LONGHOI-BENCH: BENCHMARKING AND IMPROVING NON-CONTACT HUMAN-OBJECT INTERACTION UNDERSTANDING IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding non-contact human-object interactions requires identifying a person's target from gaze, pointing, or contextual evidence. While multimodal large language models (MLLMs) support diverse visual tasks, existing benchmarks largely evaluate these cues separately, limiting systematic assessment of person-target understanding. To address this gap, we introduce **LongHOI-Bench**, containing 6,240 images and 56,880 questions spanning localization, attribute discrimination, spatial relations, and contextual intention. Human-verified targets and typed negative alternatives connect task performance with three grounding errors: drift, distractor, and ambiguity. We benchmark 13 proprietary and open-source MLLMs on LongHOI-Bench. Results reveal substantial limitations, with the strongest model reaching 68.5% mean accuracy across interaction types while humans exceed 92% in each interaction category. A balanced error audit identifies distractors as the most frequent failure. Motivated by these findings, we propose **CLASP**, which combines error-type diagnosis with contrast-based layer selection and weighted suppression through analytic weight editing. On paired localization queries, CLASP corrects 69.8% of distractor errors while introducing errors in 4.2% of initially correct predictions. Direction controls show that diagnostic relevance matters beyond captured contrast energy; the intervention also improves external gaze and pointing performance without recalibration. The benchmark and code will be publicly released upon publication.

1 INTRODUCTION

Understanding which object a person is looking at, pointing toward, or acting in relation to matters for assistive robots, human-computer interaction, and social scene understanding. A system that follows a gesture toward the wrong object can give an irrelevant answer even when it recognizes every object correctly. Such failures motivate the diagnostic intervention outlined in Figure 1. Consider the shared-attention scene in Figure 2b: identifying the laptop as the attended target requires relating both people to it, rather than selecting the nearby bottle. Confusing these objects would change answers about the target's category, attributes, and spatial relations. Non-contact interaction understanding therefore requires more than object recognition. Models must use directional and contextual evidence to connect the person with a target before interpreting its properties or role.

Existing resources leave three gaps. (1) **Fragmented interaction coverage:** gaze following (Recasens et al., 2015; Tu et al., 2022), embodied reference (Chen et al., 2021), and intention recognition (Zhong et al., 2021) employ separate datasets and protocols, preventing direct comparison across cues. (2) **Disconnected task evaluation:** gaze and pointing benchmarks (Wang et al., 2025b; 2026; Li et al., 2026) develop cue-specific evaluations, while broader perception benchmarks diagnose visual correspondence and discrimination (Fu et al., 2024; Li et al., 2024). Their different annotations limit joint assessment of localization and subsequent target understanding. (3) **Missing grounding-error distinctions:** action and object-role annotations in established HOI datasets (Chao et al., 2018; Pratt et al., 2020) do not distinguish displaced predictions from salient distractors or plausible alternative targets. Aggregate scores leave these failures combined.

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.

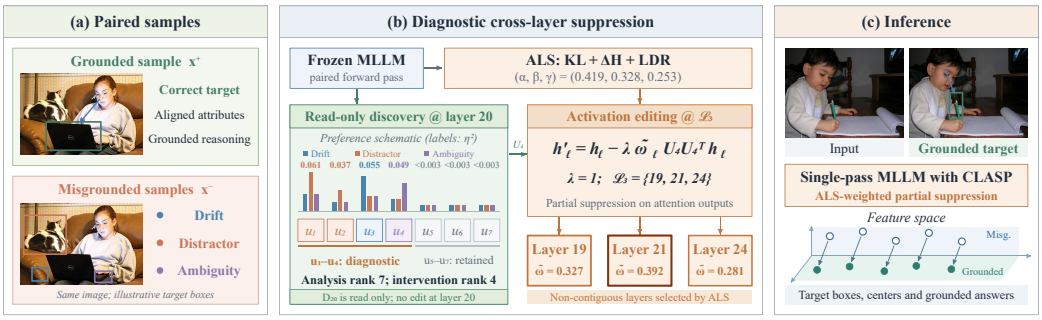

Figure 1: **From grounding errors to CLASP.** (a) Grounded/misgrounded pairs. (b) Diagnostic direction selection and ALS layer allocation. (c) Single-pass inference with edited weights. Colors indicate relative error enrichment and labels report η^2; bar heights and geometry are schematic.

To address these gaps, we introduce **LongHOI-Bench**, a benchmark for non-contact person-target understanding across gaze, pointing, and contextual intention. It contains 6,240 images and 56,880 questions drawn from MS COCO, Objects365, and SWiG-HOI (Lin et al., 2014; Shao et al., 2019; Pratt et al., 2020). The same verified targets support localization, attribute discrimination, spatial relations, and intention questions, connecting several task demands within a common collection. Human annotators establish the targets and answers before model-assisted negative mining and question wording. This shared visual reference allows evaluation to ask whether a model can locate a target, distinguish its properties from those of competing objects, and explain its role in the interaction. Question-family differences can thus be examined alongside the person-target relation.

LongHOI-Bench also supplies verified incorrect targets with grounding-error labels: drift denotes displaced localization, distractor a salient irrelevant target, and ambiguity a plausible incorrect alternative. Benchmarking 13 MLLMs without task-specific adaptation (base models) reveals that the strongest reaches 68.5% interaction-macro accuracy, whereas humans exceed 92% in every interaction category. Distractors account for 40.3% of the balanced error audit, identifying target competition as a central challenge. Motivated by this finding, we propose **CLASP**, which selects diagnostically informative feature directions and distributes their suppression across layers (Figure 1). CLASP corrects 69.8% of distractor errors on paired localization queries and improves external gaze and pointing performance without recalibration. Controlled comparisons further show that selecting the right directions matters more than suppressing a larger fraction of contrast energy.

Our contributions are summarized as follows:

- We introduce **LongHOI-Bench**, linking three non-contact interaction types and four question families through verified targets and typed negatives. Across 13 MLLMs, the strongest base model reaches only 68.5% interaction-macro accuracy.
- We identify distractors as the most frequent grounding failure, accounting for 40.3% of audited errors. Matched image and caption conditions further show that scene descriptions leave substantial person-target information unresolved.
- We propose **CLASP**, combining diagnostic direction selection with contrast-based layer allocation. It exceeds image-budget-matched DoRA by 1.4 pooled AUC points and 1.8 pooled mAP points and transfers unchanged to external gaze and pointing datasets.

2 Related Work

2.1 Interaction Understanding Benchmarks

Visual perception and discrimination are assessed by BLINK (Fu et al., 2024) and NaturalBench (Li et al., 2024), respectively. Neither organizes these tasks around shared person-target relations with typed grounding errors. HICO-DET (Chao et al., 2018) localizes action-object pairs, while

Table 1: Interaction datasets and evaluation frameworks. Cue columns indicate target-understanding tasks; HOT3D supplies eye tracking. B-RIGHT counts split-2 training and test images, HOT3D counts frames, and YouRefIt counts clips. EyeVLM evaluates existing datasets. Category and annotation coverage follow each resource; N/A marks unavailable comparable totals.

Dataset	Setting	Gaze	Point	Intent	Visual count	QA pairs	Verified QA	Typed negatives
Action / hand-object interaction								
HICO-DET (Chao et al., 2018)	Contact HOI	✗	✗	✗	47,776	✗	✗	✗
B-RIGHT (Jang et al., 2025)	Balanced HOI	✗	✗	✗	8,397	✗	✗	✗
VG-HOI (Wu et al., 2024b)	Large-vocabulary HOI	✗	✗	✗	43,118	✗	✗	✗
SWiG (Pratt et al., 2020)	Role-based HOI	✗	✗	✗	126,102	✗	✗	✗
HOT3D (Banerjee et al., 2025)	3D hand-object tracking	✗	✗	✗	>3.7M frames	✗	✗	✗
Non-contact cues								
GazeFollow (Recasens et al., 2015)	Gaze target	✓	✗	✗	122,143	✗	✗	✗
YouRefIt (Chen et al., 2021)	Embodied reference	✗	✓	✗	4,195	✗	✗	✗
VL4Gaze (Wang et al., 2025b)	Gaze VQA	✓	✗	✗	≈124,000	≈489,000	✗	✗
EyeVLM (Wang et al., 2026)	Gaze / social gaze	✓	✗	✗	reused	N/A	N/A	✗
EgoPoint-Bench (Li et al., 2026)	Egocentric pointing	✗	✓	✗	N/A	11,729	✓	✗
LongHOI-Bench	**Non-contact HOI**	✓	✓	✓	**6,240**	**56,880**	✓	✓

SWiG (Pratt et al., 2020) grounds semantic roles; these annotations do not connect non-contact cues to target-centered question families. Cue-specific resources include GazeFollow's gaze localization (Recasens et al., 2015) and YouRefIt's language-and-gesture reference resolution (Chen et al., 2021), leaving comparison across interaction cues unresolved. VL4Gaze (Wang et al., 2025b) extends gaze to VQA and EyeVLM (Wang et al., 2026) includes social gaze, yet both retain that cue focus. EgoPoint-Bench (Li et al., 2026) studies egocentric pointing, placing third-person cue comparison outside its scope. In contrast, LongHOI-Bench connects gaze, pointing, and contextual intention through common targets, question families, and typed alternatives (Table 1).

2.2 Representation Editing

Removing dominant hallucination directions, as in Nullu (Yang et al., 2025), does not establish their relevance to distinct grounding errors. VTI (Liu et al., 2025) steers visual and textual features for stability; ICT (Chen et al., 2025) strengthens image- and object-level evidence through attention-head interventions. Neither uses target-error diagnosis to distinguish competing objects. Inference-time adaptation includes ASD's bidirectional contrastive decoding (Su et al., 2025) and DMAS's context-dependent vector retrieval (Yin et al., 2026), requiring dual branches and input-dependent retrieval, respectively. In contrast, CLASP selects directions by grounding-error diagnosis and allocates suppression through layer contrasts, then incorporates the fixed intervention into model weights.

3 LongHOI-Bench

3.1 Overview

LongHOI-Bench connects gaze, pointing, and contextual intention through verified person-target relations. Its 6,240 images and 56,880 questions support evaluation across four question families and error diagnosis through verified, typed negative targets.

3.2 Image Collection

We select scenes with identifiable persons, verifiable targets, and plausible competitors from MS COCO (Lin et al., 2014), Objects365 (Shao et al., 2019), and SWiG-HOI (Pratt et al., 2020), excluding contact-dominated interactions. Five graduate annotators independently verify target boxes before negative mining; two senior annotators adjudicate disagreements. The collection covers 318 categories and 42.0% multi-person scenes (Appendix J).

Images follow a 4:1:1 training/validation/test split, with 1,040 per held-out split. Questions, boxes, and derivatives inherit their image's assignment. Method preparation accesses training images only.



Figure 2: **Representative LongHOI-Bench questions.** Panels (a) to (d) illustrate gaze, pointing, and contextual scenes. Highlights identify queried evidence; check marks indicate answers. Bounding-box coordinates are source-image pixels.

3.3 Fine-grained Task Definition

Interaction labels specify the evidence identifying the target (Figure 2). Sufficient pointing cues take precedence, followed by sufficient gaze cues. Contextual intention requires broader action, object, or scene evidence when directional cues are insufficient. Senior annotators adjudicate cases where gaze and pointing cues indicate different interaction targets.

Four question families examine each person-target relation. Localization tests target selection before description. Attribute discrimination checks whether properties belong to the target rather than a competitor. Spatial questions assess placement relative to people and objects, separating relational grounding from category recognition. Intention questions test the target's role in the activity, requiring context beyond position. Typed negatives distinguish drift, distractor, and ambiguity.

3.4 Question-answering Annotation

Verified targets and answers anchor question construction. Localization and spatial questions follow templates; GPT-5 (OpenAI, 2025) phrases attribute and intention questions from captions and structured annotations without images or control over answer keys. Florence-2-large (Xiao et al., 2024)

Table 2: Interaction-specific results. Base models are evaluated zero-shot; CLASP uses training-calibrated InternVL3.5-8B. CLASP reports three-seed means (SD 0.002 to 0.004; BERTScore 0.003); the rank-seven single-layer row reports point estimates (BERTScore SD 0.003). Human localization pools 1,080 items (†). Blue/red mark the best base/model result, excluding humans.

Method	Gaze Interaction				Pointing Interaction				Contextual Intention Interaction				BERTScore↑
	Min.D↓	AUC↑	mAP↑	ACC↑	Min.D↓	AUC↑	mAP↑	ACC↑	Min.D↓	AUC↑	mAP↑	ACC↑	
CogVLM (Wang et al., 2023)	0.484	0.571	0.204	0.384	0.462	0.588	0.219	0.389	0.499	0.572	0.196	0.365	0.166
LLaVA-v1.5-7B (Liu et al., 2024a)	0.285	0.649	0.376	0.495	0.272	0.661	0.388	0.507	0.297	0.635	0.391	0.483	0.316
VILA-8B (Lin et al., 2024)	0.296	0.613	0.352	0.441	0.316	0.631	0.349	0.425	0.328	0.618	0.338	0.431	0.297
Llama3.2-Vision-11B (Meta, 2024)	0.381	0.624	0.398	0.516	0.365	0.641	0.396	0.531	0.402	0.612	0.402	0.505	0.193
InternLM-XComposer2-7B (Dong et al., 2024)	0.427	0.518	0.386	0.503	0.473	0.529	0.361	0.428	0.451	0.523	0.372	0.497	0.128
mPLUG-Owl-7B (Ye et al., 2023)	0.566	0.412	0.194	0.229	0.561	0.431	0.206	0.244	0.442	0.395	0.181	0.219	0.229
Qwen2.5-VL-3B (Bai et al., 2025)	0.478	0.497	0.214	0.493	0.463	0.515	0.228	0.459	0.488	0.508	0.241	0.491	0.294
Qwen2.5-VL-7B (Bai et al., 2025)	0.489	0.484	0.297	0.487	0.446	0.502	0.309	0.503	0.481	0.498	0.290	0.549	0.319
InternVL3.5-2B (Wang et al., 2025c)	0.352	0.529	0.195	0.496	0.341	0.492	0.206	0.508	0.437	0.494	0.191	0.454	0.173
InternVL3.5-8B (Wang et al., 2025c)	0.329	0.510	0.278	0.561	0.334	0.526	0.309	0.574	0.349	0.515	0.283	0.579	0.224
Gemini-2.0-Flash (Google DeepMind, 2025)	**0.238**	0.662	0.399	0.641	**0.231**	0.652	0.414	0.622	**0.246**	0.662	0.401	0.658	0.324
GPT-4o (OpenAI, 2024)	0.241	**0.692**	0.416	**0.682**	0.232	**0.702**	0.427	**0.691**	0.247	**0.693**	0.454	**0.683**	**0.374**
GPT-5 (OpenAI, 2025)	0.264	0.647	**0.475**	0.615	0.261	0.659	**0.486**	0.628	0.268	0.645	**0.472**	0.619	0.362
CLASP (single-layer)	0.182	0.736	0.532	0.716	0.159	0.754	0.549	0.698	0.184	0.714	0.492	0.703	0.411
CLASP	**0.167**	**0.761**	**0.559**	**0.732**	**0.143**	**0.779**	**0.574**	**0.721**	**0.169**	**0.735**	**0.521**	**0.724**	**0.429**
Human	*0.058*†	*0.946*†	*0.917*	*0.951*	*0.058*†	*0.946*†	*0.943*	*0.974*	*0.058*†	*0.946*†	*0.886*	*0.928*	*0.904*

supplies human-revised captions for Objects365 and SWiG. Reviewers retain or correct questions against verified answers; multiple-choice options are shuffled.

LLaVA-v1.5-7B, Qwen2.5-VL-7B, and InternVL3.5-8B propose negative targets for two three-annotator panels, achieving pre-adjudication Fleiss' $\kappa = 0.781$. Multiple-choice alternatives use spatial, saliency, and applicable directional criteria. A separate annotation check yields $\kappa = 0.740$ to 0.860, with contextual intention least consistent. Appendix J reports sample counts, unanimous and box agreement, construction prompts, and candidate rules.

4 Experiments

4.1 Evaluation Setup

We evaluate 13 base MLLMs, including proprietary GPT-4o, GPT-5, and Gemini-2.0-Flash, and open-source CogVLM, LLaVA, Qwen2.5-VL, and InternVL3.5 models. Each receives the designated person, interaction type, and question through a common task interface; Appendix A specifies prompts and parsing. Five independent human evaluators assess 1,080 test items.

Heatmap ROC-AUC and normalized localization distance assess where predictions fall; category-aware mAP additionally tests target identity. Answer accuracy measures categorical understanding, while BERTScore assesses explanation similarity. Interaction-macro scores average the three interaction-specific results equally; pooled scores evaluate the combined query set, with mAP recomputed over combined predictions. Table 2 reports each interaction separately. Language-only and caption controls test how much visual information textual descriptions retain. Appendix A gives the scoring and confidence-handling rules for these evaluations.

4.2 Main Results

Table 2 compares base models with the human reference across interaction categories.

4.2.1 Overall Performance

GPT-4o leads the 13 base models in interaction-macro accuracy, but its 68.5% remains well below human performance. Open-source LLaVA-v1.5-7B exceeds the larger Llama3.2-Vision-11B on gaze and pointing AUC, showing that parameter count alone does not determine localization quality.

Textual descriptions recover only part of the visual evidence. On matched questions, GPT-4o gains 23.8 accuracy points from images over independently generated captions. Blind Qwen2.5-7B-Instruct (Yang et al., 2024a) reaches 27.8%, barely above four-way chance (Appendix J). Language priors and scene summaries therefore leave substantial target-understanding gaps.

4.2.2 TASK-WISE PERFORMANCE

GPT-4o leads base-model answer accuracy in every interaction, whereas GPT-5 leads category-aware mAP. GPT-5 and humans attain their highest mAP on pointing, followed by gaze and contextual intention. The latter requires context beyond directional cues. The strongest base-model explanation score, GPT-4o's BERTScore of 0.374, remains far below the human reference of 0.904. Answering correctly, identifying targets, and explaining interactions remain distinct challenges.

4.3 BENCHMARK VALIDATION AND ERROR ANALYSIS

Construction changes largely preserve the 13-model ordering: MiniCPM-V 2.6 (Yao et al., 2024) re-mining yields Spearman $\rho = 0.940$, and matched human-only wording yields $\rho = 0.960$ (Appendix J). Rule-selected alternatives nevertheless reduce accuracy relative to uniform same-image candidates (Appendix Table A13), distinguishing ranking stability from absolute difficulty. Because both conditions draw alternatives from the same image, this gap reflects the difficulty of choosing among plausible targets while scene content remains fixed.

A balanced audit samples 200 zero-shot errors per interaction from each of LLaVA-v1.5-7B, Qwen2.5-VL-7B, InternVL3.5-8B, and GPT-4o. Two annotators label responses independently, with a third adjudicating. Distractors constitute 40.3% of audited errors and are especially frequent for contextual intention; pointing has the largest drift share.

Targeted edits obscure head/eye or forearm regions on 600 test images, evenly divided between gaze and pointing; controls match area and strength elsewhere (Table 6b). Humans and GPT-4o show larger targeted-versus-control AUC losses than base InternVL3.5-8B. Under contextual masking, GPT-4o instead incurs a smaller accuracy loss than base InternVL3.5-8B (Appendix Table A8), showing interaction-dependent cue use. Together with distractor prevalence, these results identify a relational grounding problem: models can recognize scene objects without reliably associating them with the designated person. This motivates targeting error-relevant representations.

5 CLASP: IMPROVING NON-CONTACT TARGET UNDERSTANDING

Motivated by the relational grounding failures in Section 4.3, we propose CLASP, which combines diagnostic direction selection, contrast-based layer allocation, and analytic weight editing (Figure 3). We first describe these components, then evaluate their contributions and external transfer.

5.1 FRAMEWORK OVERVIEW

5.1.1 DIAGNOSTIC SUBSPACE SELECTION

To isolate target-conditioned differences, estimation pool E supplies grounded/misgrounded pairs that share an unmarked image and query, differing only in normalized candidate-box text. Pre-answer states $\mathbf{x}^{\pm}_{\ell,i} \in \mathbb{R}^d$ form columns of $\mathbf{X}^{\pm}_{\ell}$, giving

$$\mathbf{D}_\ell = \mathbf{X}^{+}_{\ell} - \mathbf{X}^{-}_{\ell} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}, \qquad \mathbf{u}_j \in \mathbb{R}^d. \tag{1}$$

SVD uses $\mathbf{D}_\ell$ without centering or column normalization. Spectral concentration, split-half stability, and reconstruction error select the readout using E (Appendix F). InternVL3.5-8B (Wang et al., 2025c) uses layer 20 to estimate the contrast subspace.

To identify error-relevant directions, diagnosis pool G supplies one pair per image, with $E \cap G = \varnothing$. Normalized projection energy is $s_{ij} = (\mathbf{u}_j^{\top}\mathbf{d}_i)^2/\|\mathbf{d}_i\|_2^2$. Effect size η_j^2 divides between-error-group variation by total variation. Stratified permutations assess whether directions separate error types (Appendix B). Seven directions capture 95.3% of contrast energy, but only u_1 through u_4 exceed the permutation P99 for error-type separation. CLASP selects $\mathbf{U}_4 = [\mathbf{u}_1, \ldots, \mathbf{u}_4]$: high contrast energy

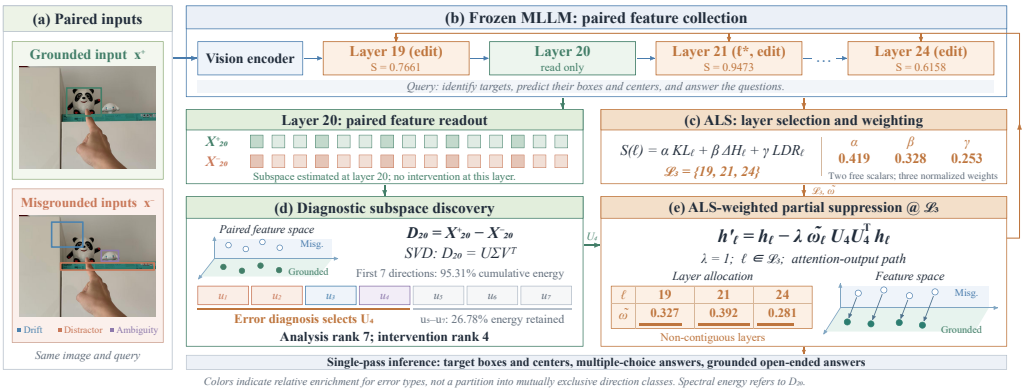

Figure 3: **CLASP estimation and intervention.** Layer 20 supplies the basis; diagnosis selects directions and ALS weights suppression at layers 19, 21, and 24. Colors indicate relative error enrichment; the feature-space geometry is schematic.

alone is insufficient grounds for suppression. Calibration and testing omit target boxes. Directions u_1 and u_2 show relative enrichment for distractor errors, u_3 for drift, and u_4 for ambiguity, linking the selected basis to the three diagnosed types of grounding failure.

5.1.2 Contrast-based Layer Allocation

ALS emphasizes layers with stronger grounded/misgrounded contrasts. Final normalization and the shared LM head produce full-vocabulary distributions $p^{\pm}_{\ell,i}$. KL divergence measures predictive disagreement; signed entropy difference measures relative uncertainty; LDR measures feature separation relative to within-condition variation. Min-max normalization across candidate layers makes the signals comparable across layers (Appendix C):

$$S(\ell) = \alpha\,\widetilde{\mathrm{KL}}_\ell + \beta\,\widetilde{\Delta H}_\ell + \gamma\,\widetilde{\mathrm{LDR}}_\ell, \qquad (\alpha,\beta,\gamma) = \mathrm{softmax}(\theta). \tag{2}$$

Calibration pool C_s, with $C_s \cap (E \cup G) = \varnothing$, fits two free variables for the three normalized coefficients. Nelder-Mead minimizes answer-token NLL with the backbone and basis fixed. Softmax over the Top-3 layer scores yields weights $\tilde{\omega}_\ell$. ALS selects $\{19, 21, 24\}$, separately from the layer-20 readout; Appendix C reports pool sizes and fitted weights.

5.1.3 Analytic Weight Editing

For suppression without extra inference passes, each selected attention-output projection applies

$$\mathbf{h}'_\ell = \mathbf{h}_\ell - \lambda\tilde{\omega}_\ell \mathbf{U}_4\mathbf{U}_4^\top \mathbf{h}_\ell. \tag{3}$$

Here $\mathbf{h}_\ell = \mathbf{W}^O_\ell \mathbf{z}_\ell$ precedes residual addition, giving $\mathbf{r}_\ell + \mathbf{h}'_\ell$. Left-multiplying $\mathbf{W}^O_\ell$ and any output bias by $\mathbf{I} - \lambda\tilde{\omega}_\ell\mathbf{U}_4\mathbf{U}_4^\top$ applies the edit to all prefill and decoding tokens. Global strength $\lambda = 1$ partially suppresses selected components while preserving orthogonal components, without updating backbone parameters through gradient-based training.

5.2 CLASP Performance

5.2.1 Performance Improvement and Baseline Comparisons

CLASP improves all seven calibrated backbones and exceeds mitigation baselines in pooled accuracy and mAP (Table 3a,c). Gains span all three interactions, although explanation BERTScore remains below half the human reference (Table 2). Better grounding thus leaves explanation quality unresolved. Appendix H details baseline implementations.

Table 3b uses 42 training images for PEFT and 1,491 preparation images for CLASP. At the matched 1,491-image budget, CLASP exceeds DoRA by 1.4 pooled AUC points and 1.8 pooled mAP points (Appendix Table A16). The gain persists with matched image access.

Table 3: Mitigation and PEFT comparisons on InternVL3.5-8B; panel (c) calibrates seven backbones separately. † marks interaction-macro frozen references; other scores are pooled. PEFT uses 42 training images; CLASP uses 1,491, including 42 for calibration. CLASP calibrates two scalars and edits weights analytically, with no gradient-trained backbone parameters. Single-layer rank: seven.

(a) Hallucination Mitigation

Method	ACC↑	mAP↑
OPERA (Huang et al., 2024)	0.581	0.373
VCD (Leng et al., 2024)	0.394	0.416
DeCo (Wang et al., 2025a)	0.624	0.448
Nullu (Yang et al., 2025)	0.653	0.471
VTI (Liu et al., 2025)	0.662	0.490
ICT (Chen et al., 2025)	0.669	0.504
CausalLens (Ji et al., 2026)	0.675	0.507
KVSmooth (Jiang et al., 2026)	0.582	0.516
CLASP	0.726	0.546

(b) 1%-Data PEFT Baselines

Method	Trainable	AUC↑	mAP↑
Backbone†	0%	0.517	0.290
Prompt Tuning (Lester et al., 2021)	0.0032%	0.563	0.336
IA^3 (Liu et al., 2022)	0.0046%	0.596	0.372
Adapter (Houlsby et al., 2019)	3.7447%	0.621	0.401
LoRA (Hu et al., 2022)	1.1689%	0.646	0.438
DoRA (Liu et al., 2024b)	0.3562%	0.658	0.451
CLASP (single-layer)	0%	0.729	0.519
CLASP	**0%**	0.754	0.546

(c) Backbone-specific calibration and intervention

		CogVLM	LLaVA-v1.5-7B	Qwen2.5-VL-7B	Qwen2.5-VL-3B	VILA-8B	InternVL3.5-2B	InternVL3.5-8B
Frozen†	AUC	0.577	0.648	0.495	0.507	0.621	0.505	0.517
	mAP	0.206	0.385	0.299	0.228	0.346	0.197	0.290
+ CLASP	AUC	0.692	0.729	0.704	0.669	0.764	0.712	0.754
	mAP	0.483	0.517	0.475	0.451	0.528	0.514	0.546

Table 4: Ablations at $\lambda = 1$, with interaction-macro task scores. Frozen POPE F1 and MMBench-style ACC are 87.43 and 81.72. MMBench reports three-seed means; displayed Δ uses rounded means, and preservation uses unrounded scores with a one-point tolerance. Uniform assigns $1/3$ at fixed layers and subspace; Random-3 SD covers 30 layer sets.

Configuration	Layers	k_{edit}	AUC↑	mAP↑	ACC↑	ΔPOPE	ΔMMBench
Frozen backbone	N/A	N/A	0.517	0.290	0.571	0.00	0.00
Single-layer, spectral rank	{21}	7	0.735	0.524	0.706	−0.37	−0.38
Single-layer, diagnostic rank	{21}	4	0.739	0.531	0.710	−0.26	−0.25
Top-3 uniform	{19, 21, 24}	4	0.746	0.537	0.715	−0.45	−0.47
Top-3 ALS, spectral rank	{19, 21, 24}	7	0.752	0.544	0.719	−0.85	−0.80
Random-3 (30 subsets)	random	4	0.709 ± 0.019	0.473 ± 0.022	0.680 ± 0.014	-0.67 ± 0.11	-0.70 ± 0.14
Bottom-3	{2, 4, 7}	4	0.687	0.442	0.666	−0.75	−0.75
CLASP (Top-3 ALS)	{19, 21, 24}	4	**0.758**	**0.551**	**0.726**	−0.52	−0.53

Paired outcomes under prespecified calibration seed 1 cover 2,413 held-out localization queries scorable by both systems. CLASP corrects 69.8% of distractor, 54.5% of drift, and 54.4% of ambiguity errors; 4.2% of initially correct predictions become incorrect (Appendix Table A14). Target competition is both the most frequent audited failure and the most correctable error type. In absolute counts, 694 errors become correct while 53 correct predictions become incorrect, yielding a net gain of 641 correct localization outcomes on this paired query set.

5.2.2 Ablation Studies

ALS Top-3 exceeds single-layer, random-layer, and bottom-layer interventions at fixed subspace and strength (Table 4). Uniform allocation over the same layers lowers interaction-macro AUC from 0.758 to 0.746, isolating the contribution of weighting (Appendix Table A6). Full ALS leads single-signal and leave-one-out variants in pooled task scores (Appendix Table A15). Adjacent readouts perform similarly; a fourth layer narrows capability margins (Appendix Tables A4 and A5).

Diagnostic rank four improves interaction-macro AUC, mAP, and ACC over spectral rank seven, while reducing capability losses (Table 5a). Paired image-bootstrap intervals for these gains exclude zero (Appendix D). The rank-three spectral elbow trades lower task scores for less capability loss. Rank four captures 68.5% of contrast energy, compared with 95.3% for rank seven; better grounding therefore accompanies suppression of fewer directions capturing less total energy.

At fixed rank, the diagnostic subset leads all 35 four-direction combinations (Table 5b). The strongest alternative retains three diagnosed directions; tail substitution substantially reduces grounding. Further sweeps show that $\lambda = 2.0$ exceeds the one-percentage-point capability-loss

Table 5: Rank rules (a) and direction subsets (b) at layers $\{19, 21, 24\}$ and $\lambda = 1$. Task scores are interaction-macro; capability changes are percentage points. MMBench reports three-seed means, with displayed Δ calculated from rounded values. Panel (a) edits spectral prefixes; the elbow considers seven directions. Panel (b) reports SD over all 35 subsets, relative to CLASP.

Criterion / directions	k	AUC↑	mAP↑	ΔPOPE	ΔMMBench
(a) Rank-selection rules					
95% cumulative energy	7	0.752	0.544	−0.85	−0.80
Spectral elbow (first seven directions)	3	0.752	0.544	−0.40	−0.41
Diagnostic η^2 cliff (CLASP)	**4**	**0.758**	**0.551**	−0.52	−0.53
(b) Direction-identity controls					
All 35 four-direction subsets (mean ± SD)	4	0.699 ± 0.032	0.485 ± 0.037	−0.64	−0.64
Tail subset $\{u_4, u_5, u_6, u_7\}$	4	0.643	0.402	−0.76	−0.76
Best alternative $\{u_1, u_3, u_4, u_5\}$	4	0.752	0.546	−0.59	−0.59

Table 6: **(a)** External transfer with the LongHOI-Bench subspace, layers, and weights fixed. AUC uses the shared heatmap protocol; Avg.D and Min.D use five supplemental target-center annotations per query. **(b)** Matched cue perturbations on 600 test images, with 300 per cue. $\Delta_{\mathrm{ctrl}} = \mathrm{Control} - \mathrm{Clean}$ and $\Delta_{\mathrm{cue}} = \mathrm{Targeted} - \mathrm{Control}$ report changes calculated before rounding. All metrics and changes are displayed to three decimal places throughout.

(a) Zero-shot transfer

Dataset	Method	Metrics			ΔAUC
		AUC↑	Avg.D↓	Min.D↓	
GazeFollow (Recasens et al., 2015)	Frozen	0.722	0.214	0.127	N/A
	CLASP (single-layer)	0.744	0.202	0.119	+0.023
	CLASP	**0.759**	**0.194**	**0.112**	**+0.037**
VideoAttnTarget (Chong et al., 2020)	Frozen	0.688	0.241	0.145	N/A
	CLASP (single-layer)	0.706	0.230	0.138	+0.019
	CLASP	**0.723**	**0.218**	**0.131**	**+0.035**
YouRefIt (Chen et al., 2021)	Frozen	0.615	0.284	0.169	N/A
	CLASP (single-layer)	0.636	0.272	0.160	+0.021
	CLASP	**0.652**	**0.260**	**0.153**	**+0.038**

(b) Cue ablation

Cue	Model	AUC↑			Attributed change	
		Clean	Control	Targeted	Δ_{ctrl}	Δ_{cue}
Gaze	InternVL3.5-8B	0.519	0.512	0.394	−0.008	−0.118
	GPT-4o	0.694	0.685	0.501	−0.009	−0.184
	CLASP	**0.767**	**0.755**	**0.540**	−0.012	**−0.215**
	Human	*0.944*	*0.932*	*0.649*	−0.012	−0.283
Pointing	InternVL3.5-8B	0.535	0.526	0.408	−0.009	−0.118
	GPT-4o	0.709	0.697	0.506	−0.011	−0.191
	CLASP	**0.784**	**0.773**	**0.561**	−0.012	**−0.211**
	Human	*0.957*	*0.945*	*0.689*	−0.012	−0.256

tolerance on both measures (Appendix Tables A2 and A3). Direction identity and allocation therefore matter beyond the amount of suppressed contrast energy.

5.2.3 External Transfer

The unchanged subspace, layers, and weights improve AUC on GazeFollow, VideoAttentionTarget, and YouRefIt by 3.7, 3.5, and 3.8 points, respectively (Table 6a). Both distance metrics also decrease on all three datasets, indicating that the AUC gains accompany improved target-center localization across gaze-following and embodied-reference tasks. CLASP also shows larger targeted-versus-control cue losses than its base model on gaze and pointing (Table 6b), consistent with greater directional-cue use. Fixed-rank HalluSpace controls also favor proxy-selected directions (Appendix Table A10). Together, diagnostic direction selection and contrast-based allocation produce grounding gains that transfer beyond LongHOI-Bench to external tasks.

6 Conclusion

We introduce LongHOI-Bench to evaluate non-contact interaction understanding through shared person-target annotations and complementary question families. Evaluating 13 MLLMs exposes a relational grounding problem: recognizing scene objects does not ensure that models associate them with the designated person. Target competition produces the most frequent audited failure and the highest correction rate under CLASP, linking diagnosis to improvement. CLASP uses error-relevant directions and contrast-based layer allocation to improve grounding, with gains extending to external gaze and pointing tasks. Controlled comparisons show that diagnostic relevance, rather than captured contrast energy alone, should guide suppression. LongHOI-Bench thus provides a common test of whether models identify the intended target, interpret its role, and improve these abilities through interventions informed by their grounding failures.

AI USE STATEMENT

GPT-5 assisted attribute and intention question wording from captions and structured annotations. Human annotators established the targets and answer keys. GPT-5 received no original images and was instructed to preserve the verified answers without introducing unsupported visual details. Vision-language models proposed negative candidates for human verification, and Florence-2-large generated human-revised construction captions for Objects365 and SWiG. Separate language controls used GPT-4o and Florence-2 captions. Reviewers accepted 76.8% of generated questions unchanged, edited 16.1%, and rejected 7.1%. Appendix J documents these construction roles and prompts. AI tools also assisted manuscript revision, LaTeX editing, and figure preparation. The authors are responsible for the evidence, references, interpretations, and findings.

REPRODUCIBILITY STATEMENT

The benchmark and code will be publicly released upon publication. The appendix provides the implementation details and supporting analyses needed to reproduce the benchmark and CLASP evaluations. Appendix A specifies inputs, prompts, parsing, metrics, pooling, and uncertainty estimation; Appendix J describes data construction, annotation provenance, model-assisted wording, and limitations. Appendices B, C, and F document subspace estimation, error-type diagnosis, scalar calibration, weight editing, and backbone-specific settings. Appendices D and E report direction, rank, strength, and layer controls. Appendix G details cue perturbations and human references, while Appendices H and I cover baseline implementations, PEFT budgets, and external evaluation. Appendices K, L, and M provide detector interfaces, candidate-selection controls, paired error transitions, ALS ablations, and image-budget-matched comparisons. These sections complement the benchmark and method definitions in Sections 3 and 5.

ETHICS STATEMENT

LongHOI-Bench reuses images from MS COCO, Objects365, and SWiG, including images of people. Source-image rights remain with their owners, and reuse is subject to their applicable terms. Contextual-intention annotations record human interpretations of visible scenes. Their intended use is model evaluation, not consequential assessment of individuals or their private intentions. Sections 3.2 and 4.3 describe annotation and evaluation roles.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025.

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: Hand and Object Tracking in 3D from Egocentric Multi-View Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7061–7071, 2025.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to Detect Human-Object Interactions. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 381–389. IEEE, 2018. doi: 10.1109/WACV.2018.00048. URL https://doi.org/10.1109/WACV.2018.00048.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. ICT: Image-Object Cross-Level Trusted Intervention for Mitigating Object Hallucination in Large Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4209–4221. IEEE, 2025. doi: 10.1109/CVPR52734.2025.00398. URL https://doi.org/10.1109/CVPR52734.2025.00398.

Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, Song-Chun Zhu, Tao Gao, Yixin Zhu, and Siyuan Huang. YouRefIt: Embodied Reference Understanding with Language and Gesture. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1385–1395. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00142. URL https://doi.org/10.1109/ICCV48922.2021.00142.

Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M. Rehg. Detecting Attended Visual Targets in Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Chong_Detecting_Attended_Visual_Targets_in_Video_CVPR_2020_paper.html.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. InternLM-XComposer2: Mastering Free-form Text-Image Composition and Comprehension in Vision-Language Large Model. *arXiv preprint arXiv:2401.16420*, 2024. URL https://arxiv.org/abs/2401.16420.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552.

Joseph L. Fleiss. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, 76(5):378–382, November 1971. doi: 10.1037/h0031619. URL https://doi.org/10.1037/h0031619.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal Large Language Models Can See but Not Perceive. In *Computer Vision–ECCV 2024*, Lecture Notes in Computer Science, pp. 148–166. Springer, 2024. doi: 10.1007/978-3-031-73337-6_9. URL https://doi.org/10.1007/978-3-031-73337-6_9.

Paul A. Games and John F. Howell. Pairwise multiple comparison procedures with unequal N's and/or variances: A Monte Carlo study. *Journal of Educational Statistics*, 1(2):113–125, 1976. doi: 10.3102/10769986001002113.

Pei Geng, Jian Yang, and Shanshan Zhang. HORP: Human-Object Relation Priors Guided HOI Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Google DeepMind. Gemini 2.0 Flash Model Card, April 2025. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-0-Flash-Model-Card.pdf.

Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007. doi: 10.1109/CVPR.2007.383267.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2022.

Yupeng Hu, Changxing Ding, Chang Sun, Shaoli Huang, and Xiangmin Xu. Bilateral Collaboration with Large Vision-Language Models for Open Vocabulary Human-Object Interaction Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01274. URL https://doi.org/10.1109/CVPR52733.2024.01274.

Yoojin Jang, Junsu Kim, Ha Yeon Kim, Eun-Ki Lee, Eun-Sol Kim, Seungryul Baek, and Jaejun Yoo. B-RIGHT: Benchmark Re-evaluation for Integrity in Generalized Human-Object Interaction Testing. In *Proceedings of the British Machine Vision Conference*, 2025.

Junyang Ji, Qifan Liu, Wenming Yang, and Zhihai He. CausalLens: Sensitivity-Guided Multi-Head Causal Intervention for Hallucination Mitigation in Large Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Siyu Jiang, Feiyang Chen, Xiaojin Zhang, and Kun He. KVSmooth: Mitigating Hallucination in Multi-modal Large Language Models through Key-Value Smoothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01316. URL https://doi.org/10.1109/CVPR52733.2024.01316.

Brian Lester, Rami Al-Rfou, and Noah Constant. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. NaturalBench: Evaluating Vision-Language Models on Natural Adversarial Samples. In *Advances in Neural Information Processing Systems*, volume 37, pp. 17044–17068. Curran Associates, Inc., 2024. doi: 10.52202/079017-0542. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/1e69ff56d0ebff0752ff29caaddc25dd-Abstract-Datasets_and_Benchmarks_Track.html.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: On Pre-training for Visual Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26689–26699, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Lin_VILA_On_Pre-training_for_Visual_Language_Models_CVPR_2024_paper.html.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Computer Vision–ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A. Raffel. Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024a. doi: 10.1109/CVPR52733.2024.02484. URL https://doi.org/10.1109/CVPR52733.2024.02484.

Sheng Liu, Haotian Ye, and James Y. Zou. Reducing Hallucinations in Large Vision-Language Models via Latent Space Steering. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b4008025c2182bfe16fcc8566ee14d64-Abstract-Conference.html.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-Decomposed Low-Rank Adaptation. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024b.

Jinguo Luo, Weihong Ren, Weibo Jiang, Xi'ai Chen, Qiang Wang, Zhi Han, and Honghai Liu. Discovering Syntactic Interaction Clues for Human-Object Interaction Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967. doi: 10.1070/SM1967v001n04ABEH001994.

Meta. Llama-3.2-11B-Vision-Instruct. Model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct.

J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7 (4):308–313, 1965. doi: 10.1093/comjnl/7.4.308.

OpenAI. GPT-4o System Card. OpenAI, August 2024. URL https://openai.com/index/gpt-4o-system-card/.

OpenAI. GPT-5 System Card. OpenAI, August 2025. URL https://openai.com/index/gpt-5-system-card/.

Belinda Phipson and Gordon K. Smyth. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 39, 2010. doi: 10.2202/1544-6115.1585.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded Situation Recognition. In *Computer Vision–ECCV 2020*, volume 12349 of *Lecture Notes in Computer Science*, pp. 314–332. Springer, 2020. doi: 10.1007/978-3-030-58548-8_19. URL https://doi.org/10.1007/978-3-030-58548-8_19.

Adrià Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where Are They Looking? In *Advances in Neural Information Processing Systems*, volume 28, pp. 199–207, 2015.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437/.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A Large-Scale, High-Quality Dataset for Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8430–8439. IEEE, 2019. doi: 10.1109/ICCV.2019.00852. URL https://doi.org/10.1109/ICCV.2019.00852.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12964–12974, 2025. doi: 10.18653/v1/2025.acl-long.634. URL https://aclanthology.org/2025.acl-long.634/.

Danyang Tu, Xiongkuo Min, Huiyu Duan, Guodong Guo, Guangtao Zhai, and Wei Shen. End-to-End Human-Gaze-Target Detection with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2202–2210. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00224. URL https://doi.org/10.1109/CVPR52688.2022.00224.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. MLLM Can See? Dynamic Correction Decoding for Hallucination Mitigation. In *International Conference on Learning Representations (ICLR)*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/24079b91da7257cb78805262996152b8-Abstract-Conference.html.

Hengfei Wang, Anshul Gupta, Pierre Vuillecard, and Jean-Marc Odobez. Eyes on VLM: Benchmarking Gaze Following and Social Gaze Prediction in Vision Language Models. *arXiv preprint arXiv:2605.19859*, 2026. doi: 10.48550/arXiv.2605.19859. URL https://arxiv.org/abs/2605.19859.

Shijing Wang, Chaoqun Cui, Yaping Huang, Hyung Jin Chang, and Yihua Cheng. VL4Gaze: Unleashing Vision-Language Models for Gaze Following. *arXiv preprint arXiv:2512.20735*, 2025b. doi: 10.48550/arXiv.2512.20735. URL https://arxiv.org/abs/2512.20735.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. CogVLM: Visual Expert for Pretrained Language Models. *arXiv preprint arXiv:2311.03079*, 2023. URL https://arxiv.org/abs/2311.03079.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo,

Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025c. doi: 10.48550/arXiv.2508.18265. URL https://arxiv.org/abs/2508.18265.

B. L. Welch. On the comparison of several mean values: An alternative approach. *Biometrika*, 38 (3–4):330–336, 1951. doi: 10.1093/biomet/38.3-4.330.

Eastman Z Y Wu, Yali Li, Yuan Wang, and Shengjin Wang. Exploring Pose-Aware Human-Object Interaction via Hybrid Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Mingrui Wu, Yuqi Liu, Jiayi Ji, Xiaoshuai Sun, and Rongrong Ji. Toward Open-Set Human Object Interaction Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6066–6073, 2024b. doi: 10.1609/aaai.v38i6.28422.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4818–4829. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00461. URL https://doi.org/10.1109/CVPR52733.2024.00461.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024a. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Jie Yang, Bingliang Li, Ailing Zeng, Lei Zhang, and Ruimao Zhang. Open-World Human-Object Interaction Detection via Multi-modal Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating Object Hallucinations in Large Vision-Language Models via HalluSpace Projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14635–14645. IEEE, 2025. doi: 10.1109/CVPR52734.2025.01364. URL https://doi.org/10.1109/CVPR52734.2025.01364.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800*, 2024. doi: 10.48550/arXiv.2408.01800. URL https://arxiv.org/abs/2408.01800.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. *arXiv preprint arXiv:2304.14178*, 2023. URL https://arxiv.org/abs/2304.14178.

Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. Dynamic multimodal activation steering for hallucination mitigation in large vision-language models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2602.21704.

Hangjie Yuan, Shiwei Zhang, Xiang Wang, Samuel Albanie, Yining Pan, Tao Feng, Jianwen Jiang, Dong Ni, Yingya Zhang, and Deli Zhao. RLIPv2: Fast Scaling of Relational Language-Image

Pre-Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21592–21604. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01979. URL https://doi.org/10.1109/ICCV51070.2023.01979.

Frederic Z Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. Exploring Predicate Visual Context in Detecting of Human-Object Interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1904.09675.

Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and Gaze: Inferring Action-Aware Points for One-Stage Human-Object Interaction Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13234–13243. IEEE, 2021. doi: 10.1109/CVPR46437.2021.01303. URL https://doi.org/10.1109/CVPR46437.2021.01303.

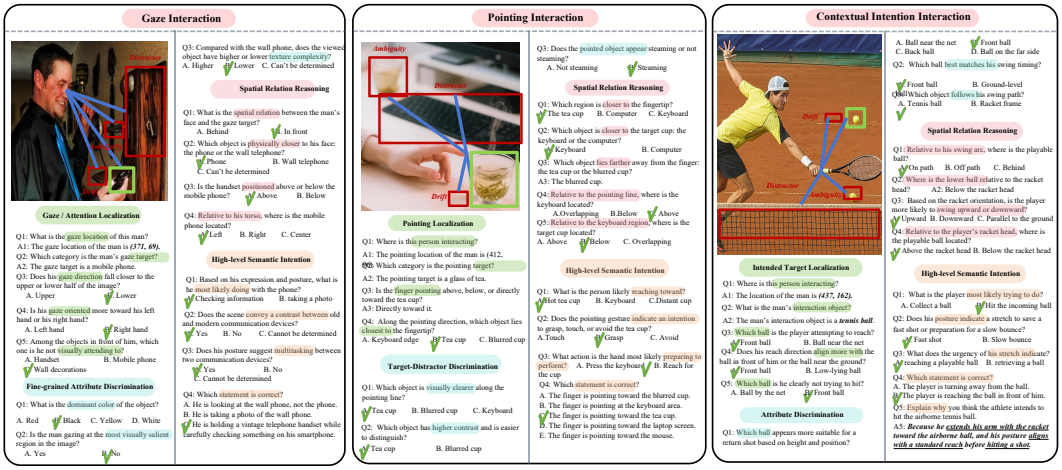

Figure A1: **Questions and typed alternatives in LongHOI-Bench.** From left to right, gaze, pointing, and contextual evidence identify the designated person's target. Each scene supports localization, attribute, spatial-relation, and intention questions. Green boxes mark verified targets, and red boxes illustrate drift, distractor, and ambiguity alternatives. Check marks identify answers.

A Experimental Scope and Evaluation Protocol

The appendix provides the construction and evaluation details for LongHOI-Bench, followed by the estimation, diagnosis, and calibration procedures used in CLASP. Figure A2 summarizes the relationship between spectral energy, error-type separation, and intervention performance. Controlled experiments examine direction identity, rank, strength, estimation depth, layer allocation, and transfer. Later sections describe annotation provenance, matched perturbations, and the evaluation interface for contact-oriented human-object interaction detectors.

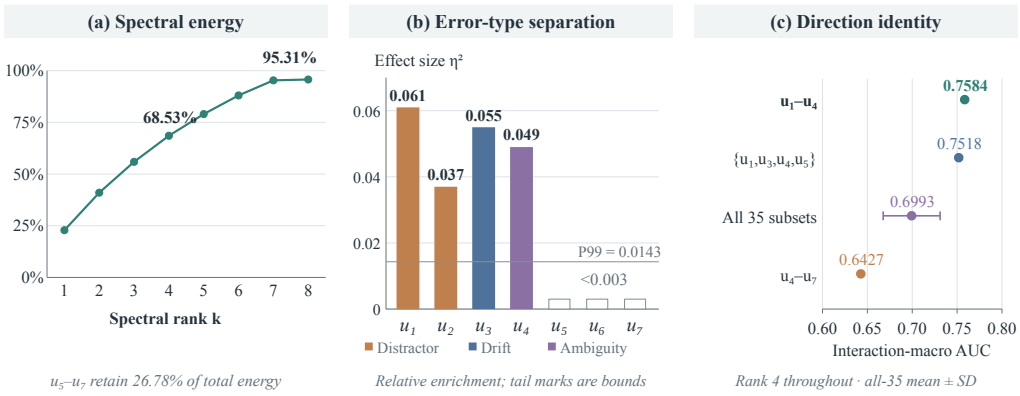

Figure A2: **Diagnostic evidence for selective suppression.** (a) Cumulative energy of the first four and seven directions. (b) Error-type effects and the permutation P99 reference; outlined tail bars indicate upper bounds. Colors represent relative error enrichment. (c) Fixed-rank controls with interaction-macro AUC; the all-35 interval shows one SD across direction subsets.

A.1 Images, Questions, Boxes, and Contrasts

LongHOI-Bench contains 6,240 images and 56,880 questions. Its 31,680 verified boxes comprise 6,240 targets and 25,440 negative alternatives. Several questions can share an image or target, and

a negative box can be paired with the correct target under a particular query. We distinguish these units when reporting dataset size, constructing experimental subsets, and estimating uncertainty.

The four question families contain 14,400 localization, 14,160 attribute, 13,920 spatial-relation, and 14,400 contextual-intention items (Figure A1). Question families specify the requested answer, while gaze, pointing, and contextual intention specify the evidence linking the person to the target. For example, an attribute question can ask about the color of a gaze target. This organization evaluates several task demands using a common person-target annotation.

We split by image into 4,160 training, 1,040 validation, and 1,040 test images. Associated questions, boxes, and image derivatives inherit this assignment. The test set contains 372 gaze, 318 pointing, and 350 contextual-intention images. The 4:1:1 ratio applies to the full collection; interaction-specific counts follow the resulting image assignment. The pointing perturbation experiment draws 300 of its 318 test images by random sampling without replacement.

CLASP uses image-disjoint training pools for estimation, diagnosis, and calibration. The fixed estimation pool E contains 832 images, 1,664 queries, and 3,287 distinct negatives, yielding 3,287 paired contrasts. Pool G contains 617 other images, with one preselected diagnostic pair per image. Each calibration pool C_s draws 42 images from the remaining training images and includes all of their questions. Thus $E \cap G = \varnothing$ and $C_s \cap (E \cup G) = \varnothing$. The calibration budget is approximately one percent of training images: $42/4,160 \simeq 1.01\%$.

A.2 Input interfaces and answer conditions

The MLLM interface supplies an image, designated human box, interaction type, and task question. Localization requests a target box; category-aware detection additionally requires a category and confidence. Other questions request a categorical or open-ended answer. The detector interface supplies the image, human box, and interaction type directly. This interface supports a separate evaluation of target detection by contact-oriented HOI systems.

Paired estimation inputs keep the image and question fixed and change the candidate box supplied as normalized text coordinates. The image remains unmarked. The resulting state difference measures the response to a correct or verified incorrect candidate under the same query. Calibration and test inputs omit this candidate target box. Direction-identity controls assess the utility of the estimated components when the model selects its own target.

Local models use greedy decoding, and API models use temperature zero where supported. Each question receives one output per condition. CLASP's three seeds resample the calibration images while retaining the estimated subspace. Comparisons across input conditions or image perturbations preserve the same question set, allowing outputs to be paired by their source image and question.

A.3 Localization prompt, parsing, and API dates

Localization uses the following output schema. It is followed by the designated person's box, interaction type, and question. Neither the target box nor the answer is supplied. Caption-only conditions replace the image with its caption while retaining the task information. The instruction reads:

```
Identify the object referred to by the
specified person and interaction.  Return
JSON only:  {"predictions":[{"category":"...",
"bbox":[x_min,y_min,x_max,y_max],"confidence":0.0}]}
Coordinates must be normalized to [0,1].  Return at most
five candidates, ordered by confidence.
```

The parser removes an enclosing Markdown code fence, checks JSON syntax, and validates each candidate's numeric values, coordinate range, and box order. Invalid candidates are discarded individually. Responses with no remaining candidate receive the invalid-prediction scores defined below. A frozen alias table normalizes category names. Parsing and alias mapping apply these frozen rules to model outputs without consulting the verified answer.

Multiple-choice options are shuffled before evaluation, with the answer index updated to match. The resulting order is held fixed across compared models and conditions. Responses must contain a

single valid option letter. Unparseable responses count as incorrect, including explanations that fail this output requirement. API evaluations took place from June 19 to June 27, 2026, excluding June 23 and June 26. The same response and parsing conventions apply across conditions.

A.4 Heatmap construction and ROC-AUC

All models use the same conversion from predicted boxes to a spatial heatmap. Let $\mathcal{B}_i = \{(b_{ij}, c_{ij})\}_j$ contain valid boxes and their explicit confidence scores for question i. In normalized coordinates $[0,1]^2$, box j has center $\boldsymbol{\mu}_{ij}$, width w_{ij}, and height h_{ij}. On the 64×64 grid, the Gaussian centered on each predicted box uses width- and height-dependent standard deviations

$$\sigma_{x,ij} = \max(w_{ij}/2, 1/64), \qquad \sigma_{y,ij} = \max(h_{ij}/2, 1/64). \tag{A1}$$

The predicted heatmap sums the confidence-weighted contributions:

$$H_i(\mathbf{z}) = \sum_{j:(b_{ij},c_{ij})\in\mathcal{B}_i} c_{ij} G_{ij}(\mathbf{z}). \tag{A2}$$

The minimum scale keeps narrow boxes represented on the grid. Heatmap construction uses predicted geometry and confidence; target annotations subsequently provide the labels for scoring.

For question i, let $\mathcal{P}_i$ contain grid centers inside the union of valid target boxes and $\mathcal{N}_i$ contain the remaining centers. ROC-AUC compares their predicted scores, assigning half credit to ties:

$$\mathrm{AUC}_i = \frac{\sum_{p\in\mathcal{P}_i}\sum_{n\in\mathcal{N}_i}\left[\mathbf{1}\{H_i(p) > H_i(n)\} + \frac{1}{2}\mathbf{1}\{H_i(p) = H_i(n)\}\right]}{|\mathcal{P}_i|\,|\mathcal{N}_i|}. \tag{A3}$$

We calculate this spatial discrimination score for each question and aggregate it using the table's stated convention. Resampling retains all questions associated with an image.

A response with no valid prediction receives a constant heatmap and AUC 0.5. Label-independent random scores have the same expected AUC. Scores below 0.5 occur when the heatmap systematically assigns higher values outside the target region. All reported AUC values use this heatmap ROC protocol, with cue effects expressed as direct differences between conditions.

A.5 Distances, detection, and answer scoring

Let $\widehat{\mathbf{c}}_i$ be the highest-confidence predicted box center in original pixel coordinates for an image of width W_i and height H_i^{img}. The valid reference centers for the same target are $\{\mathbf{c}_{ir}\}_{r=1}^{R_i}$. We define

$$d_{ir} = \frac{\|\widehat{\mathbf{c}}_i - \mathbf{c}_{ir}\|_2}{\sqrt{W_i^2 + (H_i^{\mathrm{img}})^2}}, \qquad \text{Avg. D}_i = \frac{1}{R_i}\sum_r d_{ir}, \qquad \text{Min. D}_i = \min_r d_{ir}. \tag{A4}$$

Avg.D and Min.D are then averaged across questions. Dividing pixel distance by the image diagonal preserves the aspect ratio. Invalid predictions receive distance one. With one valid reference center, Avg.D equals Min.D; with multiple centers, Min.D is at most Avg.D.

Category-aware detection orders predictions by their explicit self-reported confidence. Within each question and category, a prediction matches at most one ground-truth instance at IoU ≥ 0.5, and duplicate predictions count as false positives. We compute 101-point interpolated AP and average it over categories with positive examples. The resulting score depends jointly on the predicted location, assigned object category, and self-reported confidence. These confidence scores are not assumed to be calibrated across models; they can change AP rankings and the relative contributions of multiple predicted boxes to the resulting spatial heatmap.

Each human localization query permits one box and one normalized category. Human confidence is fixed at 1.0; all true and false positives within a tied score group are added together before updating precision and recall. This preserves invariance to ordering within a score group. The human audit and permutation check are reported in Section G.5.

Categorical accuracy is the proportion of correctly scored items in the specified question set. Open-ended answers use BERTScore F1 (Zhang et al., 2020), with `bert-score=0.3.13`, English `roberta-large`, layer 17, `idf=False`, and baseline rescaling enabled. This semantic-similarity measure complements the localization and categorical scores, allowing explanation quality to be compared with performance on selecting and identifying the target.

A.6 Pooling and units of uncertainty

Interaction-macro AUC and accuracy assign equal weight to the three interaction-specific scores. Pooled scores weight questions equally across the combined set. The two aggregations are defined over interaction subsets Q_t and individual question scores m_i as follows:

$$M_{\text{macro}} = \frac{1}{3}\sum_{t=1}^{3}\frac{1}{|Q_t|}\sum_{i\in Q_t} m_i, \qquad M_{\text{pooled}} = \frac{\sum_{t=1}^{3}\sum_{i\in Q_t} m_i}{\sum_{t=1}^{3}|Q_t|}. \tag{A5}$$

Pooled mAP is computed by rerunning confidence-ranked AP on the combined evaluation set. Tables distinguish macro frozen references from pooled edited-model scores.

Each MMBench-style configuration evaluates the same 6,380 binary-scored items under three calibration seeds. Reported accuracy averages the three correct-count fractions, with 6,380 as the denominator for each seed. Displayed capability changes subtract means rounded to two decimal places in percentage units, giving $80.96 - 81.72 = -0.76$ and $81.13 - 81.72 = -0.59$. Confidence intervals, significance decisions, and preservation checks use unrounded scores.

B Independent Subspace Estimation and Error-Type Diagnosis

B.1 Paired states and the matrix convention

A grounded input x_i^+ and misgrounded input x_i^- share the image, question, and coordinate format. They differ only in the candidate written as $[x_{\min}/W, y_{\min}/H, x_{\max}/W, y_{\max}/H]$. The original image receives no rectangle, crop, or color overlay. Both inputs therefore present the same kind of spatial information while nominating different candidate targets.

We collect $\mathbf{x}_{\ell,i}^{\pm} \in \mathbb{R}^d$ at the position immediately before the first answer token. Samples form the columns of $\mathbf{X}_\ell^+$ and $\mathbf{X}_\ell^-$, each of shape $d \times N$. At layer 20, the paired difference matrix is

$$\mathbf{D}_{20} = \left[\mathbf{x}_{20,1}^+ - \mathbf{x}_{20,1}^-, \ldots, \mathbf{x}_{20,N}^+ - \mathbf{x}_{20,N}^-\right] = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \qquad N = 3,287. \tag{A6}$$

The columns of $\mathbf{U}$ define hidden-space directions, and $\mathbf{V}$ describes variation across the paired samples. CLASP uses these left singular vectors as the candidate intervention basis. Each column of $\mathbf{D}_{20}$ corresponds to one paired contrast. We form $\mathbf{D} = \mathbf{X}^+ - \mathbf{X}^-$ directly, without additional centering or column normalization, and compute its reduced SVD in FP32 using `torch.linalg.svd(D, full_matrices=False)`. Direction j has energy $\sigma_j^2 / \sum_k \sigma_k^2$. This preprocessing is separate from the feature normalization used for LDR.

B.2 A separate diagnostic sample

Diagnosis uses 617 training images outside E, with one pair selected in advance per image: 219 drift, 204 distractor, and 194 ambiguity pairs. This pool also excludes calibration images. Each diagnostic contrast is projected onto the fixed basis estimated from E, so the direction estimates and their error-type associations are measured on disjoint training images.

For diagnostic difference $\mathbf{d}_i$ and unit direction $\mathbf{u}_j$, we compute normalized projection energy:

$$s_{ij} = \frac{(\mathbf{u}_j^\top \mathbf{d}_i)^2}{\|\mathbf{d}_i\|_2^2}. \tag{A7}$$

Squaring makes this score invariant to the arbitrary sign of a singular vector. Normalization expresses the fraction of contrast energy concentrated along that direction, allowing diagnostic comparisons across pairs with different state-difference magnitudes.

For error group g, let n_g be its size and $\bar{s}_{gj}$ its mean projection score; $\bar{s}_j$ is the overall mean. The between-group share of the total variation defines the error-type effect size:

$$\eta_j^2 = \frac{\sum_g n_g(\bar{s}_{gj} - \bar{s}_j)^2}{\sum_i (s_{ij} - \bar{s}_j)^2}. \tag{A8}$$

This quantity measures the share of score variation associated with separation between group means. Score comparisons identify the error groups with greater relative concentration.

Welch's ANOVA tests group-mean differences under unequal variances (Welch, 1951), and Games–Howell comparisons examine the group pairs (Games & Howell, 1976). Benjamini–Hochberg adjustment accounts for multiple comparisons (Benjamini & Hochberg, 1995). Effect sizes quantify the error-type separation for each of the seven analyzed directions alongside these statistical tests.

B.3 Permutation reference and the diagnostic cliff

We generate 5,000 permutations by shuffling error labels within gaze, pointing, and contextual-intention strata. Projection scores and the estimated basis remain fixed. The shuffling preserves each stratum's error-type counts while breaking their association with directional scores. Because each diagnostic image contributes one pair, the pair and image are the same permutation unit.

The empirical permutation distribution has median 2.2970×10^{-3} and P99 1.4306×10^{-2}. The observed effects for u_1 through u_4 are 0.061, 0.037, 0.055, and 0.049, all exceeding P99. Effects for the remaining three directions are below 0.003. The reference uses 617-pair stratified randomization with nonzero finite-permutation tail probabilities (Phipson & Smyth, 2010).

B.4 Spectral energy and relative error enrichment

Figure A2(a,b) compares cumulative spectral energy with error-type separation. For singular values σ_j, cumulative energy is $E_k = \sum_{j=1}^{k} \sigma_j^2 / \sum_j \sigma_j^2$. The first seven directions explain 95.31% of total contrast energy, while the first four explain 68.53%. Directions u_5 through u_7 contribute a further 26.78% of spectral energy while exhibiting weak error-type separation.

Training-side diagnosis selects $\mathbf{U}_4 = [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4]$ for intervention. The subsequent rank and subset sweeps evaluate this fixed selection against alternative directions under the same intervention layers and strength. These controls test how the diagnostic separation relates to target selection when ordinary task inputs ask the model to locate the target without a candidate box.

Directions u_1 and u_2 show relative enrichment for distractor errors, u_3 for drift, and u_4 for ambiguity. These preferences describe differences in mean projection energy across error groups. Several groups can contribute to a direction, and two leading directions favor distractors. Figure colors summarize the dominant relative association for each direction.

Direction u_3 illustrates the difference between spectral prominence and diagnostic separation. Its spectral contribution ranks third, but its effect size of 0.055 ranks second among the seven directions. It also favors drift, whereas the first two favor distractors. Spectral contribution and error-type association therefore provide distinct information for selecting an intervention.

B.5 Transfer from candidate contrasts to task inputs

Candidate-conditioned pairs hold the visual scene and query fixed while changing the nominated object. CLASP uses their state differences to estimate a basis, then evaluates the selected intervention on questions without supplied target boxes. The same-rank direction controls provide a direct test of which components support this transfer from controlled contrasts to ordinary target selection.

C ALS Calibration and Analytic Weight Editing

C.1 Intermediate-layer contrast signals

ALS scores layers from contrasts between grounded and misgrounded inputs. At the same pre-answer position, each residual state passes through the final normalization and shared language-

model head, producing full-vocabulary distributions $p^+_{\ell,i}$ and $p^-_{\ell,i}$. The three signals below measure distributional change, uncertainty, and feature separation under that shared readout.

The first signal is grounded-to-misgrounded KL divergence, averaged over paired inputs:

$$K_\ell = \frac{1}{N} \sum_i \sum_v p^+_{\ell,i}(v) \log \frac{p^+_{\ell,i}(v)}{p^-_{\ell,i}(v)}. \tag{A9}$$

We use $T = 1.0$ and cast LM-head logits to FP32 before computing $\log p^\pm = \log_\mathrm{softmax}(z^\pm/T)$. With $p^+ = \exp(\log p^+)$, the implementation sums $p^+(\log p^+ - \log p^-)$ over the full vocabulary for each pair and then averages over pairs. The log-softmax operation uses log-sum-exp stabilization; the vocabulary is not truncated. The full vocabulary contributes to the comparison at each matched position. Large values indicate a substantial readout change between the correct target and an incorrect candidate under the same image and task question, averaged over that layer's pairs.

The second signal measures the signed change in readout entropy:

$$\Delta H_\ell = \frac{1}{N} \sum_i \left[H(p^-_{\ell,i}) - H(p^+_{\ell,i}) \right], \qquad H(p) = -\sum_v p(v) \log p(v). \tag{A10}$$

A positive value indicates higher average entropy under misgrounded candidates. KL also captures shifts between vocabulary entries at similar entropy.

The third signal, LDR, scales the separation of feature means by within-condition dispersion:

$$\mathrm{LDR}_\ell = \frac{\|\boldsymbol{\mu}^+_\ell - \boldsymbol{\mu}^-_\ell\|_2^2}{\mathrm{tr}(\boldsymbol{\Sigma}^+_\ell) + \mathrm{tr}(\boldsymbol{\Sigma}^-_\ell) + 10^{-8}}. \tag{A11}$$

We L2-normalize features, center each condition at its own mean, and use sample covariance with denominator $n - 1$. The trace is computed by summing coordinate-wise sample variances, with no shrinkage or full covariance-matrix construction. The 10^{-8} term stabilizes the denominator.

LDR compares correct and incorrect candidate conditions in feature space. KL and entropy compare their language-head readouts. Together, these signals describe complementary changes induced by the same paired inputs; error-type labels enter the earlier rank diagnosis.

Each signal is min–max normalized over the candidate language layers of its backbone. Constant signals become zero, and the estimation-stage normalization constants remain fixed during evaluation. Using tildes for normalized signals, ALS computes

$$S_\ell = \alpha \widetilde{K}_\ell + \beta \widetilde{\Delta H}_\ell + \gamma \widetilde{\mathrm{LDR}}_\ell, \qquad \alpha + \beta + \gamma = 1. \tag{A12}$$

The coefficients balance the three signals. A separate softmax over the selected layer scores determines how total suppression is allocated across the three selected layers.

C.2 Calibration objective and discrete selection

For each seed, calibration uses all questions from 42 training images outside the estimation and diagnosis pools. These are ordinary task inputs, with the target box omitted. The backbone and $\mathbf{U}_{20}$ remain fixed. Two free variables parameterize the signal weights as $(\alpha, \beta, \gamma) = \mathrm{softmax}(z_1, z_2, 0)$.

Nelder–Mead (Nelder & Mead, 1965) minimizes mean answer-token negative log-likelihood. Each objective evaluation recomputes the weighted layer scores, selects Top-3 layers, and evaluates the resulting intervention. This derivative-free search incorporates discrete layer selection directly. It stops at 80 iterations or 160 objective evaluations, whichever limit is reached first.

Optimization starts at $(z_1, z_2) = (0, 0)$, giving equal signal weights, with simplex vertices $(0, 0)$, $(0.15, 0)$, and $(0, 0.15)$. The tolerances are `xatol=1e-4` and `fatol=1e-5`; `adaptive=False`, with no restarts. We return the best evaluated point. The calibration question set stays fixed across all objective evaluations within a search.

Let θ denote the two free variables and $T(q)$ the answer-token positions for calibration question q. Calibration minimizes the mean answer-token negative log likelihood:

$$\mathcal{J}_s(\theta) = -\frac{\sum_{q \in C_s} \sum_{t \in T(q)} \log p_\theta(y_{q,t} \mid x_q, y_{q,<t})}{\sum_{q \in C_s} |T(q)|}. \tag{A13}$$

The selected layers and their allocation determine the dependence on θ. Calibration adjusts these scalar settings while preserving the estimated basis and backbone during the search. Estimation and error-type diagnosis use the separate training resources described in Section A.

For the primary backbone, calibration yields $(\alpha, \beta, \gamma) = (0.41870, 0.32790, 0.25340)$ and selects layers $\{19, 21, 24\}$, with the maximum score at layer 21. The basis is estimated at layer 20. The estimation-depth controls compare this readout with alternative layers with the same intervention locations, layer weights, and suppression strength in every comparison.

C.3 From layer scores to partial suppression

A unit-temperature softmax converts the selected scores to allocation weights:

$$\widetilde{\omega}_\ell = \frac{\exp(S_\ell)}{\sum_{m \in \mathcal{L}_3} \exp(S_m)}, \qquad \ell \in \mathcal{L}_3. \tag{A14}$$

Layers 19, 21, and 24 have approximate scores 0.7661, 0.9473, and 0.6158, yielding weights 0.327, 0.392, and 0.281. Layer 21 receives the largest suppression weight.

Let $\mathbf{a}_\ell = \mathbf{W}_\ell^O \mathbf{z}_\ell + \mathbf{b}_\ell^O$ be the attention output before residual addition. With $\mathbf{P}_4 = \mathbf{U}_4 \mathbf{U}_4^\top$ and $c_\ell = \lambda \widetilde{\omega}_\ell$, the attention output and residual update are defined by

$$\mathbf{a}'_\ell = (\mathbf{I} - c_\ell \mathbf{P}_4)\mathbf{a}_\ell, \qquad \mathbf{r}'_\ell = \mathbf{r}_\ell + \mathbf{a}'_\ell. \tag{A15}$$

The main text writes this attention-output vector as $\mathbf{h}_\ell$. The residual update adds the transformed attention output to the unchanged skip contribution $\mathbf{r}_\ell$.

We fold the linear operation into the output projection:

$$\mathbf{W}_\ell^{O\prime} = (\mathbf{I} - c_\ell \mathbf{P}_4)\mathbf{W}_\ell^O, \qquad \mathbf{b}_\ell^{O\prime} = (\mathbf{I} - c_\ell \mathbf{P}_4)\mathbf{b}_\ell^O. \tag{A16}$$

When an output bias exists, it receives the same transformation. Substitution gives Eq. A15 directly. These weights act on every prefill and decoding token processed by the selected modules, using the original attention modules throughout prompt processing and answer generation.

C.4 Component attenuation and suppression strength

Decompose the attention output into components parallel and orthogonal to $\mathrm{col}(\mathbf{U}_4)$:

$$\mathbf{a}'_\ell = (1 - c_\ell)\mathbf{a}_\parallel + \mathbf{a}_\perp. \tag{A17}$$

CLASP scales the selected component by $1 - c_\ell$ and preserves the orthogonal component of this activation. At $c_\ell = 1$, the operation becomes projection onto the orthogonal complement. For $0 < c_\ell < 1$, repeated application produces further attenuation.

With $\lambda = 1$, the selected-component multipliers are approximately 0.673, 0.608, and 0.719 at layers 19, 21, and 24. All three layers therefore receive partial suppression. Nonnegative multipliers require $\lambda \leq 1/\max_\ell \widetilde{\omega}_\ell$, approximately 2.55. The strength sweep evaluates the associated grounding scores and capability losses before and near this boundary.

The component decomposition describes the immediate attention output. Subsequent residual updates and nonlinear layers propagate the edited representation through the network. We therefore evaluate behavioral preservation with POPE and MMBench-style scores alongside grounding performance, using a common one-percentage-point tolerance for the two capability measures.

D Direction Identity, Rank Selection, and Strength

Intervention settings and capability-preservation thresholds are fixed through training/validation diagnosis before testing. Estimation-layer selection uses only E, following Section F.

D.1 Direction identity at fixed rank

Figure A2(c) compares four-direction interventions under the same layer set and strength. The fixed rank isolates direction identity within the estimated basis.

Table A1 evaluates all $\binom{7}{4} = 35$ four-direction subsets, together with selected rank-three and rank-seven controls. Intervention layers remain $\{19, 21, 24\}$ and $\lambda = 1$. These controls complement the spectral-prefix comparison by testing combinations with the same number of suppressed directions.

Table A1: Direction-identity controls at $\mathcal{L}_3 = \{19, 21, 24\}$ and $\lambda = 1$. Unmarked subsets have rank four; task scores are interaction-macro. SD in the all-35 row describes direction subsets. Capability changes are percentage points from frozen; MMBench Δ subtracts rounded displayed means.

Suppressed directions	AUC↑	mAP↑	ACC↑	POPE (Δ)	MMBench (Δ)
None (frozen backbone)	0.5170	0.2900	0.5713	87.43 (+0.00)	81.72 (+0.00)
$\mathbf{u}_1$–$\mathbf{u}_4$ (CLASP)	0.7584	0.5513	0.7256	86.91 (−0.52)	81.19 (−0.53)
$\mathbf{u}_4$–$\mathbf{u}_7$	0.6427	0.4019	0.6498	86.67 (−0.76)	80.96 (−0.76)
$\mathbf{u}_3$–$\mathbf{u}_6$	0.7286	0.5067	0.6994	86.74 (−0.69)	81.03 (−0.69)
$\mathbf{u}_5$–$\mathbf{u}_7$ ($k = 3$)	0.6093	0.3686	0.6227	86.93 (−0.50)	81.21 (−0.51)
$\{\mathbf{u}_1, \mathbf{u}_3, \mathbf{u}_4\}$ ($k = 3$)	0.7542	0.5478	0.7239	87.08 (−0.35)	81.38 (−0.34)
Uniform 4 of 7 (all 35) mean ± SD	0.6993 ± 0.0317	0.4845 ± 0.0365	0.6866 ± 0.0208	86.79 ± 0.07 (−0.64)	81.08 ± 0.06 (−0.64)
Best subset $\{1, 2, 3, 4\}$	0.7584	0.5513	0.7256	86.91 (−0.52)	81.19 (−0.53)
Best alternative $\{1, 3, 4, 5\}$	0.7518	0.5457	0.7213	86.84 (−0.59)	81.13 (−0.59)
$\mathbf{u}_1$–$\mathbf{u}_7$ ($k = 7$)	0.7522	0.5435	0.7191	86.58 (−0.85)	80.92 (−0.80)

The diagnostic prefix $\{u_1, u_2, u_3, u_4\}$ reaches AUC 0.7584. The equally sized tail subset $\{u_4, u_5, u_6, u_7\}$ reaches 0.6427, and the middle subset $\{u_3, u_4, u_5, u_6\}$ reaches 0.7286. Direction identity thus affects performance at fixed rank, layers, and strength.

Across the 35 four-direction subsets, mean AUC is 0.6993 with SD 0.0317. The diagnostic prefix ranks first, followed by the best alternative $\{u_1, u_3, u_4, u_5\}$ at 0.7518. This close alternative retains three of the four selected directions. The SD covers all 35 direction subsets.

D.2 Spectral criteria select different operating points

Table A2 applies several rank-selection rules to the same spectrum. The 95% cumulative-energy rule selects the smallest prefix that explains at least 95% of squared singular-value energy, giving rank seven. The elbow within the first seven directions selects three, and the diagnostic separation selects four. Each selected rank is evaluated at the same layers and strength.

Table A2: Rank-selection rules on the same spectrum at $\mathcal{L}_3$ and $\lambda = 1$. Each rule suppresses the selected spectral prefix. Task scores are interaction-macro; capability entries give POPE F1 and MMBench-style ACC, followed by changes in percentage points.

Criterion	Typed negatives	k	AUC↑	mAP↑	ACC↑	POPE / MMBench (Δ)
95% cumulative spectral energy	✗	7	0.7522	0.5435	0.7191	86.58 (−0.85) 80.92 (−0.80)
Elbow (first seven directions)	✗	3	0.7519	0.5442	0.7203	87.03 (−0.40) 81.31 (−0.41)
Participation ratio (rounded)	✗	7	0.7522	0.5435	0.7191	86.58 (−0.85) 80.92 (−0.80)
Largest eigengap (full spectrum)	✗	7	0.7522	0.5435	0.7191	86.58 (−0.85) 80.92 (−0.80)
Marchenko–Pastur threshold	✗	8	0.7476	0.5371	0.7138	86.37 (−1.06) 80.68 (−1.04)
η^2 cliff (diagnostic)	✓	4	0.7584	0.5513	0.7256	86.91 (−0.52) 81.19 (−0.53)

For eigenvalue weights $\nu_j = \sigma_j^2$, the participation ratio is

$$r_{\mathrm{PR}} = \frac{(\sum_j \nu_j)^2}{\sum_j \nu_j^2}. \tag{A18}$$

Rounding this value selects rank seven. The largest absolute eigengap over the full spectrum also selects seven, matching the cumulative-energy rule for this basis. These rows consequently use the same intervention and report the same task and capability scores.

The Marchenko–Pastur threshold uses a random-matrix spectral reference (Marchenko & Pastur, 1967) and selects eight directions. This configuration reaches AUC 0.7476 and loses 1.06 and 1.04 percentage points on POPE and MMBench-style evaluation. Both exceed the one-point tolerance.

The elbow at rank three reaches AUC/mAP/ACC 0.7519/0.5442/0.7203, with capability losses of 0.40/0.41 points. Diagnostic rank four raises all three task scores to 0.7584/0.5513/0.7256 and incurs losses of 0.52/0.53 points. Both satisfy preservation; rank four achieves stronger grounding, while rank three leaves a larger margin on the capability measures.

D.3 Rank and strength sweeps

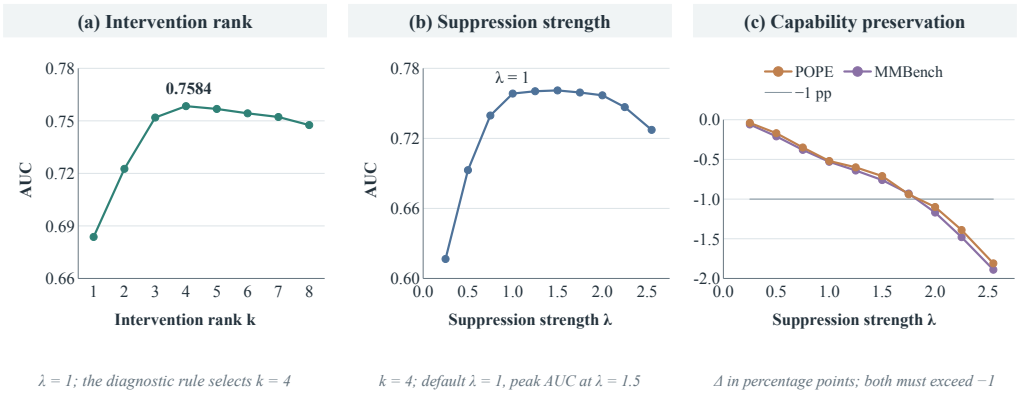

Figure A3: **Rank, strength, and capability preservation.** (a) Rank sweep at $\lambda = 1$. (b) Strength sweep at rank four, including the default $\lambda = 1$. (c) Capability changes in percentage points relative to the frozen backbone; preservation requires both changes to exceed -1. Layers remain $\{19, 21, 24\}$ throughout, with AUC averaged equally across the three interaction types.

Figure A3 and Table A3 vary rank at $\lambda = 1$ and strength at rank four. The selected layer set stays fixed. These separate sweeps show how the number of edited directions and their attenuation affect grounding and capability. Preservation requires both capability changes to exceed -1.00 points.

At fixed strength, AUC rises from 0.6837 at rank one to 0.7519 at rank three and peaks at 0.7584 at rank four. Adding directions then gives 0.7568, 0.7543, and 0.7522 at ranks five, six, and seven. Rank eight lowers AUC further and exceeds both capability-loss tolerances.

The strength sweep shows stable performance across several settings. Values from 0.75 through 1.75 satisfy preservation, with AUC between 0.7396 and 0.7611. AUC peaks at $\lambda = 1.5$, while mAP peaks at 1.25. The primary rank, direction, and layer comparisons all use the default $\lambda = 1$.

Increasing strength to $\lambda = 2$ retains AUC 0.7569 but exceeds the capability tolerance on both measures. Near the operator's nonnegative-component boundary, $\lambda = 2.55$ gives lower grounding scores and larger capability losses. The empirical preservation criterion therefore constrains intervention strength more tightly than the algebraic nonnegative-component boundary.

D.4 Paired inference for rank four versus rank seven

The rank-four versus rank-seven comparison uses 10,000 interaction-stratified, image-level paired-bootstrap resamples (Efron, 1979). Each sampled image retains its questions and both configurations' outputs. This preserves within-image dependence and the pairing between interventions while estimating uncertainty over the same set of evaluated test images.

The observed AUC/mAP/ACC gains are 0.0062/0.0078/0.0065. Their 95% intervals are [0.0023, 0.0104], [0.0031, 0.0126], and [0.0020, 0.0112]. All intervals lie above zero and the prespecified -0.01 non-inferiority margin, supporting positive task-score differences between the two fixed configurations on the common set of held-out images under image-level paired resampling.

Table A3: Rank and strength sensitivity under fixed $\mathcal{L}_3$. (a) fixes $\lambda = 1$; (b) fixes $k_{\text{edit}} = 4$. Task scores are interaction-macro and capability changes are percentage points. Preservation requires both unrounded changes to exceed -1.00; MMBench displayed differences use rounded means.

(a) Intervention rank

k	AUC↑	mAP↑	ACC↑	POPE (Δ)	MMBench (Δ)	Cum. energy
1	0.6837	0.4218	0.6489	87.28 (−0.15)	81.56 (−0.16)	22.83%
2	0.7226	0.4837	0.6834	87.16 (−0.27)	81.43 (−0.29)	41.00%
3	0.7519	0.5442	0.7203	87.03 (−0.40)	81.31 (−0.41)	55.92%
4	0.7584	0.5513	0.7256	86.91 (−0.52)	81.19 (−0.53)	68.53%
5	0.7568	0.5501	0.7242	86.82 (−0.61)	81.12 (−0.60)	79.00%
6	0.7543	0.5458	0.7214	86.69 (−0.74)	81.03 (−0.69)	88.03%
7	0.7522	0.5435	0.7191	86.58 (−0.85)	80.92 (−0.80)	95.31%
8	0.7476	0.5371	0.7138	86.37 (−1.06)	80.68 (−1.04)	95.73%

(b) Suppression strength

λ	AUC↑	mAP↑	ACC↑	POPE (Δ)	MMBench (Δ)	Pres.
0.25	0.6167	0.3648	0.6173	87.39 (−0.04)	81.66 (−0.06)	✓
0.50	0.6928	0.4473	0.6619	87.26 (−0.17)	81.51 (−0.21)	✓
0.75	0.7396	0.5197	0.7043	87.08 (−0.35)	81.34 (−0.38)	✓
1.00	0.7584	0.5513	0.7256	86.91 (−0.52)	81.19 (−0.53)	✓
1.25	0.7604	0.5551	0.7262	86.83 (−0.60)	81.08 (−0.64)	✓
1.50	0.7611	0.5546	0.7247	86.72 (−0.71)	80.96 (−0.76)	✓
1.75	0.7593	0.5528	0.7231	86.49 (−0.94)	80.79 (−0.93)	✓
2.00	0.7569	0.5492	0.7218	86.33 (−1.10)	80.55 (−1.17)	✗
2.25	0.7468	0.5347	0.7132	86.04 (−1.39)	80.24 (−1.48)	✗
2.55	0.7273	0.5126	0.6984	85.62 (−1.81)	79.83 (−1.89)	✗

E Estimation Depth, Layer Count, and Allocation

Figure A4 examines three choices in the intervention: where the basis is estimated, how many layers receive suppression, and how the selected layers share the allocation. Each comparison varies one choice and holds the other intervention settings fixed, as specified below.

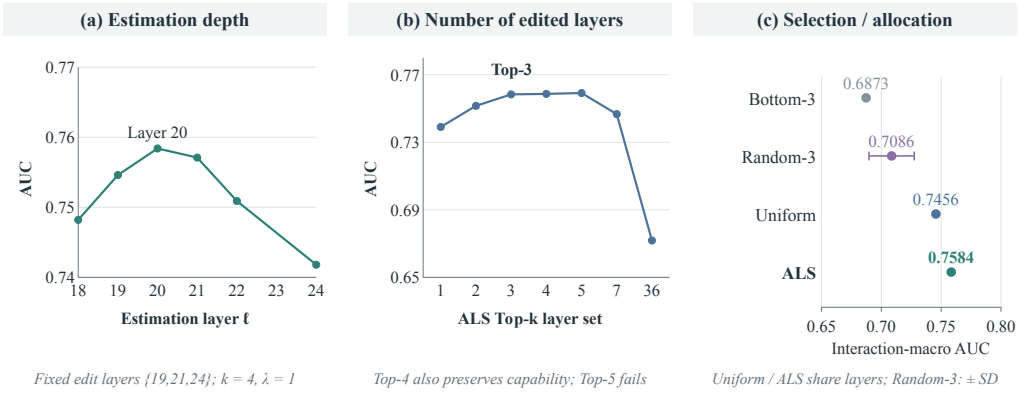

Figure A4: **Estimation depth, layer count, and allocation.** (a) Estimation-layer sweep with editing fixed at $\{19, 21, 24\}$. (b) Number of edited layers, including all 36 layers. (c) Layer selection and allocation controls. Uniform and ALS share their layers and subspace; Random-3 shows the mean and one SD across 30 sets. Rank is four, $\lambda = 1$, and AUC is interaction-macro.

E.1 ESTIMATION DEPTH AND READ–WRITE SEPARATION

The primary configuration reads paired states at layer 20 and edits attention outputs at layers 19, 21, and 24. Table A4 recomputes the basis at alternative depths while retaining this intervention set, rank four, and strength one. It reports both task performance and affinity with the layer-20 basis.

Table A4: Estimation-depth sensitivity at layers $\{19, 21, 24\}$, rank four, and $\lambda = 1$. Each estimation depth supplies its own basis; affinity measures overlap with the layer-20 basis. Task scores are interaction-macro. The final row combines the projectors estimated at layers 19 through 21.

Estimation layer	Also edited	AUC↑	mAP↑	ACC↑	Affinity with $U_4(D_{20})$
18	✗	0.7482	0.5407	0.7171	0.714
19	✓	0.7546	0.5471	0.7232	0.867
20	✗	0.7584	0.5513	0.7256	1.000
21 ($\ell^\star$)	✓	0.7571	0.5518	0.7251	0.889
22	✗	0.7509	0.5437	0.7196	0.761
24	✓	0.7418	0.5316	0.7117	0.593
Mean of projectors: 19–21	partly	0.7581	0.5522	0.7261	0.948

Estimation at layers 19, 20, and 21 produces similar task scores. Their AUCs are 0.7546, 0.7584, and 0.7571, respectively. Layer 21 gives slightly higher mAP than layer 20, at 0.5518 versus 0.5513. The useful estimation region therefore extends across adjacent layers around the primary readout.

More distant estimation depths give lower scores and smaller overlap with the layer-20 basis. Affinity is 0.714 at layer 18 and 0.593 at layer 24. Averaging the projectors from layers 19 through 21 gives AUC 0.7581, mAP 0.5522, and ACC 0.7261. This combined operator and the individual bases estimated near layer 20 yield similar task scores under the same intervention.

E.2 NUMBER OF INTERVENTION LAYERS

Table A5 varies the number of edited layers at fixed rank and strength. ALS weights are renormalized over each selected set, keeping their sum at one. Adding layers therefore redistributes the same total allocation across more attention modules. The comparison evaluates each complete configuration at the same total suppression allocation across the selected language layers.

Table A5: Layer-count and ordering controls at rank four and $\lambda = 1$. ALS weights are normalized within each selected set; task scores are interaction-macro. Layer indices are one-based. Random-3 SD covers 30 sets; capability entries give POPE F1 and MMBench-style ACC and their percentage-point changes relative to the frozen backbone for each selected set.

Configuration	Layers	$\lvert\mathcal{L}\rvert$	AUC↑	mAP↑	ACC↑	POPE / MMBench (Δ)
Top-1	$\{21\}$	1	0.7391	0.5308	0.7096	87.17 (-0.26) 81.47 (-0.25)
Top-2	$\{19, 21\}$	2	0.7516	0.5447	0.7201	87.04 (-0.39) 81.32 (-0.40)
Top-3 (CLASP)	$\{19, 21, 24\}$	3	0.7584	0.5513	0.7256	86.91 (-0.52) 81.19 (-0.53)
Top-4	$\{19, 21, 23, 24\}$	4	0.7587	0.5521	0.7252	86.63 (-0.80) 80.97 (-0.75)
Top-5	$\{18, 19, 21, 23, 24\}$	5	0.7592	0.5527	0.7249	86.31 (-1.12) 80.58 (-1.14)
Top-7	$\{17, 18, 19, 21, 23, 24, 26\}$	7	0.7467	0.5358	0.7136	85.84 (-1.59) 80.13 (-1.59)
All layers	all	36	0.6718	0.4226	0.6593	83.69 (-3.74) 77.84 (-3.88)
Bottom-3	$\{2, 4, 7\}$	3	0.6873	0.4419	0.6658	86.68 (-0.75) 80.97 (-0.75)
Random-3 (30 subsets) mean ± SD	–	3	0.7086 ± 0.0189	0.4734 ± 0.0221	0.6802 ± 0.0137	86.76 ± 0.11 (-0.67) 81.02 ± 0.14 (-0.70)

Top-1, Top-2, and Top-3 raise AUC from 0.7391 to 0.7516 and 0.7584. Top-4 and Top-5 give small further increases to 0.7587 and 0.7592. Top-4 satisfies preservation, while Top-5 exceeds the tolerance on both capability measures. Top-3 preserves a larger capability margin.

Editing all 36 layers reduces AUC to 0.6718 and incurs capability losses of 3.74 and 3.88 points. Bottom-3 reaches AUC 0.6873, while 30 random three-layer sets average 0.7086 with SD 0.0189. ALS-selected Top-3 outperforms low-score and random three-layer sets.

E.3 Allocation across selected layers

Table A6 fixes layers $\{19, 21, 24\}$, the rank-four basis, and $\lambda = 1$. Uniform allocation assigns $1/3$ to each layer, while ALS uses approximately $(0.327, 0.392, 0.281)$. Both allocations sum to one, so their comparison changes only the distribution of suppression across the selected layers.

Table A6: Allocation at fixed layers $\{19, 21, 24\}$, rank four, and $\lambda = 1$. Both rows use the same selected layers and subspace; only the allocation changes. Uniform assigns $1/3$ to each layer, matching the ALS allocation total of one. Task scores are interaction-macro.

Rule	(α, β, γ)	Layers	$\tilde{\omega}$	AUC	mAP	ACC
Uniform allocation	$(0.419, 0.328, 0.253)$	$\{19, 21, 24\}$	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	0.7456	0.5367	0.7148
ALS allocation	$(0.419, 0.328, 0.253)$	$\{19, 21, 24\}$	$(.327, .392, .281)$	0.7584	0.5513	0.7256

ALS allocation reaches AUC/mAP/ACC 0.7584/0.5513/0.7256, compared with 0.7456/0.5367/0.7148 for Uniform. Weighting the selected layers by their contrast scores therefore improves all three task measures over equal allocation within the same intervention basis.

F Backbone-Specific Calibration and Relative-Depth Transfer

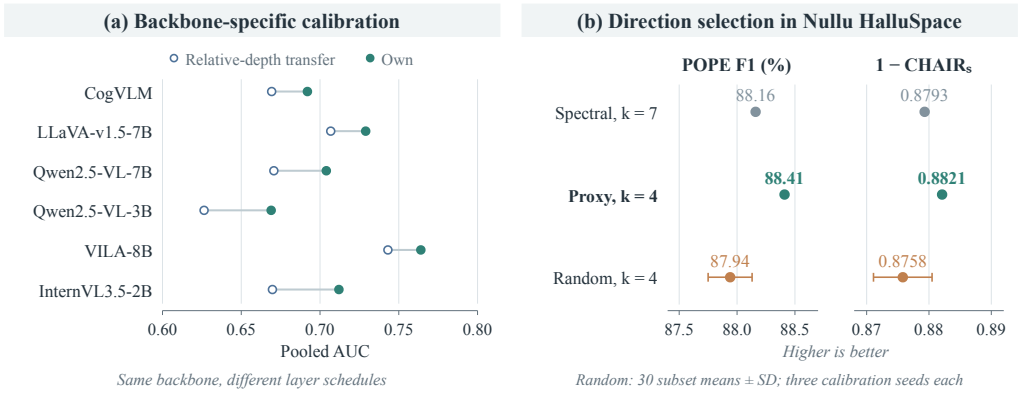

Figure A5: **Calibration transfer and HalluSpace direction selection.** (a) Pooled AUC with backbone-specific calibration or a transferred relative-depth schedule. (b) Proxy and random direction selection within Nullu HalluSpace at fixed rank four. Random intervals show one SD across 30 subset means, each averaging three calibration seeds.

Figure A5(a) and Table A7 compare backbone-specific calibration with transferring the primary model's relative layer positions and weights. The table distinguishes each model's estimation layer from its calibrated intervention set and reports depth, normalized intervention coordinates, and pooled task scores. This comparison evaluates how backbone-specific calibration changes performance relative to transferring the original layer schedule.

The estimation layer is selected separately for each backbone using only its training-side estimation pool E. At every decoder layer, we compute top-seven cumulative spectral energy E_ℓ, projector affinity A_ℓ from five interaction-stratified split-half comparisons, and rank-seven relative reconstruction error R_ℓ. We first retain layers whose E_ℓ is within 1.5 percentage points of that backbone's maximum. Among these layers, we select the largest A_ℓ; affinities within 0.01 of this maximum are treated as tied, with smaller R_ℓ preferred and any remaining tie resolved in favor of the shallower

layer. Calibration pool C_s, validation, and test data do not participate. Spectral concentration and reconstruction error summarize retained and residual energy, while split-half affinity assesses stability across samples. The selected estimation layers are reported in Table A7. The primary model estimates at layer 20 and intervenes at $\{19, 21, 24\}$, without imposing a universal $\ell^\star - 1$ rule.

Table A7: Own calibration versus transfer of the primary backbone's relative layer positions and weights. Task scores are pooled. Transferred indices round $(19, 21, 24)L/36$ to the target depth L; own depth coordinates are ℓ/L. Transfer here concerns backbone depth.

MLLM	Depth	Estimate layer	Own $\mathcal{L}_3$ (ℓ/L)	Own $\tilde{\omega}$	Transferred layers	Own AUC	Own mAP	Trans. AUC	Trans. mAP
CogVLM	32	21	$\{20, 22, 25\}$ (.625,.688,.781)	$(.316, .401, .283)$	$\{17, 19, 21\}$	0.692	0.483	0.6693	0.4618
LLaVA-v1.5-7B	32	20	$\{19, 22, 24\}$ (.594,.688,.750)	$(.341, .373, .286)$	$\{17, 19, 21\}$	0.729	0.517	0.7068	0.4906
Qwen2.5-VL-7B	28	18	$\{17, 19, 22\}$ (.607,.679,.786)	$(.302, .419, .279)$	$\{15, 16, 19\}$	0.704	0.475	0.6707	0.4389
Qwen2.5-VL-3B	36	24	$\{22, 25, 28\}$ (.611,.694,.778)	$(.291, .407, .302)$	$\{19, 21, 24\}$	0.669	0.451	0.6264	0.4087
VILA-8B	32	22	$\{19, 21, 25\}$ (.594,.656,.781)	$(.334, .386, .280)$	$\{17, 19, 21\}$	0.764	0.528	0.7431	0.5064
InternVL3.5-2B	28	19	$\{17, 20, 22\}$ (.607,.714,.786)	$(.319, .397, .284)$	$\{15, 16, 19\}$	0.712	0.514	0.6698	0.4736
InternVL3.5-8B	36	20	$\{19, 21, 24\}$ (.528,.583,.667)	$(.327, .392, .281)$	$\{19, 21, 24\}$	0.754	0.546	–	–

For a target backbone of depth L, relative-depth transfer rounds $(19, 21, 24)L/36$ to layer indices. This gives $\{17, 19, 21\}$ at depth 32 and $\{15, 16, 19\}$ at depth 28. The allocation weights are copied with the positions. External-dataset transfer keeps the same backbone and intervention.

Backbone-specific calibration improves every displayed AUC and mAP over the transferred schedule. CogVLM reaches AUC 0.692 with its own calibration versus 0.6693 with transferred positions and weights; InternVL3.5-2B reaches 0.712 versus 0.6698. Relative depth provides a usable schedule, while calibration identifies a better allocation for each model.

G Cue Perturbations, Contextual Evidence, and Error Audits

G.1 Matched perturbations for gaze and pointing

We evaluate directional evidence by comparing targeted perturbations with matched controls. Gaze perturbations blur the head/eye region, and pointing perturbations occlude the forearm. Each control applies the same operator to an irrelevant region at matched area and strength. The difference between conditions measures the additional cost of perturbing the designated cue region.

The experiment uses 300 gaze and 300 pointing images from the test set, which contains 372 and 318 images of those types. Pointing images are drawn without replacement. Clean, control, and targeted conditions retain the same questions. Clean scores refer to these subsets.

For each model and cue, we report control damage and the additional targeted-perturbation cost:

$$\Delta_{\mathrm{ctrl}} = \mathrm{AUC}_{\mathrm{control}} - \mathrm{AUC}_{\mathrm{clean}}, \qquad \Delta_{\mathrm{cue}} = \mathrm{AUC}_{\mathrm{targeted}} - \mathrm{AUC}_{\mathrm{control}}\,. \tag{A19}$$

A negative Δ_{ctrl} measures the performance loss under the control perturbation. A negative Δ_{cue} measures the further loss when the operation is placed on the designated visual cue.

Gaze cue contrasts for the frozen backbone, GPT-4o, CLASP, and humans are -0.1178, -0.1838, -0.2152, and -0.2831. The corresponding pointing contrasts are -0.1181, -0.1909, -0.2112, and -0.2562. CLASP and humans incur larger absolute cue costs than the frozen backbone.

G.2 Perturbation fidelity

Five reviewers assess 180 perturbations, comprising 90 targeted and 90 control instances and covering 15% of the 1,200 perturbed instances. Of their 900 judgments, 887 pass the fidelity check: $887/900 = 98.56\%$, reported as 98.6%. The rate aggregates five judgments per instance.

G.3 LOCAL CONTEXTUAL MASKING

For contextual intention, annotators mark local evidence regions containing supporting objects, scene signs, or interaction partners. The target and attributes needed for answering are protected with a boundary. Eligible images have supporting regions that can be masked while preserving the target and the attributes required to answer the questions.

This selection retains 287 of the 350 contextual test images, yielding 896 questions: 252 images provide three questions and 35 provide four. Clean, control, and targeted conditions share these questions. Control masks match the area and fill of targeted masks but occupy non-evidence regions. The experiment measures sensitivity to the selected local contextual evidence.

Table A8: Contextual masking on 287 test images and 896 questions ($252 \times 3 + 35 \times 4$). Accuracy is correct count divided by 896. Conditions share questions; control matches targeted mask area and fill. $\Delta_{\text{ctrl}} = \text{Control} - \text{Clean}$ and $\Delta_{\text{cue}} = \text{Targeted} - \text{Control}$ use correct counts before rounding.

Model	Clean	Control	Targeted	Δ_{ctrl}	Δ_{cue}
InternVL3.5-8B	0.6484	0.6373	0.5547	-0.0112	-0.0826
GPT-4o	0.7623	0.7500	0.6819	-0.0123	-0.0681
CLASP	0.7846	0.7734	0.6819	-0.0112	-0.0915
Human	0.9319	0.9230	0.8069	-0.0089	-0.1161

Table A8 reports target-identification accuracy. Clean/control/targeted correct counts are 581/571/497 for InternVL3.5-8B, 683/672/611 for GPT-4o, 703/693/611 for CLASP, and 835/827/723 for humans. Accuracy divides the correct count by 896; contrasts use the original counts and are rounded only for display in the table.

The contextual cue costs are -0.0826, -0.0681, -0.0915, and -0.1161 for the frozen backbone, GPT-4o, CLASP, and humans. GPT-4o's absolute decrease is smaller than the frozen backbone's, reversing their order in the directional AUC experiments. The ordering across systems consequently differs between contextual accuracy and directional localization AUC.

CLASP and GPT-4o each answer 611 questions correctly after targeted masking, while CLASP starts from higher clean and control scores. Its larger absolute cue cost reflects this change from the control condition. These aggregate counts describe performance under each masking condition.

G.4 NEGATIVE-SELECTION SENSITIVITY

A separate 600-image subset, with 200 images per interaction type, tests alternative-selection rules. It overlaps the directional-perturbation subset in 400 images. Frozen InternVL3.5-8B evaluates the same images with uniformly sampled alternatives and with rule-selected candidates. Selected options map to candidate boxes and categories under the interface in Section L.

Uniform alternatives yield AUC/mAP/ACC 0.684/0.438/0.668, while rule-selected alternatives yield 0.531/0.311/0.574. Selecting spatially, directionally, or visually plausible competitors makes this subset harder on all three metrics. Section L collects the comparison and its construction settings.

G.5 BALANCED ERROR AUDIT AND HUMAN REFERENCES

The error audit draws zero-shot responses from LLaVA-v1.5-7B, Qwen2.5-VL-7B, InternVL3.5-8B, and GPT-4o. Each model contributes 200 sampled errors per interaction type, giving $4 \times 3 \times 200 = 2,400$ responses. Two annotators label each response; a third adjudicates disagreements.

The balanced audit contains 799 drift, 968 distractor, 548 ambiguity, and 85 other errors. Distractors account for 40.3% overall and 47.6% within contextual intention. Pointing has the largest drift share, at 43.9%. These proportions describe equally sampled model-interaction cells.

Five evaluators provide the human reference on 1,080 test items. Gaze, pointing, and contextual accuracy reaches 95.1%, 97.4%, and 92.8%, respectively. Pooled localization reaches AUC 0.946 and Min.D 0.058; the main table repeats this pooled reference across interaction columns. Human and CLASP mAP share the ordering pointing, gaze, then contextual intention. Each human lo-

calization response contains one box and one normalized category, with no additional candidates. Humans do not report confidence, so their predictions receive a fixed score of 1.0. The evaluator groups predictions by confidence and adds all true and false positives in a score group before updating the precision–recall curve. Human AP is therefore invariant to ordering within the single score group. Across 200 random permutations of human predictions, the maximum absolute mAP change is below 10^{-12} for every interaction and the pooled set. The reported Human mAP values are unchanged. Model predictions remain ranked by their actual self-reported confidence, with the IoU ≥ 0.5 matching and 101-point AP convention described in Section A.5.

H Reference Comparisons and Statistical Interpretation

Table A9 summarizes interaction-macro differences between CLASP and reference systems. API comparisons average the main table's three zero-shot interaction scores; their differences are approximate because the source entries are rounded. The single-layer controls retain their respective ranks, allowing comparison with the full three-layer intervention.

Table A9: Descriptive interaction-macro differences relative to CLASP; positive values favor CLASP. API references use the zero-shot rows in Table 2, averaged equally across interactions. Their differences are approximate because the source scores are rounded. Differences use CLASP macro scores 0.7584/0.5513/0.7256 for AUC/mAP/ACC, respectively.

Comparison	Shot	ΔAUC	ΔmAP	ΔACC
vs. frozen backbone	0	+0.2414	+0.2613	+0.1543
vs. single-layer, spectral rank $k = 7$	0	+0.0237	+0.0270	+0.0199
vs. single-layer, diagnostic rank $k = 4$	0	+0.0193	+0.0205	+0.0160
vs. GPT-4o	0	$\approx +0.0627$	$\approx +0.1190$	$\approx +0.0403$
vs. GPT-5	0	$\approx +0.1081$	$\approx +0.0736$	$\approx +0.1049$
vs. Gemini-2.0-flash	0	$\approx +0.0997$	$\approx +0.1466$	$\approx +0.0853$

Single-layer spectral rank seven reaches AUC/mAP/ACC 0.7347/0.5243/0.7057, and single-layer diagnostic rank four reaches 0.7391/0.5308/0.7096. Comparing these rows tests rank within one edited layer. Comparing each with full CLASP adds the selected cross-layer allocation. The primary paired-bootstrap rank comparison holds all three intervention layers fixed.

The hallucination-mitigation baselines act on different model components. Nullu edits HalluSpace-related weight directions, VTI intervenes in visual and textual features, and ICT modifies attention heads (Yang et al., 2025; Liu et al., 2025; Chen et al., 2025). Their LongHOI-Bench runs share 832 estimation images and 42 calibration images per seed; each mechanism retains the auxiliary text inputs, perturbations, and annotations required by its implementation.

PEFT baselines optimize prompts, scaling vectors, adapters, or low-rank updates (Lester et al., 2021; Liu et al., 2022; Houlsby et al., 2019; Hu et al., 2022; Liu et al., 2024b). CLASP optimizes two scalar calibration variables and then analytically edits attention-output weights, giving zero gradient-trained backbone parameters. Its preparation uses the 832-image estimation pool and 617-image typed diagnosis pool in addition to the 42-image calibration subset.

H.1 PEFT implementation and data budgets

The PEFT comparisons in Table 3(b) use locally loaded InternVL3.5-8B configurations and 42 training images. Prompt Tuning, IA^3, and Adapter use a language-model hidden width of 4,096; LoRA and DoRA use 5,120. The five methods therefore do not share an identical backbone configuration, and their performance differences cannot be attributed solely to the adaptation method. All five methods use AdamW, bf16, an effective batch size of 32, and three epochs, with 5% warm-up, cosine learning-rate decay, and gradient clipping at 1.0. Learning rates are 3×10^{-3} for Prompt Tuning, 8×10^{-4} for IA^3, 3×10^{-4} for Adapter, 2×10^{-4} for LoRA, and 1.6×10^{-4} for DoRA. Checkpoints are selected by validation mAP, with AUC breaking ties.

LoRA and DoRA target the attention `q_proj`, `k_proj`, `v_proj`, and `o_proj` modules. LoRA uses $r = 80$ and $\alpha = 160$; DoRA uses $r = 24$ and $\alpha = 48$. Adapter inserts a bottleneck of

width 512 after attention and after the MLP in each language layer. IA^3 applies multiplicative output gates to the attention `q_proj`, `k_proj`, `v_proj`, and `o_proj` modules. Prompt Tuning uses 64 virtual tokens, initialized from the mean token embedding plus $\mathcal{N}(0, 0.02)$ noise. Trainable parameters are counted by summing `p.numel()` over parameters with `requires_grad=True`. Prompt Tuning, IA^3, Adapter, LoRA, and DoRA have 262,144, 368,640, 302,321,664, 94,371,840, and 28,753,920 trainable parameters, respectively. Reported percentages use the common reference denominator $8{,}073{,}289{,}728$: $100N_{\text{trainable}}/8{,}073{,}289{,}728$, giving 0.0032%, 0.0046%, 3.7447%, 1.1689%, and 0.3562% after rounding to four decimal places. These percentages normalize trainable counts to a shared reference rather than to each configuration's own total parameter count.

The one-percent designation refers to PEFT training or CLASP calibration, not to CLASP's complete preparation budget. CLASP additionally uses 832 estimation and 617 diagnosis images, giving $832 + 617 + 42 = 1,491$ distinct training images per seed. Consequently, these rows compare the stated training and editing protocols rather than equal total data access. Table A16 separately matches LoRA and DoRA to CLASP's 1,491-image preparation union.

H.2 Subspace geometry

For orthonormal rank-r bases, affinity is $\|\mathbf{U}_a^\top \mathbf{U}_b\|_F^2/r$. This measure lies between zero and one and is invariant to rotations within either basis. CLASP and Nullu have affinities 0.329 at rank seven and 0.275 at rank four, exceeding permutation-null P99 values 0.107 and 0.072. Their estimated subspaces therefore share structure relative to this null reference.

Cross-interaction affinities range from 0.726 to 0.764, compared with random-split medians of 0.948 to 0.959. Subspaces from different interaction types overlap substantially but agree less than random splits of the same pool. These contrasts contain shared structure together with variation specific to the visual evidence conveyed by gaze, pointing, and contextual intention.

I External Grounding and Fixed-Rank HalluSpace Controls

I.1 Unchanged transfer to external queries

We evaluate the unchanged LongHOI-Bench intervention on 400 GazeFollow, 300 VideoAttentionTarget, and 200 YouRefIt queries (Recasens et al., 2015; Chong et al., 2020; Chen et al., 2021). The subspace, selected layers, and allocation weights remain fixed on the same backbone. This tests transfer to external gaze and pointing inputs with no dataset-specific recalibration.

Five annotators independently mark the same target's center for each external query. Target-identity disagreements are adjudicated first, and the accepted centers retain their original coordinates. Avg.D and Min.D use these supplemental references under Eq. A4. For AUC, every evaluated model uses the common conversion from predicted boxes to heatmaps defined in Section A.4.

CLASP raises AUC from 0.7217 to 0.7591 on GazeFollow, from 0.6879 to 0.7228 on VideoAttentionTarget, and from 0.6146 to 0.6523 on YouRefIt. The gains are 0.0374, 0.0349, and 0.0377, respectively. Avg.D and Min.D also decrease across all three query subsets, consistent with the AUC improvements under the shared evaluation protocol.

These experiments retain the rank selected from LongHOI-Bench diagnosis. The next experiment examines direction selection in a different representation source, Nullu HalluSpace, with rank fixed at four. The two evaluations examine transfer of a fixed intervention and the value of direction identity within the separate Nullu HalluSpace representation source.

I.2 POPE and capability scoring

POPE evaluates object hallucination using object-presence questions (Li et al., 2023). Our POPE-COCO evaluation includes three sampling settings with 3,000 questions each, for 9,000 questions overall. We compute F1 separately for each setting and report the mean of the three scores using the same 9,000 questions for every evaluated intervention configuration.

MMBench-style capability evaluation uses 6,380 binary-scored items per seed. All configurations share these items, and each reported accuracy averages three seed-specific correct-count fractions.

Section A specifies display rounding. The preservation criterion checks the unrounded score against the frozen reference, with a one-percentage-point tolerance.

I.3 Direction selection inside Nullu HalluSpace

Figure A5(b) presents direction-selection controls in Nullu HalluSpace. This experiment uses truthful and hallucinated description contrasts to select four directions, then compares that selection with random four-direction subsets on external hallucination tasks.

Table A10 reports the fixed-rank comparison. Proxy selection reaches POPE F1 88.41, while random selection averages 87.94 ± 0.19. The representation source is Nullu HalluSpace, and the SD measures variation across random subset means. The proxy uses training-side description pairs.

For unit HalluSpace direction $\mathbf{u}_i$ and paired description states $\mathbf{h}_n^+$ and $\mathbf{h}_n^-$, define

$$e_{ni}^{\pm} = \frac{(\mathbf{u}_i^\top \mathbf{h}_n^{\pm})^2}{\|\mathbf{h}_n^{\pm}\|_2^2 + \epsilon}, \qquad q_i = \frac{(\bar{e}_i^- - \bar{e}_i^+)^2}{(s_{i,-}^2 + s_{i,+}^2)/2 + \epsilon}, \qquad \epsilon = 10^{-8}. \tag{A20}$$

Here $+$ denotes truthful descriptions and $-$ hallucinated descriptions. The terms $\bar{e}_i^{\pm}$ and $s_{i,\pm}^2$ are their projection-energy means and variances. The squared mean difference measures separation in either direction. We rank directions by decreasing q_i, resolve ties by direction index, and select the first four according to their scores on the fixed training-side description pool.

The random control draws 30 distinct subsets without replacement from the $\binom{7}{4} = 35$ four-direction combinations, using seed 20260917. Each subset runs under three calibration seeds. For metric m_{rs}, we first average seeds within subset r, giving $\bar{m}_r = \sum_{s=1}^{3} m_{rs}/3$, then report

$$\bar{m} = \frac{1}{30}\sum_{r=1}^{30} \bar{m}_r, \qquad s_{\text{subset}} = \sqrt{\frac{1}{29}\sum_{r=1}^{30}(\bar{m}_r - \bar{m})^2}. \tag{A21}$$

The SD therefore describes variation across 30 subset means. The LongHOI-Bench direction-identity control in Table A1 instead enumerates all 35 subsets of its own estimated basis.

Table A10: Direction selection within Nullu HalluSpace. Proxy and random controls fix rank at four; the spectral reference uses seven. POPE reports F1 (%), and CHAIR reports $1 - \text{CHAIR}_s$. Random rows give the mean and sample SD over 30 subset means, each averaging three calibration seeds. MMBench changes subtract the frozen reference of 81.72 from rounded displayed means.

Benchmark	Subspace	Direction rule	k	Task metric	MMBench (Δ)
POPE	Nullu HalluSpace	Spectral reference	7	88.16	81.11 (-0.61)
	Nullu HalluSpace	Proxy top-4 (fixed)	4	88.41	81.37 (-0.35)
	Nullu HalluSpace	Random, same rank	4	87.94 ± 0.19	81.43 ± 0.12 (-0.29)
CHAIR	Nullu HalluSpace	Spectral reference	7	0.8793	81.11 (-0.61)
	Nullu HalluSpace	Proxy top-4 (fixed)	4	0.8821	81.34 (-0.38)
	Nullu HalluSpace	Random, same rank	4	0.8758 ± 0.0047	81.43 ± 0.12 (-0.29)

The rank-seven spectral reference reaches POPE F1 88.16. Proxy-selected rank four improves this value to 88.41 and also exceeds the random rank-four mean. Random selection retains a slightly higher MMBench-style score, at 81.43 versus 81.37. The fixed-rank comparison favors proxy selection on hallucination F1 and random selection on retained MMBench-style accuracy.

CHAIR measures object hallucination in generated captions (Rohrbach et al., 2018). We report $1 - \text{CHAIR}_s$, so higher values indicate fewer sentences containing hallucinated objects. Proxy rank four reaches 0.8821, compared with 0.8793 for the rank-seven spectral reference and 0.8758 ± 0.0047 for random rank-four subsets. The result complements the POPE object-presence comparison.

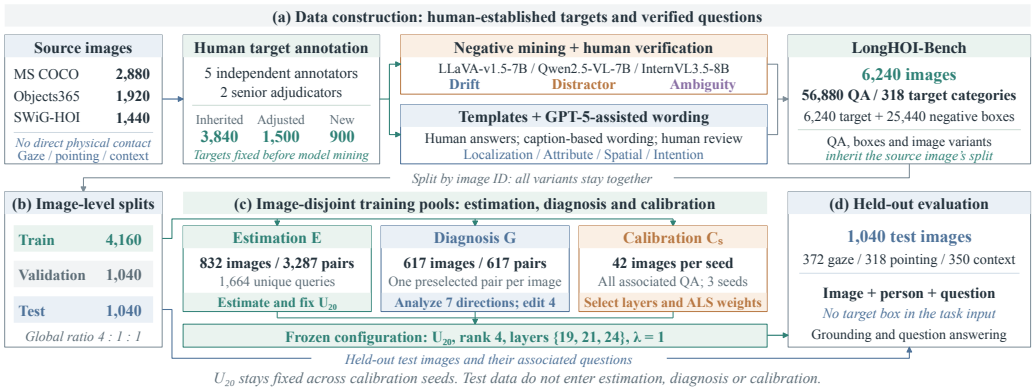

Figure A6: **Construction and experimental splits.** (a) Human target verification precedes model-assisted negative mining and question wording. (b) Questions, boxes, and image variants inherit the image-level 4:1:1 split. (c) Separate training pools estimate U_{20}, diagnose directions, and calibrate layers; the basis stays fixed across seeds. (d) Evaluation uses 1,040 test images.

J Benchmark Construction and Annotation Provenance

Figure A6 follows the process from source-image selection to held-out evaluation. Human annotation establishes targets before model-assisted negative mining and question wording. Image-level splits separate intervention design from held-out test evaluation.

J.1 Non-contact target understanding

LongHOI-Bench includes interactions without direct physical contact whose targets are identified through gaze, pointing, or contextual evidence. Selection depends on the visual person-target relation, with no fixed physical-distance threshold. Nearby and distant objects can both qualify when the scene supports a verifiable non-contact association.

Images come from MS COCO, Objects365, and SWiG-HOI (Lin et al., 2014; Shao et al., 2019; Pratt et al., 2020). Source object and role annotations provide starting points for target verification. We retain images with an identifiable person, a localizable target, and plausible alternatives (Figure A7), supporting question construction and negative-candidate mining.

Automatic screening requires an identifiable person and at least two localizable non-human objects, retaining candidates that can support a non-contact relation. It yields 13,742 COCO, 12,386 Objects365, and 7,954 SWiG-HOI images. The first manual round retains 4,961, 3,724, and 2,617, respectively. Further review excludes direct-contact-dominant scenes, unclear designated persons, targets that cannot be uniquely localized, and images without a plausible competitor. The final counts are 2,880, 1,920, and 1,440, totaling 6,240 images.

Interaction labels follow a cue-sufficiency rule. An explicit deictic arm or hand cue whose direction identifies the target takes precedence and is labeled pointing. Otherwise, eye or head direction sufficient to constrain the target yields gaze. Contextual intention requires full-body action, object affordance, scene semantics, or a partner or supporting object when neither directional cue suffices. Cases in which gaze and pointing indicate different targets are referred to senior adjudication.

Table A11 reports source composition and target-box provenance. MS COCO, Objects365, and SWiG-HOI contribute 2,880, 1,920, and 1,440 images. Gaze, pointing, and contextual-intention counts are 2,520, 1,560, and 2,160. The 2,620 multi-person images overlap these interaction categories and require the designated person to be resolved before selecting a target.

The collection uses 318 normalized interaction-target categories. Source-specific counts of 96, 281, and 203 overlap, giving a cross-source union of 318. The COCO-derived count follows this target taxonomy, which labels the interaction objects verified for LongHOI-Bench. The distribution is long-tailed: 82 categories account for 267 target instances at frequencies 1–5; 74 for 598 at 6–10; 72



Figure A7: **Representative source scenes in LongHOI-Bench.** The gallery includes gaze, pointing, and contextual-intention interactions in indoor and outdoor environments. Scenes vary in the number of people, target size, and visual clutter. These images support person-target verification, question construction, and selection of incorrect alternatives.

Table A11: Source composition and target-box provenance for MS COCO (Lin et al., 2014), Objects365 (Shao et al., 2019), and SWiG-HOI (Pratt et al., 2020). Multi-person counts overlap interaction categories. Target categories follow the normalized LongHOI-Bench taxonomy, with a cross-source union of 318. Box-provenance percentages are calculated within each source.

(a) Interaction Composition

Source	Images	Gaze	Pointing	Contextual	Multi-person	Categories
MS COCO	2,880	1,280	720	880	1,180	96
Objects365	1,920	780	500	640	800	281
SWiG-HOI	1,440	460	340	640	640	203
Total	**6,240**	**2,520**	**1,560**	**2,160**	**2,620**	**318**

(b) Bounding-box Annotations

Source	Total	Inherited	Adjusted	Newly Annotated
MS COCO	2,880	2,042 (70.9%)	602 (20.9%)	236 (8.2%)
Objects365	1,920	1,125 (58.6%)	501 (26.1%)	294 (15.3%)
SWiG-HOI	1,440	673 (46.7%)	397 (27.6%)	370 (25.7%)
Total	**6,240**	**3,840 (61.5%)**	**1,500 (24.0%)**	**900 (14.4%)**

for 1,074 at 11–20; 54 for 1,605 at 21–40; and 36 for 2,696 above 40. These grouped counts cover 318 categories and 6,240 target instances. Each target is counted once, regardless of question count.

J.2 Establishing targets before model-assisted mining

Five graduate annotators independently identify each target and inspect its box, retaining, adjusting, or creating geometry as needed. Two senior annotators adjudicate disagreements. Model predictions are hidden during target establishment and are introduced later as negative candidates. Verified person-target relations (Figure A8) supply the answer keys for question construction.

Of the 6,240 target boxes, 3,840 are inherited from source annotations, 1,500 are adjusted, and 900 are newly drawn. Their rounded proportions are 61.5%, 24.0%, and 14.4%. Annotators verify the person-target relation in every case, including targets whose source geometry can be retained.

A 20% stratified subset assesses agreement on target annotation. Fleiss' κ ranges from 0.74 to 0.86, with the lowest value for contextual intention; median pairwise box IoU is 0.89. Labels and boxes capture complementary aspects of annotation agreement (Fleiss, 1971).



Figure A8: **Verified target annotations across scenes and scales.** Green boxes identify the designated targets in gaze, pointing, and contextual-intention examples. Directional rays appear where supplied by the source annotation. The gallery includes small distant objects, larger nearby targets, and crowded scenes, illustrating the geometry used for question construction and target evaluation.

J.3 Question construction and visual access

Localization and spatial-relation questions use 12 and 16 deterministic templates. GPT-5 (OpenAI, 2025) assists attribute and intention wording from textual captions and structured annotations, while human verification supplies the targets and answer keys. The wording model receives text-only inputs; the preceding caption-generation stage has visual access.

For Objects365 and SWiG, Florence-2-large uses `<MORE_DETAILED_CAPTION>` to generate construction captions that humans revise (Xiao et al., 2024). These captions provide descriptive material for question wording alongside structured annotations. Reviewers then inspect the questions and their answers. Construction captions are stored separately from GPT-4o captions generated for the paired image-versus-text comparison in the evaluation.

Reviewers accept 76.8% of generated questions unchanged, retain 16.1% after editing, and reject 7.1%. Their checks cover wording and answer correctness. The resulting 56,880 retained questions form the four task families used in the final benchmark.

The GPT-5 construction prompt uses the following system instruction and text fields. The verified answer is supplied by annotation, and the model receives no original image.

> **System:** You assist with wording benchmark questions from verified annotations. You do not determine the answer key and you must not introduce visual facts absent from the supplied caption or structured annotations.
>
> **Input:** `caption`: {caption}; `person_reference`: {person_ref}; `interaction_type`: {`gaze|pointing|contextual_intention`}; `verified_target_category`: {target_category}; `verified_target_attributes`: {attributes}; `verified_relation`: {relation}; `verified_answer`: {answer}; `negative_candidates`: {negative_categories_and_attributes}.
>
> **Task:** Generate one natural question of type {`attribute|intention`} about the designated person and verified target. Preserve the verified answer exactly in meaning. For multiple-choice questions, produce one correct answer and three plausible but incorrect alternatives grounded only in the supplied negative candidates or structured annotations. Do not mention bounding boxes, annotation labels, model predictions, or the construction process. Avoid wording that reveals the answer.
>
> **Output JSON:** `{"question": "...", "options": ["...", "...", "...", "..."], "correct_index": 0}`

The construction output places the correct option first. Evaluation shuffles the options and updates the answer index before presenting them to a model, as described in Section A.

J.4 Scope and limitations

The collection depicts static third-person interactions. Contextual-intention labels represent culturally situated interpretations of visible evidence, rather than direct observations of private intentions. The benchmark does not evaluate temporal dynamics or establish egocentric generalization.

J.5 Typed negative candidates

LLaVA-v1.5-7B, Qwen2.5-VL-7B, and InternVL3.5-8B propose negative candidates through zero-shot prompting. Two panels of three annotators verify the proposed alternatives and assign error types. This mine-then-verify construction (Li et al., 2024) uses model predictions to identify plausible mistakes after humans establish the correct target. The three candidate-generating systems are also evaluated alongside the other models in the benchmark.

The negative-mining instruction is:

> Given the image, designated person box and interaction type, identify the most plausible interaction target. Return normalized category, box [x1,y1,x2,y2], and confidence only.

Human verification checks whether a proposal is incorrect and assigns its error type.

Error-type agreement is measured before adjudication on 4,913 negative candidates sampled by source $\times$ interaction stratum. Each candidate is independently labeled as drift, distractor, or ambiguity by three annotators from the same panel. Overall Fleiss' κ is 0.781; the two panels yield 0.789 and 0.772. All three annotators agree on 4,161 of 4,913 candidates (84.7%). This is the three-annotator unanimous agreement rate. Adjudicated labels do not enter these agreement calculations.

Drift labels localization displacement from the intended target. Distractor labels a salient irrelevant object, and ambiguity labels a plausible incorrect alternative under the verified interpretation. Figure A9 shows these error types through competing objects and displaced boxes. CLASP uses their groups to compare normalized directional projection scores.

Deduplication and quality screening retain 25,440 negatives, comprising 32.1% drift, 34.4% distractor, and 33.5% ambiguity. Construction uses no equal-type quota. The separate diagnosis pool samples 219, 204, and 194 pairs of these types, respectively, from 617 training images.

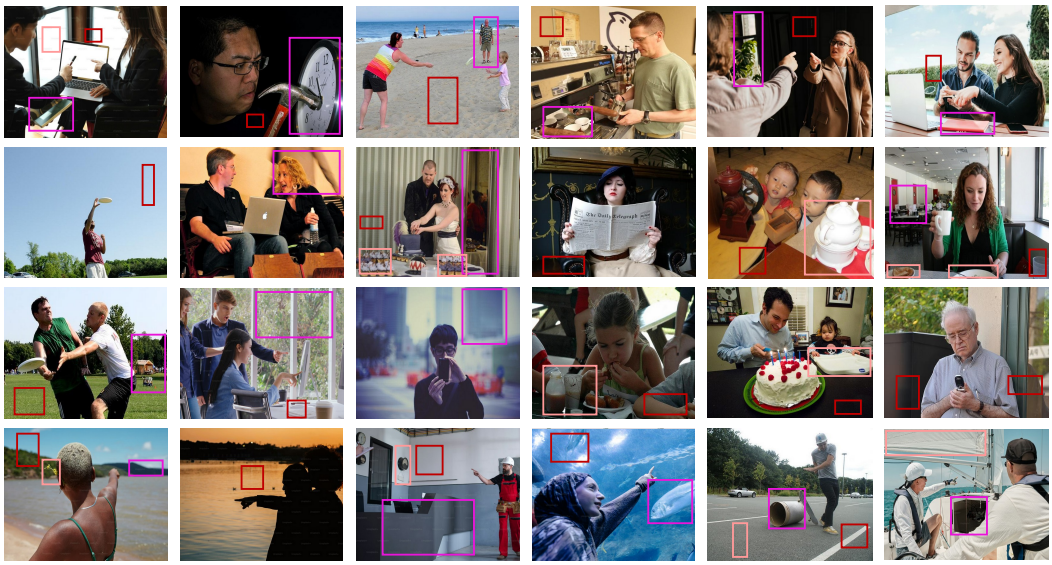

Figure A9: **Verified negative alternatives across scenes.** Rectangles identify candidate misgroundings labeled as drift, distractor, or ambiguity after the correct target is established. Several alternatives can accompany the same person-target relation. The examples illustrate competing objects and displaced boxes used to construct challenging questions and diagnose target-selection errors.

J.6 Rule-selected alternatives and spectral saliency

Multiple-choice alternatives come from non-target candidates in the same image. Spatial-near candidates have center distance at most 0.20 image diagonals and IoU below 0.10 with the target. Salient candidates occupy the top 25% by mean within-box saliency among that image's candidates. Maps are computed with the spectral-residual algorithm (Hou & Zhang, 2007).

Directional candidates lie within $\pm 20^\circ$ of the target ray and at 0.5–1.5 times its radial distance. Gaze and pointing alternatives satisfy two of the spatial, saliency, and directional criteria. Contextual-intention alternatives satisfy both non-directional criteria. These rules select competing objects using the geometry and visual structure of the same scene.

All ray coordinates use original image pixels. For gaze, the origin o is the midpoint of the two eye keypoints; with one eye missing, it is the midpoint of the valid eye and nose; with both missing, it is the head-box center. For pointing, o is the wrist on the pointing side, falling back to the same-side elbow. If both are unavailable, the directional criterion is marked unavailable. The endpoint is always the verified target-box center c_t. Writing $r = c_t - o$, a candidate center c satisfies the directional criterion when its angle to r is at most 20° and $\|c - o\|/\|r\| \in [0.5, 1.5]$. The ray criterion is not applied to contextual-intention candidates.

J.7 Language controls and construction dependence

Blind Qwen2.5-7B-Instruct reaches 27.8% accuracy on four-option questions, against 25% chance. Fine-tuning on training-split QA raises accuracy to 31.6%. Providing Florence-2-large captions yields 35.1%, adding a textual description of the scene to the language-only evaluation.

The controlled GPT-4o experiment uses the same 1,040 questions, option order, and decoding settings. Image inputs yield 687 correct answers (66.1%), while independent captions yield 440 (42.3%), a 23.8-point difference. Captions are generated from images alone, without receiving any of the evaluation questions, corresponding choices, or verified answer keys.

Localization under blind fine-tuning and GPT-4o captions gives heatmap AUC 0.5004 and 0.5019, respectively. Both values lie close to the label-independent random-score expectation of 0.5. To-

gether with four-way accuracy, these controls measure target understanding under language-only inputs, with the task questions and scoring rules retained across conditions.

Two further comparisons rank the 13 unedited models by interaction-macro AUC. Replacing the original candidates with MiniCPM-V 2.6 re-mined candidates gives Spearman $\rho = 0.94$. Comparing 600 human-only questions from 300 test images with model-assisted questions on those images gives $\rho = 0.96$. Model rankings remain broadly consistent under both construction changes.

K Detector Interface and Contact-Supervision Transfer

The detector evaluation supplies an image, designated human box, and interaction type, and requests a target box and category. It tests non-contact target detection with systems developed for HOI recognition. All rows in this comparison, including CLASP, use the detector interface.

Table A12: HOI evaluation with an image, designated human box, and interaction type as input and a target box and category as output. Asterisks identify zero-shot transfer; unmarked detectors train on LongHOI-Bench. All rows, including CLASP, use this detector interface. Backbone and interaction-specific AUC and mAP are reported for each configuration.

Method	Backbone	Gaze		Pointing		Contextual Intention	
		AUC	mAP	AUC	mAP	AUC	mAP
MP-HOI* (Yang et al., 2024b)	Swin-L+ViT	0.251	0.142	0.277	0.149	0.194	0.162
SICHOI* (Luo et al., 2024)	R101+ViT-L	0.332	0.209	0.319	0.192	0.317	0.167
RLIPv2* (Yuan et al., 2023)	Swin-L	0.270	0.135	0.286	0.132	0.229	0.096
PAFR* (Wu et al., 2024a)	Swin-L	0.358	0.264	0.354	0.271	0.326	0.258
BC-HOI* (Hu et al., 2025)	ResNet-50	0.316	0.171	0.357	0.164	0.326	0.179
PVIC* (Zhang et al., 2023)	Swin-L	0.294	0.158	0.384	0.162	0.275	0.143
HORP* (Geng et al., 2025)	Swin-L+CLIP	0.362	0.247	0.375	0.260	0.298	0.212
BC-HOI (Hu et al., 2025)	ResNet-50	0.649	0.496	0.658	0.487	0.594	0.461
PVIC (Zhang et al., 2023)	Swin-L	0.552	0.419	0.491	0.407	0.486	0.384
HORP (Geng et al., 2025)	Swin-L+CLIP	0.682	0.513	0.699	0.517	0.597	0.394
CLASP	**InternVL3.5-8B**	**0.731**	**0.542**	**0.782**	**0.548**	**0.713**	**0.534**

Table A12 separates zero-shot transfer, marked with an asterisk, from detectors trained on LongHOI-Bench. Backbone information is included for each method. The two groups show performance before and after access to the benchmark's training split under the same target-detection interface.

Zero-shot detector AUC ranges from 0.194 to 0.384 across the displayed interactions. Training on LongHOI-Bench improves several conditions substantially. HORP has the highest trained-detector AUC for each interaction, while BC-HOI gives higher contextual mAP. Localization and category-aware detection therefore distinguish different aspects of performance within this interface.

Under the detector interface, CLASP reaches gaze, pointing, and contextual AUCs of 0.731, 0.782, and 0.713. The ordinary MLLM evaluation additionally provides a task question and answer format. Its results are reported separately in the main interaction table.

L Candidate-Selection Control

Table A13 evaluates frozen InternVL3.5-8B on the same 600 test images, with 200 per interaction type, without CLASP intervention. The control samples alternatives uniformly from the same images; the benchmark applies its spatial, saliency, and directional rules. Candidate boxes instantiate the answer alternatives internally, but their coordinates are not exposed to the model. Inputs contain the image, designated person, interaction type, and candidate option text.

For this control, the model reports its selected option and confidence. The selected option maps to its associated candidate box and category, which supply the localization and detection prediction. AUC uses the common heatmap protocol, and mAP uses the selected category and model-reported confidence with IoU ≥ 0.5 and 101-point AP. ACC measures option correctness. This candidate-selection interface differs from the free-form box-generation interface in Section A.

Table A13: Alternative-selection control for frozen InternVL3.5-8B on 600 fixed test images, with 200 per interaction type. Selected options map to candidate boxes and categories for AUC and mAP; ACC measures option correctness. Uniform alternatives are sampled from same-image candidates; rule-selected alternatives satisfy the benchmark's spatial, saliency, and directional criteria as applicable to each interaction.

Alternative selection	AUC↑	mAP↑	ACC↑
Uniform same-image candidates	0.684	0.438	0.668
Rule-selected same-image candidates	0.531	0.311	0.574

Rule-selected alternatives lower all three metrics relative to uniform alternatives. The comparison shows that selecting plausible competing objects increases the target-understanding challenge on the same visual scenes. Spatial, directional, and saliency criteria are applied jointly in this experiment.

M Paired Error Transitions and Additional Controls

M.1 Paired outcomes before and after CLASP

Table A14 follows the same localization queries before and after editing. The matrix uses 2,413 valid held-out localization queries from the prespecified calibration seed 1 for which both the frozen InternVL3.5-8B model and CLASP produce a scorable outcome. It includes initially correct predictions and does not pool queries across calibration seeds. This paired query set differs from the balanced cross-model error audit described in Section G.

Table A14: Paired transitions from frozen InternVL3.5-8B to CLASP on 2,413 scorable held-out localization queries, using prespecified calibration seed 1. Entries count queries; rows show original outcomes and columns show edited outcomes. Queries are not pooled across seeds.

Before CLASP	After CLASP					Total
	Correct	Drift	Distractor	Ambiguity	Other/invalid	
Correct	1203	14	12	18	9	1256
Drift	174	94	13	29	9	319
Distractor	331	19	86	30	8	474
Ambiguity	168	13	19	101	8	309
Other/invalid	21	6	4	9	15	55
Total	1897	146	134	187	49	2413

For each error type, the correction rate is its Correct-column count divided by its row total. CLASP corrects 331/474 distractor errors (69.8%), 174/319 drift errors (54.5%), and 168/309 ambiguity errors (54.4%). Across all error types, 694 queries become correct, while 53 of the 1,256 initially correct queries become incorrect (4.2%). The net gain is 641 correct localization outcomes on this paired set. These outcome counts describe localization correctness on the selected queries, separately from confidence-ranked mAP and multiple-choice ACC.

M.2 ALS signal contributions

Table A15 compares single-signal and leave-one-out ALS variants using the same fixed subspace, rank, and calibration protocol. Task metrics use pooled benchmark scoring; POPE F1 and MMBench-style ACC are percentages. For KL-only, ΔH-only, and LDR-only, error bars are image-cluster bootstrap SDs over fixed held-out predictions (1,000 resamples). Each single-signal configuration is deterministic once E, G, and the signal are fixed. Leave-one-out and full ALS rows report mean and SD over three matched calibration seeds.

Full ALS attains the highest reported AUC, mAP, and ACC among these variants. KL plus LDR is the strongest two-signal variant, trailing full ALS by 0.0022 AUC and 0.0023 mAP. The capability

Table A15: ALS signal ablations on InternVL3.5-8B. The subspace, rank, and calibration protocol are shared. AUC, mAP, and ACC use pooled benchmark scores; POPE F1 and MMBench-style ACC are percentages. Single-signal rows report fixed-prediction scores ± image-cluster bootstrap SD (1,000 resamples); other edited rows report three-calibration-seed mean ± SD.

ALS variant	AUC ↑	mAP ↑	ACC ↑	POPE F1 ↑	MMBench ACC ↑
Frozen backbone	0.5167	0.2896	0.5692	87.43	81.72
KL only	0.7427±0.0033	0.5329±0.0041	0.6826±0.0032	86.97±0.07	81.14±0.09
ΔH only	0.7318±0.0047	0.5184±0.0048	0.6747±0.0033	86.83±0.09	81.37±0.08
LDR only	0.7469±0.0033	0.5373±0.0035	0.6868±0.0031	87.04±0.04	81.26±0.07
Without KL (ΔH + LDR)	0.7492±0.0034	0.5407±0.0033	0.6897±0.0030	86.93±0.07	81.24±0.05
Without ΔH (KL + LDR)	0.7516±0.0028	0.5438±0.0029	0.6933±0.0027	86.88±0.06	81.17±0.07
Without LDR (KL + ΔH)	0.7477±0.0039	0.5392±0.0043	0.6881±0.0036	86.95±0.07	81.28±0.10
Full ALS	**0.7538**±0.0031	**0.5461**±0.0032	**0.7261**±0.0028	86.91±0.06	81.19±0.08

metrics favor different reduced variants: LDR alone has the highest edited POPE F1, while ΔH alone has the highest edited MMBench-style ACC. Thus the full combination leads on the target tasks without dominating every capability metric. All ALS variants remain within one percentage point of the frozen references on both capability metrics.

M.3 Image-budget-matched LoRA and DoRA

Table A16 expands LoRA and DoRA from 42 training images to the same 1,491-image training union used in CLASP preparation. For CLASP this count includes 832 estimation, 617 diagnosis, and 42 calibration images; it is not a gradient-training count. The comparison matches image access while retaining each method's supervision and update mechanism. The local PEFT configurations are specified in Section H; matching image access does not establish identical backbone configurations.

Table A16: Image-budget-matched PEFT comparison on InternVL3.5-8B. The 1,491-image rows use the same training-image union as CLASP preparation. Task scores are pooled, capability scores are percentages, and edited rows report three-seed mean ± SD.

Method	Train images	AUC ↑	mAP ↑	ACC ↑	POPE F1 ↑	MMBench ACC ↑
Frozen backbone	0	0.5167	0.2896	0.5692	87.43	81.72
LoRA	42	0.6463±0.0046	0.4382±0.0041	0.6208±0.0034	87.11±0.08	81.43±0.09
DoRA	42	0.6584±0.0047	0.4513±0.0038	0.6317±0.0036	87.17±0.07	81.46±0.08
LoRA	1,491	0.7278±0.0084	0.5159±0.0072	0.6757±0.0051	86.68±0.11	80.94±0.13
DoRA	1,491	0.7397±0.0061	0.5279±0.0049	0.6848±0.0046	86.77±0.09	81.03±0.11
CLASP	1,491	**0.7538**±0.0031	**0.5461**±0.0032	**0.7261**±0.0028	86.91±0.06	81.19±0.08

Increasing image access improves both PEFT methods. Under the matched image budget, CLASP exceeds DoRA by 0.0141 AUC, 0.0182 mAP, and 0.0413 ACC, while retaining POPE F1 and MMBench-style ACC 0.14 and 0.16 percentage points higher, respectively. The 42-image PEFT variants retain higher capability scores but reach lower task scores. These comparisons describe the observed performance trade-offs under the stated image budgets.