SEEING WHAT CHANGED: EMBODIED CHANGE GROUNDING WITH SELECTIVE CROSS-VIEW PERCEPTION OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied agents often revisit the same environments, where objects may appear, disappear, move, or change state between visits. To study how an agent can tell *what has changed and where*, we formalize **Embodied Change Grounding (ECG)**: given a reference image, a query image captured after the scene has changed, and a language instruction, a model must identify the referred change and segment the changed object, which may lie in either image. As the two egocentric views are unaligned, this requires cross-view correspondence before localization. We introduce **ECG-Bench**, a benchmark of 10K samples from 1,859 simulated, industrial, and real revisited scenes with six change types, and **ECG-50K**, a training set with Change-CoT traces. Our evaluation shows that current MLLMs, reasoning segmentation frameworks, and RL post-training methods all fall short on ECG, and that models often produce fluent reasoning that is not grounded in the cross-view evidence. We therefore propose **ECG-Grounder**, whose **CoT-Guided Cross-View Decoder** selects the view containing the target, aligns the two views, and decodes the mask under the guidance of the reasoning chain. We further introduce Selective Cross-View Perception Optimization (SCPO) to train the model, rewarding reasoning chains whose outputs are more sensitive to removing the change-critical region than to a matched corruption elsewhere, as measured at high-entropy tokens within GRPO. ECG-Grounder with SCPO achieves the best results on ECG-Bench and also generalizes to other multi-image reasoning benchmarks such as MIG-Bench.

1 INTRODUCTION

Embodied agents often revisit the same environments. A warehouse robot patrols the same aisles every shift, while a household robot returns to the same rooms day after day. Between visits, objects may appear, disappear, move, be replaced, or change state. Understanding such changes is fundamental to long-term robot autonomy and has been explored through object-level change detection (Park et al., 2021; Friedlander et al., 2026), evolving 3D scene understanding (Steiner et al., 2026), and visual rearrangement (Liu et al., 2024; 2025c). More broadly, recent embodied and multimodal systems are beginning to extend spatial understanding from static scenes to environments that evolve over time (Zhou et al., 2025; Li et al., 2025b; Yuan et al., 2025). This raises a basic question for an agent returning to a familiar scene: *what has changed, and where is it* (Figure 1)? Answering it requires relating the current observation to prior experience and distinguishing genuine physical changes from differences caused by viewpoint, illumination, and occlusion.

We formalize this capability as **Embodied Change Grounding (ECG)**. Given a reference image from an earlier visit, a query image captured after the scene has changed, and a language instruction, a model must identify the referred change and segment the changed object. The target can lie in either image: an added object appears only in the query image, whereas a missing one appears only in the reference image. The key challenge is that the two images are captured by a moving agent and are therefore unaligned, so many unchanged regions also look different, and a model must establish cross-view correspondence before it can localize the referred change.

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.

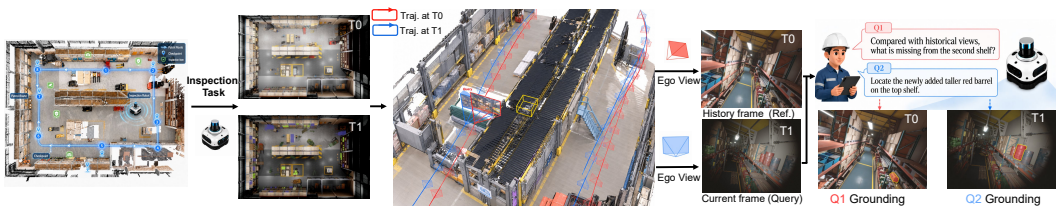


Figure 1: **Motivation and task overview of Embodied Change Grounding (ECG).** During repeated inspection, a robot revisits the same environment at T_0 and T_1 under different trajectories and viewpoints. ECG enables the robot to compare historical and current observations and ground the changed object specified by a natural-language query, supporting queries about missing, added, relocated, or state-changed objects.

We introduce **ECG-Bench**, a benchmark for evaluating language-guided change grounding in embodied environments, comprising 10K samples from 1,859 scenes across procedurally generated indoor environments, photorealistic industrial warehouses, and real revisited scenes. It covers residential, public, commercial, and industrial spaces, with six change types grouped into appearance (existence, missing, and replace), geometric (rotation and relocation), and functional (state). ECG-Bench further varies viewpoint, illumination, and same-category clutter, while its free-form instructions combine spatial, semantic, and attribute-based cues, requiring models to jointly understand spatial relations, object semantics, and cross-view changes. For post-training, we release ECG-50K, a scene-disjoint set of 50K simulated samples, each paired with a Change-CoT trace that anchors the instruction to a landmark, describes the relevant region in each view, compares the two observations, and identifies the changed object; every trace is verified against simulator ground truth.

Evaluating a broad range of models and post-training methods on ECG-Bench reveals three findings. **First**, open-source MLLMs perform poorly without task-specific training, and scaling them up helps little, suggesting that their cross-image understanding remains limited. **Second**, existing reasoning segmentation frameworks, such as LISA (Lai et al., 2024) and VisionReasoner (Liu et al., 2026), transfer poorly to ECG. They decode a mask from a single image and prompt, whereas ECG requires reasoning across both views and selecting the view that contains the target. **Third**, RL post-training such as GRPO (Shao et al., 2024) brings limited gains and is prone to CoT hallucination, producing fluent reasoning that is weakly grounded in the visual evidence. Recent perception-aware methods (Wang et al., 2026b; Huang et al., 2026) aimed at this issue also show limited improvement on ECG, suggesting that generic visual dependence is insufficient for cross-view change reasoning.

To address these issues, we propose **ECG-Grounder**, which grounds the change with a **CoT-Guided Cross-View Decoder**, and train it with **Selective Cross-View Perception Optimization (SCPO)**. Unlike SAM-like decoders (Ravi et al., 2025), the CoT-Guided Cross-View Decoder is built for cross-view change grounding: a view head selects the anchor view that contains the target, the decoder aligns the other view to it to extract change features, and the reasoning states are injected into decoding through a CoT grounding map and the `[change]` embedding. The mask is thus decoded from both views and the full reasoning chain, so that an object present only in the reference image can also be segmented. To mitigate CoT hallucination, SCPO rewards reasoning that is driven by the change-critical region rather than by change-irrelevant content. Starting from CoT-SFT on the Change-CoT traces, it compares, for each sampled chain, the output shift when the change-critical region is removed with the shift under a matched corruption elsewhere, and adds this contrast, measured at high-entropy reasoning tokens, to GRPO as a trajectory-level reward. ECG-Grounder with SCPO achieves the best results on ECG-Bench across both 7B and 9B backbones and all six change types, generalizes to MIG-Bench, and keeps its reasoning selective under large illumination and viewpoint differences.

Our contributions are threefold. (1) We introduce **Embodied Change Grounding**, which extends dynamic spatial understanding to grounding a referred change between two unaligned egocentric views with a pixel-level mask. (2) We build **ECG-Bench** (10K samples, six change types) and the **ECG-50K** training set with Change-CoT traces, and reveal the limitations of current MLLMs, reasoning segmentation frameworks, and RL post-training on ECG. (3) We propose **ECG-Grounder**, whose CoT-Guided Cross-View Decoder grounds the change across both views under the guidance of the reasoning chain, and **SCPO**, which rewards reasoning driven by the change-critical region.

Table 1: **Comparison of ECG-Bench with representative multi-image reasoning and scene change benchmarks.**

Benchmark	Venue	Domain	Egocentric	Cross-time	Change	Language Reasoning	Task (Output)
MC-Bench (Xu et al., 2025)	ICCV'25	Natural	✗	✗	✗	✓	Grounding (Box)
MIG-Bench (Li et al., 2025a)	ACL-F'25	Natural	✗	✗	✓	✓	Grounding (Box)
MedSG-Bench (Yue et al., 2025)	NeurIPS'25	Medical	✗	✓	✓	✓	Grounding (Box)
OpenEQA (Majumdar et al., 2024)	CVPR'24	Embodied	✓	✗	✗	✓	VQA (Text)
MindCube (Wang et al., 2026a)	ICLR'26	Embodied	✓	✗	✗	✓	VQA (Text)
MV-RoboBench (Feng et al., 2026)	ICLR'26	Embodied	✓	✗	✗	✓	VQA (Text)
Ego3D-Bench (Gholami et al., 2026)	ICLR'26	Driving	✓	✗	✗	✓	VQA (Text)
SceneDiff (Wu et al., 2026)	ECCV'26	Embodied	✓	✓	✓	✗	Change Det. (Mask)
VSCD (Yoon & Kim, 2026)	ICML'26	Embodied	✓	✓	✓	✗	Change Det. (Mask)
ECG-Bench (Ours)	—	**Embodied**	✓	✓	✓	✓	**Grounding (Mask)**

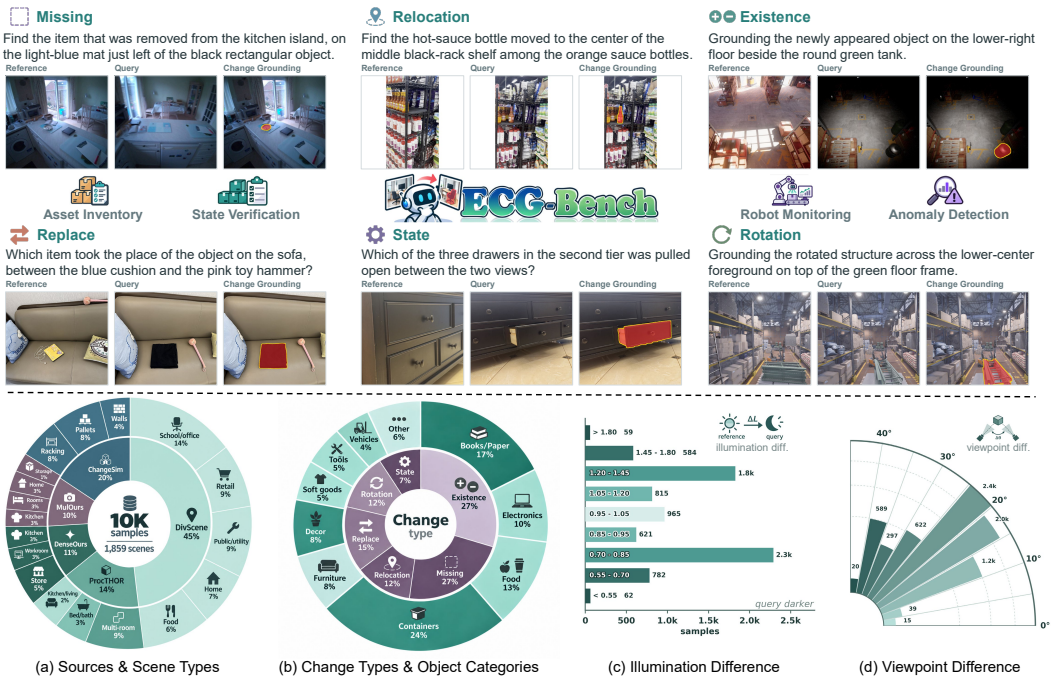
2 RELATED WORK

Egocentric scene change understanding. Perceiving what has changed when an agent revisits a scene is central to long-term robot autonomy, underlying object-level change detection (Friedlander et al., 2026), evolving-scene tracking (Steiner et al., 2026), visual rearrangement (Liu et al., 2024; 2025c), and 3D change reasoning (Mao et al., 2025). From the agent's own viewpoint, ChangeSim, SceneDiff, and VSCD detect changes in revisited indoor scenes (Park et al., 2021; Wu et al., 2026; Yoon & Kim, 2026), and Situat3DChange describes and answers questions about them over paired point clouds (Liu et al., 2025a). The former segment *every* change from videos or sequences with no referring expression, while the latter answers in text without localizing. ECG instead grounds the change an instruction refers to with a pixel-level mask from two unaligned egocentric views (Appendix A).

Language-guided reasoning grounding. MLLM-based grounding couples language reasoning and pixel-level localization in one model. LISA decodes masks from a `[SEG]` token (Lai et al., 2024), and follow-ups extend it to grounded conversation, multiple or absent referents, and unified image–video segmentation (Rasheed et al., 2024; Ren et al., 2024; Xia et al., 2024; Wei et al., 2025); all of them ground within a single image. Multi-image methods carry grounding across images, by combining CoT reasoning with direct answering in hybrid reinforcement fine-tuning (Zheng et al., 2026b), learning cross-view consistency from image triplets with GRPO (Chen et al., 2025), or injecting cross-image reasoning through staged SFT-then-RL training (Zhang et al., 2026a; Zheng et al., 2026a). None of them is built to separate a physical change from the viewpoint and illumination differences between two egocentric views, the first step that ECG demands.

Multi-image reasoning benchmarks. Multi-image grounding benchmarks such as MIG-Bench, MC-Bench, and MedSG-Bench evaluate correspondence, comparison, and difference grounding across images (Li et al., 2025a; Xu et al., 2025; Yue et al., 2025), while embodied benchmarks such as OpenEQA, MindCube, MV-RoboBench, and Ego3D-Bench evaluate multi-view spatial reasoning from an agent's own camera (Majumdar et al., 2024; Wang et al., 2026a; Feng et al., 2026; Gholami et al., 2026). As Table 1 shows, the former are not egocentric, and the latter observe a scene at a single instant and answer in text; ECG-Bench is the only one that pairs egocentric observations across time and asks for pixel-level grounding of the referred change.

Visual reinforcement and policy learning. RL with verifiable rewards was brought to visual perception by Visual-RFT (Liu et al., 2025d) and to dense and compositional grounding by Seg-Zero, VisionReasoner, and Rex-Thinker (Liu et al., 2025b; 2026; Jiang et al., 2026); built on GRPO and DAPO (Shao et al., 2024; Yu et al., 2025), later work refines the reward or anchors intermediate reasoning to the target (Sun et al., 2026; Yang et al., 2026). Closest to us, PAPO, VPPO, and VAPO model perceptual dependence during optimization, by contrasting outputs on original and corrupted images, reweighting perceptually pivotal tokens, or inserting visual anchors (Wang et al., 2026b; Huang et al., 2026; Tian et al., 2026). On ECG, rewarding such *generic* visual dependence proves too loose; SCPO instead rewards *selective* sensitivity, reacting only when the true cross-view change evidence is removed.



(a) Sources & Scene Types (b) Change Types & Object Categories (c) Illumination Difference (d) Viewpoint Difference

Figure 2: **Overview of ECG-Bench:** change-type examples (top) and statistics (bottom).

3 THE ECG BENCHMARK

ECG-Bench contains 10K samples collected from 1,859 scenes, covering three complementary settings: about 7K from procedurally generated indoor environments, 2K from photorealistic industrial warehouses, and 1K from real revisited scenes. As summarized in Fig. 2, the benchmark spans residential, public, professional, commercial, and industrial environments, with broad coverage of object categories and six types of physical changes: existence, missing, replace, rotation, relocation, and state change. The reference and query observations also differ substantially in viewpoint and illumination, so the benchmark is explicitly cross-view rather than pixel-aligned, and the changed object lies in the query view in 7,300 samples and in the reference view in 2,700. Appendices B.1 and B.3 give per-source statistics, change-type counts, and instruction diversity.

Figure 3 illustrates the data collection procedure. For controllable simulated environments, built on scenes from ProcTHOR (Deitke et al., 2022), DivScene (Wang et al., 2025), and our own dense scenes (DenseOurs), we record a scene in its reference state, modify a single object, and revisit it from a different trajectory and camera pose to obtain the query observation. Viewpoint and illumination can vary independently between the two visits, while the simulator provides accurate change labels and pixel-level masks by construction. For industrial and real revisited scenes, the image pairs come from ChangeSim (Park et al., 2021), SceneDiff (Wu et al., 2026), 3RScan (Wald et al., 2019), PASLCD (Galappaththige et al., 2025), and our own captures; we keep pairs with clearly observable and uniquely referable changes and take the changed-object masks from the source annotations or annotate them manually. Since none of these sources provides instructions that refer to a change, we add a free-form referring instruction to every change, describing the target through spatial, semantic, or attribute cues (details in Appendix B.1).

4 METHOD

4.1 PROBLEM SETUP AND CROSS-VIEW REASONING

We write an ECG instance as $x = (I_r, I_q, q)$, with I_r the reference observation, I_q the query observation, and q the instruction; the target is a binary mask $M \in \{0,1\}^{H\times W}$ over whichever of the two views contains the referred change. Since the two views are unregistered, a model cannot recover

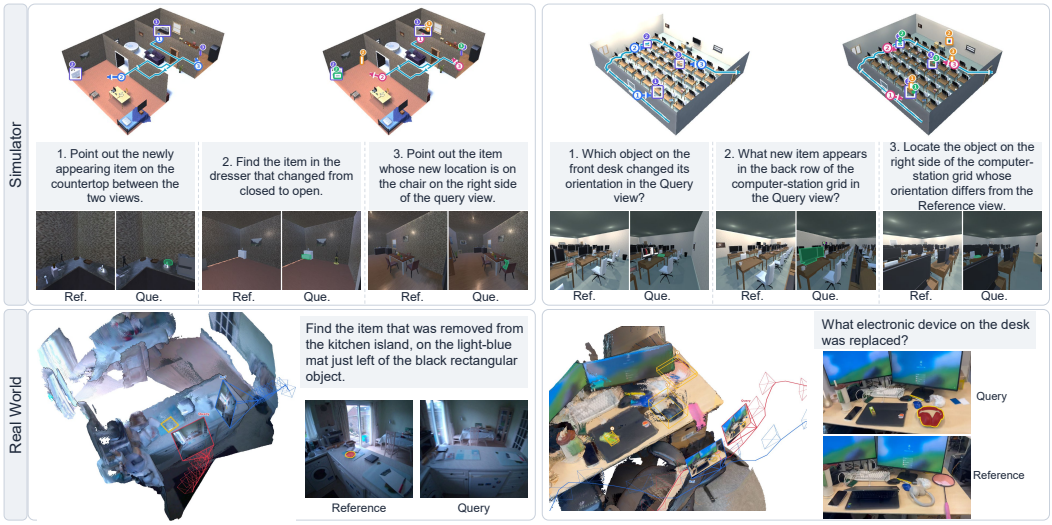


Figure 3: **ECG-Bench construction and examples.** *Top:* in the simulator, objects are edited between two visits recorded along different trajectories, and each change is paired with a referring instruction; masks are rendered directly. *Bottom:* real revisited scenes, shown with the camera trajectories of both visits; masks are manually annotated and verified.

M by differencing pixels: it must establish cross-view correspondence, decide which discrepancy is the referred change, bind the instruction to it, and localize it densely.

We call our architecture **ECG-Grounder**, shown in Figure 4. Given x, the policy π_θ autoregressively generates a reasoning trajectory $\tau = (y_1, \ldots, y_T)$ that follows the cross-view comparison order *reference evidence* $\rightarrow$ *query evidence* $\rightarrow$ *comparison* $\rightarrow$ *changed object*, and terminates in a special `[change]` token. A lightweight *CoT-Guided Cross-View Decoder* then turns the two views and the reasoning states into a mask.

CoT-Guided Cross-View Decoder. SAM-like decoders used in reasoning segmentation (Lai et al., 2024; Liu et al., 2025b; Ravi et al., 2025) decode a single image from a single prompt token. They ignore the relation between the reference and query views that defines an ECG target, and perform poorly on ECG (Table 3). The CoT-Guided Cross-View Decoder instead decodes from both views and the whole chain. With c an attention-pooled summary of the CoT hidden states h_{CoT} and $h = h_{\texttt{[change]}}$, a *view head* on (c, h) first selects the *anchor* view $a \in \{r, q\}$ predicted to contain the target; $\bar{a}$ is the other view, and Z_a, $Z_{\bar{a}}$ are their visual tokens from the MLLM's own vision encoder. The decoder then computes

$$F = \mathrm{Fuse}\big(Z_a, \mathrm{Attn}(Z_a, Z_{\bar{a}})\big), \quad G = \sigma\big(\mathrm{sim}(F, c)\big), \quad M = \mathrm{Dec}\big(F + G \odot \phi(c),\, h + \psi(c)\big). \tag{1}$$

Anchor-guided cross-view alignment lets each anchor token retrieve its correspondence in the other view despite the viewpoint change, and *change fusion* combines the anchor, the aligned view, and their difference terms into a change-aware map F on the anchor grid, so a missing object, present only in the reference view, is segmented there. The *CoT grounding* map G, supervised by the target mask, ties the chain to image regions, and *CoT injection* conditions decoding on the chain through a spatial path gated by G and a semantic path on the `[change]` embedding, both zero-initialized. Appendix D gives the full formulation.

We first warm up this behaviour by supervised fine-tuning on the Change-CoT traces of ECG-50K (Appendix C), $\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{T} \log \pi_\theta(y_t \mid x, y_{<t})$, jointly with the mask and grounding losses of the CoT-Guided Cross-View Decoder (Appendix D), and use the resulting checkpoint to initialize the RL policy. As Section 5 shows, this warmup is not optional: without it, reinforcement learning on ECG underperforms plain supervised fine-tuning.

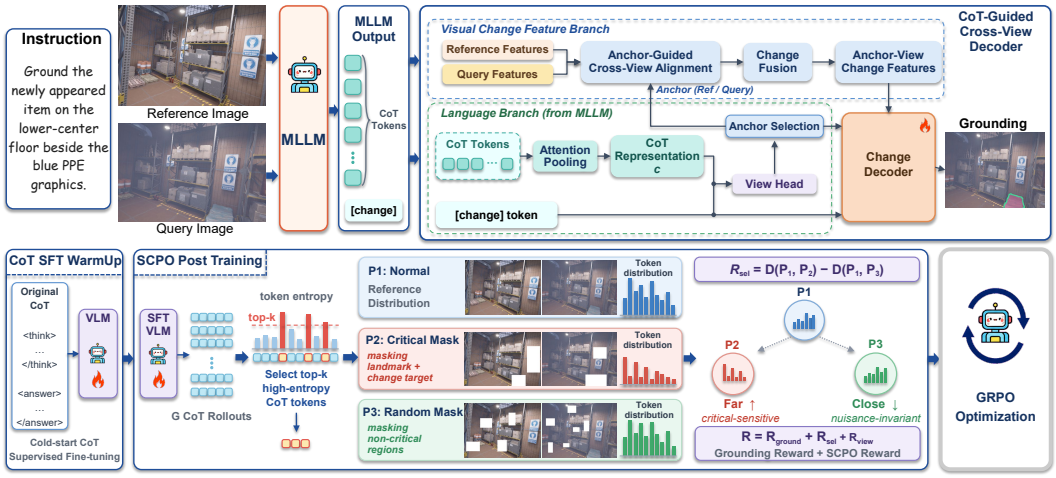


Figure 4: **Overview of ECG-Grounder and SCPO.** The policy writes a cross-view chain ending in a `[change]` token; a view head selects the anchor view, and the CoT-Guided Cross-View Decoder aligns the other view to it and decodes the mask with the reasoning states. During RL, each chain is re-scored under the intact pair, a critical intervention, and a matched nuisance corruption, and the gap between the two divergences at entropy-selected tokens is added to the GRPO reward.

4.2 Selective Cross-View Perception Optimization

CoT warmup teaches the model to *write* cross-view reasoning, but not to *depend* on it: hiding the change-critical region shifts the outputs of the CoT-SFT and GRPO policies only slightly more than hiding a matched region elsewhere (critical share 0.53–0.58, where 0.5 means no selectivity; Appendix E.3). An outcome reward cannot see this failure, since a chain that ignores the change can still land on the right region, and a generic visual-dependence reward (Wang et al., 2026b) cannot isolate it, since almost any corruption of a cluttered two-view scene perturbs the output. SCPO therefore rewards the *gap* between sensitivity to the change evidence and sensitivity to everything else.

Group rollouts and three visual conditions. For each instance we sample a group of G trajectories $\tau_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid I_r, I_q, q)$ and re-score each by teacher forcing under three visual conditions, holding the chain fixed so that only the visual evidence varies:

$$P^k_{i,t} = \pi_{\theta_{\text{old}}}(\cdot \mid X_k, q, \tau_{i,<t}), \qquad X_1 = (I_r, I_q), \quad X_2 = X_{\text{crit}}, \quad X_3 = X_{\text{nuis}}. \tag{2}$$

X_1 is the intact pair. The critical intervention X_{crit} (*critical mask*) destroys the change evidence by masking the *change-critical region*: the changed object and the landmark named in the instruction, whose masks the simulator provides during data synthesis. The matched nuisance X_{nuis} (*random mask*) applies a corruption of comparable area and strength that deliberately *avoids* this region.

Entropy-guided token selection. Not every reasoning token carries evidence; function words and formatting are insensitive to the image by construction. We therefore restrict the measurement to the positions where the policy is least certain, and where hallucination consequently originates:

$$\mathcal{S}_i = \text{TopK}_{t \in \tau_i} \; H_{i,t}, \qquad H_{i,t} = -\sum_v P^1_{i,t}(v) \log P^1_{i,t}(v). \tag{3}$$

Selective perception reward. A trustworthy trajectory should react more to the change evidence than to matched clutter, $D(P^1, P^2) > D(P^1, P^3)$, rather than become globally sensitive. We turn this into a scalar score,

$$R_{\text{sel}}(\tau_i) = \frac{1}{|\mathcal{S}_i|} \sum_{t \in \mathcal{S}_i} \Big[D\big(P^1_{i,t}, P^2_{i,t}\big) - D\big(P^1_{i,t}, P^3_{i,t}\big) \Big], \tag{4}$$

where D is the Jensen–Shannon divergence.

Policy update. The selective score is combined with the task rewards and normalized within the group,

$$R_i = R_{\text{ground}}(\tau_i) + \lambda_v R_{\text{view}}(\tau_i) + \lambda R_{\text{sel}}(\tau_i), \qquad A_i = \frac{R_i - \text{mean}(R)}{\text{std}(R) + \epsilon}, \tag{5}$$

where $R_{\text{view}} = \mathbb{1}[a = a^*]$ is a small-weight auxiliary reward for selecting the ground-truth anchor view a^* with the view head of Section 4.1, and the advantages drive the standard GRPO objective (Shao et al., 2024). R_{sel} is a *stop-gradient* scalar computed under $\pi_{\theta_{\text{old}}}$, not a differentiable loss: rollouts whose reasoning was not driven by the cross-view change lose advantage even when their mask is correct, unlike PAPO, which optimizes a perception term directly and without a nuisance control.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Implementation. All post-trained models are trained on ECG-50K, whose simulated scenes are disjoint from ECG-Bench; since it contains no real-world pairs, the real split of ECG-Bench measures synthetic-to-real transfer. We instantiate ECG-Grounder on two backbones, Qwen2.5-VL-7B (Bai et al., 2025) and Qwen3.5-9B (Qwen Team, 2026), and keep the grounding reward fixed across all post-trained methods. Post-training runs on 8 NVIDIA A100 GPUs with a learning rate of 1×10^{-5} in two stages: a 5-epoch CoT-SFT warmup on the ECG-50K traces, then 2 epochs of RL from that checkpoint with $G = 2$ CoT rollouts per instance. SCPO uses $\lambda = 0.1$, $K = 16$ entropy-selected tokens, and Jensen–Shannon divergence for D; the critical branch masks the changed object and the landmark named in the instruction, and the nuisance branch applies a corruption of matched area and strength elsewhere in the pair. The CoT-Guided Cross-View Decoder is trained in both stages, with grounding-loss weight $\lambda_g = 0.25$ and its loss scaled by 0.25 during RL; the view reward uses $\lambda_v = 0.1$. We report **mIoU** between the predicted and the ground-truth mask for each of the six change types, and **AVE**, their macro-average. Baselines include training-free MLLMs such as InternVL3 (Zhu et al., 2025) and LLaVA-OV (Li et al., 2024), SFT-based reasoning segmentation methods, and RL post-training methods; Appendix E.1 details how each is run.

5.2 MAIN RESULTS

Table 2 supports the three findings stated in Section 1. **First**, training-free MLLMs perform poorly and scale helps little: the best, InternVL3-78B, reaches only 17.3 AVE, and scaling Qwen2.5-VL from 7B to 72B raises it from 9.3 to 13.3; only *State* is relatively easy, as it is often recognizable from the query view alone. **Second**, reasoning segmentation frameworks transfer poorly: trained on ECG-50K, LISA, PixelLM, and GLaMM reach 20.5–22.3, and even with our reasoning traces LISA reaches 26.3, against 30.3 for ECG-Grounder with the same CoT-SFT, a gap attributable to LISA's single-image, single-token mask interface (see also Table 3). **Third**, RL brings limited gains: from the base policy, Visual-RFT and VisionReasoner fall below every SFT baseline (13.1 and 15.2) and nearly collapse on *Missing* (2.1 and 2.9), as an IoU reward favors a plausible mask in the query view even when the object has disappeared; from our CoT-SFT checkpoint, VisionReasoner† gains only $+2.3$ (Appendix E.2), and GeM-VG, PAPO, and VPPO stay within 1.3 points of it, so generic visual dependence does not suffice.

SCPO is the only objective that leaves this cluster. It reaches 35.6 AVE on Qwen2.5-VL-7B and 38.7 on Qwen3.5-9B ($+1.7$ and $+2.1$ over VPPO), is best on all six change types on both backbones, and gains $+5.3$ over its CoT-SFT checkpoint, more than twice VisionReasoner†. Its margin over VPPO is largest on *Rotate* ($+3.6$) and smallest on *State* ($+0.5$), the type most often decidable from the query view alone. Still, *Relocate* (26.5) and *Missing* (30.3) remain far below *State* (61.3), so ECG-Bench is far from solved.

5.3 ABLATION STUDIES

All ablations use the Qwen2.5-VL-7B backbone and the ECG-50K training data of Section 5.1, and report mIoU on ECG-Bench; the ablation of the two post-training stages is in Appendix E.2.

Table 2: **Main results on ECG-Bench** (mIoU, %). **CoT Init.**: trained with CoT-SFT or, for RL rows, initialized from its checkpoint; †: re-run from that checkpoint instead of the base policy. Baseline settings are in Appendix E.1.

Model	Base MLLM	CoT Init.	Appearance			Geometric		Function	AVE
			Existence	Missing	Replace	Rotate	Relocate	State	
Training-Free									
InternVL3-8B	–	–	9.8	8.9	10.7	9.2	6.8	23.6	11.5
Qwen2.5-VL-7B	–	–	7.2	6.3	8.6	7.0	5.4	21.5	9.3
Qwen3.5-9B	–	–	11.5	12.5	12.8	11.1	12.9	35.2	16.0
LLaVA-OV-72B	–	–	12.1	8.6	11.4	9.8	10.1	26.9	13.2
Qwen2.5-VL-72B	–	–	13.8	6.6	12.1	10.5	8.7	28.1	13.3
InternVL3-78B	–	–	13.2	19.3	11.8	10.1	17.9	31.4	17.3
Post-Training SFT									
LISA (CVPR'24)	Qwen2.5-VL-7B	✗	18.7	17.4	21.2	19.4	15.8	41.1	22.3
PixelLM (CVPR'24)	Qwen2.5-VL-7B	✗	17.4	13.6	19.8	17.9	12.7	41.4	20.5
GLaMM (CVPR'24)	Qwen2.5-VL-7B	✗	19.2	18.2	22.0	20.1	14.6	39.9	22.3
LISA + CoT	Qwen2.5-VL-7B	✓	22.7	19.8	24.5	22.8	18.9	49.0	26.3
ECG-Grounder (CoT-SFT)	Qwen2.5-VL-7B	✓	**26.5**	**22.7**	**31.5**	**25.5**	**22.3**	**53.2**	**30.3**
Post-Training RL									
Visual-RFT (ICCV'25)	Qwen2.5-VL-7B	✗	12.6	2.1	14.3	12.2	8.0	29.4	13.1
VisionReasoner (ICLR'26)	Qwen2.5-VL-7B	✗	14.8	2.9	16.7	14.5	10.4	31.9	15.2
VisionReasoner†	Qwen2.5-VL-7B	✓	28.6	26.9	32.0	29.7	21.9	56.2	32.6
GeM-VG (AAAI'26)	Qwen2.5-VL-7B	✓	30.2	23.1	31.4	30.8	19.2	56.3	31.8
PAPO (ICLR'26)	Qwen2.5-VL-7B	✓	31.0	26.4	33.2	31.5	22.1	56.6	33.5
VPPO (ICLR'26)	Qwen2.5-VL-7B	✓	31.5	25.9	34.1	32.0	22.8	57.2	33.9
ECG-Grounder + SCPO (Ours)	Qwen2.5-VL-7B	✓	**33.8**	**27.1**	**36.2**	**35.6**	**23.4**	**57.7**	**35.6**
Visual-RFT (ICCV'25)	Qwen3.5-9B	✗	15.4	3.7	16.8	14.6	10.0	32.4	15.5
VisionReasoner (ICLR'26)	Qwen3.5-9B	✗	17.5	4.5	19.2	17.1	12.6	34.8	17.6
VisionReasoner†	Qwen3.5-9B	✓	31.2	26.2	34.5	32.0	24.0	58.4	34.4
DAPO (ICLR'26)	Qwen3.5-9B	✓	32.7	27.9	35.4	33.5	24.7	58.8	35.5
PAPO (ICLR'26)	Qwen3.5-9B	✓	33.4	29.0	36.0	34.2	25.0	59.4	36.2
VPPO (ICLR'26)	Qwen3.5-9B	✓	33.8	28.6	36.8	34.7	25.5	60.0	36.6
ECG-Grounder + SCPO (Ours)	Qwen3.5-9B	✓	**36.6**	**30.3**	**39.3**	**38.4**	**26.5**	**61.3**	**38.7**

Table 3: **Ablation on the CoT-Guided Cross-View Decoder** (mIoU, %; same SCPO policy). *Ref.*: also sees I_r; *Align*: cross-view alignment and change fusion; *CoT*: CoT grounding and injection.

Mask decoder	Ref.	Align	CoT	mIoU
Predicted box → frozen SAM 2	✗	✗	✗	22.1
`[change]` → SAM 2 dec., query	✗	✗	✗	26.9
`[change]` → SAM 2 dec., both	✓	✗	✗	28.2
Ours w/o alignment	✗	✗	✓	33.6
Ours w/o CoT cond.	✓	✓	✗	34.9
Ours (full)	✓	✓	✓	**35.6**

Table 4: **Ablation inside SCPO** (mIoU, %). Each row removes one ingredient of the selective reward from the same CoT-SFT checkpoint, under an identical training budget.

Variant	AVE	ΔAVE
Plain GRPO ($\lambda = 0$)	32.6	−3.0
$+R_{\text{sel}}$ w/o critical mask	33.0	−2.6
$+R_{\text{sel}}$ w/o random mask	34.4	−1.2
$+R_{\text{sel}}$ w/o entropy sel.	35.1	−0.5
SCPO (full, Ours)	**35.6**	—

CoT-Guided Cross-View Decoder. Table 3 fixes the SCPO policy and varies only the mask interface. SAM-like interfaces stay at or below 28.2: a frozen SAM 2 prompted with a predicted box gives 22.1, a SAM 2 decoder driven by the `[change]` embedding 26.9, and showing it the reference view as well adds only 1.3, since nothing relates the two views. Every variant of our decoder clears 33.6; within it, removing cross-view alignment, and with it the reference view, costs 2.0, and removing CoT conditioning 0.7.

Inside SCPO. Table 4 removes one ingredient at a time. Without the selective reward the policy is plain GRPO at 32.6. Dropping the critical intervention recovers almost nothing (33.0): knowing what a chain should *not* depend on is not enough. Dropping the matched nuisance branch reduces the reward to a PAPO-style $D(P^1, P^2)$ and reaches 34.4, so roughly 60% of the gain survives — but the last 1.2 comes precisely from the subtraction. Entropy-based token selection is a smaller refinement (−0.5). Measured directly, SCPO's reasoning also stays selective under large illumination and viewpoint differences, whereas CoT-SFT and GRPO show little selectivity (Appendix E.3).

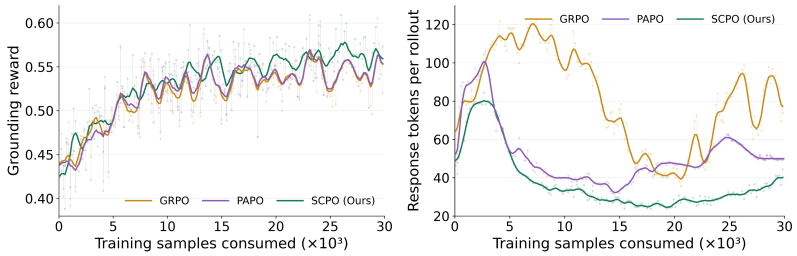


Figure 5: **RL training dynamics.** Grounding reward (left) and response length (right) of the training rollouts; faint markers are raw values, solid lines smoothed curves.

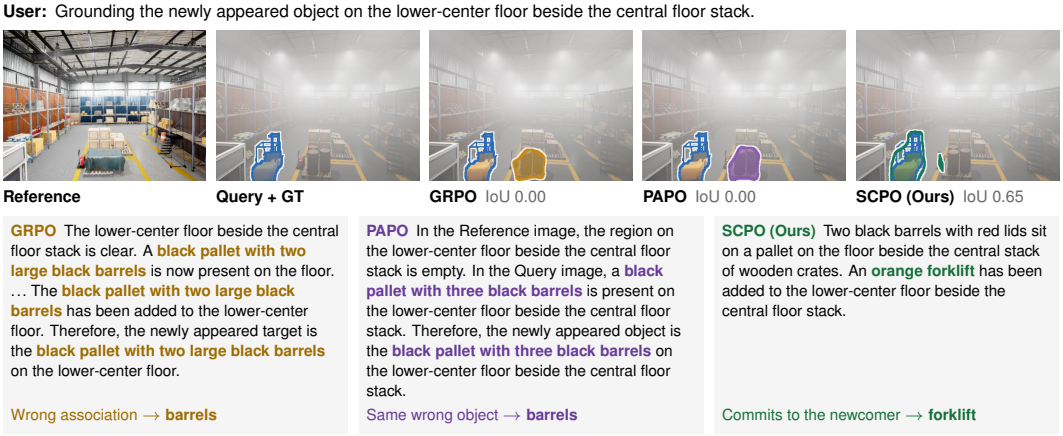


Figure 6: **Predicted chains and masks of GRPO, PAPO, and SCPO** for a newly appeared object in a hazed warehouse; highlighted text marks the referent each chain names.

5.4 ANALYSIS

Training dynamics. Figure 5 tracks the rollouts during RL. Both perception-aware objectives shorten the chain early: after an initial rise, the response length of SCPO and PAPO falls from a peak of 80–100 tokens to about 40 within the first 10K samples, and SCPO settles lowest, at 25–40 tokens. Inspecting the rollouts, the shortened chains skip the exhaustive per-view description of the CoT-SFT traces and state the change directly, suggesting that the policy learns to find the decisive evidence rather than enumerate it. GRPO stays at 100–120 tokens for most of the first 12K samples and rises again to 65–95 late in training. Shorter chains do not cost accuracy: SCPO keeps the highest grounding reward on the training rollouts for most of the second half of training.

Qualitative comparison. In Figure 6, the chains of GRPO and PAPO are long but uninformative: they restate the instruction's location several times, latch onto the black barrels beside the stack as the visual cue, and repeat that referent until the conclusion, even disagreeing on the barrel count (IoU 0.00 for both). SCPO's chain takes two sentences: it describes the barrels only as the surroundings of the referred location, then names the orange forklift as the added object, and its mask follows (IoU 0.65). Its reasoning thus goes straight to the change-critical evidence, in line with the shorter chains of Figure 5. Appendix G shows twelve further cases covering all six change types.

6 CONCLUSION

We introduced Embodied Change Grounding, which asks a model to segment the referred change between two unaligned egocentric views, together with ECG-Bench and the ECG-50K training set. Current MLLMs, reasoning segmentation frameworks, and RL post-training methods all fall short on this task. ECG-Grounder grounds the change across both views under the guidance of the reasoning chain, and SCPO rewards reasoning driven by the change-critical region; together they achieve the best results on ECG-Bench. Relocated and missing objects nonetheless remain far from solved, and we hope ECG-Bench will inspire further work on embodied agents that understand how their environments change over time.

Ethics Statement

ECG-Bench and ECG-50K are built from simulated scenes and real indoor scenes. The third-party sources, ProcTHOR, DivScene, ChangeSim, SceneDiff, 3RScan, and PASLCD, are used under their respective licenses; for samples derived from them we will release only our annotations, and users obtain the source data from the original datasets. The 124 real scene pairs we captured ourselves show common laboratory, office, and residential interiors; they contain no people or other personally identifiable information, and we additionally applied blurring as a privacy safeguard. All manual annotation was performed by the authors.

Reproducibility Statement

We will publicly release the code, ECG-Bench, and ECG-50K. Section 3 and Appendix B describe how each data source is converted into ECG samples, including the manual annotation workflow, and Appendix C describes how the Change-CoT traces are generated and verified. Section 4 and Appendix D give the architecture and training objectives of ECG-Grounder and SCPO, Section 5.1 gives the backbones, training stages, and hyper-parameters, and Appendix E.1 gives the configuration of every baseline.

AI Use Statement

We used generative AI tools in three limited roles. First, for language editing, improving the clarity, concision, grammar, and academic style of author-written text. Second, to assist in preparing figures. Third, within the semi-automatic pipeline that annotates part of the Change-CoT traces used for CoT warmup: GPT-5 drafts a trace from the two views and the instruction, and every draft is verified against the simulator ground truth and regenerated if it fails (Appendix C). The research questions, task and benchmark design, method, experiments, analyses, and scientific conclusions were developed and verified by the authors. All AI-assisted text, figures, and annotations were reviewed by the authors, who take full responsibility for the final content of the paper.

References

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*, 2025.

Xi Chen, Mingkang Zhu, Shaoteng Liu, Xiaoyang Wu, Xiaogang Xu, Yu Liu, Xiang Bai, and Hengshuang Zhao. Mico: Multi-image contrast for reinforcement visual reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-scale embodied AI using procedural generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Zhiyuan Feng, Zhaolu Kang, Qijie Wang, Zhiying Du, Jiongrui Yan, Shubin Shi, Chengbo Yuan, Huizhi Liang, Yu Deng, Qixiu Li, Rushuai Yang, Ruichuan An, Leqi Zheng, Weijie Wang, Shuang Chen, Sicheng Xu, Yaobo Liang, Jiaolong Yang, and Baining Guo. Seeing across views: Benchmarking spatial reasoning of vision-language models in robotic scenes. In *The Fourteenth International Conference on Learning Representations*, 2026.

Almog Friedlander, Ariel Shamir, and Ohad Fried. GOLDILOCS: General object-level detection and labeling of changes in scenes. In *International Conference on Learning Representations (ICLR)*, 2026.

Yuqian Fu, Runze Wang, Bin Ren, Guolei Sun, Biao Gong, Yanwei Fu, Danda Pani Paudel, Xuanjing Huang, and Luc Van Gool. ObjectRelator: Enabling cross-view object relation understanding across ego-centric and exo-centric perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6530–6540, 2025.

Chamuditha Jayanga Galappaththige, Jason Lai, Lloyd Windrim, Donald G. Dansereau, Niko Sünderhauf, and Dimity Miller. Multi-view pose-agnostic change localization with zero labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11600–11610, 2025.

Chamuditha Jayanga Galappaththige, Jason Lai, Lloyd Windrim, Donald G. Dansereau, Niko Sünderhauf, and Dimity Miller. Changes in real time: Online scene change detection with multi-view fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32246–32256, 2026.

Mohsen Gholami, Ahmad Rezaei, Zhou Weimin, Sitong Mao, Shunbo Zhou, Yong Zhang, and Mohammad Akbari. Spatial reasoning with vision-language models in ego-centric multi-view scenes. In *The Fourteenth International Conference on Learning Representations*, pp. 101989–102021, 2026.

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, 2024.

Qi Guo, Jue Wang, Yinhe Liu, and Yanfei Zhong. OpenDPR: Open-vocabulary change detection via vision-centric diffusion-guided prototype retrieval for remote sensing imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20399–20409, 2026.

Yicheng He, Chengsong Huang, Zongxia Li, Jiaxin Huang, and Yonghui Yang. Visplay: Self-evolving vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Qing Jiang, Xingyu Chen, Zhaoyang Zeng, Junzhi Yu, and Lei Zhang. Rex-thinker: Grounded object referring via chain-of-thought reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jae-Woo Kim and Ue-Hwan Kim. Towards generalizable scene change detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24463–24473, 2025.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

You Li, Heyu Huang, Chi Chen, Kaiyu Huang, Chao Huang, Zonghao Guo, Zhiyuan Liu, Jinan Xu, Yuhua Li, Ruixuan Li, and Maosong Sun. Migician: Revealing the magic of free-form multi-image grounding in multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9845–9867, 2025a.

Yun Li, Yiming Zhang, Tao Lin, Xiangrui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. STI-Bench: Are MLLMs ready for precise spatial-temporal world understanding? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5622–5632, 2025b.

Ruiping Liu, Junwei Zheng, Yufan Chen, Zirui Wang, Kunyu Peng, Kailun Yang, Jiaming Zhang, Marc Pollefeys, and Rainer Stiefelhagen. Situat3dchange: Situated 3d change understanding dataset for multimodal large language model. In *Advances in Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2025a.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025b.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified reasoning-integrated visual perception via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yuyi Liu, Xinhang Song, Weijie Li, Xiaohan Wang, and Shuqiang Jiang. A category agnostic model for visual rearrangment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16457–16466, 2024.

Yuyi Liu, Xinhang Song, Tianliang Qi, and Shuqiang Jiang. Trial-oriented visual rearrangement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8022–8031, 2025c.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2034–2044, 2025d.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul McVay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Alexander Sax, and Aravind Rajeswaran. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Ye Mao, Weixun Luo, Junpeng Jing, Anlan Qiu, and Krystian Mikolajczyk. Hypo3D: Exploring hypothetical reasoning in 3D. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 43011–43034, 2025.

Lorenzo Mur-Labadia, Maria Santos-Villafranca, Jesus Bermudez-Cameo, Alejandro Perez-Yus, Ruben Martinez-Cantin, and Jose J. Guerrero. O-MaMa: Learning object mask matching between egocentric and exocentric views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6892–6903, 2025.

OpenAI. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Jiancheng Pan, Runze Wang, Tianwen Qian, Mohammad Mahdi, Yanwei Fu, Xiangyang Xue, Xiaomeng Huang, Luc Van Gool, Danda Pani Paudel, and Yuqian Fu. V^2-SAM: Marrying SAM2 with multi-prompt experts for cross-view object correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16910–16919, 2026.

Jin-Man Park, Jae-Hyuk Jang, Sahng-Min Yoo, Sun-Kyung Lee, Ue-Hwan Kim, and Jong-Hwan Kim. Changesim: Towards end-to-end online scene change detection in industrial indoor environments. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.

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

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13009–13018, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, 2025.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26374–26383, 2024.

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

Emily Steiner, Jianhao Zheng, Henry Howard-Jenkins, Chris Xie, and Iro Armeni. ReScene4D: Temporally consistent semantic instance segmentation of evolving indoor 3D scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10710–10720, 2026.

Haoxiang Sun, Tao Wang, Chenwei Tang, Li Yuan, and Jiancheng Lv. Dr. seg: Revisiting grpo training for visual large language models through perception-oriented design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24320–24329, 2026.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, Mengqi He, Fabian Waschkowski, Lukas Wesemann, Peter Tu, and Jing Zhang. More thought, less accuracy? on the dual nature of reasoning in vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. RIO: 3D object instance re-localization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Li Fei-Fei, and Manling Li. Spatial mental modeling from limited views. In *The Fourteenth International Conference on Learning Representations*, pp. 82559–82622, 2026a.

Zhaowei Wang, Hongming Zhang, Tianqing Fang, Ye Tian, Yue Yang, Kaixin Ma, Xiaoman Pan, Yangqiu Song, and Dong Yu. DivScene: Towards open-vocabulary object navigation with large vision language models in diverse scenes. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 9666–9686, 2025.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. Perception-aware policy optimization for multimodal reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Cong Wei, Yujie Zhong, Haoxian Tan, Yingsen Zeng, Yong Liu, Hongfa Wang, and Yujiu Yang. Instructseg: Unifying instructed visual segmentation with multi-modal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20193–20203, 2025.

Size Wu, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, and Chen Change Loy. F-lmm: Grounding frozen large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24710–24721, 2025.

Yuqun Wu, Chih-Hao Lin, Henry Che, Aditi Tiwari, Chuhang Zou, Shenlong Wang, and Derek Hoiem. SceneDiff: A benchmark and method for multiview object change detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3858–3869, 2024.

Yunqiu Xu, Linchao Zhu, and Yi Yang. Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Shannan Yan, Leqi Zheng, Keyu Lv, Jingchen Ni, Hongyang Wei, Jiajun Zhang, Guangting Wang, Jing Lyu, Chun Yuan, and Fengyun Rao. Learning cross-view object correspondence via cycle-consistent mask prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6653–6663, 2026.

Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. Lisa++: An improved baseline for reasoning segmentation with large language model. *arXiv preprint arXiv:2312.17240*, 2023.

Tao Yang, Qing Zhou, Yanliang Li, and Qi Wang. Discriminative perception via anchored description for reasoning segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Jiae Yoon and Ue-Hwan Kim. VSCD: Video-based scene change detection in unaligned scenes. In *International Conference on Machine Learning (ICML)*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yuqian Yuan, Ronghao Dang, Long Li, Wentong Li, Dian Jiao, Xin Li, Deli Zhao, Fan Wang, Wenqiao Zhang, Jun Xiao, and Yueting Zhuang. EOC-Bench: Can MLLMs identify, recall, and forecast objects in an egocentric world? In *Advances in Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2025.

Jingkun Yue, Siqi Zhang, Zinan Jia, Huihuan Xu, Zongbo Han, Xiaohong Liu, and Guangyu Wang. Medsg-bench: A benchmark for medical image sequences grounding. In *Advances in Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2025.

Qi Zang, Dong Zhao, Shuang Wang, Dou Quan, and Zhun Zhong. Feature spectrum learning for remote sensing change detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12647–12657, 2025.

Yufei Zhan, Ziheng Wu, Yousong Zhu, Rongkun Xue, Guanghua Zhou, Ruipu Luo, Zhenghao Chen, Can Zhang, Yifan Li, Zhentao He, Zheming Yang, Ming Tang, Minghui Qiu, and Jinqiao Wang. Gthinker: Towards general multimodal reasoning via cue-guided rethinking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Bob Zhang, Haoran Li, Tao Zhang, Jianan Li, Cilin Yan, Xikai Liu, Jiayin Cai, and Yanbin Hao. Improving the reasoning of multi-image grounding in mllms via reinforcement learning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026a.

Xu Zhang, Danyang Li, Xiaohang Dong, Tianhao Wu, Hualong Yu, Jianye Wang, Qicheng Li, and Xiang Li. UniChange: Unifying change detection with multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 42169–42179, 2026b.

Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. PSALM: Pixelwise segmentation with large multi-modal model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 74–91, 2024.

Lihao Zheng, Zhenwei Shao, Yu Zhou, Yan Yang, Xintian Shen, Jiawei Chen, Hao Ma, and Tao Wei. Cgc: Compositional grounded contrast for fine-grained multi-image understanding. *arXiv preprint arXiv:2604.22498*, 2026a.

Shurong Zheng, Yousong Zhu, Hongyin Zhao, Fan Yang, Yufei Zhan, Ming Tang, and Jinqiao Wang. Gem-vg: Towards generalized multi-image visual grounding with multimodal large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(34):28857–28865, 2026b.

Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. VLM4D: Towards spatiotemporal awareness in vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8600–8612, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

APPENDIX

Contents

A Extended Related Work 16

Expands Section 2 with egocentric and remote-sensing change detection, cross-view object correspondence without language, MLLM-based reasoning grounding, and further visual RL methods, and states how each differs from ECG.

B ECG-Bench Data 18

Details how each source of Section 3 — ProcTHOR, DivScene, DenseOurs, ChangeSim, SceneDiff, 3RScan, PASLCD, and our own captures — is converted into ECG samples, with the manual annotation workflow and per-change-type counts (B.1); introduces DenseOurs, the dense procedural scenes, and compares their objects per scene, floor area, and floor-plan shape with ProcTHOR-10k and DivScene (Figure A1); analyzes the instruction vocabulary and the referring and change types of each source (Figures A2–A3); then shows thirty real-world samples across the six change types and diverse scenes, including samples from PASLCD (Figures A4–A6), and ten high-fidelity simulated warehouse samples from ChangeSim under low light and dust (Figure A7).

C Change-CoT Annotation 22

Describes how the cross-view reasoning traces for CoT warmup are generated with GPT-5 and verified against the simulator ground truth (Figure A8), with example traces for five change types (Figure A9).

D CoT-Guided Cross-View Decoder: Design and Analysis 22

Gives the full formulation of the CoT-Guided Cross-View Decoder of Section 4.1 — view selection, anchor-guided alignment and change fusion, CoT grounding and injection, and mask decoding (D.1) — its training objective in SFT and RL (D.2), a comparison with SAM-like decoders (Table A1), and a qualitative analysis of the fused cues on simulated and real-world cases (Figures A10–A11).

E Experimental Details and Additional Results 32

Details how each baseline of Table 2 is run — the ECG reasoning pipeline for training-free MLLMs, the SFT re-implementations, and the RL objectives (E.1); ablates the two post-training stages, answer-only versus CoT supervision followed by no RL, GRPO, or SCPO (Table A2); and shows that SCPO's reasoning stays selective across illumination and viewpoint differences (Figure A12).

F Token Entropy and Visual Dependence along the Reasoning Chain 33

Defines Vision-PMI and tracks it together with token entropy along the reasoning chain, showing that high-entropy tokens concentrate on target naming and change decisions, the evidence behind the entropy-based token selection of SCPO (Figure A13).

G Additional Qualitative Results 33

Extends the comparison of Figure 6 with predicted chains and masks of GRPO, PAPO, and SCPO on twelve further cases from procedural simulation, high-fidelity simulation, and real-world scenes, covering all six change types (Figures A14–A17).

H Transfer to Other Multi-Image Reasoning Benchmarks 40

Evaluates the method on the ten box-grounding tasks of MIG-Bench against general MLLMs and multi-image grounding methods (Table A3), and compares its chains and boxes with those of VisionReasoner[†] (Figure A18).

A EXTENDED RELATED WORK

This section expands Section 2, with a detailed analysis of scene change understanding and a discussion of cross-view object correspondence, which the main text does not cover.

A.1 Scene Change Understanding

Egocentric scene change understanding. The works closest to ECG in setting detect changes in indoor scenes revisited by a moving camera. ChangeSim provides online scene change detection in industrial indoor scenes (Park et al., 2021). SceneDiff targets objects that have been added, removed, or moved between two captures of the same scene, motivated by applications such as robotic tidying; its benchmark contains 350 video pairs with object instance annotations, and its training-free method aligns the two captures in 3D with pretrained 3D, segmentation, and image-encoding models before comparing spatial and semantic region features (Wu et al., 2026). VSCD formulates video-based scene change detection: given a reference and a query RGB video of the same indoor space, recorded at different times under unconstrained camera motion and not temporally synchronized, it predicts a pixel-wise change mask for every query frame (Yoon & Kim, 2026). Its benchmark contains over 1.1 million frames with pixel-accurate change masks and a real-world test set, and its model aligns candidate reference features to the query through local patch correspondence and fuses them with frame- and patch-level confidence; the authors deploy it on a mobile robot for visual surveillance and object incremental learning. Other ground-level work generalizes change detection across street-level domains with frozen SAM (Kim & Kim, 2025), or learns a change channel in a 3D Gaussian Splatting model of the reference scene, offline from a few post-change images (Galappaththige et al., 2025) or online from a live image stream (Galappaththige et al., 2026).

These works establish change perception from robot viewpoints as an important problem, and they share ECG's difficulty of separating genuine change from viewpoint-induced differences. They differ from ECG in three respects. First, they have no referring expression: the output is a category-agnostic mask of *every* change, whereas an embodied agent is typically asked about one specific change among many, which a model must resolve from language before localizing. Second, SceneDiff, VSCD, and the multi-view methods draw on dense observations, video pairs or a reconstructed reference scene, to align the two visits, whereas ECG gives a single reference view and a single query view, so correspondence has to be established from two unaligned images. Third, their change vocabulary is narrower: SceneDiff covers added, removed, and moved objects and VSCD objects that appear or disappear, while ECG additionally covers replacement, rotation, and state changes. Situat3DChange brings language to scene change: built from human observations of environmental changes, it provides 121K question–answer pairs, 36K change descriptions, and 17K rearrangement instructions that span egocentric and allocentric perspectives, and pairs them with SCReasoner, a 3D MLLM that compares point clouds (Liu et al., 2025a). It thus extends change understanding beyond localization to captioning and question answering, but its inputs are point-cloud reconstructions and its outputs are text rather than pixel-level masks.

Remote-sensing change detection. Bi-temporal remote sensing is the most mature form of change detection. FeaSpect suppresses pseudo-changes caused by imaging conditions through a trainable feature spectrum transformation in the latent space (Zang et al., 2025); UniChange unifies binary and semantic change detection in an MLLM, using special tokens and text prompts that specify the change categories (Zhang et al., 2026b); and OpenDPR performs training-free open-vocabulary change detection by classifying class-agnostic change proposals with diffusion-generated visual prototypes (Guo et al., 2026). All three produce pixel-level masks, but none involves a referring expression: they operate on co-registered overhead image pairs, and the text that UniChange and OpenDPR accept names a change category rather than referring to a particular object.

A.2 Cross-View Object Correspondence

A related line locates the same object across views without language. Ego-Exo4D defines a correspondence benchmark in which an object is given by a mask in one of two synchronized egocentric and exocentric views and must be segmented in the other if visible (Grauman et al., 2024). PSALM, an LMM-based segmenter that accepts visual prompts, is evaluated on it zero-shot (Zhang et al., 2024); ObjectRelator builds on PSALM, fusing the query mask with a generated text description of the object and aligning object features across views (Fu et al., 2025); O-MaMa recasts the task as matching candidate masks in the target view against the query mask (Mur-Labadia et al., 2025); V^2-SAM adapts SAM2 with anchor and appearance prompt experts derived from the source mask and selects among them by cyclic consistency (Pan et al., 2026); and Yan et al. (2026) train mask

prediction with a cycle-consistency objective that also enables test-time training. This line is closest to ECG in establishing object-level correspondence under large viewpoint change, and its tools, such as mask matching and cycle consistency, target the same difficulty of unaligned views. Its query, however, is *visual*: the object is already specified by a mask, and the task is to find the *same* object elsewhere. Absence is at most a side case: ObjectRelator keeps only samples visible in both views (Fu et al., 2025), V^2-SAM reports temporarily absent targets as a failure case (Pan et al., 2026), and Yan et al. (2026) add a separate visibility classifier. In ECG the referent is specified only through language, and the change is itself the target: a missing object has no counterpart in the query view, a replaced one has a counterpart that looks different, and the model must decide which view to ground in. Correspondence is thus a sub-step of ECG rather than its goal.

A.3 Reasoning Grounding with MLLMs

LISA drives a segmentation decoder from a special `[SEG]` token (Lai et al., 2024), and LISA++ extends this embedding-as-mask idea to instance-level masks in dialogue (Yang et al., 2023). GLaMM produces interleaved text and masks in a grounded conversation (Rasheed et al., 2024), PixelLM handles several targets at once with a light pixel decoder and a mask codebook (Ren et al., 2024), and GSVA emits multiple segmentation tokens plus a rejection token for multiple or missing referents (Xia et al., 2024). F-LMM reads masks from a frozen MLLM's word–pixel attention (Wu et al., 2025), and InstructSeg unifies referring and reasoning segmentation over images and videos (Wei et al., 2025). All of them ground within a single image, where cross-image cues and viewpoint bias do not arise. Across images, GeM-VG combines CoT reasoning with direct answering in a hybrid reinforcement fine-tuning under a rule-based reward (Zheng et al., 2026b), MiCo builds image triplets and adapts GRPO to learn cross-view consistency (Chen et al., 2025), and other methods inject cross-image reasoning through staged SFT-then-RL training (Zhang et al., 2026a; Jiang et al., 2026; Liu et al., 2026).

A.4 Visual Reinforcement and Policy Learning

Beyond the methods discussed in Section 2, Dr. Seg adopts a perception-oriented distribution-ranked reward (Sun et al., 2026), DPAD anchors intermediate reasoning to the referred target (Yang et al., 2026), and VisPlay and GThinker restructure training data and sampling (He et al., 2026; Zhan et al., 2026). These refine how an outcome or perception reward is shaped for single-image or generic multi-image grounding, whereas SCPO changes what the reward measures: dependence on the true cross-view change evidence relative to a matched nuisance corruption.

B ECG-Bench Data

B.1 Sources and Construction

This subsection gives, source by source, how the samples summarized in Section 3 are obtained.

Procedural simulation. ProcTHOR (Deitke et al., 2022) provides a large collection of interactive indoor residential scenes. DivScene (Wang et al., 2025) substantially expands the diversity of scene types: its 4,614 scenes cover 82 categories, including classrooms, libraries, laboratories, locker rooms, greenhouses, and retail spaces. DenseOurs, our own set of high-density scenes, contributes 1,140 samples from 259 scenes and is described in Appendix B.2. All three are rendered in interactive Unity-based environments: we record a scene as the reference state, programmatically edit one object, and capture the scene again from a different camera pose, varying viewpoint and illumination independently between the two captures. Because the edit is scripted, the changed object, its change type, and its pixel-level mask come directly from the simulator, without manual annotation.

High-realism industrial scenes. ChangeSim (Park et al., 2021) was designed for online scene change detection and records repeated first-person traversals of industrial warehouses that naturally include dust, haze, illumination change, and large viewpoint variation. It provides no language annotation and is therefore not directly compatible with ECG. We first construct reference–query pairs from its repeated trajectories, then retain changed objects that are sufficiently visible and can be uniquely referred to, discarding targets that are too small, heavily occluded, or visually ambiguous;

for each retained target we generate a free-form referring instruction. This converts the change-detection data into the language-guided format of ECG and contributes approximately 2K samples (examples in Appendix B.5).

Real-world revisited scenes (MulOurs). The 1K real-world samples come from four sources. (i) From the multi-temporal videos of SceneDiff (Wu et al., 2026), covering kitchens, supermarkets, living rooms, and bedrooms, we align the two video sequences of each revisited scene, select clear keyframes for every changed object, and transform the provided dense change mask onto the selected RGB frame; we retain only image pairs with a meaningful viewpoint difference and remove near-duplicate observations. (ii) From 3RScan (Wald et al., 2019), which contains repeated RGB-D scans of the same indoor environments in which objects may move or change between scans, we manually annotate the changed instances between revisited scans. After filtering and manual verification, and after removing 100 pairs with large viewpoint differences, these two sources contribute 630 samples. (iii) We select 100 high-quality samples from PASLCD (Galappaththige et al., 2025), which revisits workplace interiors such as lounges, meeting rooms, and cantinas. (iv) We collect 124 pairs of indoor scenes ourselves and manually annotate 270 changed instances. Examples from all four sources are shown in Appendix B.4.

Manual annotation. For samples requiring manual annotation, annotators first identify the changed object and draw a bounding box around it, then write a referring instruction that uniquely identifies the target through its spatial relation to a stable landmark, for example *"the newly placed object on the table to the right"*. We avoid using the target category directly whenever possible, so that a model must resolve the spatial reference rather than rely on simple category matching. We convert the annotated boxes into pixel-level masks with SAM 2 (Ravi et al., 2025), followed by manual visual verification, and discard pairs with ambiguous changes, severe occlusion, or unreliable correspondence between the two views.

Change-type distribution. The 10K samples divide into existence (2,658), missing (2,700), replace (1,478), rotation (1,229), relocation (1,216), and state change (719).

B.2 Dense Simulated Scenes

To make referring grounding harder than in existing procedural scene collections, we build DenseOurs (Section 3), 1,140 samples from 259 dense 3D scenes such as call centers, computer rooms, electronics stores, grocery stores, laboratories, restaurant kitchens, and warehouses, where many objects of similar appearance or of the same category are placed close together. Figure A1 compares it with ProcTHOR-10k (Deitke et al., 2022) and DivScene (Wang et al., 2025). DenseOurs contains far more objects per scene: the median is 120, against 64 for ProcTHOR and 77 for DivScene, and 20% of our scenes contain more than 400 objects, against almost none in the other two. Its scenes are also larger single spaces, with a median floor area of 115 m^2 against 78 m^2 and 48 m^2, and its floor plans include L-shaped and complex outlines in addition to rectangular rooms, giving more varied outlines than DivScene. In such scenes category-level cues rarely single out the referred object, so a model has to resolve the spatial relation in the instruction.

B.3 Instruction Diversity

Figures A2 and A3 analyze the instructions of ECG-Bench: its 10K samples carry 6,656 distinct instructions over a vocabulary of 1,330 words.

Vocabulary. Figure A2 sizes each word by the number of distinct instructions that contain it and colors it by cue type. The instructions state where a change happens and what it is, rather than merely naming an object: the largest word is the spatial term *beside*, followed by change-event terms such as *newly appeared*, *missing*, *replaced*, *removed*, and *rotated*, whereas most object names are small. A model must therefore understand both the change and the spatial relation; recognizing the object category is not enough. The same change is also phrased in many ways, e.g., *missing*, *removed*, *disappeared*, *no longer*, and *absent* for a missing object, and *replaced*, *substituted*, and *took the place* for a replacement, so keyword matching cannot solve the task. Spatial and attribute cues



Figure A1: **DenseOurs compared with existing procedural scene collections.** (a–c) Example scenes from ProcTHOR-10k, DivScene, and DenseOurs (*Ours*). (d–f) Distributions of objects per scene, floor area per scene, and floor-plan shape.

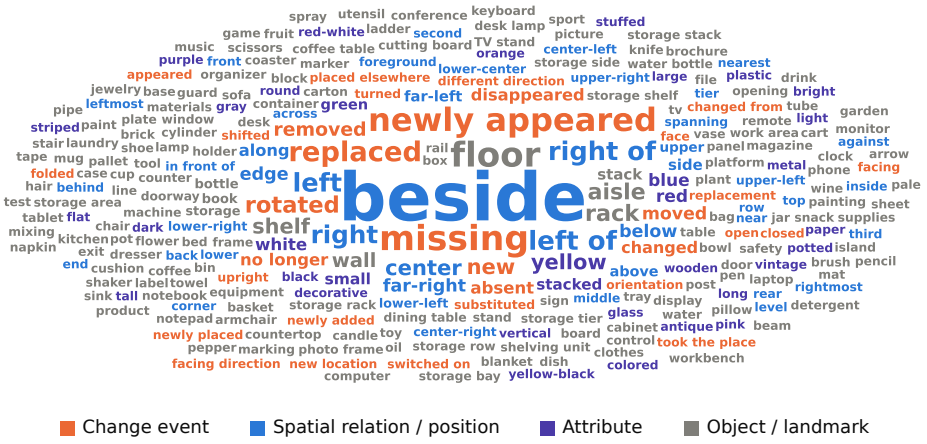


Figure A2: **Instruction vocabulary of ECG-Bench.** Word size is the number of distinct instructions that contain the word; color marks the cue type: change event, spatial relation or position, attribute, or object and landmark.

are fine-grained: beyond *left of* and *right of*, the instructions use positions in the frame (*far-right*, *lower-left*), ordinals (*second*, *third*, *leftmost*), and appearance attributes (*white*, *blue*, *wooden*).

Referring types. Every instruction names the change event and gives a spatial anchor, so Figure A3 distinguishes referring types by how the target is singled out: the spatial reference (a single anchor, an anchor plus a region of the image, multiple anchors, or ordinal counting) crossed with the semantic cue (a generic reference, an appearance attribute, or the target category), giving twelve types. The sources are complementary. Each simulated source concentrates on one ability: 91% of ProcTHOR samples use a single anchor with a generic reference, 94% of DenseOurs samples use ordinal counting, and 65% of ChangeSim samples combine an anchor with an image region, with 76% carrying an attribute such as color. The four real-world sources of MulOurs are more compositional: multiple anchors account for 47–87% of their samples, and the target category is named in 29–50%. Across the benchmark, single-anchor references make up 57%, and the remaining 43% require compositional reasoning over several candidates: anchor plus image region (16%), multiple anchors (16%), and ordinal counting (11%). Change types are balanced overall, with existence 27%, missing 27%, replace 15%, relocation 12%, rotation 12%, and state change 7%, while individual sources lean toward particular types, e.g., state change is the largest type in ProcTHOR and relocation in SceneDiff.

B.4 Real-World Examples

Figures A4–A6 show thirty real-world samples of ECG-Bench, ten per figure, each figure covering all six change types. The first two span residential, public, and commercial scenes such as kitchens, offices, a bus, a grocery store, a gym, and a pool table; the third shows samples from PASLCD (Galappaththige et al., 2025), in workplace interiors such as lounges, meeting rooms, and a cantina. Each sample shows the reference and query views, the ground-truth mask on the view that holds the target, the change type, the scene, and the instruction. The instructions locate the target relative to landmarks ("between the red water bottle and the black carton", "just to the right of the leftmost blue chair"), and missing targets are annotated in the reference view.

B.5 High-Fidelity Simulated Examples

Figure A7 shows ten samples converted from ChangeSim (Park et al., 2021) as described in Section 3, covering existence, missing, rotation, and replace changes in industrial warehouses. Several pairs are captured under low light or dust, as marked in the scene label, and the instructions anchor the target to warehouse landmarks such as floor stacks, shelf rows, partitions, and floor markings.

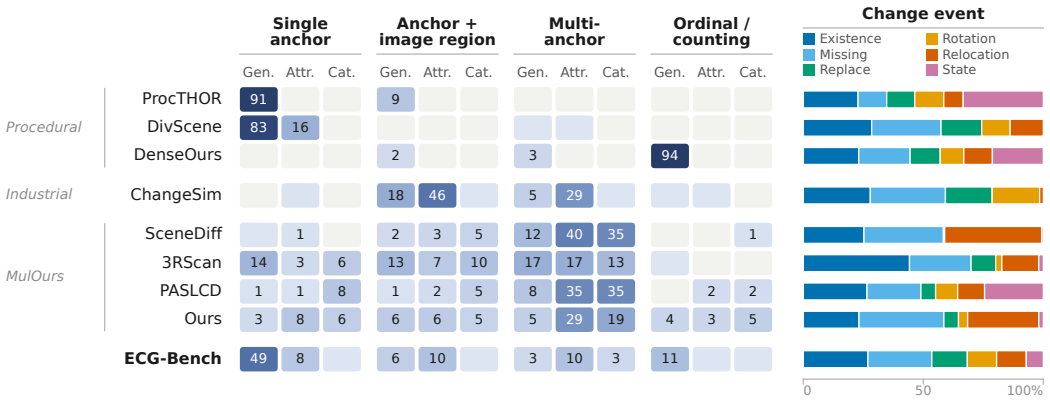


Figure A3: **Referring types and change types by source.** *Left:* share (%) of each source's samples in each of the twelve referring types, which cross the spatial reference (single anchor, anchor + image region, multi-anchor, ordinal/counting) with the semantic cue (Gen.: generic, Attr.: attribute, Cat.: category). Sources are grouped into procedural, industrial, and real-world (MulOurs) scenes, and the bottom row aggregates all samples of ECG-Bench. *Right:* change-type distribution of the same samples.

C CHANGE-COT ANNOTATION

The cross-view reasoning traces used for CoT warmup (Section 4.1) are annotated automatically. Figure A8 shows the pipeline. For each sample we give GPT-5 (OpenAI, 2025) the reference image, the query image, and the referring instruction, and prompt it in four steps that mirror the comparison order the policy is meant to learn. GPT-5 first parses the instruction into the spatial relation or semantic landmark that anchors the target rather than into the target category; it then captions that anchored local region separately in each of the two views; and it compares the two captions to infer what changed there, from which the final Change-CoT answer is produced. Anchoring both captions to the same region before comparing them is what keeps a trace grounded: each view is described on its own terms, so the chain reasons over an observed difference instead of narrating a comparison it never made.

Every generated trace is then verified against the simulator ground truth on three criteria — whether the named object matches the annotated open-set category, whether the target is grounded in the correct view and region, and whether the inferred change is consistent with the scripted edit. Traces failing any criterion are regenerated rather than kept, which removes the hallucinated chains that supervised fine-tuning would otherwise reinforce.

Example traces. Figure A9 shows traces produced by this pipeline on eight warehouse and household samples, covering missing, replace, existence, rotation, and state changes. Every trace follows the order of Section 4.1: it describes the anchored region in the reference view, then the same region in the query view, states the comparison, and closes with one sentence that names the changed object. The traces reuse the landmark of the instruction, such as the white door panel, the sofa, or the desk, to tie the region across the two views, and several explicitly discount observation differences — a darker query view, a viewpoint shift, or camera motion — before attributing the discrepancy to the object. For the two missing cases, the conclusion names an object present only in the reference view, and the mask is annotated there.

D COT-GUIDED CROSS-VIEW DECODER: DESIGN AND ANALYSIS

This section details the CoT-Guided Cross-View Decoder of Eq. 1 and its training objective, contrasts it with SAM-like decoders, and visualizes the cues it fuses.

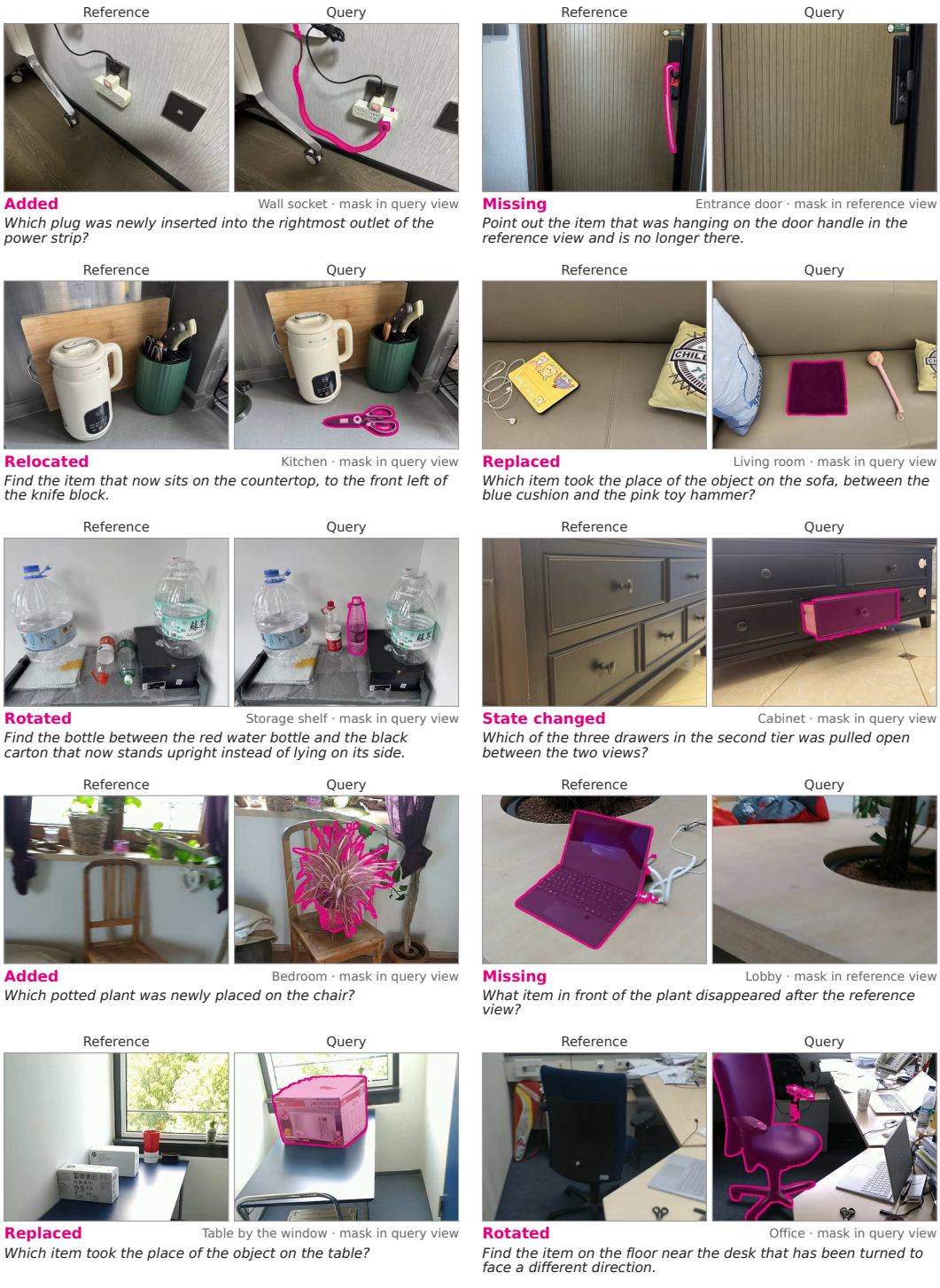


Figure A4: **Real-world samples of ECG-Bench (1/3).** Each sample shows the reference and query views with the ground-truth mask on the view that holds the target, the change type, the scene, and the instruction.

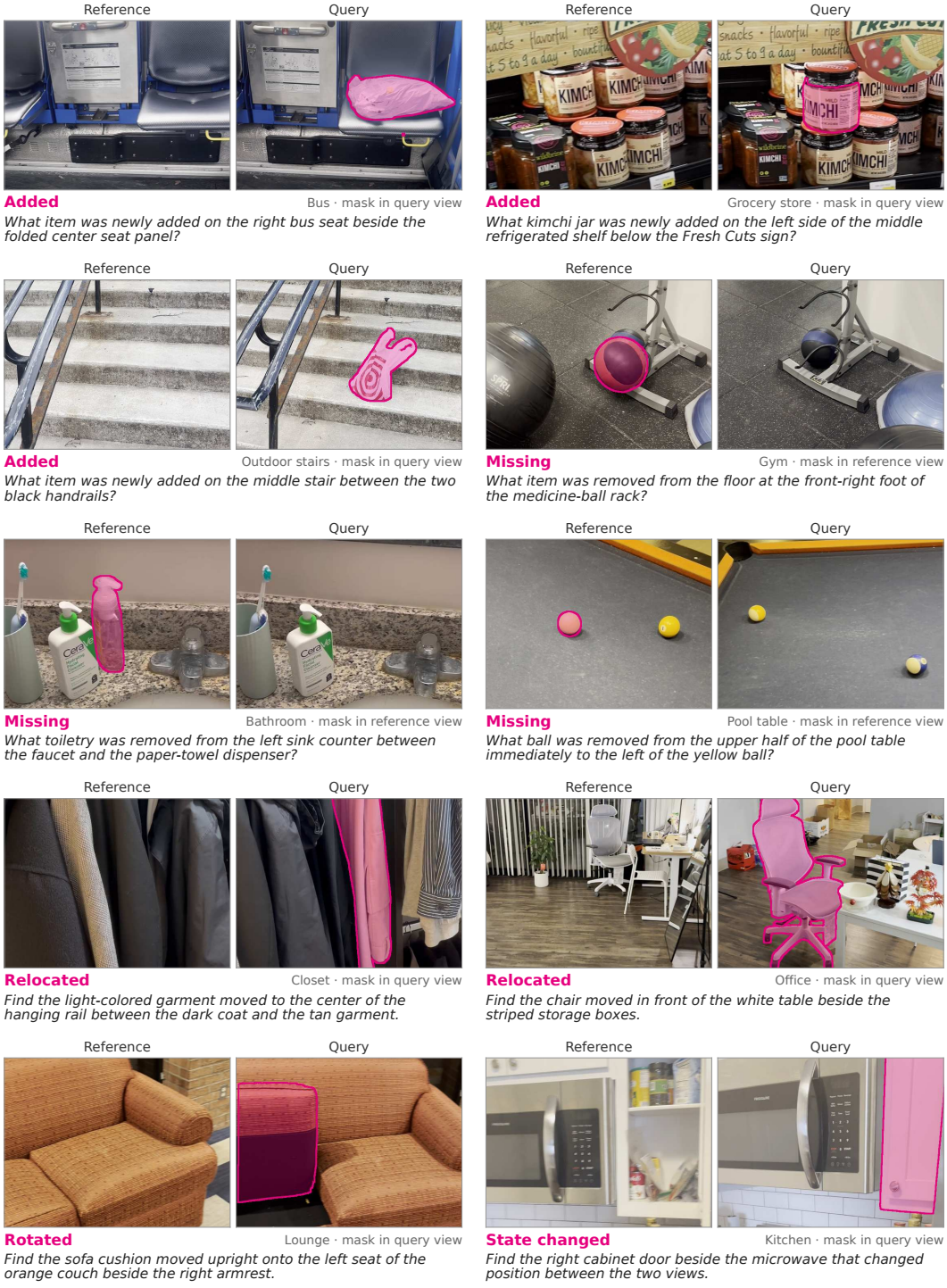


Figure A5: **Real-world samples of ECG-Bench (2/3).** Continued from Figure A4.

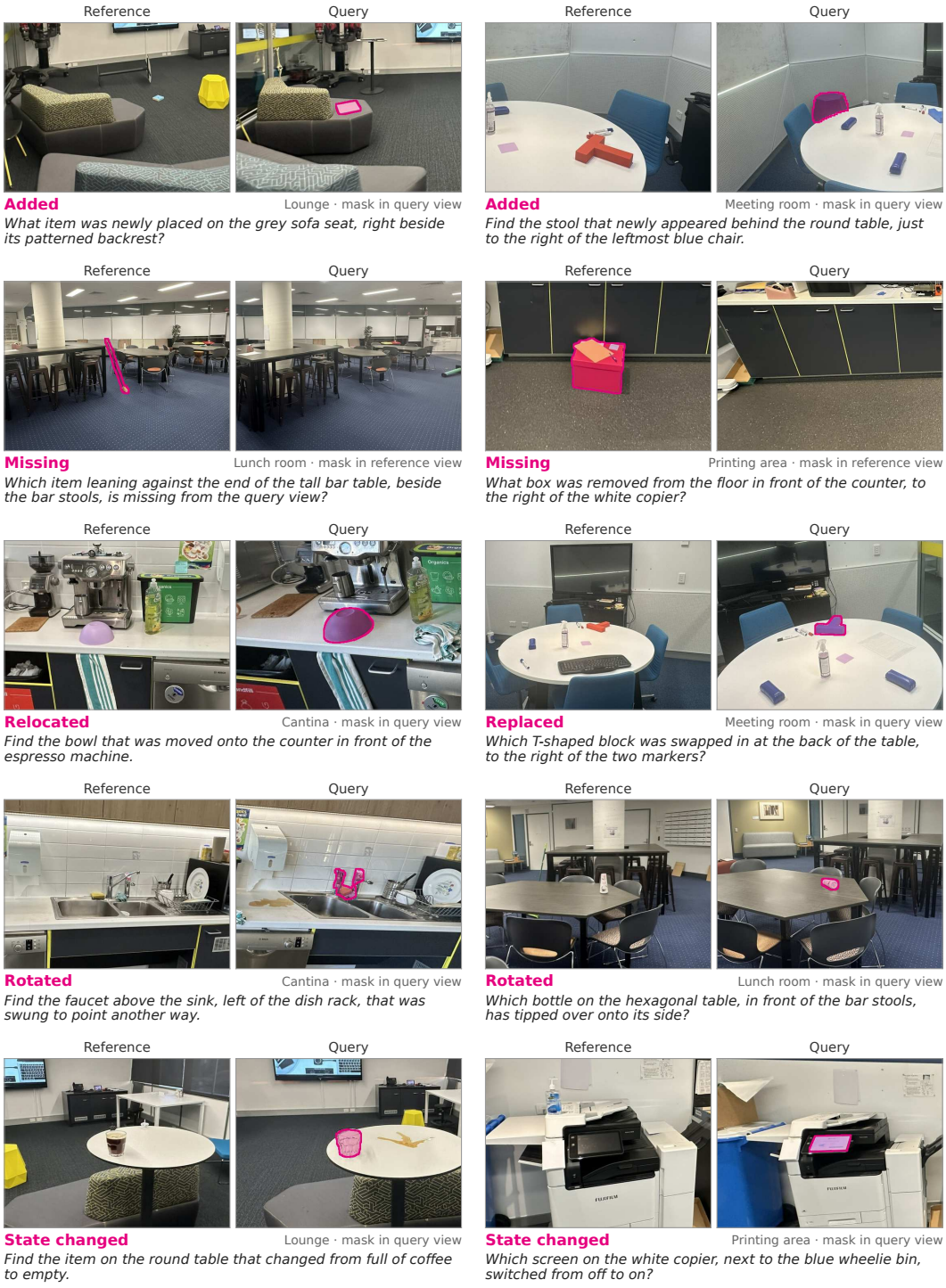


Figure A6: **Real-world samples of ECG-Bench (3/3): PASLCD.** Samples drawn from PASLCD (Galappaththige et al., 2025), in the format of Figure A4.

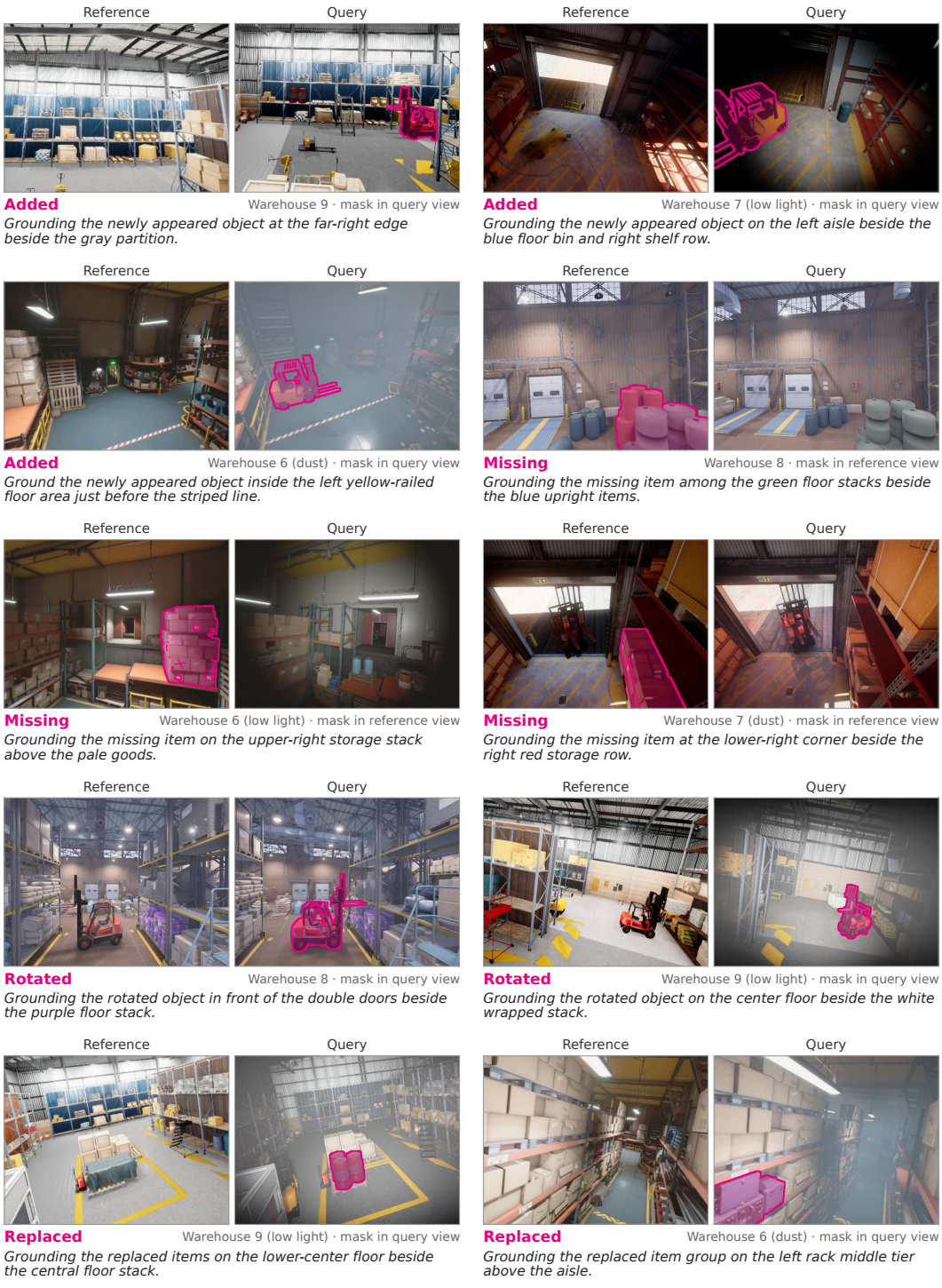


Figure A7: **High-fidelity simulated samples of ECG-Bench from ChangeSim.** Samples in the format of Figure A4; the scene label marks the warehouse and, where present, the low-light or dust condition.

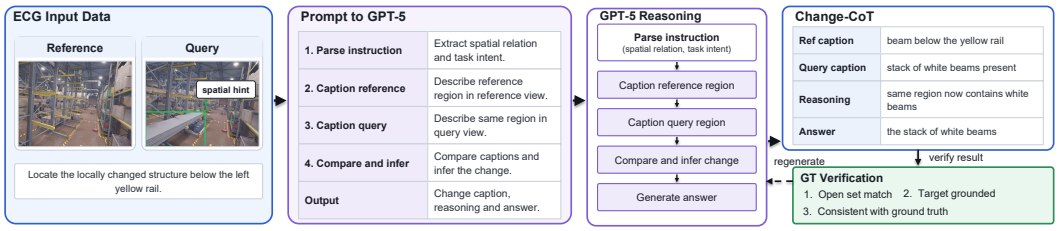


Figure A8: **GPT-5-assisted Change-CoT annotation pipeline for ECG.** Given a reference image, a query image, and a referring instruction, GPT-5 first parses the spatial relation or landmark in the instruction, then captions the anchored local region in the two views, compares the two descriptions to infer the change, and generates the final Change-CoT answer. The generated result is further verified against simulator ground truth to ensure consistency in change type and target grounding.

D.1 Architecture

Inputs and view selection. $Z_r, Z_q \in \mathbb{R}^{N\times D}$ are the visual tokens of I_r and I_q from the MLLM's vision encoder, so no separate SAM image encoder is needed. A *view head* reads the CoT summary c (defined below) and the `[change]` embedding s and predicts $p = P(\text{target in } I_r)$; *anchor selection* then sets the anchor view $a = r$ if $p > 0.5$ and $a = q$ otherwise, and the mask is decoded on the token grid of I_a. The predicted anchor is used in training as well as at inference, so the decoder is trained under the view errors it meets at test time. With $\bar{a}$ the other view, a projection shared by both views maps the tokens to the decoder width $d = 256$, $x_a = \mathrm{Lin}(\mathrm{LN}(Z_a))$ and $x_{\bar{a}} = \mathrm{Lin}(\mathrm{LN}(Z_{\bar{a}}))$.

Anchor-guided cross-view alignment and change fusion. Token coordinates in $[-1, 1]^2$ pass through a Linear–Tanh–Linear encoder to give positional encodings $\mathrm{PE}_a, \mathrm{PE}_{\bar{a}}$. A single 8-head cross-attention layer (head width 32, without FFN or residual) aligns the other view to the anchor grid,

$$\tilde{x} = \mathrm{MHA}\big(\mathrm{LN}(x_a + \mathrm{PE}_a),\ \mathrm{LN}(x_{\bar{a}} + \mathrm{PE}_{\bar{a}}),\ x_{\bar{a}}\big), \tag{6}$$

with arguments in (query, key, value) order; the value carries neither positional encoding nor normalization, so $\tilde{x}$ mixes features of the other view only. The anchor, the aligned view, and three difference terms are then fused with an anchor residual, and a learned view embedding $e_v \in \{e_{\text{ref}}, e_{\text{que}}\}$ marks which view is the anchor:

$$F = \mathrm{MLP}_{\text{fuse}}\big([\,x_a;\ \tilde{x};\ x_a - \tilde{x};\ |x_a - \tilde{x}|;\ x_a \odot \tilde{x}\,]\big) + x_a + e_v, \tag{7}$$

where $[\cdot;\cdot]$ is channel concatenation and $\mathrm{MLP}_{\text{fuse}}$ maps $5d \to 2d \to d$ with GELU and ends in LayerNorm. Because $\tilde{x}$ lives on the anchor grid, F compares corresponding content rather than identical pixel positions, which is what makes it usable on unregistered views.

CoT grounding. The hidden states $h_1, \ldots, h_T$ of the response tokens before `[change]`, teacher-forced during training, are pooled by a static learned query $u \in \mathbb{R}^D$ that scores normalized states but averages raw ones:

$$\alpha_i = \mathrm{softmax}_i\ \big(u^\top \mathrm{LN}(h_i)\big), \qquad c = \textstyle\sum_i \alpha_i h_i. \tag{8}$$

The grounding logit of anchor token j is a scaled cosine similarity with this summary,

$$g_j = \sqrt{d}\,\cos\big(\rho_v(F_j),\ \rho_c(c)\big) + b, \tag{9}$$

where ρ_v, ρ_c are LN–Linear–GELU–Linear projections, the scale $\sqrt{d} = 16$ is fixed, and b is a learnable scalar initialized to 0. The map $G = \sigma(g)$ is supervised by the target mask (Appendix D.2), so it localizes the referent the chain has committed to, independently of the difference signal in F.

CoT injection and mask decoding. The summary conditions decoding through a spatial path gated by G and a semantic path on the `[change]` embedding, ϕ and ψ, whose last layers are zero-initialized so that training starts from a plain `[change]`-token decoder:

$$F'_j = F_j + \sigma(g_j)\,\phi(c), \qquad s' = s + \psi(c). \tag{10}$$

Missing barrel in the aisle

Instruction Grounding the missing item in the center aisle beside the blue floor cylinder.



CoT In the central cluster of barrels located in the aisle, there is a large blue barrel with white horizontal stripes situated between a green barrel and a brown barrel. The specific spot in the central cluster of barrels between the green and brown barrels is empty, revealing the gray concrete floor underneath. Comparing the two images shows that the blue and white striped barrel visible in the Reference image has been removed in the Query image, leaving that space vacant. **Therefore, the missing target is a blue barrel with white horizontal stripes.**

Rack item replaced by a sack stack

Instruction Grounding the replaced item on the upper-right shelf.



CoT An upper rack item is visible at the target position in the reference. The query shows the white wrapped sack stack at that position. The upper-right item is replaced by the sack stack. **Therefore, the replacement target is the white sack stack.**

Newly appeared red containers

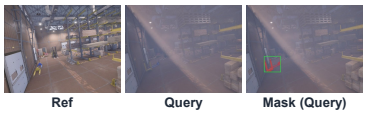
Instruction Grounding the newly appeared item on the left floor below the wall fire-safety panel.



CoT The blue floor below the wall fire-safety panel is empty. Two red rectangular containers stand on that floor below the panel. The fire-safety panel and rear wall anchor the floor position; the red containers are newly added. **Therefore, the newly appeared target is the pair of red rectangular containers below the safety panel.**

Rotated yellow pallet jack

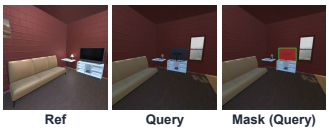
Instruction Grounding the rotated item on the lower-left floor beside the white door panel.



CoT In the lower-left area near the white door panel, there is a yellow pallet jack standing upright with a blue object resting on its forks. In the corresponding lower-left area near the white door panel, the yellow pallet jack is now lying flat on the ground surface, oriented horizontally against the wall. The yellow pallet jack has moved from an upright operational position to a horizontal, tipped-over position on the floor. Stable shelving, wall, container, and floor landmarks align the corresponding area, ruling out a camera-only explanation. **Therefore, the rotated target is the yellow pallet jack lying flat on the floor.**

Rotated television beside the sofa

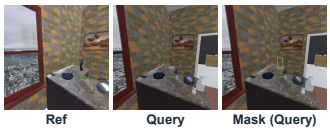
Instruction Please mark the item to the right of the sofa that was turned to a different orientation between the two views.



CoT In the Reference image, the target television to the right of the sofa has orientation 90.0 degrees. The nearby sofa is unchanged. In the Query image, the same television to the right of the sofa has orientation 270.0 degrees. The unchanged sofa identifies the same object across the shifted views. The query view is darker, but the stable landmark still aligns the referred region. Its location is maintained while its orientation changes, so the television was rotated. **Therefore, the rotated object is the television to the right of the sofa.**

Wine bottle replaced by a statue

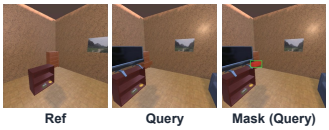
Instruction What item on the countertop replaced the object shown there in the Reference image?



CoT In the Reference image, a wine bottle occupies the location on the countertop, identified by the unchanged countertop. In the Query image, a statue occupies that same anchored location instead. The unchanged countertop aligns the location despite the viewpoint shift. The query view is darker, but the stable landmark still aligns the referred region. The wine bottle is absent from its former position and the statue has taken its place, so this is a replacement. **Therefore, the replacement object is the statue on the countertop.**

Opened drawer in the dresser

Instruction Please highlight the item in the dresser that was opened between the two views.



CoT In the Reference image, the target drawer in the dresser is closed. The surrounding dresser remains the stable anchor. In the Query image, the same drawer in the dresser is open. The unchanged dresser links the referred region across the two views, so the comparison is between the same spatially anchored object. The drawer changes state from closed to open. **Therefore, the changed object is the drawer in the dresser.**

Missing laptop on the desk

Instruction Please mark the item on the desk that is missing from the Query image.



CoT In the Reference image, the target laptop is visible on the desk, and the unchanged desk fixes the referred location. In the Query image, the same region on the desk remains identifiable from the desk, but the laptop is no longer present. The unchanged desk aligns the referred region across the two views. The laptop is present only in the Reference image, so it is the missing item. **Therefore, the missing object is the laptop on the desk.**

Figure A9: **Examples of Change-CoT traces** generated by the pipeline of Figure A8. Each case shows the instruction, the reference and query views, the annotated mask labelled with the view that contains it, and the trace, whose final sentence (red) names the changed object.

s' then modulates F' through a bounded FiLM that acts per channel and is identical at every location,

$$[\gamma, \beta] = \mathrm{MLP}\big(\mathrm{LN}(s')\big), \qquad F'' = F' \odot \big(1 + 0.5 \tanh \gamma\big) + \beta, \tag{11}$$

which keeps the scale in $[0.5, 1.5]$ and leaves β unbounded. F'' is refined on the token grid by Conv3×3–GN–GELU and a ResBlock, then upsampled by three $\times 2$ stages, each a bilinear upsampling followed by Conv3×3–GN–GELU and a ResBlock, with channels $256 \to 128 \to 64 \to 32$; each ResBlock is two Conv3×3+GN layers with a residual connection and a final GELU. A Conv3×3–GELU–Conv1×1 head outputs the mask logits, which are bilinearly resized to a fixed 192×256.

D.2 Training Objective

With M^* the ground-truth mask, the CoT-SFT warmup minimizes

$$\mathcal{L} = \mathcal{L}_{\text{SFT}} + \mathcal{L}_{\text{dec}}, \qquad \mathcal{L}_{\text{dec}} = \mathcal{L}_{\text{mask}}(M, M^*) + \lambda_g \, \mathcal{L}_{\text{ground}}(g, M^*) + \lambda_p \, \mathcal{L}_{\text{view}}\big(p, \mathbb{1}[a^* = r]\big), \tag{12}$$

where a^* is the ground-truth anchor view, i.e., the view that contains the target. $\mathcal{L}_{\text{mask}}$ is a binary cross-entropy plus Dice loss on the mask logits; $\mathcal{L}_{\text{ground}}$ is a class-balanced binary cross-entropy plus Dice loss on the grounding logits g, bilinearly upsampled to 192×256, with $\lambda_g = 0.25$; and $\mathcal{L}_{\text{view}}$ is a binary cross-entropy on the view prediction, class-weighted to compensate for the imbalance between reference- and query-view targets, with $\lambda_p = 0.1$ since the view is comparatively easy to predict. The mask and grounding terms back-propagate through c and s into the CoT hidden states, so the chain is trained to be *groundable*, not only fluent. During RL the decoder is not frozen: it keeps being trained with the same loss, scaled to $0.25\,\mathcal{L}_{\text{dec}}$, and the view prediction also enters the policy reward through R_{view} (Section 4.2).

D.3 Comparison with SAM-like Decoders

SAM-like decoders used in reasoning segmentation (Lai et al., 2024; Liu et al., 2025b; 2026; Ravi et al., 2025) segment one image from one prompt, a box or a single mask-token embedding (Table A1). They do not model the relation between the reference and the query view, which is what defines an ECG target, and they perform poorly on ECG: in Table 3 every SAM-like interface stays at or below 28.2 mIoU, including the one that sees both views, against 35.6 for the CoT-Guided Cross-View Decoder.

Table A1: **SAM-like decoders vs. the CoT-Guided Cross-View Decoder.**

	SAM-like decoder	**CoT-Guided Cross-View Decoder (Ours)**
Visual input	one image	reference and query views
Cross-view relation	not modelled	alignment and difference features
Condition	a prompt or one token	`[change]` and CoT (semantic and spatial paths)
Reasoning → mask	through one token only	also through the supervised map G
Target view	the input image	either view, chosen by a view head
Image encoder	separate SAM encoder	MLLM vision tokens

D.4 Qualitative Analysis

Figures A10 and A11 visualize the two cues that the CoT-Guided Cross-View Decoder fuses (Appendix D.1), on simulated and real-world cases spanning existence, missing, replace, rotation, and relocation changes. The cross-view change response carries usable evidence when the two views share enough appearance structure, but it degrades exactly where ECG is hard: when the referred object is small or blurred, and when the scene contains several instances of the same category, its strongest activations spread over unchanged clutter and no longer isolate the target. The CoT grounding response is complementary. Because the chain has already committed to a referent and to the landmark that anchors it, it supplies the semantics the difference signal lacks and stays concentrated on the referred object in precisely those cluttered cases. The last two columns make the effect explicit: without CoT injection the mask drifts to a neighbouring object or covers only part of the target, while with it the prediction snaps to the referred object — the qualitative counterpart of removing CoT conditioning in Table 3.

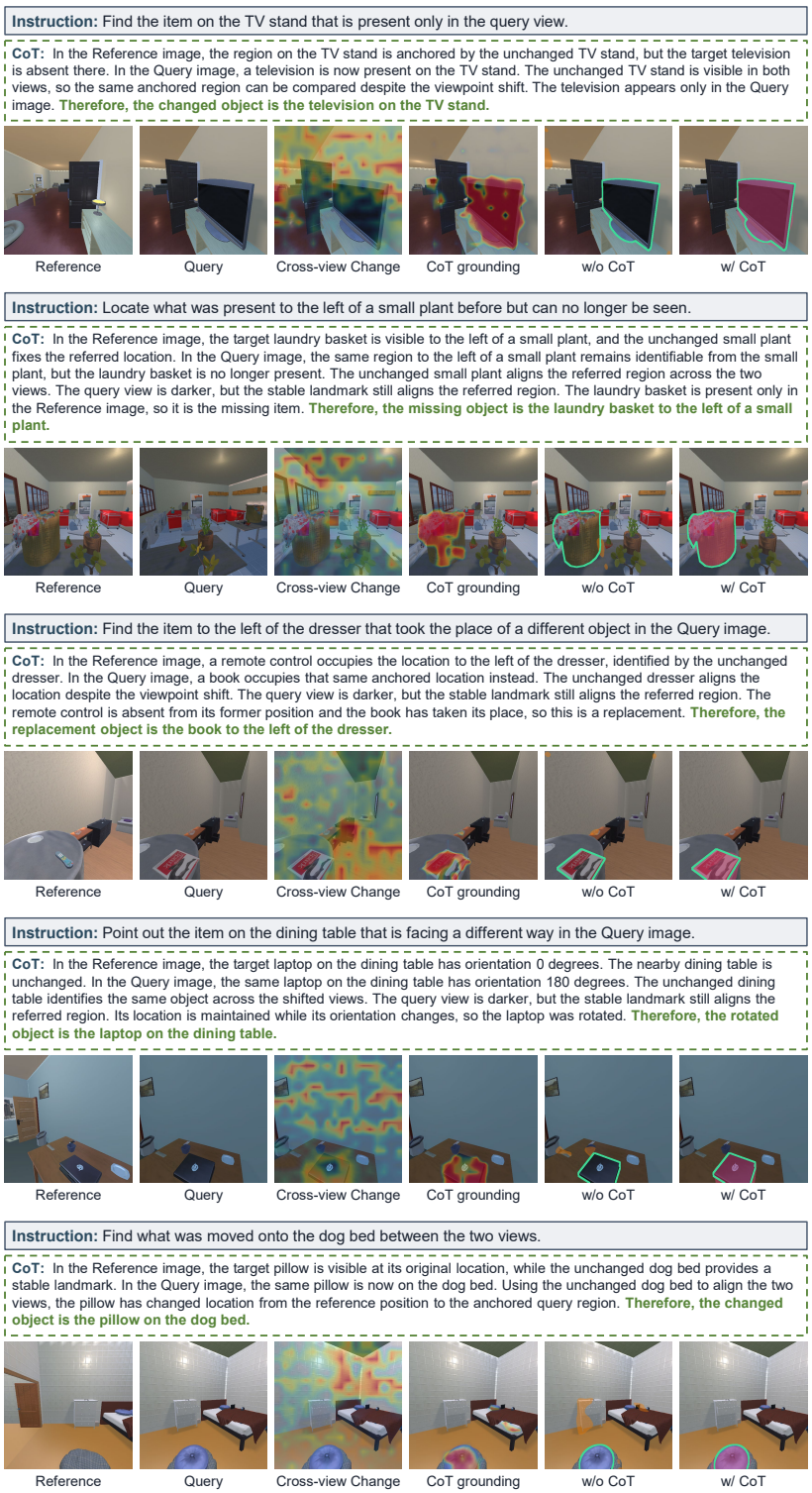


Figure A10: **Grounding cues used by the CoT-Guided Cross-View Decoder on simulated cases.** Each case shows the instruction and the generated chain, followed by six panels: the reference view, the query view, the cross-view change response, the CoT grounding response, and the decoded mask *w/o CoT* and *w/ CoT* injection. The change response alone is reliable only when the two views are visually close, while the CoT response stays centred on the referred object under clutter and repeated categories.

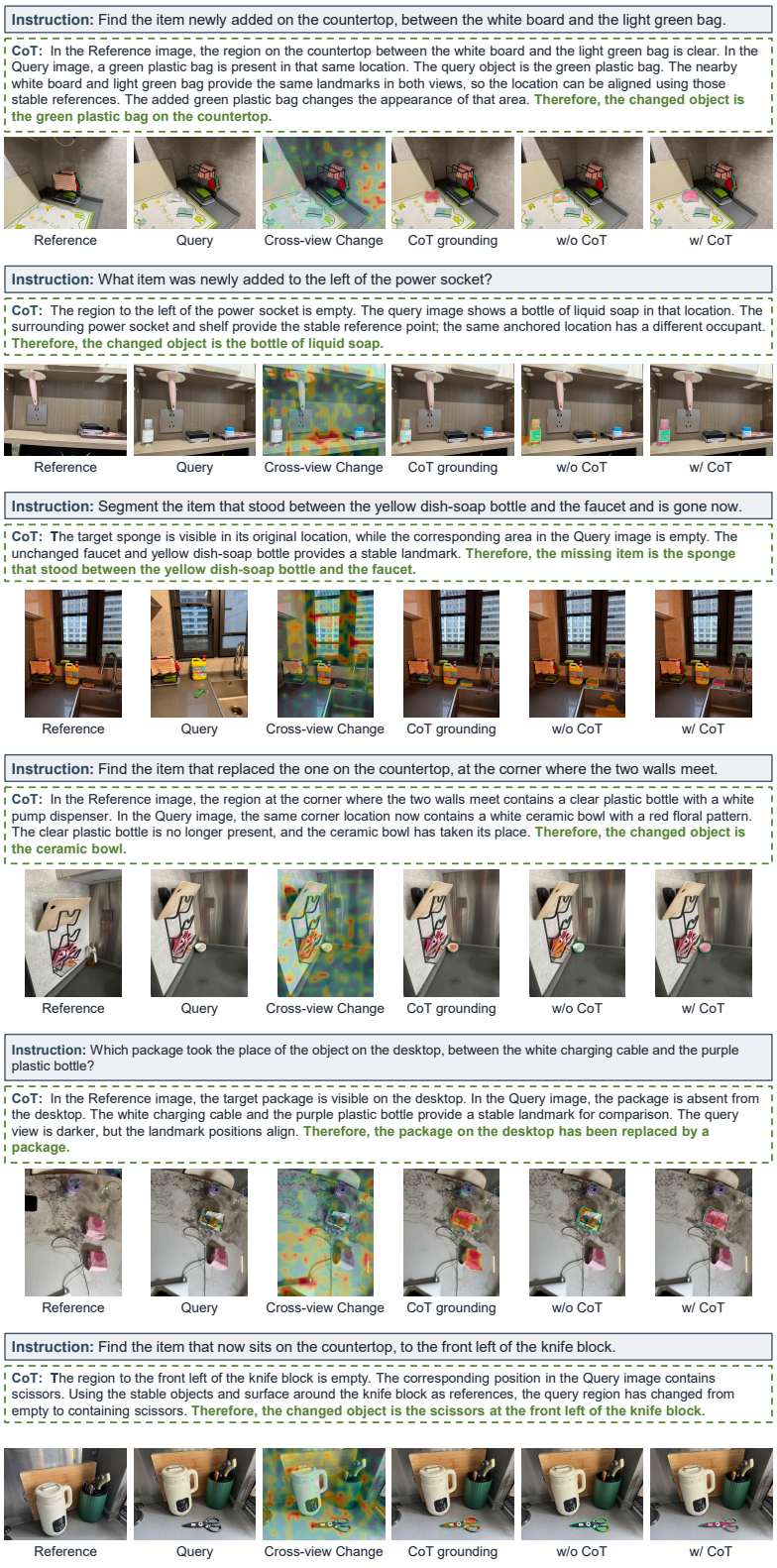


Figure A11: **Grounding cues used by the CoT-Guided Cross-View Decoder on real-world cases.** Columns are as in Figure A10.

E EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

E.1 BASELINES

This subsection details the three groups of baselines in Table 2. All post-trained baselines are trained on ECG-50K.

Training-free. No off-the-shelf MLLM emits masks for ECG, so rather than a naive prompt we give each model an ECG-specific reasoning pipeline: it describes the instruction-relevant content of each view separately, compares the two descriptions into an explicit change statement, and then names the target together with the view that contains it; SAM 2 (Ravi et al., 2025) segments the named target in that view. We run it with InternVL3-8B/78B (Zhu et al., 2025), Qwen2.5-VL-7B/72B (Bai et al., 2025), LLaVA-OV-72B (Li et al., 2024), and Qwen3.5-9B (Qwen Team, 2026).

Post-training SFT. LISA (Lai et al., 2024), PixelLM (Ren et al., 2024), and GLaMM (Rasheed et al., 2024) keep their own mask decoders, are re-implemented on the same backbone with two-image input, and are trained on ECG-50K with grounding supervision only; LISA + CoT additionally consumes our reasoning traces.

Post-training RL. Visual-RFT (Liu et al., 2025d) and VisionReasoner (Liu et al., 2026) apply free-form CoT RL with verifiable rewards directly to the base policy, while VisionReasoner† is the same recipe restarted from our CoT-SFT checkpoint. On top of that checkpoint we further compare GeM-VG (Zheng et al., 2026b), which optimizes for multi-image grounding, DAPO (Yu et al., 2025), which stabilizes the optimizer, and the two perception-aware objectives closest to ours, PAPO (Wang et al., 2026b) and VPPO (Huang et al., 2026). These four modify only the RL objective, so they run on ECG-Grounder and share its initialization, grounding reward, rollout budget, and schedule; the comparison therefore isolates the learning signal.

E.2 ABLATION ON THE POST-TRAINING PIPELINE

Table A2: **Ablation on the post-training pipeline** (mIoU, %). Every row uses ECG-Grounder and ECG-50K and differs only in the two post-training stages. *Stage 1*: answer-only supervision (*Direct*) or supervision on the cross-view reasoning traces (*CoT*); *Stage 2*: the RL objective.

Stage 1 (SFT)	Stage 2 (RL)	Appearance			Geometric		Functional	AVE
		Existence	Missing	Replace	Rotate	Relocate	State	
–	GRPO	14.3	3.1	15.8	14.6	12.2	31.3	15.2
Direct	–	19.4	18.3	21.5	19.8	15.4	43.2	22.9
Direct	GRPO	20.1	19.7	22.1	21.8	17.1	46.3	24.5
CoT	–	26.5	22.7	31.5	25.5	22.3	53.2	30.3
CoT	GRPO	28.6	26.9	32.0	29.7	21.9	56.2	32.6
CoT	**SCPO (Ours)**	**33.8**	**27.1**	**36.2**	**35.6**	**23.4**	**57.7**	**35.6**

Table A2 separates the two post-training stages, on the Qwen2.5-VL-7B backbone. RL on the base policy reaches 15.2, below answer-only SFT (22.9), and adding GRPO on top of that SFT gives 24.5. Replacing answer-only supervision with reasoning traces is worth far more than either: CoT-SFT alone reaches 30.3, +7.4 over *Direct*. The revealing comparison is what an RL stage adds from this checkpoint. GRPO gains +2.3 (30.3 → 32.6), whereas SCPO gains +5.3 from the identical checkpoint, more than twice as much, with the widest margin on *Rotate* (+10.1 against +4.2). An outcome reward can refine a policy that already writes a cross-view chain, but most of the remaining headroom lies in *how* the chain uses the visual evidence — which is what the selective reward acts on.

E.3 SELECTIVITY UNDER OBSERVATION SHIFTS

We probe each checkpoint with the three conditions of Section 4.2: the intact pair (P^1), the change-critical region, i.e., the ground-truth changed object and the landmark named in the instruction, filled

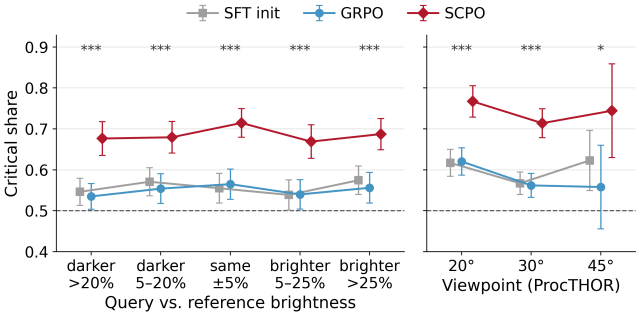


Figure A12: **Selectivity under observation shifts.** Critical share of three checkpoints, by query–reference brightness ratio (left; 150 samples per bin, preserving the data-source mix) and by yaw difference on ProcTHOR (right; 20 samples at $45°$). The dashed line at 0.5 marks no selectivity.

with the image mean colour (P^2), and a fill of the same shape moved elsewhere with at most 5% overlap with the change-critical region (P^3). Averaging Jensen–Shannon divergence over the 20% highest-entropy tokens of the P^1 chain, the *critical share* $\mathrm{JS}(P^1, P^2)/[\mathrm{JS}(P^1, P^2) + \mathrm{JS}(P^1, P^3)]$ equals 0.5 when hiding the change matters no more than hiding anything else. Figure A12 bins samples by the brightness ratio between the two views and, on ProcTHOR, where viewpoint is not tied to the data source, by yaw difference. The CoT-SFT and GRPO checkpoints sit just above chance (0.53–0.58 across brightness bins) and their curves overlap: an outcome reward adds no selectivity to the chain. SCPO lifts the share to 0.67–0.71 in every brightness bin and to 0.71–0.77 across yaw differences, and it does not erode as the shift grows: every brightness bin stays within 0.05 of the matched-lighting one, and the $45°$ bin is no lower than $30°$, although its interval is wide with only 20 samples. Since R_{sel} optimizes this contrast, a higher share is expected; what the figure adds is that it holds where illumination and viewpoint differences, the nuisances ECG is built around, are largest.

F TOKEN ENTROPY AND VISUAL DEPENDENCE ALONG THE REASONING CHAIN

Vision-PMI. Token entropy says where a chain is uncertain, but not whether that uncertainty is resolved by looking. We therefore also measure, for each reasoning token, the pointwise mutual information between the token and the visual input, in nats,

$$\mathrm{PMI}_t = \log \pi_\theta(y_t \mid y_{<t}, q, I) - \log \pi_\theta(y_t \mid y_{<t}, q, \varnothing)\,, \qquad I = (I_r, I_q), \tag{13}$$

where $\varnothing$ is a blank image of the same shape. The two terms come from two teacher-forced passes over an identical token sequence — the prompt followed by the greedily decoded chain, with the `[change]` token excluded — so the passes differ only in the pixel values, which are zeroed in the second while the image grid is left unchanged. A token with $\mathrm{PMI}_t \approx 0$ is therefore produced from the language context alone, whatever its entropy.

What the chains look like. Figure A13 tracks both quantities over the reasoning chain. High-entropy positions are not spread uniformly: they concentrate where the chain names the target and decides the change attribute — the two decisions that can only be settled from the images. Once the referent is fixed, both curves fall sharply, marking the transition from exploration to convergence. Vision-PMI refines this picture by separating, among the uncertain positions, those whose prediction actually moves with the visual input. Entropy therefore locates the decision points of a chain and Vision-PMI identifies which of them are visually grounded, which is the empirical basis for measuring R_{sel} at the top-K highest-entropy tokens rather than averaging over the whole chain, and is consistent with the -0.5 AVE cost of removing entropy selection in Table 4.

G ADDITIONAL QUALITATIVE RESULTS

Figures A14–A17 extend the comparison of Figure 6 to twelve further cases, three from procedural simulation and nine from real-world and high-fidelity simulated scenes, that together cover all six change types. The failures of GRPO and PAPO recur across change types and domains: they name another object at the referred location instead of the changed one (the bread instead of the missing

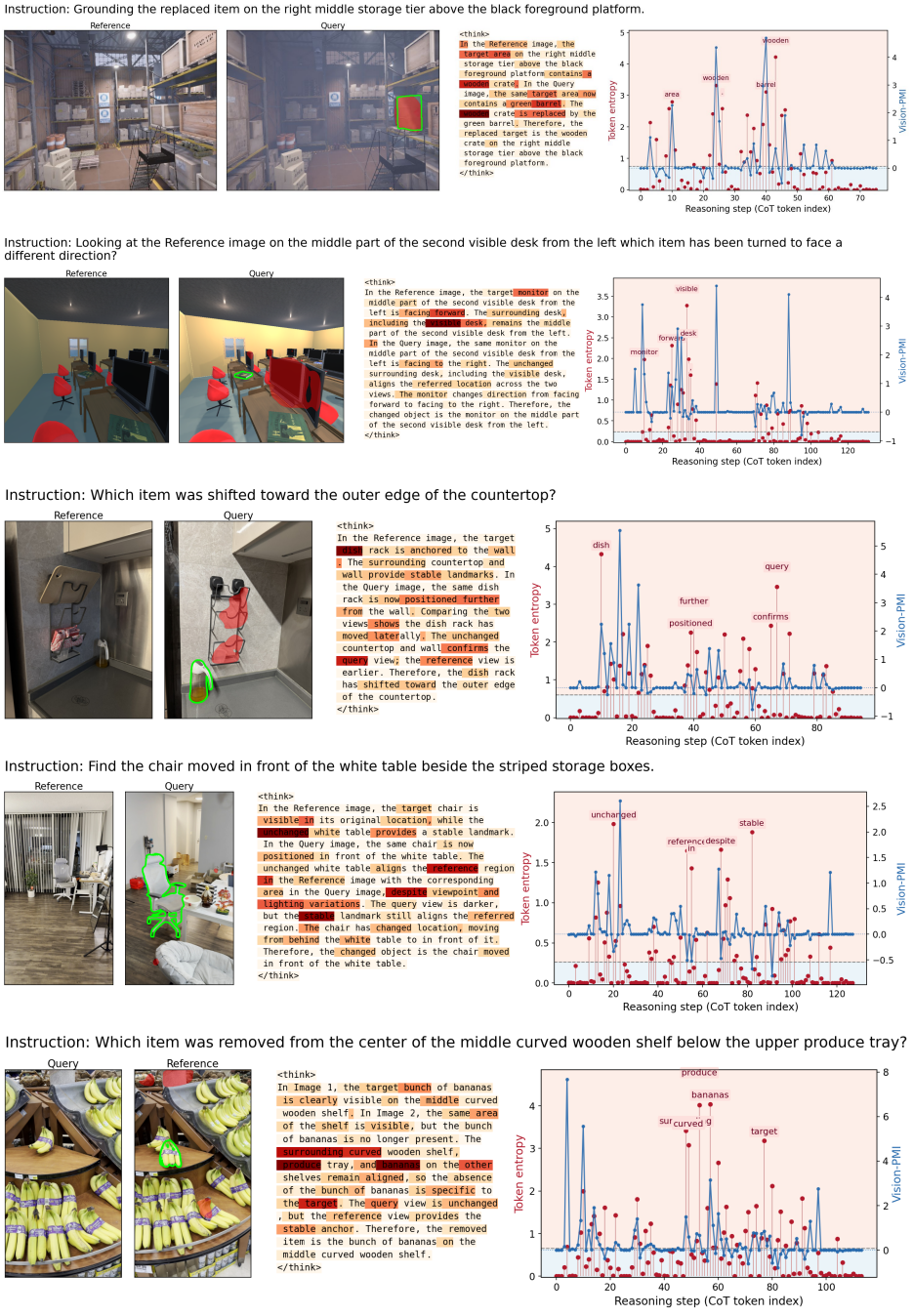


Figure A13: **Token entropy and Vision-PMI along the chain.** For each case we show the reference and query views with the predicted mask, the generated chain with tokens shaded by entropy, and the per-token entropy (left axis) together with Vision-PMI (right axis, Eq. 13), against the reasoning step. Peaks are annotated with the corresponding token and fall on target naming and change-attribute decisions.

bottle, the pillow instead of the massager that now stands upright), return the landmark that the instruction uses as an anchor (the house plant, the pump bottle, the round table), or restate the question without committing to a referent, and GRPO's chains again carry filler such as "the query view is darker". Chain and mask can also come apart: GRPO names the rotated sign but masks elsewhere, while PAPO reaches 0.67 IoU on the closed microwave with a chain that reverses the change, the kind of right-mask, wrong-reasoning rollout that a grounding reward alone would reinforce (Section 4.2). SCPO keeps the landmark as an anchor, contrasts the two views explicitly ("visible in Image 1 but absent in Image 2"), and names the changed object in all twelve cases, with the highest IoU in each. On real-world scenes its masks remain partial for thin or small objects, such as the massager (0.29) and the glass (0.31).

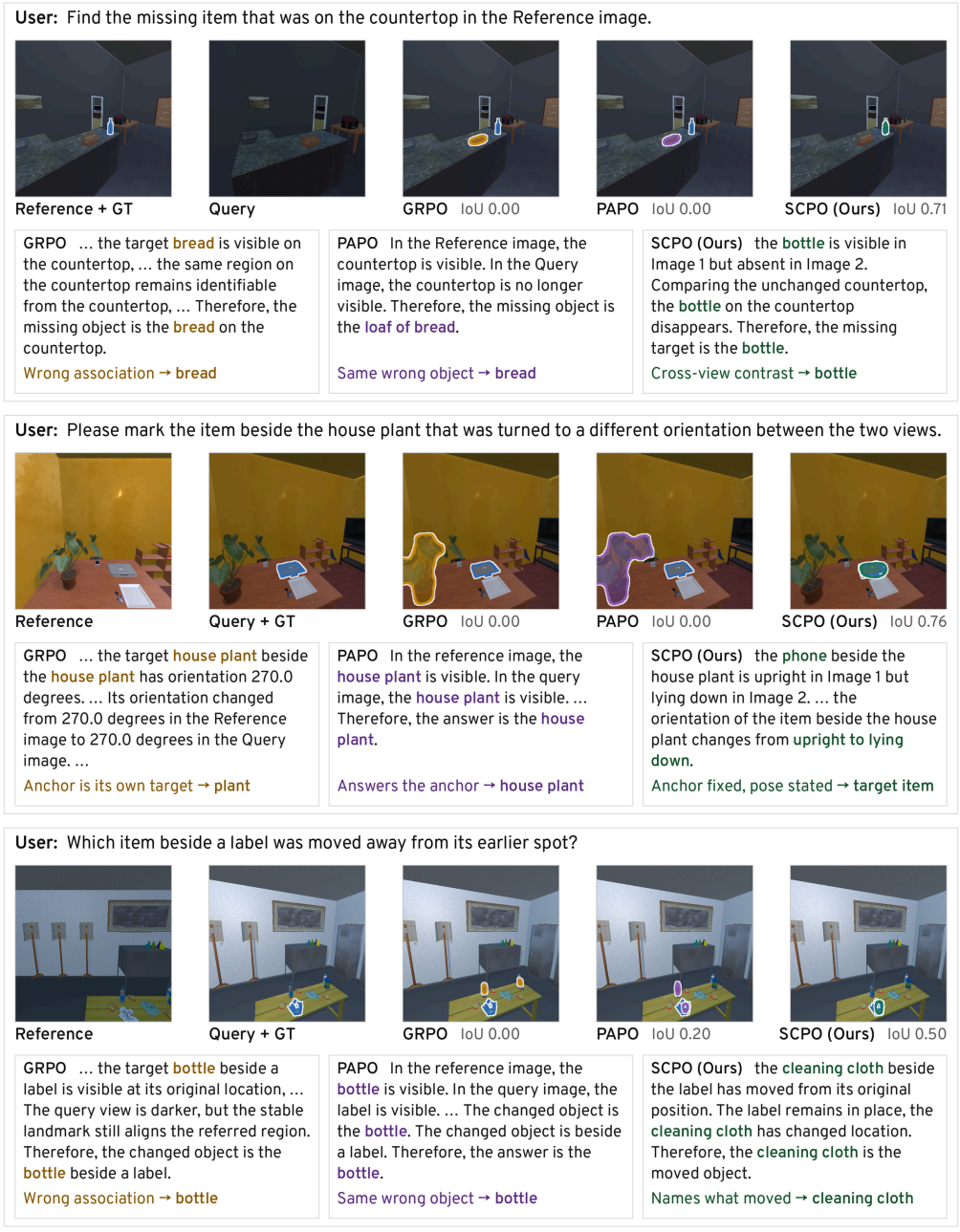


Figure A14: **Additional predictions on procedurally simulated scenes.** Missing (top), rotation (middle), and relocation (bottom) cases, in the format of Figure 6; for the missing case the ground-truth mask is shown on the reference view.

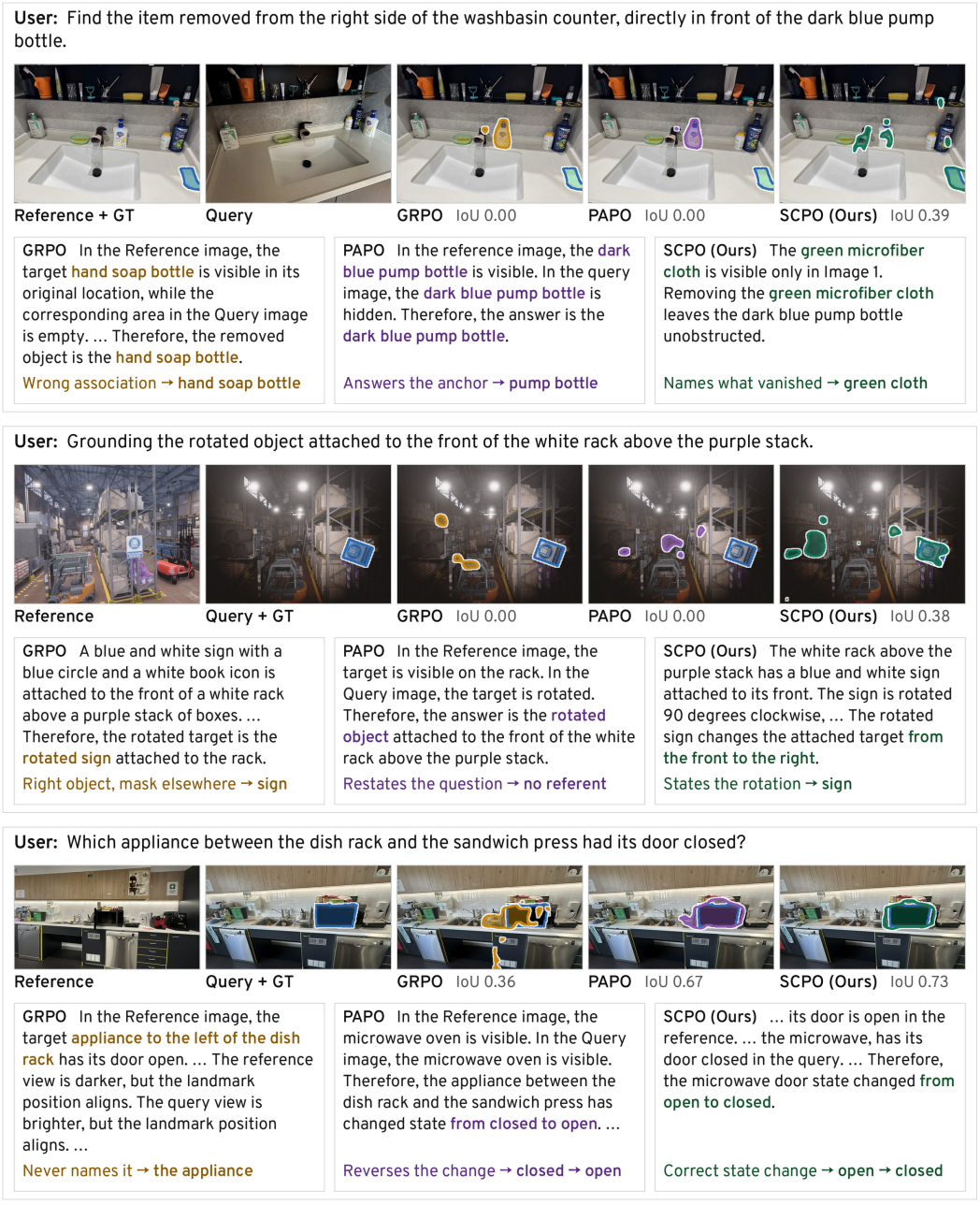


Figure A15: **Additional predictions on real-world and high-fidelity simulated scenes (1/3).** Missing (top), rotation (middle), and state-change (bottom) cases, in the format of Figure 6; for the missing case the ground-truth mask is shown on the reference view.

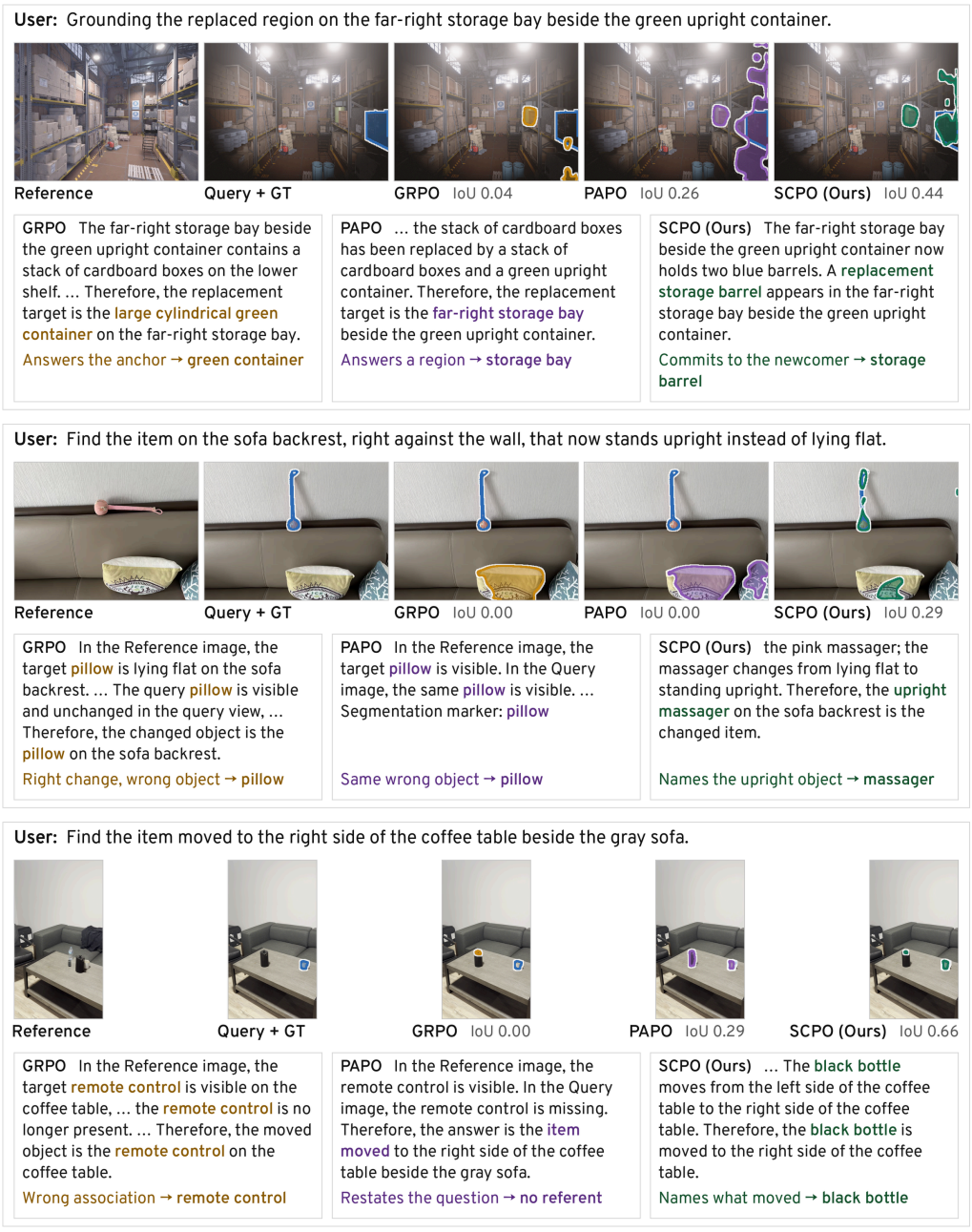
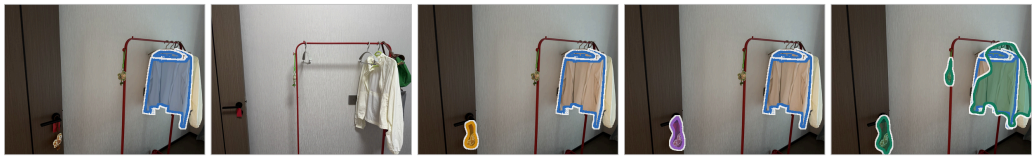


Figure A16: **Additional predictions on real-world and high-fidelity simulated scenes (2/3).** Replace (top), rotation (middle), and relocation (bottom) cases.

User: What garment between the gray hanger and the light yellow one disappeared after the reference view?



Reference + GT | Query | GRPO IoU 0.00 | PAPO IoU 0.00 | SCPO (Ours) IoU 0.52

GRPO In the Reference image, a light yellow garment is visible … The query view is darker, but the stable landmark still aligns the referred region. … the absent target is the light yellow garment between the gray hanger and the light yellow one.

Answers the anchor → light yellow one

PAPO … Therefore, the hypothesis "What garment between the gray hanger and the light yellow one disappeared after the reference view" is supported by the absence of the light yellow garment in the query image.

Restates the question → no referent

SCPO (Ours) The light yellow shirt is visible between the gray hanger and the light yellow one in Image 1. Removing that shirt leaves only the gray hanger in Image 2. Therefore, the target shirt has disappeared.

Cross-view contrast → the shirt

User: Grounding the newly appeared item near the lower-left platform edge below the wall ledge



Reference | Query + GT | GRPO IoU 0.04 | PAPO IoU 0.23 | SCPO (Ours) IoU 0.44

GRPO The lower-left platform edge below the wall ledge is clear. … Therefore, the newly appeared target is two blue cylindrical containers near the lower-left platform edge below the wall ledge.

Wrong association → blue containers

PAPO … In the Query image, a stack of cardboard boxes is present. Therefore, the newly appeared target is the stack of cardboard boxes near the lower-left platform edge below the wall ledge.

Different wrong object → box stack

SCPO (Ours) Two blue cylindrical containers are on the floor near the lower-left platform edge below the wall ledge. An open cardboard box is on the floor near the lower-left platform edge below the wall ledge.

Names the addition → open box

User: Which item at the near edge of the round table is now empty after holding coffee before?



Reference | Query + GT | GRPO IoU 0.00 | PAPO IoU 0.30 | SCPO (Ours) IoU 0.31

GRPO … The round table, its position, and the nearby area remain visible across the two views, so the query view is darker. … Therefore, the round table at the near edge of the round table has lost the coffee cup.

Names the table → round table

PAPO In the Reference image, the round table is visible. The query image excludes the round table. Therefore, the query is empty.

Incoherent conclusion → query is empty

SCPO (Ours) The clear glass on the round table has no coffee. Therefore, the clear glass at the round table has lost coffee.

Names the emptied glass → clear glass

Figure A17: **Additional predictions on real-world and high-fidelity simulated scenes (3/3).** Missing (top), existence (middle), and state-change (bottom) cases; for the missing case the ground-truth mask is shown on the reference view.

H TRANSFER TO OTHER MULTI-IMAGE REASONING BENCHMARKS

ECG-Bench tests one form of cross-image reasoning, grounding a referred change between two views. To see whether our method carries over to multi-image reasoning more broadly, we apply it to MIG-Bench (Li et al., 2025a), a multi-image grounding benchmark whose answers are bounding boxes rather than masks. The model is post-trained only on multi-image grounding data from MGrounding-630k (Li et al., 2025a), without ECG-50K, on the Qwen3-VL-8B backbone, which differs from both backbones of Table 2; as the task provides no change mask, the critical branch of SCPO masks the ground-truth box instead. VisionReasoner† (Liu et al., 2026) is trained by us in the same setting, so it differs from ours in the training recipe rather than in data or backbone. Results of the other models are taken from Li et al. (2025a) and Zheng et al. (2026a).

Benchmark. MIG-Bench contains 4.3K test instances over 5.9K images, in ten tasks. In *spontaneous grounding* the model must discover the target itself: the single difference between two similar images (*Static*), the primary difference between two images taken from slightly different viewpoints, ignoring the variations the viewpoint shift causes (*Robust*), or the object common to a group of images (*Common*). In *referential grounding* the target is specified by a visual reference, in object tracking (OT), multi-view grounding (MV), region locating (*Region*), and visual referring grounding (*Refer*); by text, in group grounding (GG); or by both, in reasoning grounding (*Reason*) and correspondence (Co-Re). Following the benchmark, a predicted box is correct when its IoU with the ground-truth box exceeds 0.5 (Acc@0.5), and AVE averages the ten tasks.

Table A3: **Results on MIG-Bench** (Li et al., 2025a) (Acc@0.5, %). *Base MLLM*: the backbone of post-trained models; †: trained by us in the same setting as ours. Results of the other models are from Li et al. (2025a) and Zheng et al. (2026a). Best result per column in bold.

Model	Base MLLM	Spontaneous Grounding			Referential Grounding							AVE
		Difference		Similarity	Visual Reference				Textual	Visual+Textual		
		Static	Robust	Common	OT	MV	Region	Refer	GG	Reason	Co-Re	
Large-Scale Models (≥70B)												
LLaVA-OV-72B	–	13.3	5.3	26.8	12.9	7.6	2.1	17.8	21.6	11.9	8.6	13.7
InternVL2-76B	–	15.9	10.6	36.4	30.7	20.8	5.7	46.5	41.3	32.7	26.5	26.7
InternVL3-78B	–	10.0	9.6	24.1	27.1	14.6	10.4	50.5	38.1	45.5	17.1	24.7
Qwen2-VL-72B	–	46.1	46.8	64.5	26.7	22.6	18.6	33.3	62.5	50.5	17.1	38.9
Qwen2.5-VL-72B	–	43.8	46.8	70.0	34.3	29.2	8.3	62.6	59.9	66.3	41.0	46.2
7–8B Models												
Mantis-8B	–	1.5	0.0	3.3	12.2	2.1	1.0	1.0	10.0	0.0	0.9	3.2
LLaVA-OV-7B	–	6.1	3.2	3.4	0.2	1.0	1.1	9.1	15.4	6.9	0.9	4.7
MiniCPM2.6-8B	–	14.6	2.1	14.3	9.8	6.3	1.8	11.1	10.0	3.0	2.6	7.6
mPLUG-Owl3-8B	–	18.6	6.4	34.9	8.6	7.6	2.4	7.1	22.9	9.1	6.0	12.4
InternVL2-8B	–	6.9	7.5	25.5	20.7	9.7	3.5	28.3	30.3	17.8	9.4	16.0
InternVL3-8B	–	23.7	14.9	48.0	14.8	6.9	12.1	7.1	34.9	16.8	2.6	18.2
Qwen2-VL-7B	–	27.8	38.3	19.4	20.7	11.8	26.0	23.2	58.5	48.5	12.0	28.6
Qwen2.5-VL-7B	–	9.6	21.6	22.4	10.6	5.9	1.3	10.0	33.5	3.0	2.6	12.0
Qwen3-VL-8B	–	41.1	23.4	68.6	34.6	31.9	38.7	52.0	60.2	63.4	22.7	43.7
Visual-RFT-7B	Qwen2-VL	34.0	31.3	50.0	15.4	12.2	15.0	24.3	62.7	41.9	10.4	29.7
Migician-7B	Qwen2-VL	65.2	46.8	**84.2**	70.7	60.1	**74.3**	76.8	66.5	59.4	34.2	63.8
CGC-8B	Qwen3-VL	55.3	**54.3**	78.3	**78.2**	66.3	70.2	80.8	76.4	73.3	42.7	67.6
VisionReasoner†	Qwen3-VL	**67.8**	44.2	82.7	73.8	63.9	72.5	79.4	68.7	61.2	36.8	65.1
Ours	Qwen3-VL	58.4	52.9	81.7	74.9	**69.8**	72.6	**83.5**	**77.1**	**73.8**	**43.3**	**68.8**

Results. Table A3 shows that our method transfers: it reaches 68.8 AVE, the best in the table, $+25.1$ over its Qwen3-VL-8B base, $+1.2$ over CGC-8B (Zheng et al., 2026a), the strongest prior model on the same backbone, and $+5.0$ over Migician-7B (Li et al., 2025a), and it is best on five of the ten tasks, all in referential grounding. Against VisionReasoner†, trained in the same setting, it gains $+3.7$ AVE and leads on eight tasks, most on *Reason* ($+12.6$), *Robust* ($+8.7$), and GG ($+8.4$). The two difference tasks split: the gain appears on *Robust*, whose images differ in viewpoint as ECG pairs do, while on *Static*, whose two images differ only in the target, VisionReasoner† stays ahead

(67.8 against 58.4). This is consistent with what SCPO rewards on ECG: separating the true change from the differences an observation shift introduces, a distinction that does not arise when the two images are otherwise identical.

Qualitative comparison. Figure A18 compares the two models on two difference-grounding samples. In the first, the woman's hair is what differs; VisionReasoner† claims that she is smiling in the second image, an expression change that neither image shows, and boxes her face, whereas our chain names the hair and its box aligns with the ground truth. In the second, the baseline chain describes books moved from the desk to a shelf, a difference that does not exist, and boxes the book stack; ours names the black pen that appears in the second image and boxes it. In both cases the box follows whatever difference the chain names, so a fluent but unfounded chain leads directly to a wrong box — the failure SCPO targets on ECG (Section 4.2). Grounding the chain in the actual difference therefore helps even when the output is a box rather than a mask.

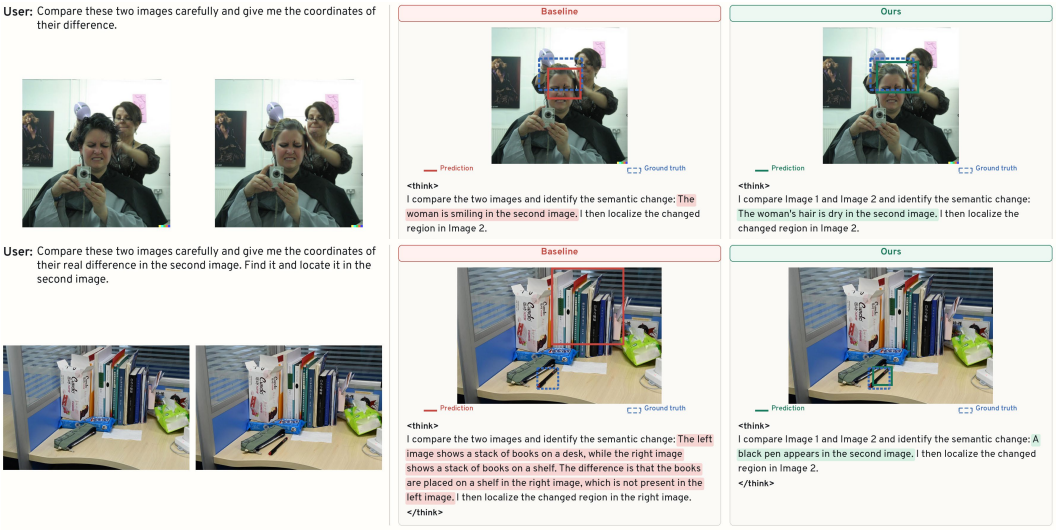


Figure A18: **Predictions of VisionReasoner† (Baseline) and ours on MIG-Bench.** Each row shows the instruction and the two input images (left), followed by each model's chain and its predicted box (solid) against the ground-truth box (dashed) on the second image; highlighted text marks the difference each chain names.