unspecified by the words. Recognizing the intended object is therefore a prerequisite for answering the question (Chen et al., 2021; Lee et al., 2024; Islam et al., 2024). Understanding pointing thus requires more than recognizing objects in an image;

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.

it requires identifying which object the user intends to refer to among multiple alternatives (Mane et al., 2025; Cao et al., 2025).

However, current multimodal large language models (MLLMs) still tend to confuse intended targets with distractors in complex scenes (Choi et al., 2026; Li et al., 2026). Overlapping objects and similar pointing directions introduce spatial ambiguity (Jirak et al., 2021; Eyiokur et al., 2026), while hand motion and viewpoint changes further complicate video understanding (Choi et al., 2026). Existing benchmarks have advanced gesture-grounded question answering (Islam et al., 2024; Choi et al., 2026; Li et al., 2026), but systematic evaluation of these ambiguities remains limited. We therefore aim to characterize how different pointing ambiguities affect target identification and explore how gesture cues can provide effective spatial guidance.

To this end, we introduce **PointConfusion**, an egocentric video understanding benchmark with 5,000 synthetic videos spanning five spatial and dynamic pointing ambiguities. Controlled target–distractor configurations and a unified target-selection task enable category-wise comparisons, revealing capability differences beyond aggregate accuracy.

How can models make better use of pointing cues? Returning to the opening example, a natural response is to follow the gesture and look for the intended object in the indicated region. The gesture suggests "where to look" (Langton & Bruce, 2000), while object appearance and linguistic context help determine "which object is meant" (Wu & Xie, 2024). Inspired by this process, we turn gesture cues into spatial prompts that guide models toward visual regions relevant to the pointing intent.

Building on this idea, we propose **PointMod** (Pointing-Guided Visual Modulation). A learned locator maps reconstructed hand geometry (Potamias et al., 2025) to spatial reference heatmaps that modulate visual features for target selection. PointMod preserves the full scene input without adding gesture tokens.

Our main contributions are as follows:

- We introduce PointConfusion, the first egocentric video understanding benchmark organized around a taxonomy of spatial and dynamic pointing ambiguities, systematically evaluating five ambiguity categories within a unified target-selection task.
- We propose PointMod, which maps hand geometry to learned spatial reference heatmaps and provides MLLMs with explicit spatial guidance through visual feature modulation, without introducing additional gesture tokens.
- We demonstrate state-of-the-art pointing target-selection performance on PointConfusion, outperforming advanced MLLMs and improving task-fine-tuned models across multiple backbones and scales.

2 Related Work

2.1 Egocentric Video Understanding

Egocentric datasets capture everyday activities from the wearer's perspective. EPIC-KITCHENS supports fine-grained action recognition and anticipation (Damen et al., 2022), while Ego4D covers diverse daily scenarios through episodic memory, hand–object interaction, and forecasting tasks (Grauman et al., 2022). Ego-Exo4D links skilled activity understanding to synchronized first- and third-person observations (Grauman et al., 2024). Beyond recognizing activities, VidEgoThink (Cheng et al., 2024) combines video question answering, hierarchical planning, visual grounding, and reward modeling to connect video understanding with embodied decision-making. These benchmarks emphasize the wearer's activities and environment; we focus on identifying intended referents under spatial and dynamic ambiguity.

2.2 Gesture-Based Reference Understanding

Pointing complements underspecified language by narrowing possible referents (Jirak et al., 2021; Lee et al., 2024), supporting reference understanding in images and videos (Chen et al., 2021) and 3D scenes (Mane et al., 2025). Representing the gesture is central to this process. CAPE (Eyiokur

et al., 2026) combines predictions from head-to-fingertip and wrist-to-fingertip ray heatmaps to accommodate variation in pointing geometry. Recent benchmarks extend gesture grounding to question answering (Choi et al., 2026; Li et al., 2026); HINT encodes 3D hand keypoints as frame-aligned tokens, allowing the language model to use gesture information alongside video. Whereas ray heatmaps encode directional priors and hand tokens expose joint geometry, we learn target-region heatmaps to modulate visual features.

2.3 Spatial Understanding in Multimodal Models

Vision–language pretraining and instruction tuning enable broad multimodal understanding (Radford et al., 2021; Li et al., 2023; Liu et al., 2023), while fine-grained reference resolution requires explicit region correspondence. Kosmos-2 (Peng et al., 2024) links text spans to image regions, incorporating localization into language generation. Set-of-Mark (Yang et al., 2023) instead overlays region identifiers, providing spatial references without retraining the model. Ferret (You et al., 2024) combines discrete coordinates with continuous regional features to retain both position and appearance across input shapes and granularities. These approaches encode spatial information through locations, visible marks, or region content; we apply spatial weights directly to existing visual tokens.

3 PointConfusion Dataset

PointConfusion isolates competing spatial and dynamic cues through controlled target–distractor configurations. Table 1 situates the benchmark among related reference-understanding datasets.

Table 1: Comparison with related reference-understanding datasets in scale, visual input, task, annotations, and evaluation structure. PointConfusion combines egocentric video with a unified target-selection task and five balanced spatial and dynamic ambiguity subsets. Ego. indicates egocentric visual input.

Dataset	Scale	Input	Ego.	Task	Key annotations	Evaluation organization
ReferAt (Schauerte & Fink, 2010)	242 acts	Video	✗	Referent identification	Referring expressions	Multimodal referring acts
YouRefIt (Chen et al., 2021)	4,195 clips	Image / video	✗	Reference grounding	Expressions; target boxes	Image- and video-level grounding
GestureTarget (Cao et al., 2025)	∼20K instances	Image	✗	Gesture target estimation	Boxes; masks; target existence	Multiple deictic gesture types
ImputeRefer (Mane et al., 2025)	35,581 samples	3D points	N/A	3D grounding	Expressions; 3D boxes; human poses	Gesture-assisted 3D reference
VRA-Ego (Yang et al., 2026)	1,000 samples	Image	✓	Clarification	Answers; corrective guidance	Visual and referential ambiguity
EgoPoint-Bench (Li et al., 2026)	11,729 QAs	Image	✓	Pointing QA	QA; target names and boxes	5 capability dimensions; 3 deixis levels
EgoPointVQA (Choi et al., 2026)	4,400 videos	Video	✓	Pointing QA	QA; target boxes	6 deictic reasoning tasks
EcoG-Bench (Zhou et al., 2026)	811 clips	Video	✓	Spatio-temporal grounding	Target labels; masks / points; stroke times	4 levels of speech–gesture binding
PointConfusion (Ours)	**5,000 videos; 15,000 QAs**	**Video**	✓	**Target selection**	**QA; hold-frame boxes; ambiguity labels**	**5 balanced spatial/dynamic ambiguity subsets**

3.1 Task Definition

Each sample comprises a pointing video, a natural-language question, and ten object-name options labeled A–J. Given the original video and text, without target boxes or crops, the model identifies the referred object and returns its option letter. Expressions such as "What is this?" require resolving the gesture rather than merely recognizing visible objects.

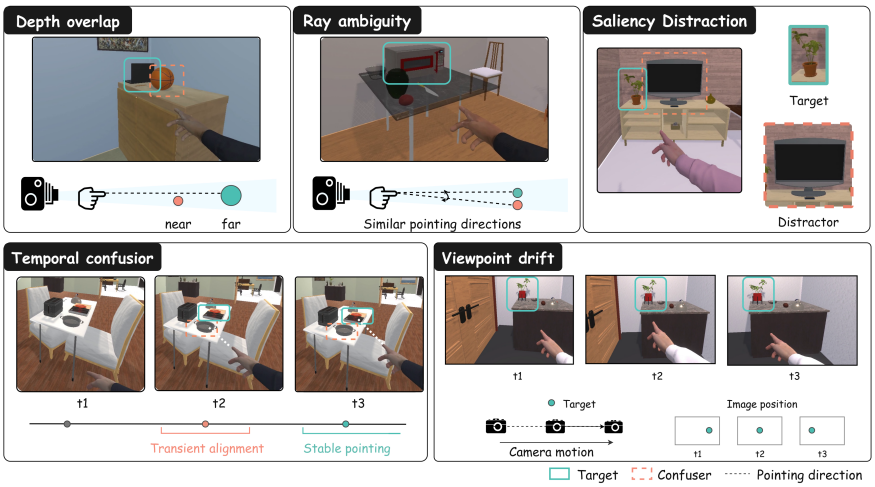

Figure 2: Five categories of pointing ambiguity in PointConfusion. Depth overlap, ray ambiguity, and saliency distraction introduce spatial competition; temporal confusion and viewpoint drift vary the pointing cues or observation over time.

3.2 Ambiguity Taxonomy

Our taxonomy draws on object-layout ambiguity (Jirak et al., 2021), geometric pointing cues (Yang et al., 2026), and the saliency, temporal, and viewpoint-related failures reported in HINT (Choi et al., 2026). Figure 2 illustrates the five conditions. Below, g, d denote the target and primary distractor, B a projected box, and τ a construction threshold; Appendix A provides the implementation details.

Depth overlap. Objects at different distances have overlapping projections. We constrain both projected overlap and depth separation, $\Omega(B_g, B_d) \geq \tau_{\text{overlap}}$ and $\Delta_{\text{depth}}(g, d) \geq \tau_{\text{depth}}$. The target can be nearer or farther, preventing proximity alone from identifying it.

Ray ambiguity. A distractor lies close to the viewpoint-to-target reference ray R_g, with $\text{dist}(p_d, R_g) \leq \tau_{\text{ray}}$. Similar directions can be ambiguous even without overlapping boxes. The rendered finger is subsequently aligned toward the designated target.

Saliency distraction. A distractor is more visually salient than the target, $S(d) - S(g) \geq \tau_{\text{saliency}}$. The construction score combines image area, position, and context, placing appearance-based prominence in conflict with pointing intent.

Temporal confusion. Target–distractor competition changes during pointing. We combine variation in relative ray distance and projected separation, $D_{\text{temp}} = D_{\text{ray}} + D_{\text{proj}}$, with path-proximity and visibility constraints. A locally plausible directional cue can therefore be misleading.

Viewpoint drift. The intended target stays fixed while the camera moves. We combine changes in projected position, overlap, and depth, $D_{\text{view}} = D_{\text{pos}} + D_{\text{overlap}} + D_{\text{depth}}$, testing referent consistency as the observation changes.

3.3 Dataset Construction

Target-constrained synthesis. We configure ProcTHOR-10K scenes (Deitke et al., 2022) in AI2-THOR (Kolve et al., 2017) and animate Mixamo characters in Unity (Adobe, n.d.). Given the proxi-

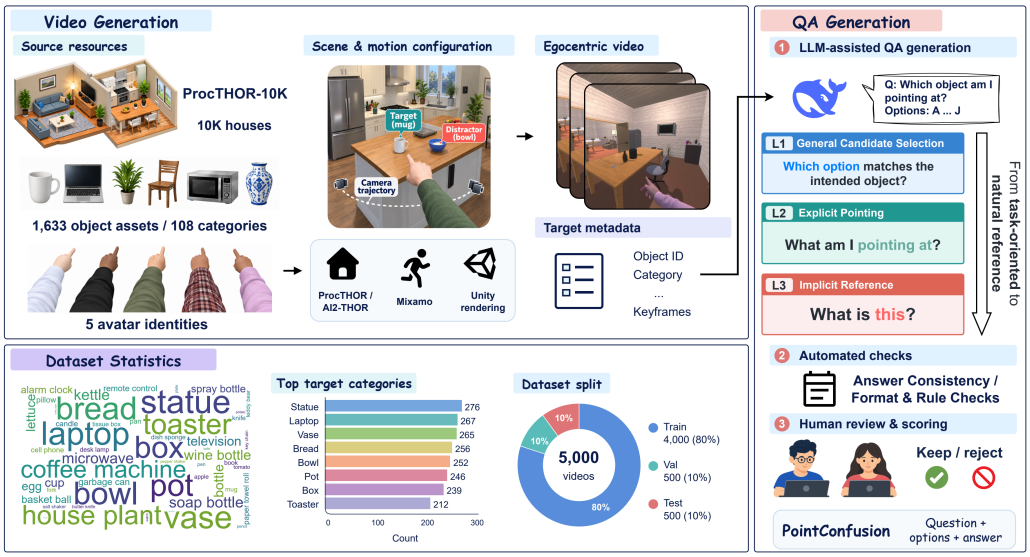

Figure 3: PointConfusion construction and statistics. Scene and motion synthesis produces videos with target metadata for QA generation and review. The lower-left panels show the target-category word cloud, category frequencies, and dataset split.

mal index-finger position p_h and target center p_g, the desired direction is

$$u_{\text{target}} = \frac{p_g - p_h}{\|p_g - p_h\|_2}. \tag{1}$$

We rotate the arm and wrist toward this direction, then combine the corrected pose with raising, holding, and lowering phases and camera motion. Target identifiers, spatial projections, and action phases accompany each video (Figure 3).

QA generation and validation. DeepSeek-V4-Flash (DeepSeek-AI, 2026) generates questions covering general referent selection, explicit pointing, and implicit reference. Scene metadata determines answers, and candidates are assembled programmatically. Automated checks validate geometry and QA records; human review verifies gesture discernibility and target–distractor correspondence. Questions exclude internal identifiers and ambiguity labels. Appendices A.4 and D detail generation, checks, and input templates.

The resulting dataset contains 5,000 synthetic videos and 15,000 QA pairs from 3,382 house layouts, with 52 target categories, five avatars, and 1,000 videos per ambiguity category. The video-level train/validation/test split is 4,000/500/500 (12,000/1,500/1,500 QA pairs); each video's three questions share its target, candidates, and split.

4 PointMod

PointMod combines a learned spatial prior with visual-token modulation for target selection (Figure 4).

4.1 Gesture-Derived Spatial Heatmaps

Pretrained WiLoR (Potamias et al., 2025) extracts 107-dimensional hand features per frame (Appendix B.1). The Locator processes them independently of questions and candidates.

A shared MLP encodes each frame's hand features as h_t, which feeds two prediction branches. The spatial branch broadcasts h_t over a 16×16 grid, concatenates the 2D coordinates at each location, and applies a convolutional head to predict spatial logits Z_t. The temporal branch assigns each frame a scalar score s_t through a linear head and normalizes scores across frames,

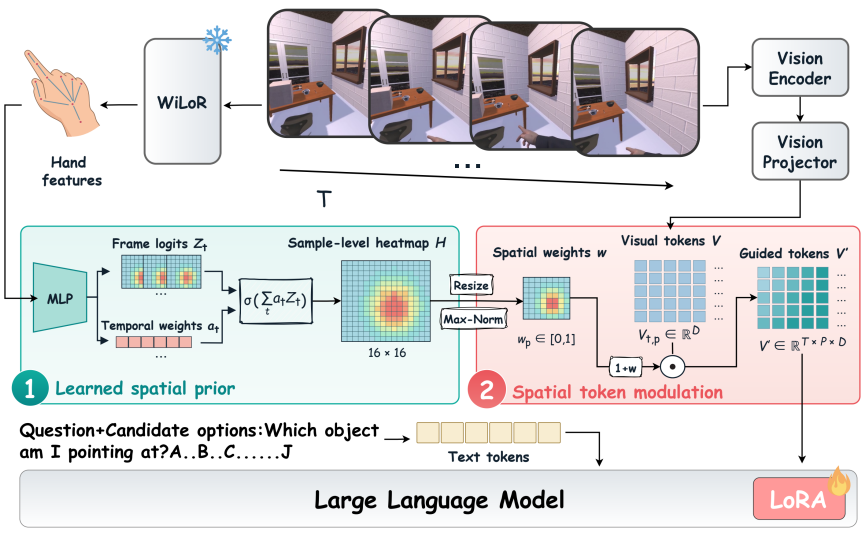

Figure 4: Overview of PointMod. (1) The Locator aggregates WiLoR hand features into a learned spatial prior. (2) Spatial weights modulate existing visual tokens, providing gesture guidance without additional tokens.

$a_t = \exp(s_t)/\sum_j \exp(s_j)$. These weights aggregate the frame-level spatial predictions into one heatmap:

$$H = \sigma\left(\sum_{t=1}^{T} a_t Z_t\right), \qquad \sum_{t=1}^{T} a_t = 1. \tag{2}$$

where σ is the sigmoid and H is shared across frames.

4.2 Learning the Spatial Prior

With WiLoR frozen, we train the Locator against a soft heatmap centered on and covering the target box in the middle pointing-hold frame:

$$\mathcal{L}_{\text{loc}} = \mathcal{L}_{\text{KL}} + \lambda_c \mathcal{L}_{\text{center}} + \lambda_s \mathcal{L}_{\text{SoftIoU}}. \tag{3}$$

KL matches spatial distributions, the center term constrains expected position, and Soft-IoU supervises regional overlap. Appendix B.2 gives the target construction and loss implementation.

4.3 Pointing-Guided Visual Modulation

With the trained Locator fixed, we use its heatmap to guide visual features. Following feature conditioning (Perez et al., 2018; Hu et al., 2018; Woo et al., 2018), we modulate projected feature $V_{t,p}$ at position p of frame t as

$$w = \text{MaxNorm}(\text{Resize}(H)), \qquad V'_{t,p} = (1 + w_p)V_{t,p}. \tag{4}$$

Resizing aligns the heatmap with the token grid; max-normalization gives $w_p \in [0, 1]$. Applied once after the projector and broadcast over channels, the weights amplify high-response regions without cropping the scene or adding tokens.

Downstream QA then trains only language-side LoRA (Hu et al., 2022), freezing WiLoR, the Locator, the vision encoder, projector, and base language weights. Autoregressive cross-entropy applies only to answer tokens; training configurations appear in Appendix B.3.

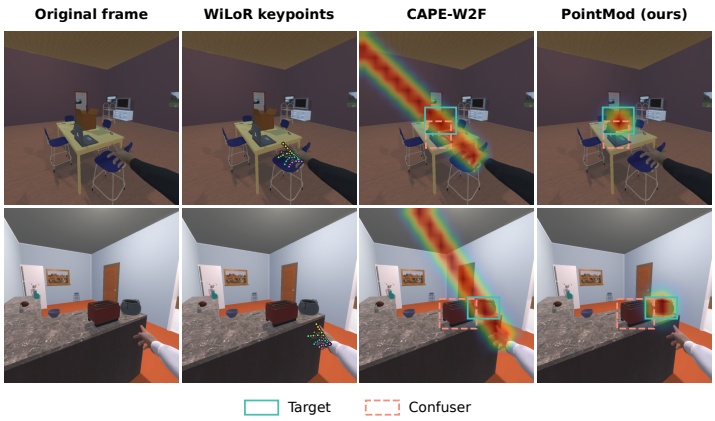

Figure 5: Comparison of gesture representations in two scenes. Columns show original frames, WiLoR keypoints, CAPE-W2F ray heatmaps, and PointMod heatmaps.

5 Experiments

We evaluate existing MLLMs against human responses, compare PointMod with task-specific fine-tuning across backbones and scales, and analyze gesture representations, localization quality, and heatmap dependence.

5.1 Experimental Setup

Protocol. Unless otherwise specified, we report overall and category-wise accuracy on the same 1,500 PointConfusion test questions. Models receive 16 uniformly sampled frames at source resolution 448 with native preprocessing; unparseable outputs count as incorrect. Humans answer the same questions directly from video.

Models. Our 13 variants include Qwen2.5-VL-3B/7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), InternVL3.5-2B/4B/8B (Wang et al., 2025), LLaVA-OneVision-1.5-4B/8B (An et al., 2025), Phi-3.5-Vision (Microsoft, 2024), GLM-4.1V-9B-Thinking (GLM-V Team, 2025), and MiniCPM-V-2.6 (Yao et al., 2024). We also evaluate SWIM-7B for fine-grained video object understanding (Sun et al., 2026) and EgoGPT-0.5B-Demo for egocentric understanding (Yang et al., 2025).

Comparisons. InternVL3.5-8B is our main backbone. RGB fine-tuning and PointMod share videos, questions, and downstream training budgets; PointMod additionally trains its Locator. Each cross-backbone comparison trains both variants. We also adapt HINT and CAPE to target-selection QA, with configurations in Appendix B.

5.2 Main Results

Existing model capabilities. Zero-shot accuracy ranges from 23.00% to 34.53%, leaving even the strongest model, LLaVA-OneVision-1.5-8B, 60.20 points behind humans (Table 2). Larger models are not consistently better, and neither SWIM nor EgoGPT surpasses the strongest general-purpose model. Human accuracy exceeds 92% in every category, highlighting a gap between general visual-language capabilities and fine-grained pointing disambiguation.

Benefits of spatial guidance. Task-specific fine-tuning raises InternVL3.5-8B from 24.93% to 80.20%. PointMod further reaches 88.33%, a gain of 8.13 points under the same downstream budget and the highest model accuracy in Table 2. It reduces the human–model gap from 14.53 to 6.40 points. Spatial guidance therefore adds value beyond exposure to pointing-related QA data.

Ambiguity-specific performance. PointMod improves all five categories, with the largest gains in ray ambiguity (+10.67 points) and viewpoint drift (+9.67). Saliency-distraction accuracy reaches 91.33%, compared with 83.00% for RGB fine-tuning and at most 29.67% zero-shot. Depth overlap

Table 2: Target-selection accuracy (%) on the PointConfusion test set. Human performance reveals a substantial gap in zero-shot model understanding. PointMod achieves the highest model accuracy and improves over RGB fine-tuning in all five ambiguity categories. Bold denotes the best model result within each evaluation setting, and green indicates gains over RGB fine-tuning.

Model	Depth	Ray	Saliency	Temporal	Viewpoint	Overall
Zero-shot						
Qwen2.5-VL-3B	33.33	28.00	**29.67**	33.00	37.00	32.20
Qwen2.5-VL-7B	33.00	31.33	22.67	35.33	**40.33**	32.53
Qwen3-VL-8B	30.33	24.67	22.00	32.67	39.00	29.73
InternVL3.5-2B	31.33	26.33	24.00	26.33	34.00	28.40
InternVL3.5-4B	28.33	31.00	22.67	29.33	34.00	29.07
InternVL3.5-8B	27.00	24.00	18.67	22.33	32.67	24.93
LLaVA-OneVision-1.5-4B	34.00	**35.67**	20.00	42.33	35.00	33.40
LLaVA-OneVision-1.5-8B	**37.33**	30.67	22.33	**44.67**	37.67	**34.53**
Phi-3.5-Vision	26.33	19.33	21.33	25.00	23.00	23.00
GLM-4.1V-9B-Thinking	27.00	25.33	22.67	35.33	33.67	28.80
MiniCPM-V-2.6	32.67	32.00	22.33	32.67	35.33	31.00
SWIM-7B	23.67	25.67	19.67	27.33	28.33	24.93
EgoGPT-0.5B-Demo	17.67	23.00	23.67	31.67	19.00	23.00
Human	95.67	93.00	97.67	92.67	94.67	94.73
Task-fine-tuned						
InternVL3.5-8B	77.00	79.00	83.00	81.33	80.67	80.20
PointMod-8B	**83.00**	**89.67**	**91.33**	**87.33**	**90.33**	**88.33**
	+6.00	+10.67	+8.33	+6.00	+9.67	+8.13

remains hardest at 83.00%, indicating persistent difficulty separating objects with similar projections.

Table 3: Transfer of PointMod across model backbones and sizes. Compared with RGB fine-tuning, PointMod improves target-selection accuracy across the evaluated configurations, with gains varying by backbone. Green indicates absolute gains in percentage points.

Backbone	RGB (%)	PointMod (%)	Gain (pp)
InternVL3.5-2B	77.60	**87.40**	+9.80
InternVL3.5-4B	80.93	**87.53**	+6.60
InternVL3.5-8B	80.20	**88.33**	+8.13
Qwen2.5-VL-3B	74.73	**74.87**	+0.13
Qwen3-VL-8B	77.93	**84.87**	+6.93
LLaVA-OneVision-1.5-8B	80.13	**83.20**	+3.07

5.3 Results across Backbones and Model Sizes

Model sizes. PointMod improves InternVL3.5-2B/4B/8B by 9.80/6.60/8.13 points (Table 3). The 2B and 4B RGB baselines differ by 3.33 points, whereas their PointMod counterparts reach 87.40% and 87.53%. PointMod improves pointing understanding without scaling up the base model.

Backbones. PointMod gains 6.93 points on Qwen3-VL-8B and 3.07 on LLaVA-OneVision-1.5-8B, extending its benefits beyond InternVL (Table 3). Qwen2.5-VL-3B improves only marginally, showing that effectiveness depends on the backbone rather than model size alone.

5.4 Ablation Studies and Method Analysis

Gesture modeling. On InternVL3.5-8B, PointMod reaches 88.33%, compared with 77.00% for HINT keypoint tokens and 79.93% for CAPE ray heatmaps (Table 4). The learned spatial inter-

face gives the strongest target-selection performance among these implementations. Appendix B.4 describes the baseline adaptations.

Table 4: Comparison of gesture-guidance methods on InternVL3.5-8B. HINT* encodes hand keypoints as additional tokens, CAPE* provides fixed-ray heatmaps, and PointMod learns spatial heatmaps for visual token modulation. Removed and Shuffled evaluate guidance removal and shuffling, respectively. * denotes task-adapted implementations.

Method	Normal (%)	Removed (%)	Shuffled (%)
CAPE*	79.93	58.00	46.47
HINT*	77.00	77.33	77.13
PointMod (8B)	**88.33**	33.47	26.73

Spatial localization. Independent of QA, learned heatmaps outperform fixed rays in peak hit, target–confuser discrimination, and in-box response mass (Table 5). The largest contrast is in peak hit, 80.60% versus 12.80%, while in-box mass rises from 7.95% to 30.78%. Thus, following a pointing direction alone is less effective than concentrating responses on the intended region. Appendix B.5 defines the metrics; Appendices C.1 and C.2 analyze Locator inputs and supervision.

Table 5: Spatial localization of learned heatmaps versus fixed wrist-to-fingertip (W2F) rays. Learned heatmaps outperform fixed rays across all three metrics.

Spatial representation	Peak hit (%)	Discrimination (%)	In-box mass (%)
Fixed W2F ray	12.80	69.26	7.95
Learned heatmap	**80.60**	**76.43**	**30.78**

Qualitative comparison. In the two scenes in Figure 5, fixed-ray responses extend over competing objects, whereas learned heatmaps concentrate more strongly on the target. Residual spread beyond target boundaries illustrates the remaining localization errors.

Removing spatial guidance. With model weights fixed, inference-time heatmap removal or shuffling reduces PointMod-8B to 33.47% or 26.73%, with similarly large drops for 2B and 4B (Table 6). Shuffling disrupts heatmap–video correspondence, revealing reliance on sample-specific guidance. CAPE also degrades, whereas HINT remains near 77% (Table 4); gesture utilization varies substantially with the representation interface.

Table 6: Effect of removing or shuffling heatmaps on PointMod across model sizes.

Model	Normal (%)	Removed (%)	Shuffled (%)
PointMod-8B	**88.33**	33.47	26.73
PointMod-4B	**87.53**	14.53	24.67
PointMod-2B	**87.40**	9.40	19.47

6 Conclusion

We introduced PointConfusion, an egocentric video benchmark that systematically evaluates five spatial and dynamic pointing ambiguities through a unified target-selection task. We further proposed PointMod, which learns spatial reference heatmaps from hand geometry and uses them to guide visual features without adding gesture tokens. PointMod achieves state-of-the-art performance on the benchmark and improves multiple backbones and model sizes while preserving general video question-answering capabilities. These results highlight the value of converting gesture geometry into explicit spatial guidance for resolving references. Further improving robustness in real-world pointing scenarios is a promising direction for egocentric assistants.

AI USE STATEMENT

We used LLMs and other AI tools to assist with manuscript drafting, translation, and language polishing; literature retrieval and discovery; and research ideation and execution, including experiment planning, code development, and result analysis. For dataset construction, LLMs assisted with question-pool generation, while ground-truth answers were determined by scene metadata. AI-generated visual material was also used in figure preparation. The authors take responsibility for the final manuscript, references, code, figures, and reported results.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A describe dataset construction and splits. Section 4 defines the Locator and modulation interface. Appendix B records implementation details and evaluation protocols, and Appendix C presents supplementary results. We plan to release the dataset and code.

REFERENCES

Adobe. Mixamo, n.d. URL https://www.mixamo.com/. Character and animation platform. Accessed September 26, 2026.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. LLaVA-OneVision-1.5: Fully Open Framework for Democratized Multimodal Training, 2025. URL https://arxiv.org/abs/2509.23661.

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.

Xu Cao, Pranav Virupaksha, Sangmin Lee, Bolin Lai, Wenqi Jia, Jintai Chen, and James Rehg. Toward Human Deictic Gesture Target Estimation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 172399–172424. Curran Associates, Inc., 2025. doi: 10.52202/085713-5736. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/fb4a7e3522363907b26a86cc5be627ac-Paper-Conference.pdf.

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 (ICCV)*, pp. 1385–1395, October 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Chen_YouRefIt_Embodied_Reference_Understanding_With_Language_and_Gesture_ICCV_2021_paper.html.

Sijie Cheng, Kechen Fang, Yangyang Yu, Sicheng Zhou, Bohao Li, Ye Tian, Tingguang Li, Lei Han, and Yang Liu. VidEgoThink: Assessing Egocentric Video Understanding Capabilities for Embodied AI, 2024. URL https://arxiv.org/abs/2410.11623.

Yura Choi, Roy Miles, Rolandos Alexandros Potamias, Ismail Elezi, Jiankang Deng, and Stefanos Zafeiriou. Do You See What I Am Pointing At? Gesture-Based Egocentric Video Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18534–18544, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Choi_Do_You_See_What_I_Am_Pointing_At_Gesture-Based_Egocentric_CVPR_2026_paper.html.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision*, 130(1):33–55, 2022. doi: 10.1007/s11263-021-01531-2. URL https://doi.org/10.1007/s11263-021-01531-2.

DeepSeek-AI. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-Scale Embodied AI Using Procedural Generation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 5982–5994. Curran Associates, Inc., 2022. doi: 10.52202/068431-0433. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/27c546ab1e4f1d7d638e6a8dfbad9a07-Paper-Conference.pdf.

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 (WACV)*, pp. 3939–3950, March 2026. URL https://openaccess.thecvf.com/content/WACV2026/html/Eyiokur_CAPE_A_CLIP-Aware_Pointing_Ensemble_of_Complementary_Heatmap_Cues_for_WACV_2026_paper.html.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

GLM-V Team. GLM-4.5V and GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning, 2025. URL https://arxiv.org/abs/2507.01006.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18995–19012, June 2022. URL https://openaccess.

thecvf.com/content/CVPR2022/html/Grauman_Ego4D_Around_the_World_in_3000_Hours_of_Egocentric_Video_CVPR_2022_paper.html.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C.V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19383–19400, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Grauman_Ego-Exo4D_Understanding_Skilled_Human_Activity_from_First-_and_Third-Person_Perspectives_CVPR_2024_paper.html.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html.

Md Mofijul Islam, Alexi Gladstone, Riashat Islam, and Tariq Iqbal. EQA-MX: Embodied Question Answering using Multimodal Expression. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 48079–48102, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/d2bdcd4f51eea138365af22b50f3bf0a-Paper-Conference.pdf.

Doreen Jirak, David Biertimpel, Matthias Kerzel, and Stefan Wermter. Solving visual object ambiguities when pointing: an unsupervised learning approach. *Neural Computing and Applications*, 33 (7):2297–2319, 2021. doi: 10.1007/s00521-020-05109-w. URL https://link.springer.com/article/10.1007/s00521-020-05109-w.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Aniruddha Kembhavi, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI, 2017. URL https://arxiv.org/abs/1712.05474.

Stephen R. H. Langton and Vicki Bruce. You must see the point: Automatic processing of cues to the direction of social attention. *Journal of Experimental Psychology: Human Perception and Performance*, 26(2):747–757, 2000. doi: 10.1037/0096-1523.26.2.747. URL https://doi.org/10.1037/0096-1523.26.2.747.

Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S. Rodriguez, and Jon E. Froehlich. GazePointAR: A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in

Wearable Augmented Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–20. ACM, 2024. doi: 10.1145/3613904.3642230. URL https://doi.org/10.1145/3613904.3642230.

Chentao Li, Zirui Gao, Mingze Gao, Yinglian Ren, Jianjiang Feng, and Jie Zhou. Do MLLMs Understand Pointing? Benchmarking and Enhancing Referential Reasoning in Egocentric Vision. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 17000–17019, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.838. URL https://aclanthology.org/2026.findings-acl.838/.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Atharv Mahesh Mane, Dulanga Weerakoon, Vigneshwaran Subbaraju, Sougata Sen, Sanjay E. Sarma, and Archan Misra. Ges3ViG : Incorporating Pointing Gestures into Language-Based 3D Visual Grounding for Embodied Reference Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9017–9026, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Mane_Ges3ViG__Incorporating_Pointing_Gestures_into_Language-Based_3D_Visual_Grounding_CVPR_2025_paper.html.

Microsoft. Phi-3.5-vision-instruct, 2024. URL https://huggingface.co/microsoft/Phi-3.5-vision-instruct. Official model card.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding Multimodal Large Language Models to the World. In *International Conference on Learning Representations*, pp. 51575–51598, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e112a4671e8779aa9f640a0e3f81bd26-Abstract-Conference.html.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. doi: 10.1609/aaai.v32i1.11671. URL https://ojs.aaai.org/index.php/AAAI/article/view/11671.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3D Hand Localization and Reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12242–12254, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Potamias_WiLoR_End-to-end_3D_Hand_Localization_and_Reconstruction_in-the-wild_CVPR_2025_paper.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Boris Schauerte and Gernot A. Fink. Focusing computational visual attention in multi-modal human-robot interaction. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pp. 1–8. ACM, 2010. doi: 10.1145/1891903.1891912. URL https://doi.org/10.1145/1891903.1891912.

Boyuan Sun, Bo-Wen Yin, Yuan-Ming Li, Xihan Wei, and Qibin Hou. See What I Mean: Aligning Vision and Language Representations for Video Fine-grained Object Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36814–36827, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Sun_See_What_I_Mean_Aligning_Vision_and_Language_Representations_for_CVPR_2026_paper.html.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Sanghyun_Woo_Convolutional_Block_Attention_ECCV_2018_paper.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 (CVPR)*, pp. 13084–13094, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_V_Guided_Visual_Search_as_a_Core_Mechanism_in_Multimodal_CVPR_2024_paper.html.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V, 2023. URL https://arxiv.org/abs/2310.11441.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Bo Li, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, and Ziwei Liu. EgoLife: Towards Egocentric Life Assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28885–28900, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_EgoLife_Towards_Egocentric_Life_Assistant_CVPR_2025_paper.html.

Sicheng Yang, Yukai Huang, Weitong Cai, Shitong Sun, You He, Jiankang Deng, Hang Zhang, Jifei Song, and Zhensong Zhang. Plug-and-Play Clarifier: A Zero-Shot Multimodal Framework for Egocentric Intent Disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(21):17921–17929, 2026. doi: 10.1609/aaai.v40i21.38851. URL https://ojs.aaai.org/index.php/AAAI/article/view/38851.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM-V: A GPT-4V Level MLLM on Your Phone, 2024. URL https://arxiv.org/abs/2408.01800.

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*, pp. 57153–57180, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/fd6613131889a4b656206c50a8bd7790-Abstract-Conference.html.

Weijie Zhou, Xuantang Xiong, Zhenlin Hu, Xiaomeng Zhu, Chaoyang Zhao, Honghui Dong, Zhengyou Zhang, Ming Tang, and Jinqiao Wang. Listening with the Eyes: Benchmarking Egocentric Co-Speech Grounding across Space and Time, 2026. URL https://arxiv.org/abs/2603.07966.

This supplementary material provides dataset construction constraints, implementation details, additional experiments, and prompt examples supporting the main text.

A Dataset Construction and Annotation

A.1 Scene Resources and Data Organization

The ProcTHOR-10K resource library provides 10,000 house layouts and 1,633 object assets spanning 108 asset categories. Figure S1 illustrates scene and avatar diversity; Figure S2 shows target sizes, answer positions, and category frequencies across all videos.



Figure S1: Scene and avatar appearance diversity. The upper panel shows room layouts and object arrangements from eight different scenes; the lower panel shows hand and sleeve appearances for five avatar identities. All images correspond to final dataset samples, and hand images are cropped from their source frames.

A.2 Construction of Ambiguity Conditions

We implement the taxonomy in Section 3 through the following geometric selectors and scores. Let g and d denote the target and primary distractor, B a projected box, and p a 3D center.

Depth overlap. To account for objects of different sizes, we define projected overlap as the maximum of box IoU and relative coverage along both image axes. Axis-wise coverage is normalized by the smaller box width or height, avoiding missed cases where a large object covers a smaller object but their IoU remains low.

$$\Omega(B_g, B_d) = \max\left\{\mathrm{IoU}(B_g, B_d), \frac{I_x}{\min(w_g, w_d)}\frac{I_y}{\min(h_g, h_d)}\right\}. \tag{A.1}$$

Here, I_x and I_y are the horizontal and vertical intersection lengths, with box dimensions lower-bounded by one pixel. The retention selector uses an overlap threshold of 0.10 and a depth-difference threshold of 0.25 meters, together with changes in object relations over the observation sequence and visibility requirements.

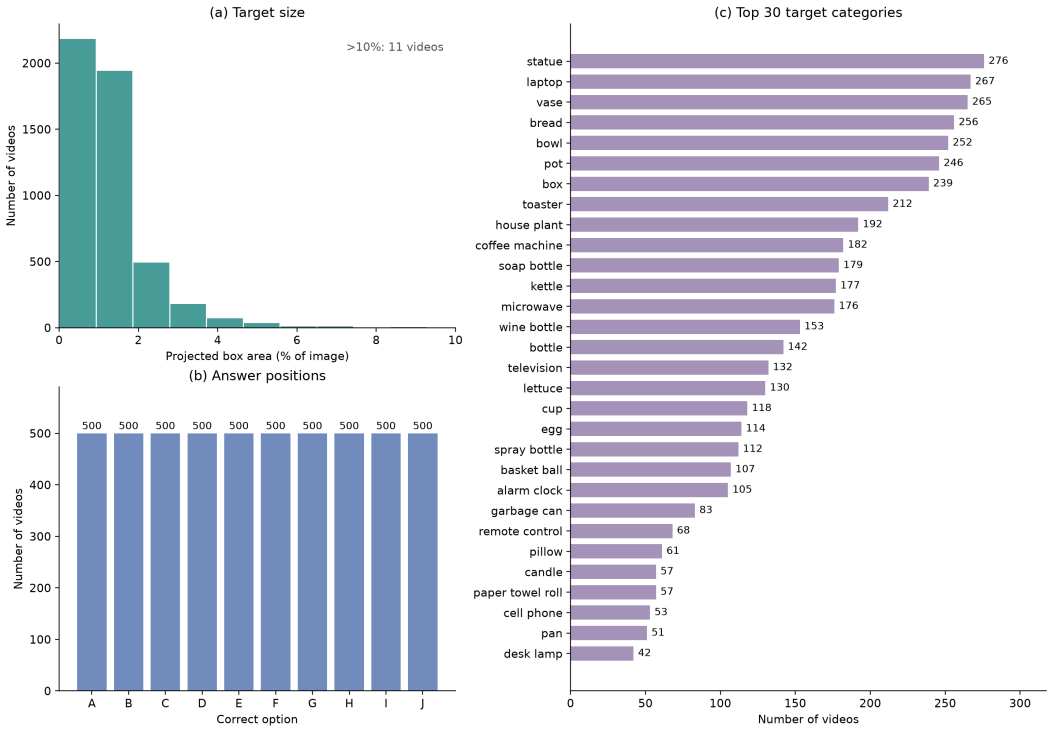

Figure S2: Target sizes, answer positions, and the top 30 target categories across all 5,000 videos. Target size is the projected target-box area relative to image area at the midpoint of the pointing-hold phase. Categories are counted by the correct target, once per video. The size plot displays the 0–10% range; another 11 videos have target-area ratios above 10%.

Ray ambiguity. For ray origin o, unit direction v, and object center p, directional proximity is measured by the distance to the forward half-ray:

$$d_{\text{ray}}(p) = \left\| p - o - \max\{0, (p-o)^\top v\} v \right\|_2 . \tag{A.2}$$

Candidate selection uses the viewpoint-to-target reference axis and a distractor-to-ray distance threshold of 0.22 meters, together with projected separation and depth difference. Finger alignment is handled by the pose-correction procedure below.

Saliency distraction. We construct a heuristic saliency score from projected area, centrality, local contrast, spatial isolation, and category priors, creating competition between a visually prominent object and the actual pointing target:

$$S(o) = 2.6 \min\{1, A(o)/0.08\} + C(o) + 1.1L(o) + 0.9D(o) + 0.35P(o). \tag{A.3}$$

A is the box-to-image area ratio, while C, L, D, and P denote center bias, local contrast, spatial isolation, and category prior, respectively. Components are normalized or clipped. During cached candidate filtering without RGB images, $L = 0$. Candidates satisfy $S(d) - S(g) \geq 0.30$, and the area term saturates at 0.08.

Temporal confusion. Let r_t be the difference between distractor and target distances to the current reference ray, and c_t their projected-center separation. The construction score combines the start-to-end change in relative ray distance with the range of projected separations:

$$S_{\text{temp}} = \min\left\{1.5, \frac{|r_{\text{end}} - r_{\text{start}}|}{0.14}\right\} + \min\left\{1.5, \frac{\max_t c_t - \min_t c_t}{90}\right\}. \tag{A.4}$$

Ray distances are measured in meters, while projected distances use pixel coordinates in candidate-generation images. Filtering also checks distractor proximity to the motion path, final pointing alignment, and keyframe visibility. This creates competition in directional cues at individual moments while preserving the designated target identity.

Viewpoint drift. Target identity remains fixed while camera motion changes the relative projections of target and distractor. We combine changes in horizontal and vertical projected separation, box overlap, and distractor depth:

$$S_{\text{view}} = \frac{|\Delta\delta_x| + |\Delta\delta_y|}{90} + 2|\Delta J| + \frac{|\Delta z_d|}{1.0}. \tag{A.5}$$

Δ denotes the last-frame minus first-frame difference, and J is box IoU. Image displacements and depths are measured in pixels and meters, respectively. Camera displacement, keyframe visibility, and target screen position jointly determine candidate retention.

A.3 Pointing Poses and Video Generation

The current pointing direction is $(e - a)/\|e - a\|_2$, where a and e are the proximal and distal index-finger bone positions. We compute its rotation toward the target direction in Equation 1 and distribute the increment across the upper arm, forearm, and wrist.

The canonical pose undergoes six correction iterations, with upper-arm, forearm, and wrist weights of 0.42, 0.30, and 0.08. Local rotations relative to the original animation are capped at 60°, 55°, and 18°, respectively. Each iteration applies the corresponding fraction of the rotation increment through spherical linear interpolation; alignment is recorded as the dot product of the current and desired directions. The arm performs most of the correction, with smaller adjustments at the wrist.

Figure S3 presents sampled motion sequences and their QA records for all five conditions.

A.4 Question Generation and Quality Control

The question pool covers general referent selection, explicit pointing, and implicit reference. For example, "Which object is being referred to?" emphasizes task-level target selection, "Which object am I pointing at?" explicitly mentions the gesture, and "What is that?" omits the target name and relies on the visual context. These formulations vary the explicitness of pointing information in language without changing the video's target.

Candidate options preferentially use distinct object names from the scene, supplemented from a predefined name bank when necessary. Answer positions are balanced, and question repetition is controlled across samples and ambiguity categories. The generation constraints and final input format are provided in Appendix D.

Automated checks cover target projection, hand framing, ambiguity conditions, candidate count, name uniqueness, and answer validity. Human review additionally checks target identifiability. Figure S4 illustrates the correspondence between video, scene metadata, and QA records.

B Model Implementation and Evaluation Protocols

B.1 Locator Architecture

The 107-dimensional WiLoR input comprises 21 2D keypoints, 21 3D joint coordinates, detection confidence, and a validity flag. Standardization statistics are estimated from valid training frames; missing-detection features are zeroed after standardization.

The hidden dimension is 96. Concatenating two grid coordinates produces a 98-channel input to the convolutional head. Table B1 specifies the layers implementing the frame encoding, scoring, and spatial prediction described in Section 4.

The scoring branch applies a common temporal softmax, including frames whose standardized features have been zeroed.

B.2 Heatmap Supervision and Training Objectives

For the target heatmap in Equation 3, each cell of the 16×16 grid spans 28 pixels in a 448×448 image. We center a Gaussian on the target box and take its pointwise maximum with 0.7 times the

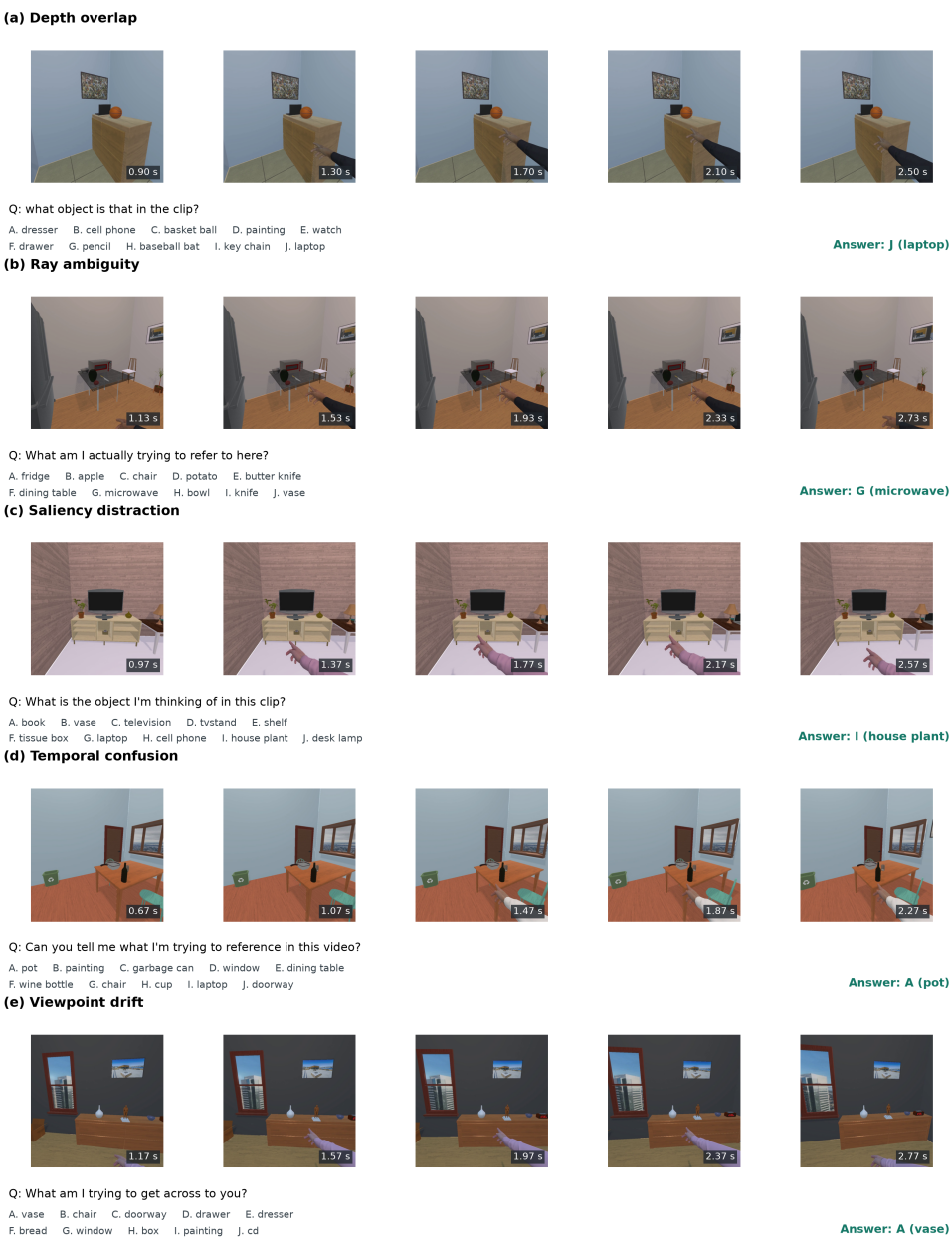

Figure S3: Video question-answering examples for the five ambiguity categories. Each group shows five sampled moments from one video, the actual question, ten candidates, and the reference answer.

in-box mask. The Gaussian is peak-normalized, with scale

$$\sigma = \max\left\{0.85, \min\left\{2.2, \frac{0.5\max(w_{\text{box}}, h_{\text{box}})}{28}\right\}\right\}. \tag{B.1}$$

Here w_{box} and h_{box} are the target-box width and height in pixels. In Equation 3, $\lambda_c = 0.20$ and $\lambda_s = 0.50$. KL uses a spatial softmax over logits; the center term applies Smooth L1 to the resulting expected coordinates. Soft-IoU separately max-normalizes the predicted sigmoid heatmap and target, using sums of pointwise minima and maxima as intersection and union.

We train the Locator on cached hand features from 4,000 training videos, using 16 uniformly sampled frames per video and one target heatmap per video. All Locator parameters are optimized with AdamW for 60 epochs, with a batch size of 128, a constant learning rate of 10^{-3}, weight decay of

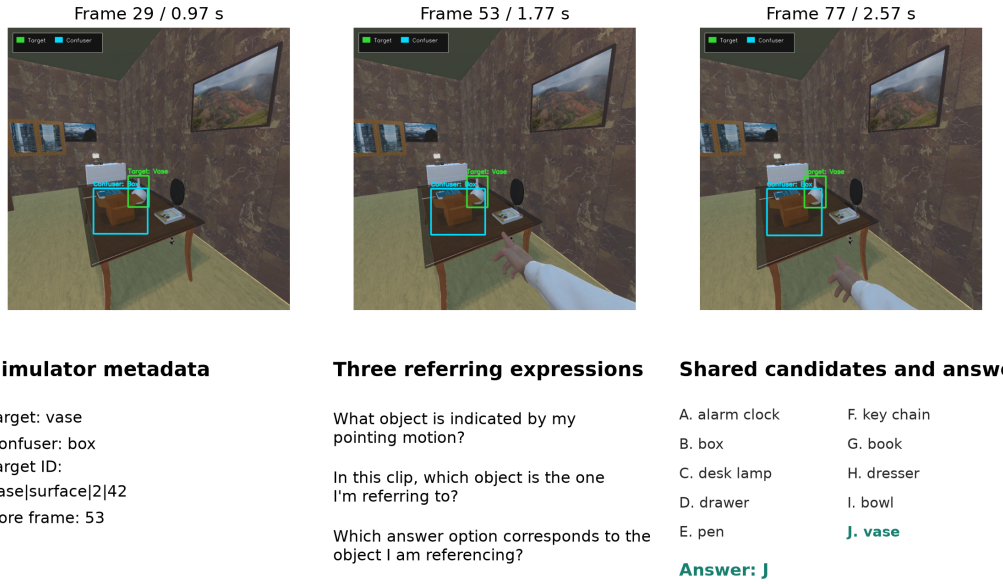

Figure S4: An example mapping video and scene metadata to QA records. Three question formulations share the same candidates and reference answer.

Table B1: Locator architecture.

Component	Architecture	Output
Frame encoder	Linear(107,96) -> LayerNorm -> GELU -> Linear(96,96) -> GELU	96 dimensions per frame
Frame scoring	Linear(96,1), softmax over frames	One weight per frame
Spatial input	Broadcast hidden vector concatenated with 2D coordinates	98x16x16
Spatial head	Conv1x1(98,96) -> GELU -> Conv3x3(96,96) -> GELU -> Conv3x3(96,48) -> GELU -> Conv1x1(48,1)	16x16 logits per frame
Temporal aggregation	Weighted sum of logits, then sigmoid	Sample-level 16x16 heatmap

10^{-4}, and gradient-norm clipping at 5.0. Training samples are shuffled each epoch. WiLoR remains frozen. We evaluate on 500 validation videos after the first epoch and every ten epochs, using an evaluation batch size of 256. Within this training run, the checkpoint is selected by validation Soft-IoU, with correlation and peak hit used to resolve ties; the retained checkpoint is from epoch 60. The trained Locator is then frozen for downstream QA adaptation.

B.3 Spatial Modulation and Downstream Training

Equation 4 uses bilinear interpolation, nonnegative clipping, and a max-normalization denominator lower-bounded by 10^{-8}. Figure S5 visualizes the shared spatial weights on different frames.

Table B2 lists the downstream configurations for PointMod, RGB fine-tuning, and the gesture baselines. The LoRA-only optimization described in Section 4 applies to PointMod; HINT also trains its hand-token adapter. Each backbone uses its native input processor. Experiments use NVIDIA V100 32GB GPUs. For the main PointMod model and its matched RGB baseline, we use AdamW with one question–answer example per optimizer update. The learning rate remains constant throughout training, without warmup or decay. Model weights and image inputs are loaded in FP16, and the model is partitioned across GPUs rather than replicated for data-parallel training.

Table B2: Training configurations of the main model and comparison methods on InternVL3.5-8B.

Setting	RGB SFT	PointMod	HINT	CAPE
Input frames	16	16	16	16
Input resolution	448	448	448	448
LoRA rank	64	64	63	64
LoRA alpha	128	128	126	128
LoRA dropout	0.05	0.05	0.05	0.05
Learning rate	1e-05	1e-05	1e-05	1e-05
Training steps	12000	12000	12000	12000
Gradient clipping norm	1.0	1.0	1.0	1.0
Weight decay	0.0	0.0	0.0	0.0
Language-side LoRA layers	q_proj, v_proj, o_proj	q_proj, v_proj, o_proj	q_proj, v_proj, o_proj	q_proj, v_proj, o_proj
Gesture interface	None	Learned heatmap	Additional tokens	Ray heatmap

B.4 Gesture Baseline Implementations

The adapted HINT baseline encodes WiLoR 3D hand keypoints through a trainable adapter into additional tokens, which are fed to the language model alongside visual and textual representations. The adapted CAPE baseline constructs a Gaussian ray heatmap along the forward wrist-to-index-fingertip direction, with a width parameter of 25 pixels. This heatmap is resized into a spatial field for visual feature modulation.

To accommodate HINT's trainable hand-token adapter with hidden dimension 185, we slightly reduce its LoRA rank while preserving the alpha-to-rank ratio. This gives a nearly identical downstream trainable-parameter budget to RGB fine-tuning, CAPE, and PointMod. Shared training settings and method-specific configurations are listed in Table B2; PointMod's separate Locator training is described in Section 4.2.

CAPE and PointMod use the same heatmap resizing, max-normalization, and boost modulation in Equation 4, applied after the visual projector. They differ in the source of the spatial guidance: CAPE uses fixed wrist-to-fingertip ray heatmaps, whereas PointMod uses learned heatmaps. Both employ the language model for candidate target selection.

For the guidance interventions in Tables 4 and 6, Removed disables heatmap modulation for PointMod and CAPE, leaving the visual tokens unchanged; for HINT, it omits the additional hand tokens. Shuffled replaces the heatmap or hand features with those from another video while retaining the current video, question, and candidate options. All interventions are applied at inference with the trained model parameters fixed.

B.5 Evaluation Metrics and Data Splits

Spatial-field and Locator analyses use 500 validation videos. Peak hit measures whether the maximum heatmap response falls inside the target box. Soft-IoU and correlation assess agreement with the reference heatmap.

Target–confuser discrimination is the fraction of samples for which the total heatmap response inside the target box exceeds that inside the primary-confuser box. It is evaluated on 488 videos with an eligible confuser; the other spatial metrics use all 500 videos. In-box mass is the response summed within the target box divided by the total heatmap response.

Target-selection accuracy follows the test-set and parsing protocol in Section 5.1.

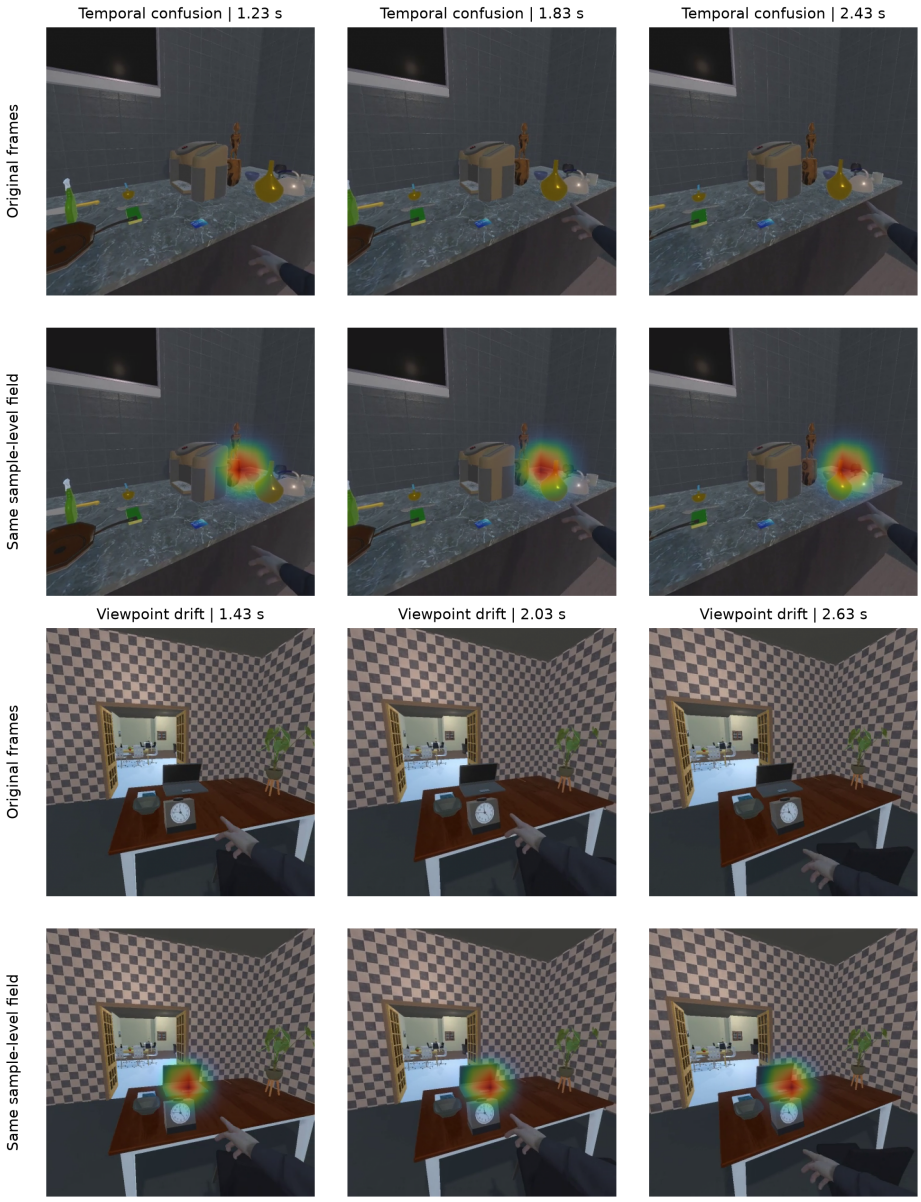

Figure S5: Sharing a sample-level spatial field across frames. Each group shows three moments from the same source video, with the same heatmap overlaid at fixed image coordinates.

C Supplementary Experiments and Analyses

C.1 Locator Input Representations

We compare direction and depth features, raw hand keypoints, and raw keypoints augmented with depth statistics, using identical center and Soft-IoU weights. Table C1 shows that the 107-dimensional raw hand representation improves peak hit by 3.80 percentage points over the 17-dimensional input and also increases regional overlap. Adding five depth statistics changes Soft-IoU by only 0.0003 and decreases peak hit by 0.20 percentage points. We therefore use the 107-dimensional hand geometry representation as Locator input.

Table C1: Locator input representations on 500 validation videos.

Input	Dim.	Peak hit (%)	Soft-IoU
Direction and depth	17	76.80	0.6110
Raw hand keypoints	107	80.60	0.6253
Raw keypoints + depth statistics	112	80.40	0.6256

C.2 Effect of Regional Supervision

During preliminary loss-function screening, we hold the 17-dimensional direction-and-depth representation fixed and independently add Dice, BBox BCE, and Soft-IoU constraints to the base KL and center objective. This experiment isolates the effect of regional supervision; the final Locator uses the 107-dimensional hand representation described in Appendix B.1 and compared in Appendix C.1. The base center weight is 0.10; the three independent additions use weights of 0.20, 0.05, and 0.20, respectively.

Table C2: Heatmap quality with independently added regional objectives on 500 validation videos.

Training objective	Peak hit (%)	Soft-IoU	Correlation
KL + Center	74.80	0.2291	0.6535
Base + Dice	74.40	0.4667	0.8616
Base + BBox BCE	75.20	0.5800	0.9025
Base + Soft-IoU	74.40	0.5927	0.9039

Regional constraints primarily change the overall heatmap distribution rather than whether its peak falls inside the target. Adding Soft-IoU raises overlap from 0.2291 to 0.5927 and correlation from 0.6535 to 0.9039, without a corresponding increase in peak hit. The maximum-response location alone is therefore insufficient to characterize a spatial field used for visual modulation: peak localization and response coverage capture different aspects of its quality.

C.3 Textual Shortcut Analysis

To examine whether questions or candidate names alone suffice to infer answers, we remove all visual and hand inputs and retain questions with options, options only, questions only, or candidate names. Each setting is evaluated with InternVL3.5-8B on all 1,500 test questions; all outputs are parseable.

Table C3: Target-selection accuracy without video input.

Input	Correct / questions	Accuracy (%)
Random selection	Expected	10.00
Question and options	127 / 1,500	8.47
Options only	138 / 1,500	9.20
Question only	157 / 1,500	10.47
Candidate names	159 / 1,500	10.60

All four text-input settings perform close to the 10% random-choice level. A name-prior heuristic achieves approximately 19.8%. These results indicate that language expressions or name statistics alone are insufficient for target selection; models need to combine them with pointing cues in the video to distinguish objects.

C.4 Number of Input Frames

We compare four base models under uniform sampling of 16 and 32 frames, with input image size set to 448 and evaluation on the same 1,500 test questions.

Table C4: Zero-shot comparison of 16-frame and 32-frame inputs.

Model	16 frames (%)	32 frames (%)	Gain (pp)
Qwen2.5-VL-7B	32.53	32.07	-0.47
Qwen3-VL-8B	29.73	31.07	+1.33
InternVL3.5-8B	24.93	26.13	+1.20
LLaVA-OneVision-1.5-8B	34.53	34.00	-0.53

Increasing the frame count does not yield consistent improvements. Qwen3-VL and InternVL improve, while the other two models decline slightly; all changes are within 1.34 percentage points. Base models differ in their use of additional frames, and increasing sampling density alone offers limited improvement for pointing target selection. Differences are computed from correct counts before rounding.

C.5 Human Performance

Five volunteers answered the questions for the 500 test videos. Each volunteer was assigned a disjoint set of 100 videos and answered all three questions per video, yielding 300 responses per volunteer. Participants viewed the videos with the same question text and ten candidate options used for model evaluation. Video replay was permitted, but target annotations and reference answers were not shown. We pooled the five volunteers' responses to compute accuracy over all 1,500 questions and within each ambiguity category. Each question was answered by one volunteer, so no voting or consensus aggregation was used.

Table C5 provides the correct-answer counts underlying the human baseline in Table 2, with 300 questions per category.

Table C5: Human performance across the five ambiguity categories.

Category	Correct / questions	Accuracy (%)
Depth overlap	287 / 300	95.67
Ray ambiguity	279 / 300	93.00
Saliency distraction	293 / 300	97.67
Temporal confusion	278 / 300	92.67
Viewpoint drift	284 / 300	94.67
Overall	1,421 / 1,500	94.73

Temporal confusion and ray ambiguity have the lowest human accuracies, while saliency distraction has the highest, suggesting that the categories differ in difficulty even for human viewers.

C.6 General Video Question-Answering Capabilities

We sample 500 questions from Video-MME short (Fu et al., 2025) and evaluate the base model, RGB-fine-tuned model, and PointMod on the same questions using eight frames at an input resolution of 448×448, without audio or subtitles. Predictions are obtained by selecting the candidate answer letter with the highest model probability at the answer position. PointMod uses its normal heatmap-guided input pipeline during evaluation.

PointMod achieves 71.60% accuracy, compared with 69.20% for the base model and 70.80% for RGB fine-tuning. The comparable performance shows that our method retains general video

Table C6: General video question-answering performance on Video-MME short.

Model	Accuracy (%)
Base InternVL3.5-8B	69.20
RGB SFT	70.80
PointMod	71.60

question-answering capabilities with heatmap guidance enabled. Since Video-MME contains real-world videos, these results also support transfer from synthetic pointing training to real-world video question answering.

D Prompts and Data Examples

We present question-generation constraints, an actual QA record, and the model evaluation input. Generation constraints retain their original numbering. The extracted question pool is combined with programmatically constructed candidates. Reference answers are stored separately and are not supplied to the model.

D.1 Question-Generation Constraints

The following excerpt preserves the original wording of the relevant entries in `STYLE_INSTRUCTION`.

```
1. Use first-person egocentric wording ("I", "me", "my"), not "the person
    ".
2. q_open_l3_primary should sound naturally deictic and implicit. It
    should not mention failure modes or hidden geometry.
   It should also avoid explicit gesture words such as pointing, gesture,
     hand, finger, looking, or indicating.
4. q_open_l1_explicit should ask for the object indicated by my pointing
    gesture.
5. q_mc_failure_aware_stem_only is only the MCQ stem; choices are added
    by code. It must ask which object is indicated/referred to. Do not
    ask why, whether it is clear, or whether the scene is ambiguous.
9. Do not use words such as ray, salient, overlap, front, rear, depth,
    viewpoint, drift, hold, stable, steady, canonical, keyframe, or
    middle frame in public questions.
10. Avoid collapsing to a single high-frequency template. Write natural,
    short questions, but vary wording across samples.
11. Do not copy examples verbatim unless the style profile explicitly
    allows it. Prefer small natural paraphrases.
```

D.2 An Actual QA Record

The following record and its separate reference answer correspond to Figure S4.

```
{
  "question": "What object is indicated by my pointing motion?",
  "options": [
    {"letter": "A", "name": "alarm clock"}, {"letter": "B", "name": "box
    "},
    {"letter": "C", "name": "desk lamp"}, {"letter": "D", "name": "drawer
    "},
    {"letter": "E", "name": "pen"}, {"letter": "F", "name": "key chain"},
    {"letter": "G", "name": "book"}, {"letter": "H", "name": "dresser"},
    {"letter": "I", "name": "bowl"}, {"letter": "J", "name": "vase"}
  ],
  "reference_answer": "J"
```

```
}
```

D.3 Evaluation Input

The input below is produced by the actual `build_prompt` function. Video frames are supplied through the visual input interface.

```
You are given sampled frames from an egocentric video in chronological
    order. A person is pointing at one object. Choose which candidate
    object is being referred to.

Question: What object is indicated by my pointing motion?
Candidate objects:
A. alarm clock
B. box
C. desk lamp
D. drawer
E. pen
F. key chain
G. book
H. dresser
I. bowl
J. vase
Return only one option letter from A to J. Do not explain.
```