

Same Scene, Different Story: Evaluating Robot Perception Across Scene Phases in the Wild

Jishnu Jaykumar P^{1,*}, Itay Kadosh^{2,*}, Narendhiran Vijayakumar^{3,*}, Srinanditha Kamath⁴, Sai Haneesh Allu¹, Govind Tyagi Rangappa¹, Animesh Maheshwari¹, Jikai Wang¹, Yu Xiang¹

^{*}Equal contribution ¹UT Dallas ²UF ³IIIT Hyderabad ⁴UT Austin

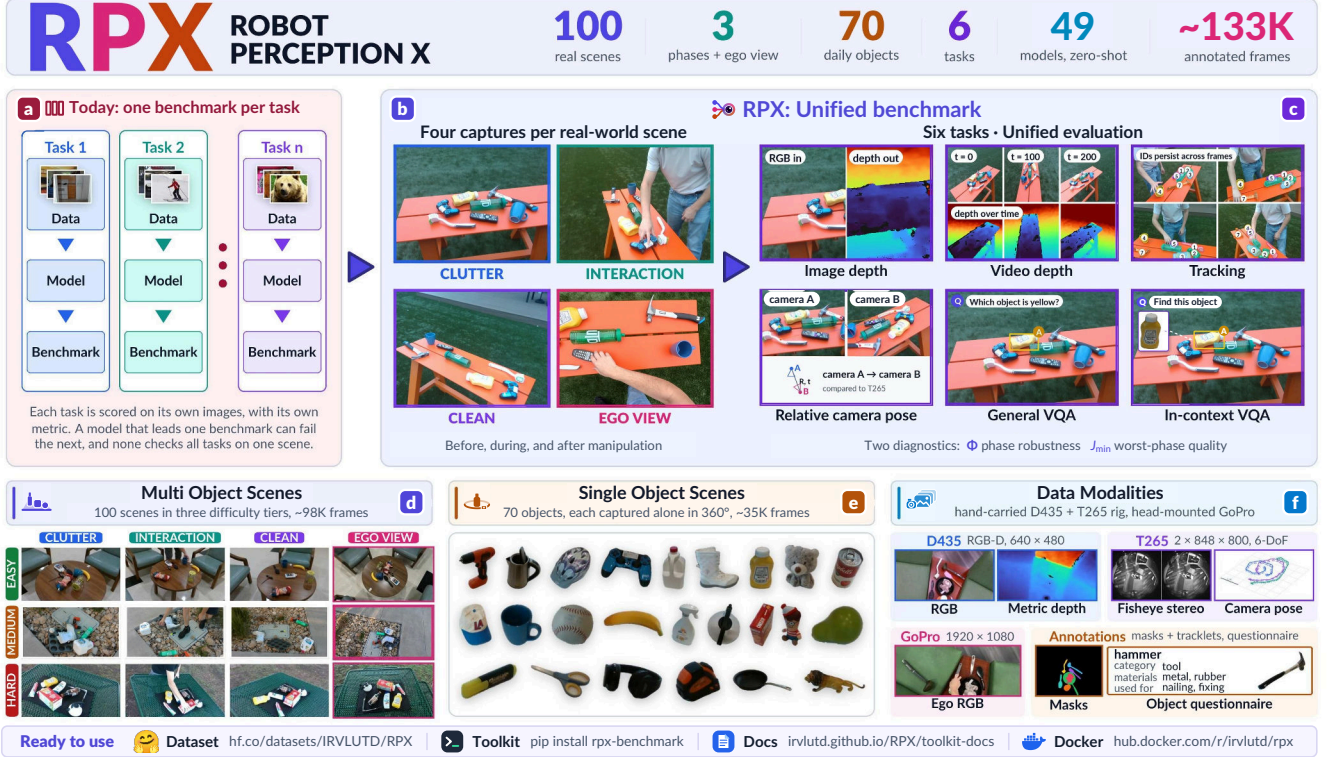


Fig. 1: **RPX** overview. (a) Today each perception task is benchmarked on its own data with its own metric. (b, c) **RPX**: (b) four captures per scene, before (Clutter), during (Interaction), and after (Clean) manipulation plus an egocentric view; (c) six tasks, one evaluation, two diagnostics (Φ phase robustness, $\mathcal{J}_{\min}$ worst-phase quality). (d) Multi-object scenes by difficulty and phase, indoor and outdoor. (e) Each object captured alone in 360° [1]. (f) Sensors and annotation modalities. Bottom: the released dataset, toolkit, and documentation. Project page: <https://irvlutd.github.io/RPX>.

Abstract—Task-specific leaderboards evaluate perception models on different data. Roboticians therefore lack a shared basis for selecting models by their quality and robustness as manipulation changes the scene. We present **RPX**, a unified real-world benchmark: 100 scenes, 70 objects, robot-relevant RGB-D sensors, ~133K frames, three exocentric phases (before, during, and after manipulation), an egocentric view during Interaction, six tasks, 49 models, and 65 zero-shot evaluations. Phase robustness Φ and worst-phase quality $\mathcal{J}_{\min}$ score every model on a common scale. We find: (i) depth and tracking show significant phase effects ($p < 0.001$), while relative pose shows no significant Clutter–Clean effect within difficulty tiers; (ii) model rankings shift across tasks, prompting modes, and input modalities on the same scenes; (iii) no evaluation reaches both $\mathcal{J}_{\min} \geq 0.75$ and $\Phi \geq 0.86$ under the chosen diagnostic thresholds. **RPX** localizes failures to specific phases, scenes, and difficulty tiers.

I. INTRODUCTION

Modern robots increasingly rely on pretrained perception models selected using separate task-specific benchmarks. Modular pipelines run frozen depth estimators, trackers, and segmenters before grasping and navigation [2]. VLA policies build on pretrained vision encoders, whose choice can strongly affect policy performance [3], [4], [5], [6]. Agentic systems call perception models during task execution [7], [8], [9]. Yet task-specific leaderboards rarely quantify how a shared perception stack responds when manipulation changes the same scene.

The scenes a robot works in are not static. A tabletop starts cluttered, a hand rearranges objects, and the scene ends clean. The camera may observe the scene externally or from the manipulator’s viewpoint. These scene transitions are routine in manipulation, but existing task-specific benchmarks

do not provide a shared before–during–after evaluation across perception tasks. The depth model is chosen from NYU Depth V2 [10], the tracker from DAVIS [11], the VLM from an internet QA benchmark, each scored on different data without a shared manipulation protocol (Table I). A roboticist assembling a perception stack lacks a shared basis for comparing models under these scene transitions.

We present **RPX** (Robot Perception X; Fig. 1), a unified real-world benchmark that gives roboticists a statistical basis for selecting pretrained perception models. Each of 100 scenes is captured exocentrically with an Intel D435 (RGB-D) and T265 (6-DoF pose) in three phases: **Clutter** (before manipulation), **Interaction** (during), and **Clean** (after), with an additional egocentric view during Interaction. Six perception tasks share the same scenes. Two diagnostics, phase robustness Φ and worst-phase quality $\mathcal{J}_{\min}$ (Sec. III-B), score every model on a common scale, so a practitioner can compare models across tasks and scene conditions in one place.

We evaluate 49 models zero-shot (65 evaluations across six tasks) and find three results. First, phase sensitivity differs by task: depth and tracking show significant phase effects ($p < 0.001$), while relative pose shows no significant Clutter–Clean effect within difficulty tiers. Second, model rankings change across tasks, prompting modes, and input modalities on the same scenes. Third, none of the 65 evaluations reaches both $\mathcal{J}_{\min} \geq 0.75$ and $\Phi \geq 0.86$ under the chosen diagnostic thresholds (Fig. 4). Sec. V reports the per-task evidence; Sec. VI draws cross-task lessons.

Contributions.

- (1) We present **RPX**, a unified real-world benchmark that lets roboticists evaluate and select pretrained perception models on the scene transitions robots produce: 100 scenes, six tasks, three exocentric phases, and an egocentric view during Interaction.
- (2) We evaluate 49 models (65 task evaluations) zero-shot and show that the right model depends on the task and scene condition: depth and tracking show significant phase effects ($p < 0.001$), relative pose shows no significant within-tier Clutter–Clean effect, and rankings change with prompting mode and input modality.
- (3) We release an open-source evaluation toolkit with documented task protocols and model-adaptor interfaces to support reproducible evaluation and extension to additional models.

II. RELATED WORK

Pretrained models as robot perception. Pretrained vision models serve as encoders in VLA policies [5], [4], [6], as preprocessing stages in modular pipelines [2], as few-shot object recognizers for grasping [18], and as pseudo-label generators for world models [19]. Adapting pretrained features to novel instances improves detection in manipulation scenes [20]. Li et al. [3] demonstrated substantial differences in VLA performance across backbones; on real hardware, adapting perception to the deployment scene [21] directly

TABLE I: Comparison with benchmarks relevant to robot perception. ● supported; ◐ partial; ○ absent. ^{||}Policy-learning dataset. *Static extrinsics only. [‡]Trajectories/trials. [§]Seven BOP-Classic-Core datasets. [†]Tasks, not scenes. [°]Binocular IR reported.

Capability	RPX	Perception				Robot data		
		ScanNet [12]	YCB-V [13]	BOP-Classic [§] [14]	RoboDepth [15]	DROID [16]	RH20T [17]	OK-Robot [2]
Data properties								
Real-world RGB-D	●	●	●	●	○	●	●	●
Metric depth GT	●	●	●	●	○	●	●	○
6-DoF camera pose	●	●	○	○	○	●*	●*	○
Calibrated stereo pairs	●	○	○	○	○	●	●°	○
Instance masks	●	●	●	●	○	○	○	○
Object questionnaire	●	○	○	○	○	○	○	○
Multi-task GT, same scenes	●	●	●	●	○	○	●	○
Protocol								
Three-phase capture	●	○	○	○	○	○	○	○
Manipulation recorded	●	○	○	○	○	●	●	○
Egocentric stream	●	○	○	○	○	●	●	○
Shared objects across phases	●	○	○	○	○	○	○	○
Single + multi-object scenes	●	○	●	●	○	○	○	○
Scene difficulty strata	●	○	○	○	●	○	○	○
# Scenes	100	1.5K	92	var.	n/a	564	147 [†]	10
# Frames	~133K	2.5M	134K	var.	n/a	76K [‡]	110K [‡]	171 [‡]
Focus	deploy.	recon.	pose	pose	depth	imit.	imit.	manip.

improves manipulation. **RPX** evaluates these models together on real scenes captured with consumer RGB-D sensors common in robot perception research, under scene change.

Per-task benchmarks. Depth models are developed on NYU Depth V2 [10], trackers on DAVIS [11], detectors on COCO [22], and odometry on KITTI [23], each with its own camera, scenes, and domain. RoboDepth [15] tests depth robustness by applying synthetic corruptions to existing benchmark images; the perturbations are added after capture, not caused by scene change. SenseShift6D [24] sweeps sensor parameters for 6-DoF pose but evaluates a single task on static scenes. **RPX** differs in two ways: the perturbation is the manipulation itself, recorded on robot sensors, and all six tasks share the same scenes.

Robot datasets. Table I compares **RPX** against datasets a practitioner might use instead. YCB-Video [13] and BOP-Classic [14] provide object-pose and mask GT, but do not use a shared before–during–after manipulation protocol. DROID [16] and RH20T [17] record manipulation at scale but lack per-frame perception GT. HOI4D [25] and HOT3D [26] capture hand–object interaction with dense labels. Grasp-Clutter6D [27] provides multi-sensor depth in heavy clutter, but the scenes are static. SceneReplica [28] benchmarks manipulation by replicating tabletop arrangements, focusing on grasping success rather than perception evaluation. **RPX** provides paired captures of the same scene through manipulation, with ground truth for multiple perception tasks on shared data.

III. BENCHMARK DESIGN

A. Three-Phase Protocol

Each scene is recorded in three successive phases that bracket a manipulation episode:

- **Clutter**: the scene as found, objects in their initial arrangement.

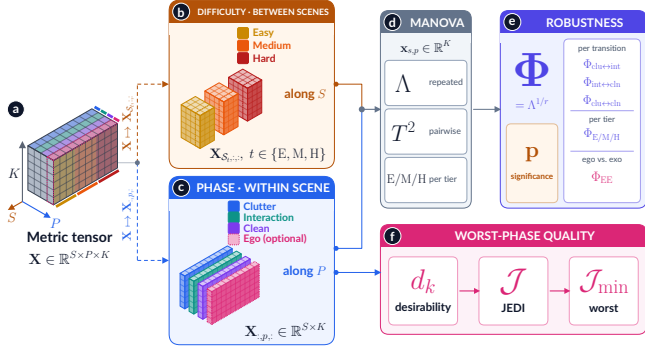


Fig. 2: Evaluation pipeline. (a) A model’s metric scores form a tensor over scenes, phases (ego dashed), and metrics. (b) Slices along S form difficulty tiers; (c) slices along P are compared within each scene and feed the quality score. (d) Repeated-measures MANOVA yields Wilks’ Λ and per-transition tests, also run per tier. (e) Phase robustness Φ : overall, per transition, per tier. (f) Per-metric desirabilities combine into $\mathcal{J}_p$; the worst phase, $\mathcal{J}_{\min}$, ranks models.

- **Interaction**: a person rearranges objects by hand, as in learning-from-demonstration collection [21], [29], [16].
- **Clean**: the scene after reorganization.

An exocentric RGB-D stream covers all three phases; an egocentric stream records Interaction from the manipulator’s viewpoint (Sec. III-C). The paired design keeps the objects, sensors, and environment shared across phases, reducing between-scene confounds while measuring the changes that accompany manipulation.

B. Evaluation Primitives

Averaging over metrics can mask the collapse of one behind the stability of another, and reveals nothing about whether model output changes when the scene does. We instead summarize each model with two unit-free diagnostics: a phase-robustness effect size Φ and a worst-phase quality index $\mathcal{J}_{\min}$.

Each task produces K metrics per (scene, phase) cell, forming a tensor $\mathbf{X} \in \mathbb{R}^{S \times P \times K}$ with $S=100$ scenes and $P=3$ phases ($P=2$ for T4), plus the ego capture for T3 and T5–T6 (Fig. 2). Slicing along P yields the per-phase $S \times K$ matrices compared within each scene; slicing along S yields the difficulty tiers $\mathcal{S}_t$, $t \in \{E, M, H\}$, on which every test below is repeated. For image depth, $K=4$ (AbsRel, RMSE, SILog, δ_1 [30]); for tracking, $K=3$ (HOTA [31], MOTA [32], IDSW); Table II lists K for every task. The metrics within a task are correlated but not redundant, motivating a joint multivariate test.

Phase robustness Φ . To test whether the K -metric vector shifts across phases, we use a one-way repeated-measures MANOVA [33]. Let $\mathbf{x}_{s,p} \in \mathbb{R}^K$ be the z-scored, direction-aligned metric vector for scene s in phase p , and $\tilde{\mathbf{x}}_{s,p} = \mathbf{x}_{s,p} - \bar{\mathbf{x}}_{s,\cdot}$, the same vector centered on the scene mean. The between-phase and within-phase scatter are captured by two

$K \times K$ SSCP matrices:

$$\mathbf{H} = S \sum_{p=1}^P \bar{\mathbf{x}}_{\cdot,p} \bar{\mathbf{x}}_{\cdot,p}^\top, \quad \mathbf{E} = \sum_{s,p} (\tilde{\mathbf{x}}_{s,p} - \bar{\mathbf{x}}_{\cdot,p})(\tilde{\mathbf{x}}_{s,p} - \bar{\mathbf{x}}_{\cdot,p})^\top \quad (1)$$

where $\bar{\mathbf{x}}_{\cdot,p}$ is the phase mean of the centered data. $\mathbf{H}$ captures the signal (between-phase variance); $\mathbf{E}$ captures the noise (within-phase variance). The effect size follows from Wilks’ Λ :

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{H} + \mathbf{E}|}, \quad \eta_p^2 = 1 - \Lambda^{1/r}, \quad \Phi = 1 - \eta_p^2, \quad (2)$$

with $r = \min(K, P-1)$. Φ near 1 indicates a small estimated phase effect; $\Phi \leq 0.86$ corresponds to a large effect under the $\eta^2 \geq 0.14$ convention [34]. Significance is assessed via Rao’s F -approximation [35], with $P-1$ hypothesis and $(S-1)(P-1)$ error degrees of freedom, assuming multivariate sphericity. This omnibus Φ tests all P phases jointly. To pinpoint which transition drives the effect (e.g., Clutter $\leftrightarrow$ Interaction vs. Interaction $\leftrightarrow$ Clean), we run the same procedure on each pair of phases via paired Hotelling T^2 tests [36] on the difference vectors $\mathbf{x}_{s,p_a} - \mathbf{x}_{s,p_b}$, Holm-corrected [37] over the $\binom{P}{2}$ pairs. We also repeat the omnibus test within each difficulty tier $\mathcal{S}_t$, yielding per-tier Φ values (E, M, H in the tables). T4’s reported summary Φ and quality scores average the three difficulty-tier results; its p entries refer to the individual tier tests.

Worst-phase quality $\mathcal{J}_{\min}$. Each metric k is mapped to a desirability $d_k = \text{clip}((v_k - v_k^{\text{worst}})/(v_k^{\text{best}} - v_k^{\text{worst}}), 0, 1)$, where $(v_k^{\text{worst}}, v_k^{\text{best}})$ is a fixed window per metric (e.g., AbsRel: 0.5 $\rightarrow$ 0; δ_1 : 0 $\rightarrow$ 1; HOTA: 0 $\rightarrow$ 1; full list in the toolkit). We compute the geometric mean of the K desirabilities in each scene–phase cell, then average over scenes:

$$\mathcal{J}_{s,p} = \left(\prod_{k=1}^K d_{k,s,p} \right)^{1/K}, \quad (3)$$

$$\mathcal{J}_p = \frac{1}{S} \sum_s \mathcal{J}_{s,p}, \quad \mathcal{J}_{\min} = \min_p \mathcal{J}_p.$$

The geometric mean ensures that a single collapsed metric drives the corresponding cell’s quality toward zero. $\mathcal{J}_{\min}$, the quality of the worst phase, is our primary ranking criterion [38]. Overall $\mathcal{J}$ averages the exocentric phase scores; Ego is reported separately and excluded from $\mathcal{J}_{\min}$. Metric windows and K are task-specific, so equal quality scores across tasks do not imply equal downstream utility.

Why two numbers? Φ and $\mathcal{J}_{\min}$ are complementary by construction. A model can score high Φ (phases look alike) despite low $\mathcal{J}_{\min}$ (poor quality across phases). The diagnostics favor both high $\mathcal{J}_{\min}$ (high quality in the worst phase) and high Φ (small estimated phase effects).

C. Capture Protocol and Dataset

The protocol compares successive scene states around the same manipulation episode. Two people record each scene using a single rig (Fig. 3). One carries the sensors:



Fig. 3: Capture setup. (a) D435 + T265 rig on a laptop mount, carried by hand. (b) Sensor detail with the T265→D435 extrinsic and world→T265 pose frames. (c) Three phases of one scene; the GoPro is visible during Interaction; faces blurred for anonymity.

a RealSense D435 (RGB-D, 640×480) and a T265 (6-DoF visual-inertial odometry) on a shared laptop mount (Fig. 3a); the calibrated T265→D435 extrinsic (Fig. 3b) gives every D435 frame a metric camera pose. The second person rearranges objects during Interaction and wears a head-mounted GoPro (1920×1080 , RGB only, $6.75 \times$ the exocentric resolution) that records the egocentric stream. The rig is hand-held to reflect the moving viewpoint of a working robot rather than a static tripod capture. All three phases are recorded with the same sensors and object set; viewpoint, framing, and lighting vary naturally, so the phase factor tested in Sec. III-B measures their combined change across the manipulation episode (Fig. 3c).

Difficulty tiers. Scenes are stratified into Easy (33), Medium (33), and Hard (34) by a weighted composite of 27 perceptual descriptors—covering annotation effort, scene complexity, occlusion, depth quality, photometric conflict, temporal stability, camera motion, and fisheye/stereo characteristics—defined with full equations in Appendix B.3.

Contents. The dataset comprises $\sim 133,000$ frames across 100 indoor and outdoor scenes (Fig. 1d) with 70 everyday objects, organized into two streams. The **multi-object stream** ($\sim 98K$ frames) contains 100×3 phase captures at ~ 250 D435 RGB-D frames per phase, plus the egocentric GoPro stream during Interaction. Each exocentric frame has metric depth, a 6-DoF camera pose from T265 visual-inertial odometry (validated against fiducial markers in Sec. V), instance-level masks (SAM 2 [39] initialized with GroundingDINO, then manually corrected), and per-object attributes used by VQA; the RGB-only ego stream has masks, tracks, and VQA labels. The **single-object stream** ($\sim 35K$ frames) captures each of the 70 objects in isolation with RGB-D, pose, and calibrated stereo pairs, providing the reference images for in-context grounding (T6).

IV. TASKS & MODELS

All six tasks are evaluated on the same 100 scenes (Tables III–V, VII; T4 results in Sec. V). Models evaluated in multiple tasks count once toward the total of 49 models;

task-specific configurations yield 65 evaluations across six tasks. Φ and $\mathcal{J}$ are computed from the standard metrics of each task’s dominant benchmark.

T1–T2: Image & Video Depth. Both tasks predict metric depth from the D435 RGB stream and are scored against the D435 depth sensor as ground truth. T1 evaluates individual frames with $K=4$ metrics (AbsRel, RMSE, SILog, δ_1). T2 evaluates the full phase clip (~ 250 frames) and adds two temporal consistency metrics (TGM, TGSE), giving $K=6$.

T3: Multi-Object Tracking. Eight semi-supervised trackers are initialized with ground-truth first-frame masks; SAM 3.1 receives bounding boxes; text-prompted GSAM2 ($\ddagger$) receives only an open-vocabulary label. The Interaction phase, where objects disappear behind a hand and reappear in new positions, is the primary stress test.

T4: Relative Camera Pose Estimation. Ten models estimate the relative pose between $\sim 20K$ fixed image pairs. Each pair is scored via $AUC@7$, the area under the pose-error curve over the larger of rotation error and translation-direction error. Because objects move during Interaction and pairwise pose assumes a static scene, T4 is scored on Clutter and Clean only ($P=2$).

T5–T6: Regular & In-Context VQA. Six VLM families at two sizes each (12 models) respond with a bounding box through their native grounding interface. A missing or unparseable box counts as a complete miss (Acc 0, Center 0, GIoU at its floor of -1). The $K=3$ metric vector is Acc@0.5 (detection), GIoU [50] (spatial quality), and Center-in-GT (coarse localization). T5 asks a model to localize an object from a text description built from its colour, material, or function ($\sim 12,000$ scored questions per model across 5 attribute types; spatial-reference questions are released but not scored). T6 asks the same 5 attribute question types conditioned on a single reference image of the target object from the single-object stream rather than its text description ($\sim 6,000$ scored questions per model).

Ego-exo evaluation. T3 and T5–T6 are additionally evaluated on the egocentric GoPro stream during Interaction, with ground-truth masks and tracklets for both viewpoints, enabling a direct comparison between the ego and exo cameras at $6.75 \times$ the spatial resolution.

Extensibility. A PyPI-distributed toolkit (`rpv-benchmark`) provides task runners, metric evaluation, and adapter interfaces for additional models. The [toolkit documentation](#) describes installation, reference evaluation workflows, and custom-model integration.

V. EXPERIMENTS

All 49 models are evaluated zero-shot (no training or fine-tuning on RPX) on 100 scenes across three phases (two for T4), with the protocols of Sec. IV. Table II summarizes the top and bottom model per task; Fig. 4 plots all 65 evaluations in the Φ – $\mathcal{J}_{\min}$ plane. Scene-level analyses cover 100 scenes; T4 significance is reported within the 33/33/34 tiers.

TABLE II: Phase-robustness summary. Top and bottom model per task by $\mathcal{J}_{\min}$; $\Phi \leq 0.86$: large effect. T4 uses 2 phases (Clutter vs. Clean) and tier-averaged summary scores; all other tasks use 3 phases. Hardest transition: lowest mean transition Φ across models. VQA p entries apply to most models; four of 24 model–task pairs do not reach $p < .001$. [‡]Text-prompted (no GT masks). [§]Tied at displayed precision with DeepSeek-VL2 1.0B.

Task	K	M	Top model	$\mathcal{J}_{\min} \uparrow$	Bottom model	$\mathcal{J}_{\min}$	Φ (range)	Hardest transition	p
T1: Image Depth	4	10	Metric3D V2 [40]	0.71	Lotus-2 [41]	0.35	0.53–0.79	Clu→Int	<.001
T2: Video Depth	6	10	VGGT [42]	0.70	RollingDepth [43]	0.20	0.45–0.77	Clu→Int	<.001
T3: Tracking	3	11	SAM 2 Long [44]	0.96	SAM 3.1 [‡] [45]	0.72	0.50–0.75	Int (both)	<.001
T4: Rel. Pose	3	10	DA3 [46]	0.38	MonST3R [47]	0.08	0.86–0.97	Clu→Cln	>.05 (tier)
T5: Regular VQA	3	12	Qwen3-VL 8B [48]	0.57	Florence-2 B [§] [49]	0.12	0.62–0.98	Clu→Cln	<.001 (most)
T6: In-Context VQA	3	12	Qwen3-VL 8B [48]	0.23	Florence-2 L [49]	0.01	0.65–0.96	Clu→Cln	<.001 (most)

TABLE III: T1: Image depth ($K=4$: AbsRel, RMSE, SILog, δ_1). Out: D depth, P pose, N normals, GS Gaussian splats. Columns follow Fig. 2; bold Φ : best non-[‡] model. [‡]Relative depth, affine-aligned per phase, which can inflate Φ .

Model specification				Desirability $\mathcal{J} \uparrow$				Transition $\Phi \uparrow$			Scene-tier $\Phi \uparrow$			Summary	
Model	Venue	Out	Params	Overall	Clu	Int	Cln	Clu↔Int	Int↔Cln	Clu↔Cln	E	M	H	$\mathcal{J}_{\min} \uparrow$	$\Phi \uparrow$
ZipDepth [‡] [51]	ECCV’26	D	6.14 M	0.56	0.58	0.51	0.60	0.42	0.37	0.54	0.63	0.59	0.46	0.51	0.60
HyDen [‡] [52]	ICLR’26	D	0.35 B	0.44	0.42	0.45	0.45	0.73	0.50	0.64	0.76	0.78	0.70	0.42	0.79
FE2E [‡] [53]	CVPR’26	D	12.47 B	0.68	0.70	0.65	0.68	0.39	0.41	0.49	0.62	0.55	0.49	0.65	0.59
DA3 Metric-L [46]	ICLR’26	D,P,N,GS	0.35 B	0.72	0.73	0.70	0.73	0.23	0.28	0.43	0.49	0.54	0.46	0.70	0.53
UniDepth V2 [54]	TPAMI’26	D	0.35 B	0.71	0.70	0.69	0.72	0.31	0.30	0.53	0.56	0.55	0.48	0.69	0.56
MoGe-2 ViT-L [‡] [55]	NeurIPS’25	D,P	0.33 B	0.52	0.48	0.54	0.53	0.43	0.51	0.51	0.66	0.66	0.61	0.48	0.66
Lotus-2 [41]	arXiv’25	D	17.58 B	0.36	0.35	0.38	0.35	0.34	0.66	0.44	0.64	0.61	0.55	0.35	0.64
Depth Pro [56]	ICLR’25	D	0.93 B	0.63	0.62	0.64	0.63	0.30	0.34	0.56	0.58	0.54	0.54	0.62	0.57
Metric3D V2 [40]	TPAMI’24	D,N	1.37 B	0.72	0.72	0.72	0.71	0.44	0.50	0.56	0.63	0.58	0.57	0.71	0.68
DA-V2 Large [57]	NeurIPS’24	D	0.34 B	0.60	0.60	0.55	0.66	0.27	0.33	0.43	0.59	0.49	0.51	0.55	0.54

A. T1: Image Depth

All ten models exhibit a significant phase effect ($\Phi \leq 0.79$, $p < 0.001$; Table III). Metric3D V2 achieves the highest worst-phase quality ($\mathcal{J}_{\min} = 0.71$) and is among the most phase-stable models ($\Phi = 0.68$). DA3 Metric-L reaches comparable accuracy ($\mathcal{J}_{\min} = 0.70$) but is the most fragile entry ($\Phi = 0.53$), with most of the degradation concentrated at the Clutter→Interaction transition. Clutter–Interaction is the lowest- Φ transition for seven of ten image-depth models. Among relative-depth models, only FE2E ($\mathcal{J}_{\min} = 0.65$) approaches the metric group; significant phase effects remain

despite affine alignment.

B. T2: Video Depth

Video depth is the one task where quality and robustness align: VGGT leads on both axes ($\mathcal{J}_{\min} = 0.70$, $\Phi = 0.77$; Table IV). DA3 matches VGGT’s overall quality but not its phase stability; most of the gap is attributable to Interaction ($\Phi_{\text{clu} \leftrightarrow \text{int}} = 0.23$). At the bottom of the ranking, Video-DA, DepthCrafter, and RollingDepth combine moderate phase stability with $\mathcal{J}_{\min} \leq 0.30$, offering neither accuracy nor robustness. A result common to all ten models: near-range AbsRel at 0.3–1.0 m, the range in which a robot grasps, increases 2.0–4.4 $\times$ during Interaction relative to Clutter.

C. T3: Tracking

Every tracker exhibits a significant phase effect ($\Phi \leq 0.75$, $p < 10^{-10}$; Table V), and the degradation is sharply localized: transitions involving Interaction yield $\Phi = 0.21$ –0.55, while Clutter↔Clean has weaker effects ($\Phi = 0.73$ –0.98). Because quality is uniformly high (even the weakest mask-initialized model reaches $\mathcal{J}_{\min} = 0.87$), it is Φ rather than $\mathcal{J}_{\min}$ that differentiates these models. SAM 2 Long leads in quality ($\mathcal{J}_{\min} = 0.96$); GSAM2, a text-prompted tracker, reaches the highest Φ (0.75) but trades quality for phase stability ($\mathcal{J}_{\min} = 0.79$), and SAM 3.1 with text prompts has the lowest quality ($\mathcal{J}_{\min} = 0.72$). SAM 3.1 with box prompts has the lowest Φ of any tracker (0.50).

Ego–exo gap. Switching to the head-mounted camera has little effect on tracking quality (SAM 2 Long and DAM4SAM maintain $\mathcal{J} \geq 0.95$), yet 7 of 11 trackers cross the $\Phi_{\text{EE}} \leq 0.86$ threshold. The camera change can reshape the metric profile while maintaining high quality; viewpoint, resolution, and sensor effects are confounded in this comparison.

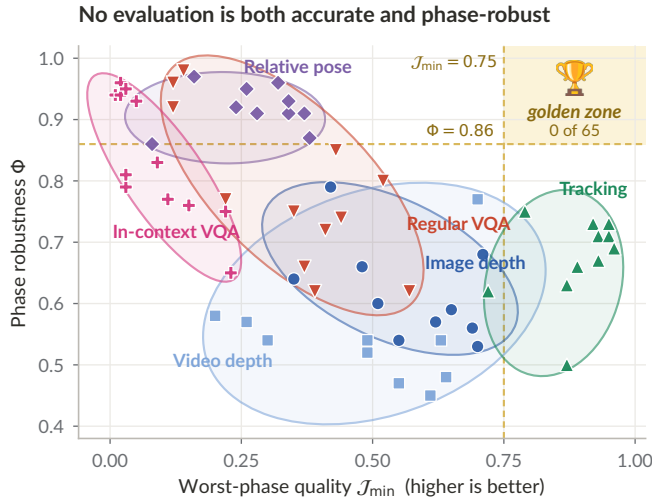


Fig. 4: Φ vs. $\mathcal{J}_{\min}$ for all 65 evaluations; each ellipse is the smallest enclosing a task’s points. Pose (♦) spans $\Phi = 0.86$ –0.97; tracking (▲) achieves high quality but degrades; depth and VQA span a broad range. Dashed: $\Phi = 0.86$ and an illustrative $\mathcal{J}_{\min} = 0.75$ bar; no evaluation reaches the golden zone.

TABLE IV: T2: Video depth ($K=6$: AbsRel, RMSE, SILog, δ_1 , TGM, TGSE). Same diagnostic format as Table III. All models evaluated on the full video of every capture (250 frames).

Model specification				Desirability $\mathcal{J} \uparrow$				Transition $\Phi \uparrow$			Scene-tier $\Phi \uparrow$			Summary	
Model	Venue	Out	Params	Overall	Clu	Int	Cln	Clu $\leftrightarrow$ Int	Int $\leftrightarrow$ Cln	Clu $\leftrightarrow$ Cln	E	M	H	$\mathcal{J}_{\min} \uparrow$	$\Phi \uparrow$
DA3 (video) [46]	ICLR'26	D,P,N,GS	0.41 B	0.72	0.78	0.64	0.74	0.23	0.26	0.46	0.46	0.42	0.37	0.64	0.48
Video-DA [58]	CVPR'25	D	0.38 B	0.33	0.35	0.35	0.30	0.21	0.55	0.32	0.57	0.49	0.44	0.30	0.54
ViGeo [59]	arXiv'26	D	1.25 B	0.64	0.63	0.63	0.67	0.23	0.42	0.35	0.53	0.54	0.48	0.63	0.54
VGGT [42]	CVPR'25	D,P,N	1.26 B	0.72	0.70	0.73	0.72	0.63	0.58	0.64	0.70	0.65	0.74	0.70	0.77
MonST3R [47]	ICLR'25	D,P	0.57 B	0.55	0.58	0.49	0.57	0.27	0.37	0.39	0.49	0.55	0.38	0.49	0.52
GemDepth [60]	ICML'26	D	0.71 B	0.69	0.76	0.61	0.71	0.17	0.31	0.34	0.46	0.39	0.37	0.61	0.45
DepthCrafter [61]	CVPR'25	D	2.25 B	0.30	0.31	0.31	0.26	0.33	0.54	0.30	0.60	0.49	0.47	0.26	0.57
DVD v1.1 [62]	arXiv'26	D	2.24 B	0.62	0.69	0.55	0.62	0.19	0.31	0.39	0.42	0.45	0.35	0.55	0.47
ChronoDepth [63]	CVPR'25	D	2.25 B	0.54	0.59	0.49	0.55	0.28	0.49	0.33	0.54	0.54	0.42	0.49	0.54
RollingDepth [43]	CVPR'25	D	1.29 B	0.22	0.23	0.24	0.20	0.27	0.59	0.34	0.60	0.54	0.47	0.20	0.58

TABLE V: T3: Multi-object tracking ($K=3$: HOTA, MOTA, IDSW; IDF1 excluded due to $|r| > 0.99$ collinearity with MOTA). GT mask/box initialization; ego = GoPro, Interaction only. [‡]Text-prompted (no GT masks). Φ_{EE} : ego vs. exo-Interaction (camera, resolution and view confounded).

Model specification				Desirability $\mathcal{J} \uparrow$					Transition $\Phi \uparrow$				Scene-tier $\Phi \uparrow$				Summary	
Model	Venue	Prompt	Params	Overall	Clu	Int	Cln	Ego	Clu↔Int	Int↔Cln	Clu↔Cln	Φ_{EE}	E	M	H	$\mathcal{J}_{\min} \uparrow$	$\Phi \uparrow$	
SAM 2 Long [44]	ICCV'25	mask	224 M	0.98	0.98	0.96	0.99	0.96	0.42	0.40	0.94	0.78	0.71	0.66	0.60	0.96	0.69	
DAM4SAM [64]	CVPR'25	mask	224 M	0.97	0.98	0.95	0.99	0.95	0.46	0.45	0.93	0.86	0.73	0.68	0.62	0.95	0.71	
EdgeTAM [65]	CVPR'25	mask	14 M	0.96	0.97	0.93	0.98	0.92	0.43	0.45	0.94	0.96	0.67	0.74	0.65	0.93	0.71	
SAM 2++ [66]	arXiv'25	mask	101 M	0.96	0.98	0.92	0.98	0.86	0.43	0.48	0.98	0.66	0.70	0.76	0.66	0.92	0.73	
SAM 2 [39]	ICLR'25	mask	224 M	0.97	0.98	0.95	0.99	0.93	0.46	0.48	0.96	0.91	0.74	0.79	0.61	0.95	0.73	
Cutie [67]	CVPR'24	mask	35 M	0.96	0.97	0.93	0.98	0.94	0.39	0.36	0.91	0.88	0.68	0.60	0.66	0.93	0.67	
SAM 3.1 [45]	2026	bbox	873 M	0.94	0.97	0.87	0.97	0.90	0.21	0.21	0.93	0.71	0.53	0.49	0.41	0.87	0.50	
MiTS [68]	ICCV'23	mask	33 M	0.93	0.95	0.87	0.97	0.86	0.44	0.31	0.73	0.80	0.62	0.67	0.54	0.87	0.63	
XMem [69]	ECCV'22	mask	62 M	0.94	0.96	0.89	0.98	0.91	0.42	0.34	0.88	0.92	0.65	0.62	0.65	0.89	0.66	
SAM 3.1 [‡] [45]	2026	text	873 M	0.79	0.83	0.72	0.82	0.74	0.29	0.35	0.96	0.55	0.66	0.58	0.52	0.72	0.62	
GSAM2 [‡] [70], [71]	2024	text	397 M	0.85	0.87	0.79	0.90	0.86	0.55	0.50	0.93	0.75	0.84	0.72	0.67	0.79	0.75	

TABLE VI: T4: Relative camera pose ($K=3$: AUC@5°, @10°, @20°). Exocentric only (no GoPro pose ground truth); Clutter vs. Clean.

Specification		$\mathcal{J} \uparrow$			Trans.	Tier $\Phi \uparrow$			Summary	
Model	Params	All	Clu	Cln	Clu $\leftrightarrow$ Cln	E	M	H	$\mathcal{J}_{\min} \uparrow$	$\Phi \uparrow$
DA3 [46]	1.36 B	0.42	0.38	0.45	0.87	0.87	0.89	0.86	0.38	0.87
VGGT [42]	1.26 B	0.37	0.34	0.41	0.91	0.96	0.94	0.85	0.34	0.91
Reloc3R [72]	0.43 B	0.40	0.37	0.42	0.91	0.92	0.89	0.91	0.37	0.91
Pi3X [73]	0.99 B	0.38	0.34	0.41	0.93	0.93	0.93	0.92	0.34	0.93
MonST3R [47]	0.27 B	0.09	0.08	0.11	0.86	0.90	0.81	0.86	0.08	0.86
MUS3R [74]	0.42 B	0.27	0.24	0.29	0.92	0.91	0.93	0.92	0.24	0.92
Fast3R [75]	0.65 B	0.28	0.26	0.30	0.95	0.95	0.95	0.95	0.26	0.95
CUT3R [76]	0.46 B	0.31	0.28	0.34	0.91	0.95	0.86	0.92	0.28	0.91
MAST3R [77]	0.69 B	0.34	0.32	0.37	0.96	0.97	0.97	0.93	0.32	0.96
DUST3R [78]	0.57 B	0.18	0.16	0.19	0.97	0.95	0.99	0.96	0.16	0.97

D. T4: Relative Camera Pose

An independent stationary-board (ARTag) audit across 70 single-object scenes yields median end-to-end inconsistency of 1.53° / 4.68 cm, including detection and synchronization noise. The result contrasts sharply with every other task: no model reaches statistical significance in any difficulty tier ($p > 0.05$; $\Phi = 0.86$ –0.97; Table VI). DA3 leads in quality ($\mathcal{J}_{\min} = 0.38$); DUST3R has the smallest estimated phase effect ($\Phi = 0.97$) but is among the weakest ($\mathcal{J}_{\min} = 0.16$). Interaction is excluded, so this does not establish invariance during manipulation.

E. T5–T6: Visual Question Answering

Qwen3-VL 8B achieves the best worst-phase quality in both VQA modes: $\mathcal{J}_{\min} = 0.57$ for regular grounding and 0.23 for in-context grounding (Table VII). The switch from text to a reference image (a ~60% drop for Qwen3-VL 8B) reshuffles the ranking: Qwen3-VL 2B rises from fourth to second, and Cosmos Reason2 2B from ninth to fifth. A cautionary pattern

emerges in the Φ column: the most phase-stable models (Florence 2, $\Phi = 0.94$ –0.98) are those closest to the quality floor in every phase. Three of the four model–mode pairs that do not reach $\alpha = 0.001$ are Florence 2 variants; their high Φ must be interpreted alongside uniformly low quality. Among models with meaningful quality, the best are also the most fragile (Qwen3-VL 8B, $\Phi = 0.62$ –0.65), reinforcing that $\mathcal{J}_{\min}$ should be inspected before Φ .

Ego–exo gap. The egocentric view consistently improves grounding: all 12 VLMs achieve higher quality on regular questions when evaluated on the GoPro stream, with the largest gains in models of moderate exocentric quality (Cosmos Reason2 8B: $\mathcal{J} = 0.52 \rightarrow 0.67$, $\Phi_{EE} = 0.23$; DeepSeek-VL2: 0.41 $\rightarrow$ 0.53). In-context grounding has mixed effects: at displayed precision, eight models improve, one is unchanged, and three decline. Cosmos Reason2 8B triples its quality (0.11 $\rightarrow$ 0.33).

VI. DISCUSSION

Phase sensitivity is task-selective. Image depth, video depth, and tracking all exhibit significant phase effects ($\Phi \leq 0.79$, ≤ 0.77 , and ≤ 0.75 respectively), with the Clutter $\rightarrow$ Interaction transition dominant in 15 of 20 depth models. VQA spans a wider range ($\Phi = 0.62$ –0.98), but the high end is occupied by models near the quality floor. Relative pose stands apart: $\Phi \geq 0.86$ with no difficulty-tier test reaching significance for Clutter–Clean; Interaction is not evaluated.

Rankings depend on the task and protocol. VGGT leads video depth but ties for third in pose at displayed precision; Qwen3-VL 8B leads both VQA modes, yet the

TABLE VII: T5–T6: VQA localization ($K=3$: Acc@0.5, GlOUI, Center-in-GT). Regular: 5 attribute question types, $\sim 12,000$ scored questions per model; in-context: the same 5 types conditioned on a reference image instead of text, $\sim 6,000$ questions. Counts include ego questions. Each block has its own per-scene MANOVA. † MoE active parameters. Φ_{EE} as in Table V.

Model specification		Desirability $\mathcal{J} \uparrow$					Transition $\Phi \uparrow$				Scene-tier $\Phi \uparrow$			Summary	
Model	Params	Overall	Clu	Int	Cln	Ego	Clu $\leftrightarrow$ Int	Int $\leftrightarrow$ Cln	Clu $\leftrightarrow$ Cln	Φ_{EE}	E	M	H	$\mathcal{J}_{\min} \uparrow$	$\Phi \uparrow$
T5: Regular grounding															
Cosmos Reason2 [79]	8 B	0.54	0.55	0.52	0.54	0.67	0.75	0.63	0.58	0.23	0.70	0.80	0.80	0.52	0.80
Cosmos Reason2 [79]	2 B	0.25	0.28	0.26	0.22	0.38	0.64	0.86	0.45	0.48	0.76	0.72	0.77	0.22	0.77
Qwen3-VL [48]	8 B	0.61	0.63	0.57	0.61	0.69	0.31	0.80	0.35	0.48	0.55	0.56	0.67	0.57	0.62
Qwen3-VL [48]	2 B	0.46	0.43	0.49	0.45	0.62	0.85	0.71	0.62	0.50	0.74	0.85	0.84	0.43	0.85
InternVL 3.5 [80]	14 B	0.48	0.52	0.44	0.49	0.55	0.59	0.74	0.45	0.53	0.70	0.70	0.76	0.44	0.74
InternVL 3.5 [80]	1 B	0.42	0.45	0.39	0.43	0.41	0.43	0.63	0.32	0.71	0.60	0.60	0.64	0.39	0.62
PaliGemma 2 [81]	10 B	0.38	0.43	0.35	0.37	0.42	0.49	0.87	0.50	0.60	0.69	0.77	0.76	0.35	0.75
PaliGemma 2 [81]	3 B	0.39	0.42	0.37	0.39	0.44	0.50	0.66	0.33	0.67	0.66	0.64	0.66	0.37	0.66
DeepSeek-VL2 [82]	4.5 B †	0.48	0.54	0.41	0.48	0.53	0.47	0.64	0.56	0.26	0.72	0.66	0.70	0.41	0.72
DeepSeek-VL2 [82]	1.0 B †	0.15	0.19	0.12	0.14	0.19	0.77	0.98	0.82	0.44	0.90	0.94	0.87	0.12	0.92
Florence 2 [49]	0.77 B	0.15	0.16	0.15	0.14	0.16	0.99	0.95	0.95	0.65	0.96	0.94	0.98	0.14	0.98
Florence 2 [49]	0.23 B	0.13	0.14	0.12	0.12	0.16	0.91	0.99	0.90	0.69	0.98	0.92	0.92	0.12	0.96
T6: In-context grounding															
Cosmos Reason2 [79]	8 B	0.13	0.15	0.11	0.14	0.33	0.50	0.96	0.47	0.29	0.79	0.73	0.73	0.11	0.77
Cosmos Reason2 [79]	2 B	0.10	0.12	0.09	0.09	0.21	0.66	0.97	0.51	0.58	0.72	0.86	0.83	0.09	0.83
Qwen3-VL [48]	8 B	0.27	0.31	0.23	0.28	0.34	0.40	0.74	0.35	0.63	0.57	0.70	0.64	0.23	0.65
Qwen3-VL [48]	2 B	0.23	0.25	0.22	0.22	0.35	0.67	0.76	0.37	0.67	0.73	0.74	0.73	0.22	0.75
InternVL 3.5 [80]	14 B	0.19	0.24	0.15	0.17	0.12	0.58	0.86	0.46	0.67	0.68	0.75	0.79	0.15	0.76
InternVL 3.5 [80]	1 B	0.05	0.07	0.03	0.04	0.02	0.64	0.82	0.52	0.61	0.76	0.83	0.72	0.03	0.79
PaliGemma 2 [81]	10 B	0.03	0.04	0.03	0.03	0.09	0.95	0.93	0.85	0.49	0.95	0.86	0.96	0.03	0.95
PaliGemma 2 [81]	3 B	0.04	0.04	0.03	0.05	0.07	0.53	0.83	0.73	0.73	0.76	0.79	0.84	0.03	0.81
DeepSeek-VL2 [82]	4.5 B †	0.07	0.10	0.05	0.05	0.10	0.83	1.00	0.83	0.75	0.89	0.96	0.88	0.05	0.93
DeepSeek-VL2 [82]	1.0 B †	0.02	0.03	0.02	0.02	0.01	0.84	0.99	0.84	0.94	0.92	0.95	0.87	0.02	0.94
Florence 2 [49]	0.77 B	0.01	0.02	0.01	0.01	0.02	0.81	0.95	0.91	0.88	0.90	0.91	0.90	0.01	0.94
Florence 2 [49]	0.23 B	0.02	0.02	0.02	0.02	0.02	0.87	0.98	0.92	0.72	0.93	0.89	0.93	0.02	0.96

order behind it reshuffles when the prompt switches from text to a reference image. These results support task- and protocol-specific model selection; pipeline-level performance requires separate evaluation.

Robustness and quality are complementary. A high Φ can accompany uniformly poor predictions (Florence 2: $\Phi \geq 0.94$, $\mathcal{J}_{\min} \leq 0.14$); VGGT offers higher quality but remains phase-sensitive ($\Phi = 0.77$, $\mathcal{J}_{\min} = 0.70$).

Prompt modality is part of the model. SAM 3.1 with boxes ($\mathcal{J}_{\min} = 0.87$, $\Phi = 0.50$) is the most phase-sensitive tracker, while the text-prompted GSAM2 ($\mathcal{J}_{\min} = 0.79$, $\Phi = 0.75$) combines lower quality with higher phase stability; changing the grounding context also changes VQA rankings. Reporting the prompt protocol alongside model identity is necessary for interpreting these scores.

Reference-based grounding remains challenging. Seven of the twelve VLMs return a syntactically valid box on $\geq 90\%$ of in-context questions, yet the best Acc@0.5 is 0.215; regular grounding reaches 0.593. Valid box output does not ensure correct localization. The gap suggests difficulty matching reference images to target instances, although this evaluation does not isolate the underlying cause. This challenge is relevant to teach-and-repeat manipulation, where a robot is shown an object once and must re-identify it under changed scene conditions.

Limitations. The rig is hand-carried rather than robot-mounted, and phase comparisons include natural viewpoint, framing, and lighting changes (Sec. III-C). D435 depth ground truth may itself be affected during Interaction; the paired design supports between-model comparison but not absolute accuracy claims. Relative pose uses two phases only. **RPX** measures zero-shot perception, not end-to-end task success.

VII. CONCLUSION

We presented **RPX**, a unified real-world benchmark that evaluates pretrained perception models on the scene transitions that manipulation produces. The benchmark captures 100 scenes with consumer RGB-D sensors in three exocentric phases (Clutter, Interaction, Clean), plus an egocentric view during Interaction, with ground truth shared across six perception tasks. Two diagnostics, phase robustness Φ and worst-phase quality $\mathcal{J}_{\min}$, provide a common statistical framework across tasks and scene conditions.

Evaluating 49 models zero-shot reveals three results. First, phase sensitivity differs by task: depth and tracking show significant phase effects ($p < 0.001$), while relative pose shows no significant within-tier Clutter–Clean effect. Second, model rankings change across tasks, VQA modes, and prompting strategies on the same scenes, motivating task- and protocol-specific model selection. Third, the diagnostic reveals phase-specific failures, including $2.0\text{--}4.4\times$ near-range AbsRel during Interaction; none of the 65 evaluations reaches both $\mathcal{J}_{\min} \geq 0.75$ and $\Phi \geq 0.86$ under the chosen thresholds.

Scores resolved by phase, scene, and difficulty tier identify conditions associated with poor performance and provide a basis for targeted model improvement. Future work will connect $\mathcal{J}_{\min}$ to downstream manipulation success rates and measure whether fine-tuning on the failure phase closes the quality gap.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation (NSF) under Grant No. 2520553, and the NVIDIA Academic Grant Program Award.

REFERENCES

- [1] J. J. P. et al., “FewSOL: A Dataset for Few-Shot Object Learning in Robotic Environments,” in *ICRA*, 2023.
- [2] P. Liu et al., “OK-Robot: What really matters in integrating open-knowledge models for robotics,” in *RSS*, 2024.
- [3] X. Li et al., “What matters in building vision-language-action models for generalist robots,” *Nature Machine Intelligence*, vol. 8, no. 2, pp. 158–172, 2026.
- [4] K. Black et al., “ π_0 : A vision-language-action flow model for general robot control,” in *RSS*, 2025.
- [5] M. J. Kim et al., “OpenVLA: An open-source vision-language-action model,” in *CoRL*, 2024.
- [6] J. Bjorck et al., “GR00T N1: An open foundation model for generalist humanoid robots,” *arXiv preprint arXiv:2503.14734*, 2025.
- [7] W. Huang et al., “VoxPoser: Composable 3D value maps for robotic manipulation with language models,” in *CoRL*, 2023.
- [8] M. Ahn et al., “Do as i can, not as i say: Grounding language in robotic affordances,” in *CoRL*, 2022.
- [9] R. Lu et al., “ASPIRE: Agentic skills discovery for robotics,” *arXiv preprint arXiv:2607.00272*, 2026.
- [10] N. Silberman et al., “Indoor segmentation and support inference from RGBD images,” in *ECCV*, 2012.
- [11] F. Perazzi et al., “A benchmark dataset and evaluation methodology for video object segmentation,” in *CVPR*, 2016.
- [12] A. Dai et al., “ScanNet: Richly-annotated 3D reconstructions of indoor scenes,” in *CVPR*, 2017.
- [13] Y. Xiang et al., “PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes,” in *RSS*, 2018.
- [14] T. Hodaň et al., “BOP challenge 2023 on detection, segmentation and pose estimation of seen and unseen rigid objects,” *arXiv preprint arXiv:2403.09799*, 2024.
- [15] L. Kong et al., “RoboDepth: Robust out-of-distribution depth estimation under corruptions,” in *NeurIPS*, vol. 36, 2023.
- [16] A. Khazatsky et al., “DROID: A large-scale in-the-wild robot manipulation dataset,” in *RSS*, 2024.
- [17] H.-S. Fang et al., “RH20T: A comprehensive robotic dataset for learning diverse skills in one-shot,” in *ICRA*, 2024.
- [18] J. J. P. et al., “Proto-CLIP: Vision-language prototypical network for few-shot learning,” in *IROS*, 2024, pp. 2594–2601.
- [19] H. Zhen et al., “Tesseract: Learning 4D embodied world models,” in *ICCV*, 2025.
- [20] Y. Lu et al., “Adapting pre-trained vision models for novel instance detection and segmentation,” in *IROS*, 2025, pp. 13 341–13 348.
- [21] J. J. P. et al., “iTeach: In the wild interactive teaching for failure-driven adaptation of robot perception,” *arXiv:2410.09072*, 2026.
- [22] T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” in *ECCV*, 2014.
- [23] A. Geiger et al., “Are we ready for autonomous driving? the KITTI vision benchmark suite,” in *CVPR*, 2012.
- [24] Y. Han et al., “SenseShift6D: Multimodal RGB-D benchmarking for robust 6D pose estimation across environment and sensor variations,” *arXiv preprint arXiv:2507.05751*, 2025.
- [25] Y. Liu et al., “HOI4D: A 4D egocentric dataset for category-level human-object interaction,” in *CVPR*, 2022.
- [26] P. Banerjee et al., “HOT3D: Hand and object tracking in 3D from egocentric multi-view videos,” in *CVPR*, 2025.
- [27] S. Back et al., “GraspClutter6D: A large-scale real-world dataset for robust perception and grasping in cluttered scenes,” *IEEE Robotics and Automation Letters*, vol. 10, no. 10, pp. 10 498–10 505, 2025.
- [28] N. Khargonkar et al., “SceneReplica: Benchmarking real-world robot manipulation by creating replicable scenes,” in *ICRA*, 2024, pp. 8258–8264.
- [29] S. H. Allu et al., “HRT1: One-shot human-to-robot trajectory transfer for mobile manipulation,” *arXiv preprint arXiv:2510.21026*, 2025.
- [30] D. Eigen et al., “Depth map prediction from a single image using a multi-scale deep network,” in *NeurIPS*, 2014.
- [31] J. Luiten et al., “HOTA: A higher order metric for evaluating multi-object tracking,” *IJCV*, vol. 129, pp. 548–578, 2021.
- [32] K. Bernardin et al., “Evaluating multiple object tracking performance: The CLEAR MOT metrics,” *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1–10, 2008.
- [33] S. S. Wilks, “Certain generalizations in the analysis of variance,” *Biometrika*, vol. 24, no. 3/4, pp. 471–494, 1932.
- [34] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum, 1988.
- [35] C. R. Rao, “An asymptotic expansion of the distribution of Wilks’ criterion,” *Bulletin of the International Statistical Institute*, vol. 33, no. 2, pp. 177–180, 1951.
- [36] H. Hotelling, “The generalization of Student’s ratio,” *Annals of Mathematical Statistics*, vol. 2, no. 3, pp. 360–378, 1931.
- [37] S. Holm, “A simple sequentially rejective multiple test procedure,” *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.
- [38] E. C. Harrington, “The desirability function,” *Industrial Quality Control*, vol. 21, no. 10, pp. 494–498, 1965.
- [39] N. Ravi et al., “SAM 2: Segment anything in images and videos,” in *ICLR*, 2025.
- [40] M. Hu et al., “Metric3D v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation,” *IEEE TPAMI*, vol. 46, no. 12, pp. 10 579–10 596, 2024.
- [41] J. He et al., “Lotus-2: Advancing geometric dense prediction with powerful image generative model,” *arXiv preprint arXiv:2512.01030*, 2025.
- [42] J. Wang et al., “VGGT: Visual geometry grounded transformer,” in *CVPR*, 2025.
- [43] B. Ke et al., “Video depth without video models,” in *CVPR*, 2025.
- [44] S. Ding et al., “SAM2Long: Enhancing SAM 2 for long video segmentation with a training-free memory tree,” in *ICCV*, 2025.
- [45] Meta AI, “SAM 3.1: Faster and more accessible real-time video detection and tracking with multiplexing and global reasoning,” Meta AI model update, Mar. 2026.
- [46] H. Lin et al., “Depth anything 3: Recovering the visual space from any views,” in *ICLR*, 2026.
- [47] J. Zhang et al., “MonST3R: A simple approach for estimating geometry in the presence of motion,” in *ICLR*, 2025.
- [48] S. Bai et al., “Qwen3-VL technical report,” *arXiv preprint arXiv:2511.21631*, 2025.
- [49] B. Xiao et al., “Florence-2: Advancing a unified representation for a variety of vision tasks,” in *CVPR*, 2024.
- [50] H. Rezatofighi et al., “Generalized intersection over union: A metric and a loss for bounding box regression,” in *CVPR*, 2019.
- [51] F. Tosi et al., “ZipDepth: Bringing lightweight zero-shot monocular depth anywhere, on any device,” in *ECCV*, 2026.
- [52] Z. Zhang et al., “Hyden: A hybrid dual-path encoder for monocular geometry of high-resolution images,” in *ICLR*, 2026.
- [53] J. Wang et al., “FE2E: From editor to dense geometry estimator,” in *CVPR*, 2026.
- [54] L. Piccinelli et al., “UniDepthV2: Universal monocular metric depth estimation made simpler,” *IEEE TPAMI*, vol. 48, no. 3, pp. 2354–2367, 2026.
- [55] R. Wang et al., “MoGe-2: Accurate monocular geometry with metric scale and sharp details,” in *NeurIPS*, vol. 38, 2025.
- [56] A. Bochkovskii et al., “Depth pro: Sharp monocular metric depth in less than a second,” in *ICLR*, 2025.
- [57] L. Yang et al., “Depth anything V2,” in *NeurIPS*, 2024.
- [58] S. Chen et al., “Video depth anything: Consistent depth estimation for super-long videos,” in *CVPR*, 2025.
- [59] Z. Yu et al., “Towards consistent video geometry estimation,” *arXiv preprint arXiv:2605.30060*, 2026.
- [60] Y. Liu et al., “Gemdepth: Geometry-embedded features for 3D-consistent video depth,” in *ICML*, 2026.
- [61] W. Hu et al., “Depthcrafter: Generating consistent long depth sequences for open-world videos,” in *CVPR*, 2025.
- [62] H. Zhang et al., “DVD: Deterministic video depth estimation with generative priors,” *arXiv preprint arXiv:2603.12250*, 2026.
- [63] J. Shao et al., “Learning temporally consistent video depth from video diffusion priors,” in *CVPR*, 2025.
- [64] J. Videnovic et al., “A distractor-aware memory for visual object tracking with SAM2,” in *CVPR*, 2025.
- [65] C. Zhou et al., “EdgeTAM: On-device track anything model,” in *CVPR*, 2025.
- [66] J. Zhang et al., “SAM 2++: Tracking anything at any granularity,” *arXiv preprint arXiv:2510.18822*, 2025.
- [67] H. K. Cheng et al., “Putting the object back into video object segmentation,” in *CVPR*, 2024.
- [68] Y. Xu et al., “Integrating boxes and masks: A multi-object framework for unified visual tracking and segmentation,” in *ICCV*, 2023.
- [69] H. K. Cheng et al., “XMem: Long-term video object segmentation with an Atkinson-Shiffrin memory model,” in *ECCV*, 2022.
- [70] S. Liu et al., “Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection,” in *ECCV*, 2024.
- [71] T. Ren et al., “Grounded SAM 2: Ground and track anything in videos,” Official IDEA-Research software release, 2024.
- [72] S. Dong et al., “Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization,” in *CVPR*, 2025.
- [73] Y. Wang et al., “Pi3X model release,” Official π^3 project repository, Dec. 2025.
- [74] Y. Cabon et al., “MUST3R: Multi-view network for stereo 3D reconstruction,” in *CVPR*, 2025.
- [75] J. Yang et al., “Fast3R: Towards 3D reconstruction of 1000+ images in one forward pass,” in *CVPR*, 2025.
- [76] Q. Wang et al., “Continuous 3D perception model with persistent state,” in *CVPR*, 2025.
- [77] V. Leroy et al., “Grounding image matching in 3D with MAST3R,” in *ECCV*, 2024.
- [78] S. Wang et al., “DUST3R: Geometric 3D vision made easy,” in *CVPR*, 2024.
- [79] NVIDIA, “Cosmos-Reason2: Physical AI common sense and embodied reasoning models,” <https://huggingface.co/nvidia/Cosmos-Reason2-8B>, 2025.
- [80] W. Wang et al., “InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency,” *arXiv preprint arXiv:2508.18265*, 2025.
- [81] A. Steiner et al., “PaliGemma 2: A family of versatile VLMs for transfer,” *arXiv preprint arXiv:2412.03555*, 2024.
- [82] Z. Wu et al., “DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding,” *arXiv preprint arXiv:2412.10302*, 2024.
- [83] J. J. P., “Robokit: A toolkit for robotic tasks,” 2024, <https://github.com/jishnujyakumar/robokit>.

APPENDIX

This supplement documents the data, procedures, and complete results behind the main paper. Appendix A describes the dataset composition, capture protocol, sensor rig, capture sites, scenes, and object catalogue (main paper Sec. III-C). Appendix B covers mask, tracklet, and egocentric annotation and the RPX difficulty score that defines the Easy, Medium, and Hard tiers. Appendix C defines every task metric and the VQA questions, and explains the MANOVA estimator and desirability windows behind Φ and $\mathcal{J}$ (Sec. III-B). Appendix D reports per-model results and hardware and latency measurements (Sec. V), and Appendix E explains how to add models and tasks.

A. DATA COLLECTION

A.1 Dataset composition

RPX contains 133,121 frames in three streams (Table VIIIa): multi-object exocentric captures (MOS) of 100 scenes in the Clutter, Interaction, and Clean phases; egocentric captures during Interaction; and single-object captures (SOS) of 70 objects. Sixty scenes are indoor and 40 outdoor, recorded at 20 sites (Fig. 8); Table VIIIb gives their distribution over the difficulty tiers.

The question banks contain 4,529,322 single-image and 749,083 in-context question-frame records. Evaluation uses fixed, seeded subsets of about 21,000 T5 and 10,500 T6 questions over Clutter, Interaction, Clean, and Ego Interaction; the published grounding scores use the attribute questions among them (about 12,000 and 6,000; Appendix C.2).

TABLE VIII: RPX dataset composition.

(a) Streams

Stream	Frames	Capture; modalities and annotations
MOS	75,000	100 scenes $\times$ 3 phases; RGB-D, stereo, pose, masks, tracks, object joins, questions
Ego	23,121	100 Interaction clips; RGB, masks, tracks, questions
SOS	35,000	70 objects $\times$ 500 frames; RGB-D, stereo, pose, masks, questionnaire
Total	133,121	

(b) MOS scenes by environment and difficulty tier

Environment	Easy	Medium	Hard	Total
Indoor	22	23	15	60
Outdoor	11	10	19	40
Total	33	33	34	100

Figure 5 shows the same counts as within-tier percentages: the outdoor share is 33% of Easy, 30% of Medium, and 56% of Hard scenes. The tiers come from the RPX difficulty score (RPX-DS; Appendix B.3), which does not use the environment label.

A.2 Objects

RPX uses 70 everyday objects (Fig. 6). Each is first recorded alone, in a single-object capture (SOS) of 500 frames (35,000 in total) while the hand-held rig orbits it; SOS

Are hard scenes more often outdoors?

Yes: 56% of Hard scenes vs. 30–33% of the others

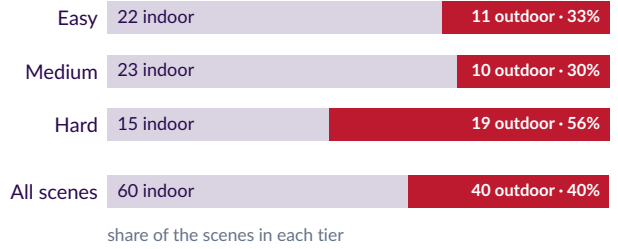


Fig. 5: Indoor and outdoor scenes in each RPX-DS difficulty tier: the outdoor share is 33% of Easy, 30% of Medium, and 56% of Hard scenes; the full dataset is 60 indoor and 40 outdoor. Labels give scene counts.

images are the reference images for in-context grounding (T6). Annotators also describe each object in a short questionnaire (Table IX), and Table XIII lists one canonical response per field for every object.

TABLE IX: SOS questionnaire schema.

Field	Annotator prompt	Use in T5/T6
Name	Names and valid aliases	Catalogue label only
Category	Semantic object class	Catalogue label only
Colour	Visible colour descriptions	Colour questions
Material	Constituent material	Material, composition, odd-one-out
Function	Common purpose or action	Function, composition

A.3 Modalities captured

One hand-held rig records every capture: a D435 for RGB and metric depth and a T265 for calibrated fisheye stereo and 6-DoF camera pose, with a head-mounted GoPro adding the egocentric view during Interaction (Secs. III-A and III-C). Figure 7 shows every modality for one SOS frame, together with the instance mask (Appendix B) and the questionnaire. The multi-object scenes record the same modalities in each exocentric phase (Table X). The exocentric (D435, T265) and egocentric (GoPro) streams run on independent clocks, and phase boundaries are set manually during recording.

TABLE X: Modalities and annotations per capture view for all 100 MOS scenes (✓: provided; N/A: not captured in this view). Pose is recorded in all three exocentric phases; relative-pose evaluation uses Clutter and Clean pairs only.

Modality	Clutter	Interaction	Clean	Ego Int.
RGB	✓	✓	✓	✓
Metric depth (D435)	✓	✓	✓	N/A
Instance masks	✓	✓	✓	✓
Tracklets (persistent IDs)	✓	✓	✓	✓
Fisheye stereo (T265)	✓	✓	✓	N/A
Camera pose (T265)	✓	✓	✓	N/A
Grounding questions	✓	✓	✓	✓
Frames per capture	250	250	250	195–250

The rig uses only commercially available parts (Table XI).



Fig. 6: All 70 SOS objects, each cut out of one released SOS frame with its instance mask and numbered as in Table XIII.

Replication requires a T265 or another calibrated 6-DoF visual-inertial pose source that is time-aligned with the D435. The bracket and the D435–T265 extrinsic follow Intel’s reference design for this camera pair.

A.4 Capture sites

The multi-object scenes were recorded at 20 sites, 12 indoor and 8 outdoor, with five consecutive scenes per site (Fig. 8). Each site is photographed without the scene objects; the environment of a scene is that of its site.

A.5 Indoor and outdoor scenes

At each site the 70 objects are combined into multi-object scenes (MOS) and captured with all of the modalities above in the Clutter, Interaction, and Clean phases, plus the egocentric view during Interaction. Figures 9 and 10 show one indoor and one outdoor scene per difficulty tier in all four views, with the released instance masks overlaid; each object keeps one colour across views (Appendix A.6).

The complete visual catalogue of all 100 scenes, one middle frame (index $\lfloor n/2 \rfloor$ of n) per view, is available on the [project page](#). Each scene contains six to eight SOS objects, seven in 98 of the 100 scenes. The object set is the same in all

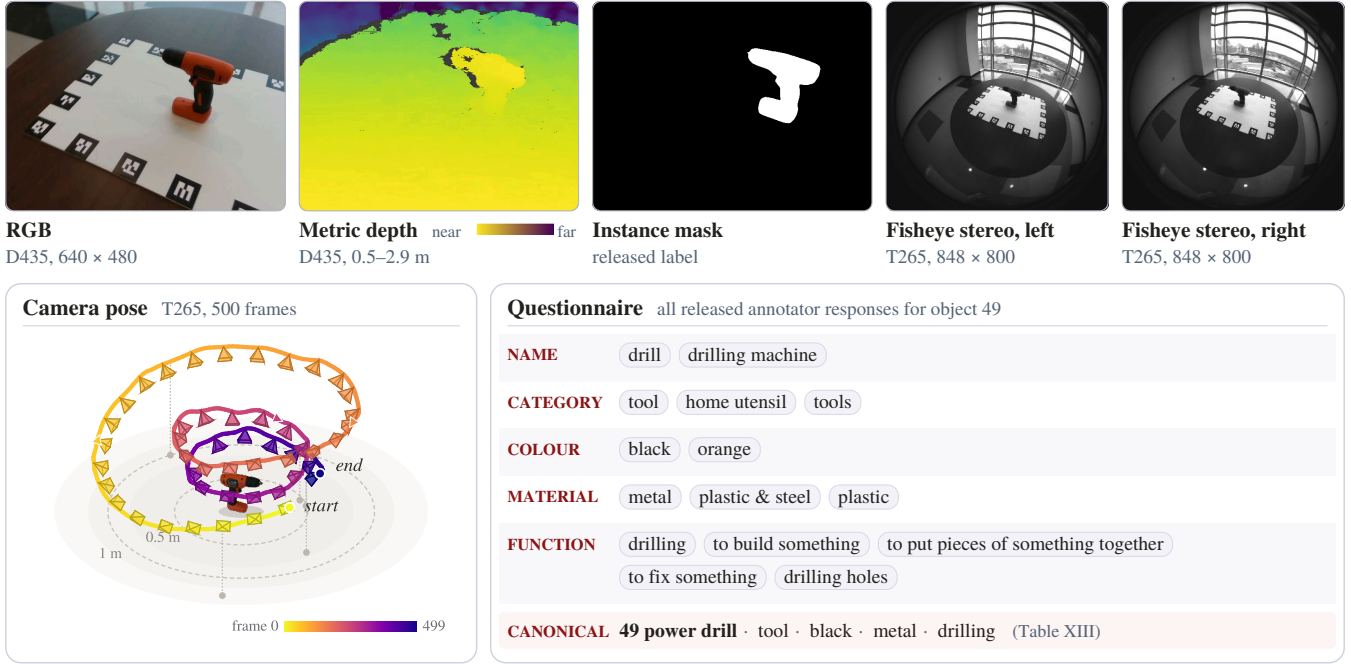


Fig. 7: SOS modalities for one frame (289 of 500) of object 49, power drill: RGB and metric depth (D435), the released instance mask, the calibrated T265 fisheye stereo pair, the T265 camera trajectory over all 500 frames (camera frustums coloured by frame, the object at the cameras’ common look-at point, and floor rings at 0.5 m and 1 m), and the questionnaire responses behind the canonical entry. The exocentric MOS captures record the same RGB-D, mask, stereo, and pose modalities; the MOS figures show only the RGB phase images.

TABLE XI: Capture rig bill of materials. Prices are approximate retail at the time of capture (USD).

Component	Role	Price
Intel D435	RGB-D (640×480, 30 fps; metric depth via active IR stereo)	\$180
Intel T265	6-DoF VIO at up to 200 Hz (30 fps fisheye stereo, IMU-interpolated); two wide-FOV cameras. Discontinued 2021; secondary market	\$250–400
GoPro (head mount)	Egocentric RGB during Interaction (independent clock)	\$350
3D-printed bracket	Couples D435 and T265 using Intel’s reference geometry	\$5
Commercial clamp	Attaches assembly to laptop screen for hand-held use	\$15
Total		~\$1K

three exocentric phases, and each of the 70 objects appears in 9–11 scenes (Table XII).

A.6 Identifiers and object joins

MOS mask IDs are local to each (`scene_id`, `phase`) pair. Object-join metadata map each local ID through `object_id` to a dataset-level `global_object_id`. The global identity connects an object across phases, views, and the SOS catalogue.

B. DATA PREPROCESSING

B.1 Mask and tracklet annotation

The annotation pipeline uses RoboKit [83] to initialize SAM2 masks from grounded detections and propagate them

bidirectionally through each sequence (Fig. 11). Annotators correct missed, merged, or drifting instances with point or box prompts; propagation restarts at the corrected frame. Quality control checks nonempty foreground, identity continuity, temporal area continuity, and the scene–phase object mapping. Tracklets inherit the corrected persistent mask identities.

B.2 Egocentric annotation

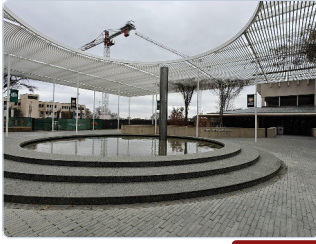
The GoPro and D435 share no timecode. Egocentric clips are sampled offline to at most 250 frames, including both endpoints; the release contains 195–250 frames per scene (23,121 in total), and `ego_frame_map.json` maps each sample to its source frame. Masks follow the MOS pipeline. Annotators give each egocentric instance the object identity of its exocentric counterpart, so identities match across cameras without pixel or temporal alignment.

B.3 Difficulty descriptors

Table XV lists 27 difficulty descriptors, computed from the aligned frame sequence of each of the 300 exocentric scene–phase cells. Two measure annotation effort; the other 25 cover scene complexity, occlusion, depth quality, photometric conflict, temporal stability, camera motion, and fisheye stereo. The release includes all descriptor values. Table XIV gives the shared notation; group-specific symbols are introduced where needed.

1) Annotation effort (2 descriptors)

Let $\mathcal{F}_i$ contain the frame indices still marked faulty after manual correction round i ($i=1, \dots, 4$). The per-frame effort



Site 01
Fountain plinth steps
Scenes 001–005



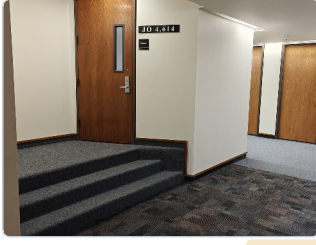
Site 02
Round table by windows
Scenes 006–010



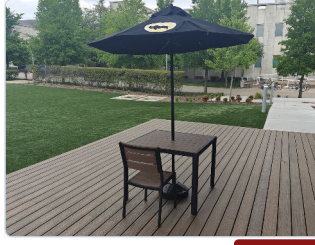
Site 03
Sofa lounge
Scenes 011–015



Site 04
Checkerboard dining table
Scenes 016–020



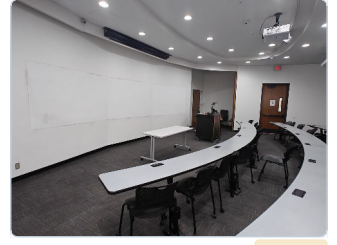
Site 05
Carpeted stairs
Scenes 021–025



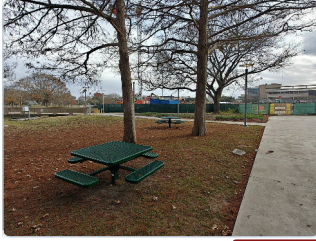
Site 06
Deck table with umbrella
Scenes 026–030



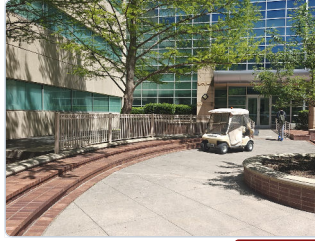
Site 07
Sitting area
Scenes 031–035



Site 08
Lecture hall table
Scenes 036–040



Site 09
Picnic bench
Scenes 041–045



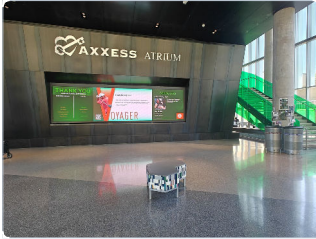
Site 10
Shaded brick steps
Scenes 046–050



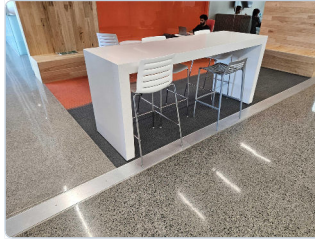
Site 11
Tennis court
Scenes 051–055



Site 12
Lounge benches
Scenes 056–060



Site 13
Atrium seating block
Scenes 061–065



Site 14
Tall white table
Scenes 066–070



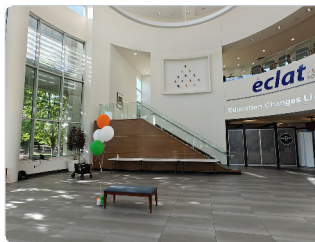
Site 15
Atrium stairs
Scenes 071–075



Site 16
Round lounge table
Scenes 076–080



Site 17
Garden planter
Scenes 081–085



Site 18
Atrium bench
Scenes 086–090



Site 19
Stone bench
Scenes 091–095



Site 20
Rock garden
Scenes 096–100

Fig. 8: The 20 capture sites, photographed without the scene objects. Each site hosts five consecutive MOS scenes.

count

$$k_t = 1 + |\{i : t \in \mathcal{F}_i\}| \quad (4)$$

captures how many review passes frame t required: $k_t=1$ if accepted on the first pass, $k_t=5$ if still faulty after all four

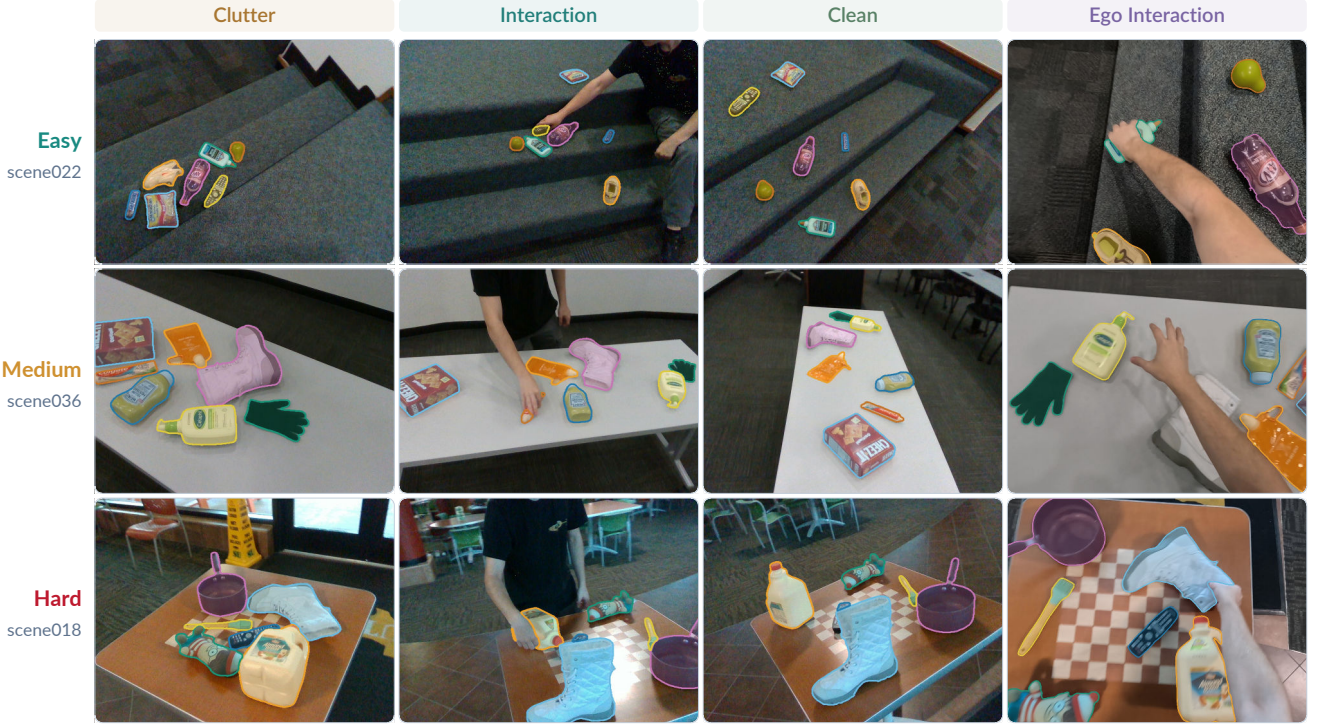


Fig. 9: Representative indoor MOS scenes, one per difficulty tier, in the four capture views; overlays are the released instance masks, and each object keeps one colour across its views. Percentiles rank a scene’s phase-averaged descriptor (Appendix B.3) among all 100 scenes. **Easy**, scene 022 (carpeted stairs; RPX-DS 0.29, 1st percentile): little object overlap (9th) and nearly complete depth (0.3% of pixels invalid). **Medium**, scene 036 (lecture-hall table; RPX-DS 0.49, 47th): typical on most descriptors, with more dark regions lacking depth (84th). **Hard**, scene 018 (checkerboard dining table; RPX-DS 0.63, 92nd): a wide depth range (82nd), objects changing size in view (93rd), glare and dark regions lacking depth (79th, 88th), and more annotation correction rounds (86th).

rounds.

$$\text{iter_mean} = \frac{1}{T} \sum_{t=1}^T k_t, \quad \text{iter_max} = \max_t k_t. \quad (5)$$

2) Scene complexity (3 descriptors)

$$\text{obj_mean} = \frac{1}{T} \sum_t n_t, \quad (6)$$

$$\text{obj_std} = \text{std}_t(n_t), \quad (7)$$

$$\text{obj_consist} = \frac{1}{T} \sum_t \mathbf{1}[n_t = N]. \quad (8)$$

obj_consist is the fraction of frames in which all N labelled instances are visible.

3) Occlusion (3 descriptors)

For each visible instance i at frame t , let B_i^t be its tight bounding box with area $|B_i^t|$. The normalized pairwise overlap is

$$o_i^t = \frac{1}{|B_i^t|} \sum_{j \in \mathcal{I}_t \setminus \{i\}} |B_i^t \cap B_j^t|. \quad (9)$$

Taking the temporal mean per instance, $\bar{o}_i = \text{mean}_t o_i^t$ (over frames where i is visible), the three features are

$$\text{occ_mean} = \text{mean}_i \bar{o}_i, \quad (10)$$

$$\text{occ_p90} = P_{90}(\{\bar{o}_i\}), \quad (11)$$

$$\text{occ_heavy} = \frac{1}{|\mathcal{I}|} \sum_i \mathbf{1}[\bar{o}_i > 0.3]. \quad (12)$$

4) Depth quality (4 descriptors)

$$\text{depth_invalid} = \frac{1}{T} \sum_t |\bar{V}_t| / (HW), \quad (13)$$

$$\text{depth_invalid_mask} = \frac{1}{T} \sum_t |\bar{V}_t \cap \mathcal{M}_t| / |\mathcal{M}_t|. \quad (14)$$

Depth spread is measured over all valid pixels and over object pixels:

$$\text{depth_std} = \frac{1}{T} \sum_t \sigma(D_t|_{V_t}), \quad (15)$$

$$\text{depth_std_mask} = \frac{1}{T} \sum_t \sigma(D_t|_{V_t \cap \mathcal{M}_t}), \quad (16)$$

where $\sigma(D_t|_S)$ is the standard deviation of depth values over pixel set S .

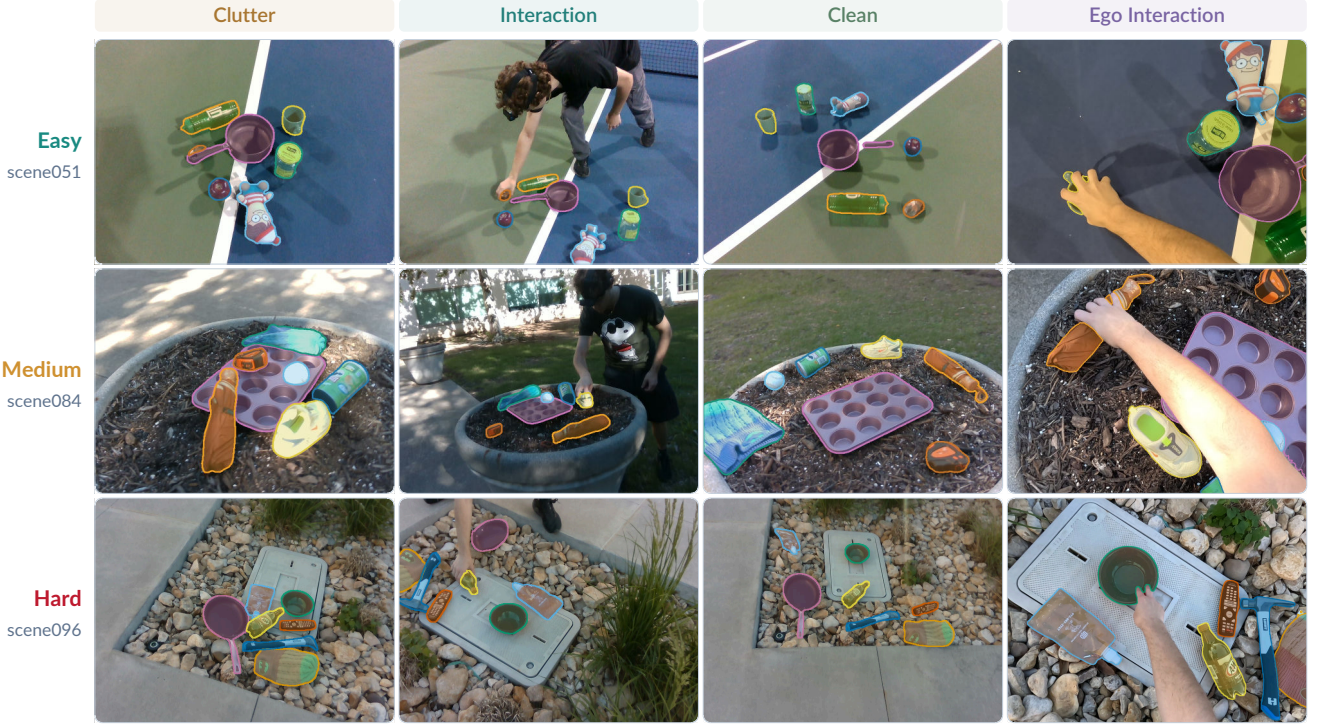


Fig. 10: Representative outdoor MOS scenes, one per difficulty tier, shown as in Fig. 9. **Easy**, scene 051 (tennis court; RPX-DS 0.43, 21st percentile): little overlap and few depth gaps, but strong glare (70th) and fast camera motion (83rd). **Medium**, scene 084 (garden planter; RPX-DS 0.53, 61st): the heaviest object overlap (88th), a wide depth range (93rd), the most glare (94th), and objects changing size in view (95th). **Hard**, scene 096 (rock garden; RPX-DS 0.60, 84th): high object overlap (74th) and among the most annotation correction rounds (99th).

Table XII lists the objects in every MOS scene by SOS catalogue ID.

TABLE XII: Objects in each MOS scene, listed by their SOS catalogue ID (Table XIII); the set is identical in all three exocentric phases. Scene numbers are coloured by difficulty tier (**Easy**, **Medium**, **Hard**); underlined numbers are outdoor scenes, following the capture sites of Fig. 8 (40 outdoor, 60 indoor). Table XVI lists the same split by tier.

Scene	Objects	Scene	Objects	Scene	Objects	Scene	Objects
<u>001</u>	1, 2, 11, 20, 23, 27, 39	<u>026</u>	6, 21, 38, 43, 48, 49, 65	<u>051</u>	3, 20, 28, 33, 44, 45, 70	<u>076</u>	11, 14, 35, 41, 42, 60, 63
<u>002</u>	28, 37, 42, 46, 52, 56, 66	<u>027</u>	1, 10, 22, 41, 50, 60, 66	<u>052</u>	11, 29, 31, 39, 43, 52, 54	<u>077</u>	3, 15, 17, 22, 23, 30, 38
<u>003</u>	3, 9, 31, 49, 50, 51, 54	<u>028</u>	8, 23, 29, 33, 54, 58, 59	<u>053</u>	7, 30, 32, 34, 56, 62, 68	<u>078</u>	18, 26, 32, 43, 44, 47, 69
<u>004</u>	4, 14, 44, 55, 62, 65, 69	<u>029</u>	11, 24, 31, 34, 46, 51, 70	<u>054</u>	4, 37, 41, 47, 49, 58, 63	<u>079</u>	2, 10, 12, 27, 33, 56, 59
<u>005</u>	7, 13, 17, 19, 25, 60, 63	<u>030</u>	3, 37, 40, 57, 67, 68, 69	<u>055</u>	18, 35, 40, 48, 51, 55, 59	<u>080</u>	1, 25, 29, 39, 46, 61, 68
<u>006</u>	12, 15, 30, 32, 34, 38, 59	<u>031</u>	8, 29, 38, 56, 58, 61, 65	<u>056</u>	1, 12, 27, 36, 53, 67, 69	<u>081</u>	3, 17, 18, 24, 35, 36, 66
<u>007</u>	16, 22, 40, 43, 58, 64, 67	<u>032</u>	5, 20, 25, 26, 33, 45, 63	<u>057</u>	2, 6, 17, 24, 25, 46, 64	<u>082</u>	26, 29, 30, 42, 44, 51, 61
<u>008</u>	5, 8, 24, 35, 47, 48, 53	<u>033</u>	1, 13, 15, 40, 53, 54, 68	<u>058</u>	9, 15, 19, 21, 26, 38, 60	<u>083</u>	10, 16, 21, 37, 45, 46, 67
<u>009</u>	6, 21, 26, 33, 36, 57, 61	<u>034</u>	12, 16, 19, 41, 55, 60, 62	<u>059</u>	5, 10, 13, 14, 22, 61, 66	<u>084</u>	4, 6, 8, 50, 56, 64, 69
<u>010</u>	10, 18, 29, 41, 45, 68, 70	<u>035</u>	2, 49, 50, 57, 59, 64, 69	<u>060</u>	8, 16, 23, 42, 50, 57, 65	<u>085</u>	2, 20, 38, 40, 41, 49, 58
<u>011</u>	3, 4, 29, 30, 46, 54, 56	<u>036</u>	11, 18, 24, 35, 46, 51, 67	<u>061</u>	5, 6, 7, 15, 19, 21, 28	<u>086</u>	9, 22, 25, 43, 47, 55, 60
<u>012</u>	2, 14, 48, 53, 57, 65, 66	<u>037</u>	3, 7, 22, 30, 34, 36, 39	<u>062</u>	14, 25, 30, 39, 41, 43, 59	<u>087</u>	1, 23, 31, 39, 62, 63, 65
<u>013</u>	6, 19, 22, 31, 39, 59, 61	<u>038</u>	4, 27, 37, 43, 47, 66, 70	<u>063</u>	18, 29, 37, 47, 60, 66, 70	<u>088</u>	14, 27, 28, 32, 34, 57, 59
<u>014</u>	5, 24, 27, 38, 42, 64, 70	<u>039</u>	6, 14, 17, 21, 31, 44, 48	<u>064</u>	13, 26, 31, 32, 40, 51, 69	<u>089</u>	7, 11, 15, 33, 52, 68, 70
<u>015</u>	9, 17, 25, 32, 34, 58, 67	<u>040</u>	9, 10, 23, 28, 32, 42, 52	<u>065</u>	3, 22, 24, 34, 35, 48, 57	<u>090</u>	5, 12, 13, 19, 48, 53, 54
<u>016</u>	10, 15, 16, 45, 47, 49, 63	<u>041</u>	13, 14, 28, 37, 44, 60, 69	<u>066</u>	2, 27, 42, 45, 53, 55, 56	<u>091</u>	23, 24, 39, 47, 51, 57, 60
<u>017</u>	23, 35, 36, 40, 41, 44, 62	<u>042</u>	42, 46, 53, 55, 65, 66, 67	<u>067</u>	11, 16, 36, 46, 50, 52, 58	<u>092</u>	11, 22, 29, 30, 31, 37, 42
<u>018</u>	7, 11, 20, 28, 43, 52	<u>043</u>	4, 15, 25, 30, 45, 56, 57	<u>068</u>	9, 23, 33, 57, 61, 62, 67	<u>093</u>	2, 3, 26, 44, 50, 63, 68
<u>019</u>	8, 21, 37, 50, 55, 68, 69	<u>044</u>	3, 5, 10, 40, 49, 59, 64	<u>069</u>	10, 12, 17, 49, 54, 64, 65	<u>094</u>	5, 9, 33, 36, 46, 55, 67
<u>020</u>	1, 12, 13, 18, 26, 33, 60	<u>045</u>	21, 22, 23, 32, 38, 41, 58	<u>070</u>	4, 8, 20, 38, 44, 63, 68	<u>095</u>	7, 10, 45, 62, 64, 66, 69
<u>021</u>	5, 7, 13, 20, 27, 28, 63	<u>046</u>	1, 6, 19, 29, 36, 43, 50	<u>071</u>	4, 8, 28, 37, 45, 49, 62	<u>096</u>	8, 12, 21, 27, 35, 52, 53
<u>022</u>	14, 25, 32, 47, 52, 53, 56	<u>047</u>	8, 9, 12, 18, 47, 52, 61	<u>072</u>	5, 9, 20, 31, 50, 66, 70	<u>097</u>	1, 14, 17, 41, 48, 58, 61
<u>023</u>	4, 12, 17, 26, 39, 42, 44	<u>048</u>	7, 20, 24, 26, 31, 63, 68	<u>073</u>	13, 16, 34, 36, 40, 52, 65	<u>098</u>	15, 18, 25, 28, 32, 34, 59
<u>024</u>	2, 9, 15, 16, 18, 19, 61	<u>049</u>	11, 17, 33, 39, 48, 51, 62	<u>074</u>	7, 19, 21, 53, 54, 58, 67	<u>099</u>	13, 16, 19, 20, 38, 56, 65
<u>025</u>	30, 35, 36, 45, 55, 62, 64	<u>050</u>	2, 16, 27, 34, 35, 54, 70	<u>075</u>	6, 24, 48, 51, 55, 57, 64	<u>100</u>	4, 6, 11, 40, 43, 49, 54, 70

5) Photometric conflict (2 descriptors)

These descriptors count pixels that are very bright or very dark *and* have no valid depth:

$$\text{specular} = \frac{1}{T} \sum_t \frac{|\{(u, v) : L_t(u, v) > 230\} \cap \bar{V}_t|}{HW}, \quad (17)$$

Table XIII lists all 70 objects with their canonical responses.

TABLE XIII: Canonical SOS questionnaire catalogue: every object with one lightly normalized canonical category, colour, material, and function response.

ID	Object	Category	Colour	Material	Function
1	air duster can	utilities	grey	plastic/metal	cleaning
2	alarm clock	clocks and watches	silver	metal	waking up
3	apple	food	red	plant	eating
4	baking tray	baking	black	metal	baking food
5	banana	fruit	yellow	plastic	eating
6	baseball ball	sports	white	rubber	playing
7	basting brush	office tools	brown and blue	plastic	clean keyboard
8	beanie	clothing	purple blue and dark purple	fabric	fashion
9	bear	toy	grey	cotton	playing
10	book	stationery	white	paper	study
11	boot	apparel	white	leather	footwear
12	bowl	utensil	red	glass	keeps food inside it
13	cable	accessories	white	rubber	connect
14	candy bar	food	brown	plastic	eat
15	cleaning brush	hygiene equipment	white	plastic	cleaning
16	coffee bottle	bottle	brown	plastic	drinking
17	coffee can	can	red	metal	to eat
18	cracker box	box	red	plastic	eating
19	cup	food tool	yellow and white	paper	drinking liquids
20	doll	toys	red	rubber	playing
21	fry pan	kitchenware	black	cast iron	cooking
22	food bag	food	yellow green	plastic	eating
23	game controller	gaming	black	plastic	gaming
24	glove	clothing accessory	black	cloth	going outside in the cold
25	glue bottle	holding	white orange and black	plastic	glue things
26	hair brush	beauty	black	plastic	styling
27	hammer	tool	black	metal	hammering
28	handy pot	utensil	black	metal & plastic	cooking
29	hat	accessory	white	cotton	for head
30	headphone	technology	black	plastic	listening
31	helmet	safety	purple	fiberglass	wearable
32	instant noodle packet	packet	red	noodles	eating
33	jar	storage	yellow	glass	storing food
34	jello box	box	white	paper	eating
35	juice pouch	container	orange	plastic	drink
36	kettle	utensil	silver	plastic and metal	tea or water
37	knife	kitchen accessories	red and silver	steel	cut food
38	ladle	utensils	green	plastic	eating
39	lion	toys	tan	plastic	play
40	marker	office tools	black and yellow	plastic	writing
41	meat can	can	blue	metal	eating
42	milk box	food	red	paper	eating
43	milk jug	container	white	plastic	containing liquid
44	mouse	computer accessories	blue	plastic	working
45	mug	kitchen accessory	blue	porcelain	drink tea
46	mustard bottle	food	yellow and white	plastic	seasoning
47	pear	fruit	green	plastic	eating
48	peeler	kitchen tool	black	plastic	peeling the fruits and vegetables
49	power drill	tool	black	metal	drilling
50	pringles can	food	green	paper	eating
51	pump bottle	beauty product	white	plastic	lotion
52	remote controller	tool	black	plastic	change channel
53	root beer bottle	drink	brown	plastic	drinking
54	scissors	tool	grey	steel	cutting
55	shampoo bottle	beauty	white	plastic	scrubbing
56	shoe	apparel	white green and blue	rubber	running
57	soda can	food	blue	metal	drink
58	soft scrub cleanser bottle	bottle	white	plastic	cleaning
59	soup can	food	red	metal	drink
60	spatula	utensil	brown	wood	cooking
61	sponge	cleaning tool	blue	glycerin	cleaning
62	spray bottle	tools	blue	plastic	spray
63	t-handle	tool	black	metal	remove something
64	tape measure	tool	black	metal	to measure
65	teapot	utensils	white	plastic	store
66	tong	utensils	silver	metal	eating
67	toothpaste	health	blue	plastic	brushing tooth
68	towel	cleaning tool	yellow	cloth	clean
69	umbrella	accessory	brown	plastic	covers and protect from rain
70	water bottle	bottle	green	plastic	storing liquids

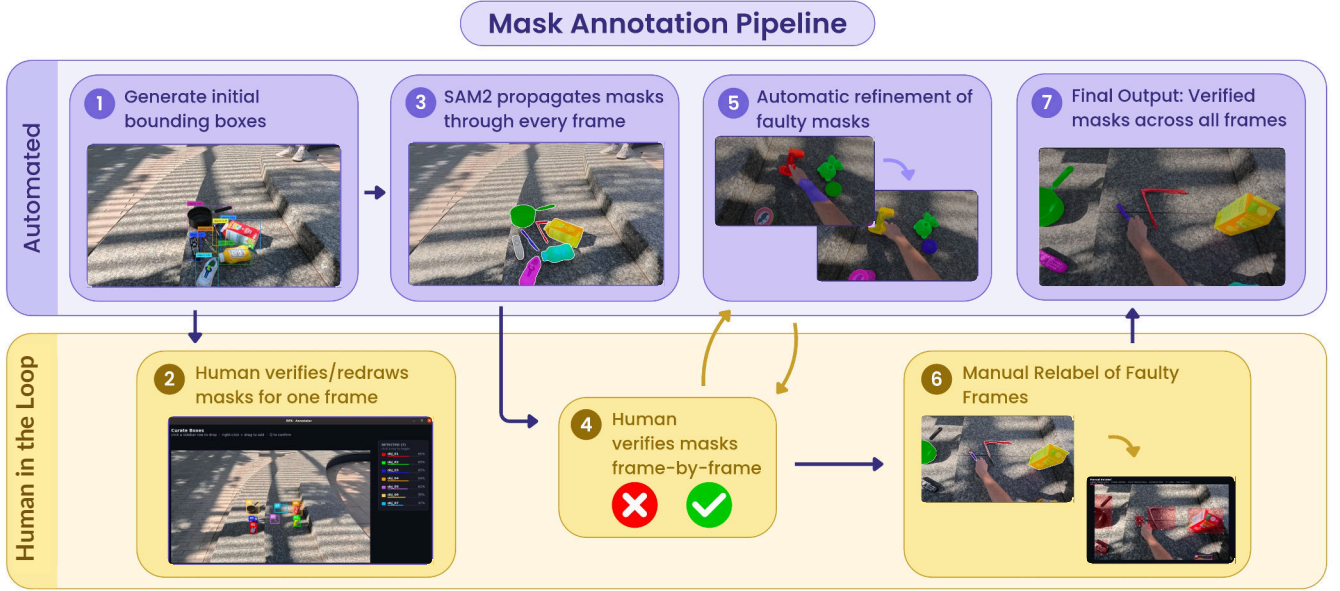


Fig. 11: Instance-mask annotation workflow using RoboKit [83]. Grounded proposals initialize SAM2 masks; bidirectional propagation alternates with human correction and identity and continuity checks.

TABLE XIV: Shared notation for the 27 difficulty descriptors.

Symbol	Definition
T	Number of aligned frames in the cell
M_t	Instance mask at frame t (pixel value = instance ID, 0=background)
D_t	Metric depth map (metres; 0=invalid)
L_t	ITU-R BT.601 luma of the RGB frame
$p_t \in \mathbb{R}^3$	T265 camera position
$q_t \in \mathbb{H}$	T265 unit orientation quaternion
$\mathcal{I}_t$	Visible instance IDs at frame t ; $n_t = \mathcal{I}_t $
$\mathcal{I}$	Union of all instance IDs across frames
N	Total labelled instance count (from mask-to-object metadata)
V_t	Valid-depth pixel set $\{(u, v) : D_t(u, v) > 0\}$
$\mathcal{M}_t$	Mask-covered pixel set $\{(u, v) : M_t(u, v) > 0\}$
$H \times W$	Frame resolution (640×480)

$$\text{dark} = \frac{1}{T} \sum_t \frac{|\{(u, v) : L_t(u, v) < 30\} \cap \overline{V}_t|}{HW}. \quad (18)$$

6) Temporal stability (3 descriptors)

Let A_i^t be the pixel area of instance i at frame t (0 if not visible), $v_i^t = \mathbf{1}[A_i^t > 0]$ its visibility indicator, and $r_i^t = (A_i^{t-1} - A_i^t) / A_i^{t-1}$ the relative area drop (defined when $A_i^{t-1} > 0$).

$$\text{area_cv} = \frac{1}{|\mathcal{I}|} \sum_i \frac{\text{std}_t(A_i^t)}{\text{mean}_t(A_i^t)}, \quad (19)$$

$$\text{area_drop} = \frac{1}{|\mathcal{I}|} \sum_i \frac{1}{T} \sum_{t=2}^T \mathbf{1}[r_i^t > \frac{1}{2}], \quad (20)$$

$$\text{vis_instability} = \frac{1}{|\mathcal{I}|} \sum_i \frac{1}{T} \sum_{t=2}^T \mathbf{1}[v_i^t \neq v_i^{t-1}]. \quad (21)$$

area_cv (Eq. 19) is computed only over frames where $A_i^t > 0$. area_drop is the fraction of frames in which an instance loses more than half of its area, and vis_instability the fraction in which it appears or disappears.

7) Camera motion (5 descriptors)

From the T265 trajectory $(p_t, q_t)_{t=1}^T$, the per-frame increments are

$$\delta_t^{\text{trans}} = \|p_{t+1} - p_t\|_2, \quad \delta_t^{\text{rot}} = 2 \arccos(|\langle q_t, q_{t+1} \rangle|), \quad (22)$$

where $\langle q_t, q_{t+1} \rangle$ is the quaternion inner product and the absolute value handles the double-cover ambiguity. Before the statistics are computed, increments more than 5×1.4826 median absolute deviations above the median are replaced by the median.

$$\text{trans_mean} = \text{mean}_t \delta_t^{\text{trans}}, \quad (23)$$

$$\text{trans_p90} = P_{90}(\delta_t^{\text{trans}}), \quad (24)$$

$$\text{rot_mean} = \text{mean}_t \delta_t^{\text{rot}}, \quad (25)$$

$$\text{rot_p90} = P_{90}(\delta_t^{\text{rot}}), \quad (26)$$

$$\text{jerk} = \frac{1}{T-2} \sum_{t=1}^{T-2} |\delta_{t+1}^{\text{trans}} - \delta_t^{\text{trans}}|. \quad (27)$$

8) Fisheye/stereo quality (5 descriptors)

Computed from the T265 wide-FOV fisheye image pairs (F_t^L, F_t^R) . Let L_t^{fe} denote the grayscale intensity of the left fisheye frame, with resolution $H_{\text{fe}} \times W_{\text{fe}}$.

$$\text{fisheye_dark} = \frac{1}{T} \sum_t \frac{|\{(u, v) : L_t^{\text{fe}}(u, v) < 30\}|}{H_{\text{fe}} W_{\text{fe}}}, \quad (28)$$

$$\text{fisheye_bright} = \frac{1}{T} \sum_t \frac{|\{(u, v) : L_t^{\text{fe}}(u, v) > 230\}|}{H_{\text{fe}} W_{\text{fe}}}. \quad (29)$$

TABLE XV: Supplementary difficulty descriptors: schema, weights in RPX-DS (Eq. (33)), and raw distribution over the 300 scene-phase cells. Statistics are raw descriptor values before percentile ranking, shown in the unit listed (e.g. 10^{-3} : value $\times 10^{-3}$).

Group	Descriptor	Weight	Unit	Mean	SD	Median	P10	P90	Min	Max
Annotation effort	iter_mean	0.125		1.40	0.56	1.13	1.00	2.12	1.00	3.90
	iter_max	0.125		2.31	1.05	3.00	1.00	4.00	1.00	5.00
Scene complexity	obj_mean	0.030		6.97	0.16	7.00	6.90	7.00	5.90	8.00
	obj_std	0.030		0.101	0.158	0	0	0.311	0	0.803
	obj_consist	0.030		0.97	0.07	1.00	0.92	1.00	0.12	1.00
Occlusion	occ_mean	0.030		0.185	0.169	0.139	0.020	0.445	0	0.818
	occ_p90	0.030		0.32	0.29	0.24	0.04	0.74	0	1.54
	occ_heavy	0.030		0.20	0.26	0	0	0.57	0	1.00
Depth quality	depth_invalid	0.030	10^{-2}	2.37	1.74	2.08	0.46	4.77	0.03	9.23
	depth_invalid_mask	0.030		0.018	0.022	0.007	<0.001	0.051	<0.001	0.108
	depth_std	0.030		2.14	1.80	1.59	0.45	4.86	0.20	8.96
	depth_std_mask	0.030		0.18	0.10	0.17	0.09	0.24	0.06	1.04
Photometric conflict	specular	0.030	10^{-2}	0.34	0.45	0.22	<0.01	0.87	0	4.42
	dark	0.030	10^{-3}	0.70	1.14	0.18	0.01	2.41	0	7.10
Temporal stability	area_cv	0.030		0.330	0.082	0.329	0.227	0.446	0.163	0.542
	area_drop	0.030	10^{-2}	0.24	0.41	0.06	0	0.63	0	2.80
	vis_instability	0.030	10^{-2}	0.24	0.54	0	0	0.58	0	4.51
Camera motion	trans_mean	0.030		0.023	0.011	0.022	0.015	0.029	0.005	0.148
	trans_p90	0.030		0.032	0.021	0.030	0.022	0.039	0.008	0.313
	rot_mean	0.030	10^{-2}	2.36	0.57	2.54	1.57	2.96	0.67	3.95
	rot_p90	0.030	10^{-2}	3.49	0.75	3.61	2.41	4.35	1.29	6.06
	jerk	0.030	10^{-2}	0.55	0.50	0.50	0.31	0.71	0.15	6.58
Fisheye/stereo	fisheye_dark	0.030		0.342	0.037	0.328	0.304	0.402	0.290	0.462
	fisheye_bright	0.030		0.062	0.024	0.059	0.029	0.093	0.008	0.107
	fisheye_sharpness	0.030		399	209	329	205	724	146	1046
	fisheye_corr	0.030		0.870	0.031	0.868	0.833	0.908	0.770	0.933
	fisheye_texture	0.030		4.75	1.24	4.52	3.44	6.33	3.04	8.22

27 descriptors = 2 effort + 25 non-effort; $\sum_i w_i = 1$

$n = 300$ scene-phase cells

Sharpness uses the variance of the Laplacian, and texture density uses the mean gradient magnitude:

$$\begin{aligned} \text{fisheye_sharpness} &= \frac{1}{T} \sum_t \text{Var}(\nabla^2 F_t^L), \\ \text{fisheye_texture} &= \frac{1}{T} \sum_t \text{mean}_{u,v} \|\nabla F_t^L(u,v)\|_2. \end{aligned} \quad (30)$$

$$(31)$$

The left-right stereo correlation is

$$\text{fisheye_corr} = \frac{1}{T} \sum_t \rho(\text{vec}(F_t^L), \text{vec}(F_t^R)), \quad (32)$$

where ρ is the Pearson correlation over all pixel values.

B.4 Feature matrix

Figure 12 shows which descriptor groups separate the tiers.

B.5 Score construction

The descriptors are combined into the RPX difficulty score (RPX-DS). Each raw descriptor is percentile-ranked across nonmissing scene-phase cells; ties receive their average rank, constant columns map to 0.5, and unavailable optional modalities receive the neutral midpoint. With effort set E and $\alpha = 0.25$:

$$S_c = \sum_i w_i \tilde{F}_{c,i}, \quad w_i = \begin{cases} \alpha/2, & i \in E, \\ (1 - \alpha)/25, & i \notin E, \end{cases} \quad (33)$$

so each effort feature has weight 0.125 and each non-effort feature 0.03. The three exocentric phase scores are averaged per scene.

B.6 Score distributions

The 33% and 67% quantiles are 0.444 and 0.559 for the 300 cells and 0.462 and 0.531 for the 100 scenes.

Figures 13a and 13b show the cell and scene distributions with their cut points, and Fig. 13c shows the per-scene scores in each tier.

B.7 Difficulty-tier membership

Table XVI lists the scene membership of the Easy, Medium, and Hard tiers used throughout RPX, split by environment; the same labels index the scene catalogue of Appendix A.5.

C. EVALUATION MECHANICS

C.1 Task metric definitions

Table XVII lists every metric in the MANOVA K -vectors. Full definitions follow, grouped by task.

1) T1-T2: depth ($K=4$ image; $K=6$ video)

Let $\hat{D}_t, D_t$ be predicted and ground-truth depth at frame t , evaluated over valid pixels $V_t = \{(u,v): D_t(u,v) > 0\}$ with

What sets the tiers apart?

Mostly annotation effort, then objects changing in view and occlusion

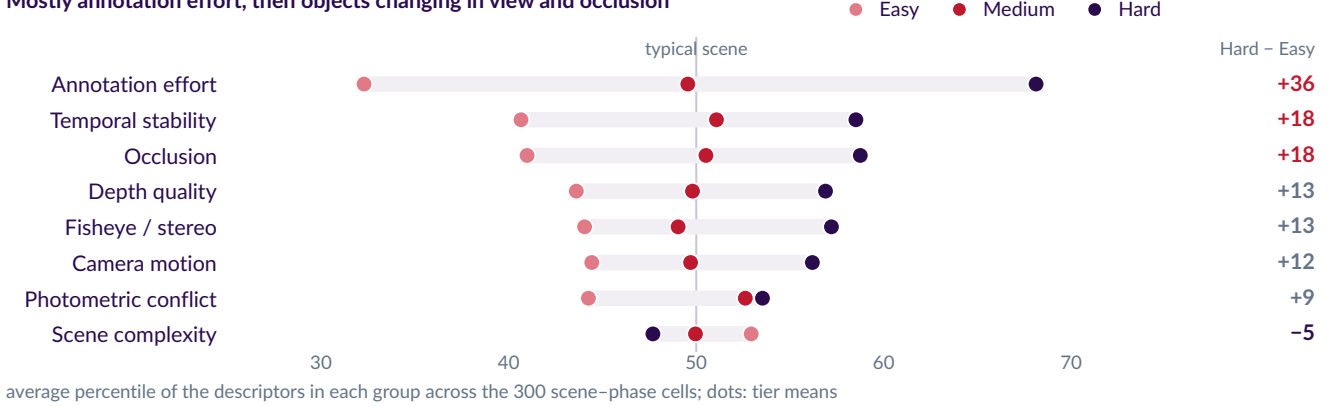
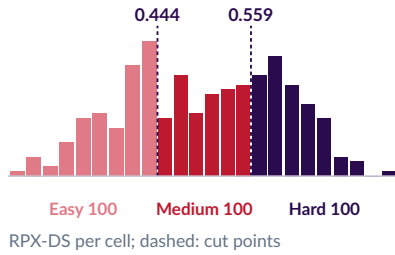


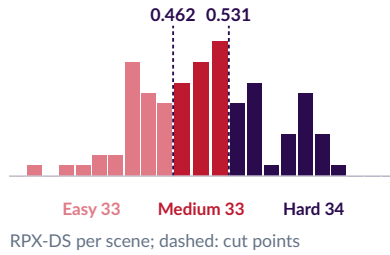
Fig. 12: What separates the difficulty tiers. Each descriptor is percentile-ranked over the 300 scene-phase cells, as in RPX-DS, averaged over a scene’s phases and over the descriptors of each group; dots are the Easy, Medium, and Hard means and the right column is the Hard minus Easy gap in percentile points. Annotation effort separates the tiers most; Hard scenes show fewer visible objects, so scene complexity runs slightly the other way.

How are scene-phase cells split?
Into equal thirds: 100, 100, 100



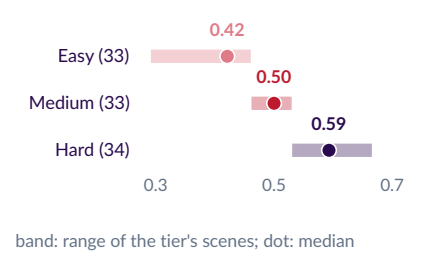
(a) Scene-phase cells ($n=300$)

How are scenes split?
Into equal thirds: 33, 33, 34



(b) Phase-averaged scenes ($n=100$)

Do the tiers overlap?
No: each is its own score band



(c) Scenes by difficulty tier

Fig. 13: RPX-DS scores. (a, b) Score histograms with the two cut points that split cells and scenes into equal thirds. (c) Score range (band) and median (dot) of the scenes in each tier; the tiers do not overlap.

TABLE XVI: Scene membership of the Easy, Medium, and Hard tiers used throughout RPX, split by environment. Scene numbers are coloured by tier as in Table XII; the environment of each scene is that of its capture site (Fig. 8).

Tier	Env.	n	Scenes
Easy	Indoor	22	008, 011, 012, 013, 015, 022, 023, 024, 031, 032, 034, 035, 037, 038, 057, 059, 060, 066, 068, 071, 073, 075
	Outdoor	11	004, 026, 029, 030, 051, 053, 054, 055, 092, 093, 094
Medium	Indoor	23	006, 007, 009, 014, 017, 021, 025, 036, 056, 058, 064, 065, 069, 070, 072, 074, 076, 077, 078, 079, 086, 089, 090
	Outdoor	10	028, 047, 048, 050, 052, 084, 091, 097, 099, 100
Hard	Indoor	15	010, 016, 018, 019, 020, 033, 039, 040, 061, 062, 063, 067, 080, 087, 088
	Outdoor	19	001, 002, 003, 005, 027, 041, 042, 043, 044, 045, 046, 049, 081, 082, 083, 085, 095, 096, 098

$n_t = |V_t|$. Write $e_i = \log \hat{D}_i - \log D_i$ and $\Delta D_t = D_{t+1} - D_t$.

$$\text{AbsRel} = \frac{1}{n_t} \sum_{V_t} |\hat{D} - D| / D, \quad (34)$$

TABLE XVII: Metrics entering each task’s MANOVA K -vector.

Task	K	Metrics	Eq.
T1 Image depth	4	AbsRel $\downarrow$, RMSE $\downarrow$, SILog $\downarrow$, $\delta_1\uparrow$	34–37
T2 Video depth	6	T1 metrics + TGM $\downarrow$, TGSE $\downarrow$	38–39
T3 Tracking	3	HOTA $\uparrow$, MOTA $\uparrow$, IDSW $\downarrow$	40–42
T4 Rel. pose	3	AUC@5 $^\circ\uparrow$, @10 $^\circ\uparrow$, @20 $^\circ\uparrow$	44
T5–T6 Grounding	3	Acc@0.5 $\uparrow$, GIoU $\uparrow$, CiGT $\uparrow$	45–47

$$\text{RMSE} = \left(\frac{1}{n_t} \sum_{V_t} (\hat{D} - D)^2 \right)^{1/2}, \quad (35)$$

$$\text{SILog} = \left(\frac{1}{n_t} \sum e_i^2 - \left(\frac{1}{n_t} \sum e_i \right)^2 \right)^{1/2}, \quad (36)$$

$$\delta_1 = \frac{1}{n_t} \left| \left\{ V_t : \max(\hat{D}/D, D/\hat{D}) < 1.25 \right\} \right|. \quad (37)$$

All four are computed per frame and averaged over a cell. Video depth (T2) adds two temporal metrics on consecutive-frame depth changes:

$$\text{TGM} = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{1}{n_t} \sum_{V_t} |\Delta \hat{D}_t - \Delta D_t|, \quad (38)$$

$$\text{TGSE} = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{1}{n_t} \sum_{V_t} (\Delta \hat{D}_t - \Delta D_t)^2. \quad (39)$$

Alignment. Relative-depth models in both T1 and T2 receive one least-squares affine fit per scene-phase cell, pooled

over all of its frames; per-frame fitting is not used. Metric-depth models are evaluated without fitting. All evaluation is restricted to (0.3, 5.0) m; invalid D435 readings are excluded.

2) T3: tracking ($K=3$)

Let TP_t , FP_t , FN_t be detection counts at frame t , GT_t the ground-truth count, and $IDSW_t$ the identity-switch count.

$$HOTA = \frac{1}{|\mathcal{A}|} \sum_{\alpha \in \mathcal{A}} \sqrt{\text{Det}A_\alpha \cdot \text{Ass}A_\alpha}, \quad (40)$$

$$MOTA = 1 - \sum_t (FP_t + FN_t + IDSW_t) / \sum_t GT_t, \quad (41)$$

$$IDSW = \sum_{t=1}^T IDSW_t, \quad (42)$$

where, at localization threshold α , $\text{Det}A_\alpha = |TP_\alpha| / (|TP_\alpha| + |FP_\alpha| + |FN_\alpha|)$ and $\text{Ass}A_\alpha$ measures long-range identity consistency of the matched detections. HOTA, DetA, and AssA are averaged over $\mathcal{A} = \{0.05, 0.10, \dots, 0.95\}$; MOTA and IDSW use the MOTChallenge IoU threshold of 0.5, and IDSW is the total number of identity switches in the clip. IDF1 is excluded because $|r(\text{MOTA}, \text{IDF1})| > 0.99$, which would inflate the condition number of $\mathbf{E}$.

3) T4: relative camera pose ($K=3$)

For a predicted relative pose $(\hat{R}, \hat{t})$ and ground truth (R, t) ,

$$e_R = \arccos \frac{\text{tr}(\hat{R}^T R) - 1}{2}, \quad e_t = \arccos \frac{|\hat{t}^T t|}{\|\hat{t}\| \|t\|}. \quad (43)$$

The combined error $e = \max(e_R, e_t)$ yields

$$R(x) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[e_i \leq x], \quad \text{AUC@}\tau = \frac{1}{\tau} \int_0^\tau R(x) dx, \quad (44)$$

with $\tau \in \{5^\circ, 10^\circ, 20^\circ\}$. R is the empirical cumulative distribution of the pose error, and $\text{AUC@}\tau$ is its normalized area up to τ . T4 evaluates Clutter and Clean only; Interaction is excluded because objects move during manipulation and pairwise relative pose assumes a static scene.

4) T5–T6: visual grounding ($K=3$)

Each model returns a box $\hat{B}$; ground truth B is the tight axis-aligned enclosure of the instance mask. Let C be the smallest box enclosing $\hat{B} \cup B$.

$$\text{Acc@0.5} = \mathbf{1}[\text{IoU}(\hat{B}, B) \geq 0.5], \quad (45)$$

$$\text{GIoU} = \text{IoU}(\hat{B}, B) - |C \setminus (\hat{B} \cup B)| / |C|, \quad (46)$$

$$\text{CiGT} = \mathbf{1}[\text{center}(\hat{B}) \in B]. \quad (47)$$

GIoU lies in $[-1, 1]$ and, unlike IoU, keeps decreasing as a non-overlapping box moves away from the target. CiGT credits a box whose centre lies in the target even when its extent is wrong. A missing, unparseable, or invalid box scores 0 on Acc@0.5 and CiGT and -1 on GIoU.

C.2 VQA question design

Questions are generated from the colour, material, and function responses of the SOS questionnaire (Table IX) and the instance masks; name and category are not used. The canonical catalogue (Table XIII) shows one normalized response per field; the release keeps all original responses and aliases. Both tasks are scored with the grounding metrics of Appendix C.1.4.

1) Regular questions (T5)

The question plan contains eight T5 types, five attribute and three spatial; the published T5 scores ($\mathcal{J}$, Φ) use the five attribute types. Three attribute types refer to the target by a single colour, material, or function (e.g. “the red object”, “the metal object”, “the object used for cutting”). The remaining two combine or negate questionnaire attributes: a composition type names both material and function (“the object made of *metal* that is used for *cutting*”), and an odd-one-out type asks for the object *not* made of a given material. The spatial types ask for the leftmost or rightmost object, the object closest in depth, and the farthest object by image-plane distance; they are asked of the exocentric views only and are released but not scored.

2) In-context questions (T6)

T6 replaces the textual subject reference with an SOS photograph followed by the target scene image. Its plan contains the five attribute types of T5 and the farthest-object spatial type; the published T6 scores use the five attribute types. The model must localize the referenced object in that scene.

C.3 MANOVA estimator choice

Φ is computed from Wilks’ Λ . Let $\lambda_1, \dots, \lambda_s$ be the nonzero eigenvalues of $\mathbf{H}\mathbf{E}^{-1}$, with $s = \min(K, P-1)$, and $\rho_i^2 = \lambda_i / (1 + \lambda_i)$ the squared canonical correlations. Wilks and Pillai yield:

$$\eta_W^2 = 1 - \text{GM}(1 - \rho_i^2), \quad \eta_P^2 = 1 - \text{AM}(1 - \rho_i^2). \quad (48)$$

Because $\text{GM} \leq \text{AM}$, these definitions give $\Phi_{\text{Wilks}} \leq \Phi_{\text{Pillai}}$. We report the Wilks-based Φ , the smaller of the two.

C.4 Model evaluation configurations

Table XVIII lists, for every model configuration, the input handling, initialization or prompt, and depth alignment used in evaluation, beyond what the main-paper result tables report.

C.5 Metric desirability windows

Each native metric v_k enters $\mathcal{J}$ through the desirability transform (Eq. (3)):

$$d_k = \text{clip}\left(\frac{v_k - v_k^{\text{worst}}}{v_k^{\text{best}} - v_k^{\text{worst}}}, 0, 1\right).$$

Windows are frozen and shared by all models: bounded metrics use their natural range, and unbounded errors use a fixed worse endpoint beyond which $d_k = 0$. A value outside its window is clipped rather than refitted, so adding a model never changes the scores of the others.

Table XIX lists the desirability window of every metric used in T1–T6.

TABLE XVIII: Evaluation configuration of every model–task evaluation, beyond what main-paper Tables III–VII list.

Task	Configurations	Input handling	Initialization / prompt	Alignment / special setting
T1	DA3 Metric-L, DA-V2 Large, Depth Pro, UniDepth V2, Metric3D V2, MoGe-2 ViT-L, HyDen	single RGB frame	N/A	metric output; no scale or shift fitting
T1	Lotus-2, FE2E, ZipDepth	single RGB frame	N/A	relative output; least-squares scale and shift to D435 depth
T2	DA3 (video), Video-DA, ViGeo, VGGT, GemDepth, DVD v1.1	phase clip in frame order	N/A	relative output; one scale and shift per clip
T2	DepthCrafter, ChronoDepth	phase clip; sliding windows	N/A	relative output; one scale and shift per clip
T2	RollingDepth	phase clip; streaming	N/A	relative output; one scale and shift per clip
T2	MonST3R	phase clip; 4-frame multi-view chunks	N/A	relative output; one scale and shift per clip
T3	SAM 2 Long, DAM4SAM, EdgeTAM, SAM 2++, SAM 2, Cutie, MiTS, XMem	RGB sequence	ground-truth masks of the first frame	persistent IDs scored one-to-one
T3	SAM 3.1 (bbox)	RGB sequence	ground-truth boxes of the first frame	persistent IDs scored one-to-one
T3	GSAM2 (text)	RGB sequence	one fixed prompt per object and clip: “(primary colour) (object name)”	persistent IDs scored one-to-one
T4	all 10 models of the T4 table in the main paper	RGB frame pair within one capture	N/A	Clutter and Clean pairs scored separately; five-frame stride; GT in OpenCV camera axes
T5	Qwen3-VL 8B / 2B	target image + text question	question text	JSON box on a 0–1000 coordinate grid; one scored generation; greedy decoding; 96-token limit
T5	Cosmos Reason2, InternVL 3.5, PaliGemma 2, DeepSeek-VL2, Florence 2 (all sizes)	target image + text question	deterministic referring expression from the question	native or architecture-specific grounding output; one scored generation; greedy decoding; 64–128-token limit
T6	same 12 configurations as T5	SOS reference image then target image	question referring to the reference	same grounding mode and decoding as T5
T5–T6	all VLMs	task-specific image input	task-specific question or referring expression	boxes converted to inclusive pixel XYYY in original image coordinates; missing or invalid box = miss

 TABLE XIX: Desirability windows of all native metrics. $d_k = 0$ at the worse endpoint and 1 at the better endpoint; range: natural bounds of the metric; fixed: frozen error cap shared by all models. Values beyond an endpoint are clipped: MOTA, which can be negative, at 0 (all observed values exceed 0.59), and SIlog at 25, which the phase means of three video-depth models exceed.

Metric	Task(s)	Worse	Better	Dir.	Basis
AbsRel	T1, T2	0.5	0	↓	fixed
RMSE (m)	T1, T2	2.0	0	↓	fixed
SIlog	T1, T2	25	0	↓	fixed
δ_1	T1, T2	0	1	↑	range
TGM	T2	0.1	0	↓	fixed
TGSE	T2	0.1	0	↓	fixed
HOTA	T3	0	1	↑	range
MOTA	T3	0	1	↑	range
IDSW (per clip)	T3	100	0	↓	fixed
AUC@5°	T4	0	1	↑	range
AUC@10°	T4	0	1	↑	range
AUC@20°	T4	0	1	↑	range
Acc@0.5	T5, T6	0	1	↑	range
GIoU	T5, T6	−1	1	↑	range
Center-in-GT	T5, T6	0	1	↑	range

D. BENCHMARKING

D.1 Findings

Eight findings from the 65 evaluations and 74 hardware profiles; each links to its detailed subsection.

F1. No model is both accurate and phase-robust. None of the 65 evaluations reaches the golden region $\mathcal{J}_{\min} \geq 0.75$, $\Phi \geq 0.86$. The closest model per task is 0.10 (VGGT, video depth) to 0.54 (Qwen3-VL 2B, in-context VQA) away from it (Sec. D.4).

F2. Quality and robustness rank models differently. Within tasks, Spearman $\rho(\mathcal{J}_{\min}, \Phi)$ ranges from -0.86 (T6) to $+0.37$ (T3), and the highest- $\mathcal{J}_{\min}$ model is also the most robust only in video depth. A single score would hide this trade-off.

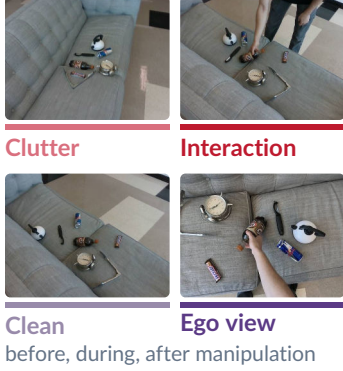
F3. Interaction drives phase sensitivity in geometric tasks. For 30 of 31 depth and tracking models, the transition with the lowest Φ involves Interaction, and Interaction is the lowest-quality phase for all 11 trackers (Sec. D.5).

F4. Harder scenes cost robustness only in geometric tasks. Hard-tier Φ is below Easy-tier Φ for 29 of 31 depth and tracking models but for only 7 of 24 VQA evaluations. Scene difficulty therefore stresses geometry and tracking more than grounding.

F5. A high Φ can mean uniformly weak predictions. Replacing text with a reference image lowers $\mathcal{J}_{\min}$ for all 12 VLMs but raises Φ for 8, and the most robust VQA configurations are Florence-2 models with $\mathcal{J}_{\min} \leq 0.14$. Φ must be read together with $\mathcal{J}_{\min}$.

F6. Reference images cut grounding quality by half or more. In-context $\mathcal{J}_{\min}$ is 49–93% lower than regular $\mathcal{J}_{\min}$.

1 Capture



2 Evaluate six tasks

T1 Image depth	AbsRel, RMSE, SLog, δ_1
T2 Video depth	+ temporal consistency
T3 Tracking	HOTA, MOTA, ID switches
T4 Relative pose	pose-error AUC
T5 Regular VQA	box accuracy, GloU
T6 In-context VQA	box accuracy, GloU

same scenes, task-specific metrics

3 Compare two scores

Worst-phase quality $\mathcal{J}_{\min}$
How good is the weakest phase?
higher is better

Phase robustness Φ
How stable is a model across phases?
higher is better

read together: high Φ can mean uniformly weak

Fig. 14: Shared evaluation framework, read left to right. The same scenes are captured before, during, and after manipulation and from an egocentric view; six tasks score them with their own metrics (Table XVII); every model is then summarized by worst-phase quality $\mathcal{J}_{\min}$ and phase robustness Φ . Higher is better for both, and the two are read together. Ego views are scored for T3 and T5–T6 and reported separately.

TABLE XX: Task-specific evaluation protocols. Metrics are listed in Table XVII.

Task	Input / ground truth	Key protocol detail
T1 Image depth	RGB / aligned D435 depth	Valid range (0.3, 5.0) m; frozen alignment for relative depth.
T2 Video depth	RGB clip / ordered D435 depth	Preserve frame order and IDs; one relative alignment per clip.
T3 Tracking	RGB sequence + initialization / masks and persistent IDs	Declare initialization; persistent IDs remain one-to-one.
T4 Relative pose	RGB pair / T265 relative pose	Clutter/Clean only; validate the pose convention.
T5 Regular VQA	RGB + attribute question / mask-derived box	Five attribute question types scored; missing or invalid boxes are misses.
T6 In-context VQA	Ordered SOS reference + query / mask-derived box	Ordered SOS reference; check global IDs and leakage.

while model rankings are largely preserved (Spearman $\rho = 0.87$; Sec. D.9).

F7. The egocentric camera changes the metric profile, not tracking quality. Ego and exo tracking quality differ by a median of 0.00, yet $\Phi_{EE} \leq 0.86$ for 7 of 11 trackers; grounding improves on the ego view for all 12 regular-VQA models (Sec. D.8).

F8. Near-best quality is often far cheaper. UniDepth V2 is $22.7\times$ cheaper than Metric3D V2 for 0.02 lower $\mathcal{J}_{\min}$, SAM 2 is $4.7\times$ cheaper than SAM 2 Long for 0.01, and Qwen3-VL 2B is $2.2\times$ cheaper than 8B in-context for 0.01. Compute does not predict quality: in image depth, parameter count and $\mathcal{J}_{\min}$ are uncorrelated ($\rho = +0.02$; Sec. D.10).

D.2 Shared evaluation framework

All six tasks share one pipeline (Fig. 14). A versioned split fixes sample identity and available modalities, and a task adapter invokes a reference model or any user-supplied local or API callable. Predictions pass task-specific checks of shape, units, coordinates, completeness, and identifiers before scoring (Table XX); native metrics follow the definitions of Appendix C. Invalid shapes, missing predictions, and inference failures stay in the denominator; every rerun records dataset, model, and evaluator versions with stable sample identifiers.

D.3 Diagnostics used in this section

Let $\mathcal{J}_{s,p}$ be the desirability of scene s in phase p (Eq. (3)) and $\mathcal{J}_p$ its scene average, so that $\mathcal{J}_{\min} = \min_p \mathcal{J}_p$. $\Phi = 1 - \eta_p^2$ is the Wilks-based phase robustness over the scored phases (Appendix C). Restricting the same MANOVA gives the variants used below:

$$\Phi_{p \leftrightarrow q} = 1 - \eta_p^2|_{\{p,q\}}, \quad (49)$$

$$\Phi_g = 1 - \eta_p^2|_{s \in \mathcal{S}_g}, \quad g \in \{E, M, H\}, \quad (50)$$

$$\Phi_{EE} = 1 - \eta_v^2|_{v \in \{\text{ego}, \text{exo-Int}\}}, \quad (51)$$

with $\mathcal{S}_g$ the scenes of tier g and η_v^2 the camera-view effect size. An evaluation a dominates b when $\mathcal{J}_{\min}^a \geq \mathcal{J}_{\min}^b$ and $\Phi^a \geq \Phi^b$ with one inequality strict. The distance to the golden region and the relative in-context drop are

$$G = \sqrt{\max(0, 0.75 - \mathcal{J}_{\min})^2 + \max(0, 0.86 - \Phi)^2}, \quad (52)$$

$$\Delta_{T6} = 100 (\mathcal{J}_{\min}^{T5} - \mathcal{J}_{\min}^{T6}) / \mathcal{J}_{\min}^{T5}. \quad (53)$$

Associations are Spearman rank correlations ρ across the models of one task.

D.4 Quality–robustness trade-offs

Main-paper Fig. 4 places all 65 evaluations against the two thresholds of the golden region; none meets both. Ten of the 11 trackers pass the quality threshold (SAM 3.1 with text

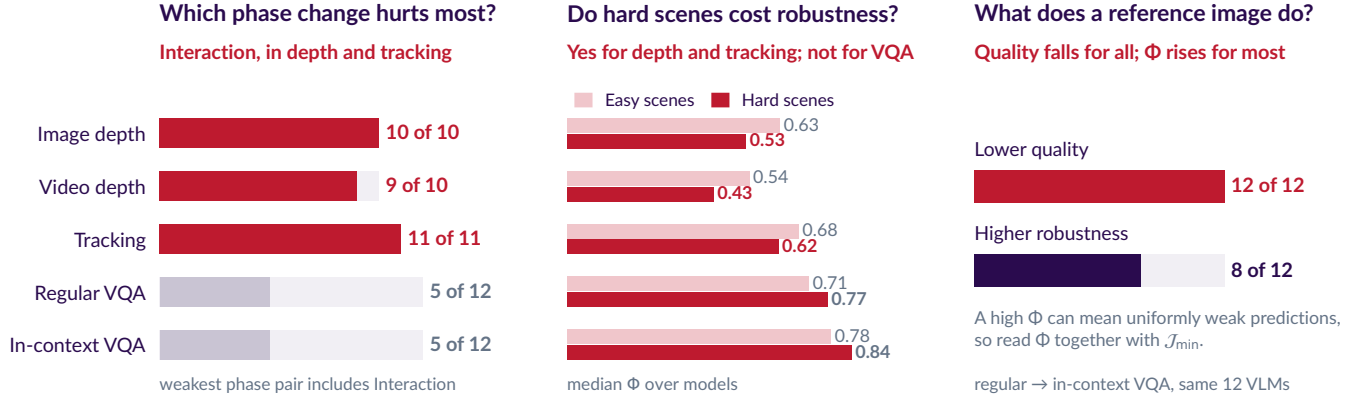


Fig. 15: Where phase robustness is lost. (a) Models whose weakest phase transition, the lowest pairwise Φ (Eq. (49)), involves Interaction. (b) Median tier robustness (Eq. (50)) on Easy and Hard scenes. (c) Regular to in-context VQA for the same 12 VLMs: all lose $\mathcal{J}_{\min}$ and 8 gain Φ . T4 is omitted (two phases).

prompts falls short at 0.72) but none passes robustness; all 10 relative-pose models and 3 regular and 5 in-context VQA configurations pass robustness but none passes quality; no depth model passes either. The closest evaluations are VGGT and GSAM2 ($G = 0.10, 0.11$; Eq. (52)), followed by SAM 2 and SAM 2++ (0.13). Within tasks the two diagnostics are weakly or negatively associated ($\rho(\mathcal{J}_{\min}, \Phi) = -0.43, -0.25, +0.37, -0.23, -0.59, -0.86$).

D.5 Where robustness is lost

Transitions. Φ follows each model’s weakest transition (Eq. (49); $\rho = 0.95, 0.81, 0.97, 0.96, 0.97$ in T1, T2, T3, T5, T6). In depth and tracking that transition involves Interaction for 30 of 31 models (Fig. 15a). In grounding it is Clutter $\leftrightarrow$ Clean for 7 of 12 models in both tasks, although Interaction is the lowest-quality phase for 8 of 12 T5 and 5 of 12 T6 models (9 and 12 counting ties).

Scene difficulty. Tier robustness (Eq. (50)) falls from Easy to Hard in the geometric tasks: $\Phi_H < \Phi_E$ for 10/10, 9/10, and 10/11 models in T1–T3, with medians falling from 0.63 to 0.53, 0.54 to 0.43, and 0.68 to 0.62 (Fig. 15b). Grounding shows no such decline ($\Phi_H < \Phi_E$ for 3/12 and 4/12 models), and better video-depth models lose less robustness from Easy to Hard ($\rho(\mathcal{J}_{\min}, \Phi_E - \Phi_H) = -0.92$).

Low-quality floor. A reference image lowers $\mathcal{J}_{\min}$ for all 12 VLMs yet raises Φ for 8 (Fig. 15c), and the highest- Φ configuration in each VQA task is a Florence-2 model with $\mathcal{J}_{\min} \leq 0.14$. Where both diagnostics vary, they agree: the gap $\mathcal{J} - \mathcal{J}_{\min}$ shrinks as Φ rises in T2 and T6 ($\rho = -0.82, -0.83$).

D.6 Detailed per-model metrics

These phase aggregates summarize scene-level measurements; applying the geometric mean to a phase-averaged metric vector does not reproduce the scene-averaged quality score.

D.7 Per-phase quality

Quality during Interaction is below the mean of Clutter and Clean for 5/10, 6/10, 11/11, 10/12, and 11/12 models in T1, T2, T3, T5, and T6 (Fig. 16). The dip is most consistent in tracking (median -0.06) and largest for individual video-depth models (up to -0.125 , GemDepth).

D.8 Ego vs. exo comparison

Tracking quality is unchanged by the head-mounted camera (ego exceeds exo for 5 of 11 trackers; median difference 0.00), whereas grounding improves on the GoPro view for all 12 regular-VQA models (median $+0.09$) and 8 of 12 in-context models (Fig. 17). Even so, $\Phi_{EE} \leq 0.86$ (Eq. (51)) for 7 of 11 trackers and 22 of 24 VLM evaluations: the camera reshapes the metric profile even when quality rises. Φ_{EE} is unrelated to phase robustness ($\rho(\Phi, \Phi_{EE}) = +0.19, -0.06, +0.32$), so camera and phase sensitivity are separate axes. Viewpoint, resolution ($6.75\times$ pixel count), and framing differ between streams, so no single factor is isolated.

D.9 Text vs. image prompt

All 12 VLM configurations score lower on in-context than on regular VQA, with relative drops Δ_{T6} (Eq. (53)) of 49–93% (Fig. 18). Rankings are largely kept (Pearson $r = 0.69$, Spearman $\rho = 0.87$), and the drop is unrelated to regular-VQA quality ($\rho = -0.40, p = 0.19$). Qwen3-VL 8B and 2B remain strongest ($\mathcal{J}_{\min} = 0.23, 0.22$); InternVL 3.5 1B and both PaliGemma 2 models fall to 0.03.

D.10 Hardware efficiency and deployment trade-offs

Profiling protocol. The release contains 74 hardware profiles: 64 of the 65 evaluated configurations (SAM 3.1 with text prompts was not profiled) and 10 egocentric tracking profiles. All profiles use NVIDIA RTX 6000 Ada Generation GPUs (48 GB each) and exclude one warmup call or query. Vision profiles use the Easy split; published quality summaries cover the full benchmark. T1 and T4 time 1,000 adapter calls, T5 and T6 1,000 queries per model, T2 1,000 frames in 20 calls of 50 frames, and T3 1,000 MOS frames in four

Depth. Tables [XXI](#) and [XXII](#) report image-depth and video-depth metrics, respectively.

TABLE XXI: T1 image depth: native metrics per phase. **Bold**: best; **shaded**: each model’s worst phase.

Model	AbsRel ↓			RMSE ↓			SILog ↓			δ_1 ↑		
	Clu.	Int.	Clean	Clu.	Int.	Clean	Clu.	Int.	Clean	Clu.	Int.	Clean
ZipDepth	0.167	0.161	0.158	0.454	0.506	0.483	15.51	18.10	15.27	0.732	0.761	0.756
HyDen	0.267	0.248	0.256	0.625	0.602	0.657	12.27	13.58	12.07	0.329	0.364	0.348
FE2E	0.142	0.140	0.145	0.382	0.447	0.439	11.77	14.63	12.53	0.815	0.833	0.807
DA3 Metric-L	0.126	0.117	0.118	0.323	0.387	0.362	10.71	13.67	11.05	0.849	0.909	0.870
UniDepth V2	0.142	0.126	0.129	0.354	0.393	0.387	11.07	13.44	10.69	0.800	0.885	0.830
MoGe-2 ViT-L	0.253	0.211	0.227	0.539	0.516	0.553	11.18	13.53	11.37	0.380	0.542	0.479
Lotus-2	0.221	0.201	0.218	0.450	0.492	0.491	23.82	22.67	24.50	0.630	0.676	0.636
Depth Pro	0.170	0.139	0.165	0.446	0.460	0.492	12.98	14.86	12.86	0.711	0.858	0.736
Metric3D V2	0.124	0.109	0.123	0.361	0.384	0.406	11.41	12.95	11.39	0.845	0.908	0.837
DA-V2 Large	0.210	0.216	0.158	0.392	0.510	0.390	12.54	15.35	12.85	0.704	0.696	0.804
Times worst	8	1	1	0	6	4	0	9	1	7	1	2

TABLE XXII: T2 video depth: native metrics per phase. **Bold**: best; **shaded**: each model’s worst phase.

Model	AbsRel ↓			RMSE ↓			SILog ↓			δ_1 ↑			TGM ↓			TGSE ↓		
	Clu.	Int.	Clean	Clu.	Int.	Clean	Clu.	Int.	Clean	Clu.	Int.	Clean	Clu.	Int.	Clean	Clu.	Int.	Clean
DA3 (video)	0.060	0.079	0.058	0.209	0.316	0.235	9.03	12.45	9.32	0.969	0.951	0.969	0.0210	0.0284	0.0233	0.0312	0.0524	0.0398
Video-DA	0.207	0.182	0.203	0.413	0.457	0.454	24.75	23.73	25.74	0.680	0.740	0.692	0.0245	0.0308	0.0285	0.0333	0.0501	0.0440
ViGeo	0.090	0.073	0.062	0.297	0.308	0.272	14.21	12.42	11.55	0.908	0.959	0.957	0.0197	0.0271	0.0224	0.0491	0.0562	0.0509
VGGT	0.067	0.062	0.057	0.259	0.272	0.255	10.93	11.05	10.02	0.955	0.963	0.963	0.0231	0.0226	0.0226	0.0439	0.0415	0.0452
MonST3R	0.191	0.169	0.179	0.413	0.459	0.446	14.44	16.98	14.26	0.693	0.769	0.722	0.0369	0.0433	0.0409	0.0355	0.0570	0.0464
GemDepth	0.098	0.096	0.088	0.289	0.365	0.311	10.26	13.41	10.70	0.917	0.927	0.935	0.0177	0.0273	0.0217	0.0319	0.0558	0.0427
DepthCrafter	0.226	0.217	0.227	0.475	0.538	0.533	25.03	24.98	26.80	0.625	0.642	0.616	0.0283	0.0337	0.0346	0.0363	0.0525	0.0483
DVD v1.1	0.101	0.098	0.094	0.337	0.390	0.351	12.58	14.48	12.31	0.892	0.914	0.912	0.0184	0.0308	0.0241	0.0365	0.0631	0.0507
ChronoDepth	0.187	0.177	0.185	0.441	0.499	0.495	16.48	18.18	16.87	0.709	0.752	0.706	0.0205	0.0285	0.0241	0.0310	0.0495	0.0398
RollingDepth	0.267	0.216	0.240	0.510	0.519	0.529	29.21	27.20	29.00	0.556	0.644	0.598	0.0253	0.0324	0.0313	0.0316	0.0446	0.0421
Times worst	8	1	1	0	9	1	2	6	2	7	1	2	1	8	1	0	9	1

Tracking and pose. Tables [XXIII](#) and [XXIV](#) report tracking across four views and relative pose on the Clutter and Clean pair sets.

TABLE XXIII: T3 multi-object tracking: native metrics per view and phase. **Bold**: best; **shaded**: each model’s worst phase or view.

Model	HOTA ↑				MOTA ↑				IDSW ↓			
	Clu.	Int.	Clean	Ego	Clu.	Int.	Clean	Ego	Clu.	Int.	Clean	Ego
SAM 2 Long	0.973	0.930	0.975	0.952	0.983	0.938	0.991	0.953	0.65	1.10	0.21	1.73
DAM4SAM	0.968	0.923	0.974	0.942	0.973	0.922	0.987	0.938	0.77	1.93	0.21	2.70
EdgeTAM	0.949	0.888	0.955	0.884	0.967	0.889	0.979	0.877	0.61	1.23	0.23	1.84
SAM 2++	0.954	0.890	0.957	0.813	0.973	0.881	0.979	0.785	0.42	1.19	0.24	0.75
SAM 2	0.967	0.921	0.970	0.904	0.977	0.923	0.985	0.888	0.60	1.28	0.21	2.00
Cutie	0.944	0.891	0.955	0.904	0.969	0.906	0.988	0.911	0.58	1.11	0.19	1.54
SAM 3.1 (bbox)	0.956	0.838	0.953	0.886	0.971	0.816	0.973	0.875	0.62	1.62	0.31	1.96
MiTTS	0.909	0.821	0.931	0.827	0.941	0.802	0.959	0.785	0.62	2.15	0.19	3.73
XMem	0.927	0.845	0.945	0.868	0.958	0.828	0.978	0.861	0.33	1.27	0.19	1.49
SAM 3.1 (text)	0.814	0.688	0.803	0.742	0.741	0.590	0.731	0.676	0.80	1.77	0.93	4.11
GSAM2 (text)	0.865	0.776	0.884	0.843	0.777	0.643	0.819	0.757	0.52	0.63	0.22	1.50
Times worst	0	8	0	3	0	7	0	4	0	1	0	10

TABLE XXIV: T4 relative camera pose: native metrics for the Clutter and Clean pair sets. **Bold**: best; **shaded**: each model’s worst pair set.

Model	AUC@5° ↑		AUC@10° ↑		AUC@20° ↑	
	Clu.	Clean	Clu.	Clean	Clu.	Clean
DA3	0.205	0.263	0.454	0.518	0.685	0.729
VGGT	0.175	0.224	0.404	0.473	0.636	0.694
Reloc3R	0.198	0.242	0.442	0.495	0.672	0.710
Pi3X	0.181	0.228	0.411	0.476	0.645	0.696
MonST3R	0.031	0.045	0.092	0.130	0.222	0.290
MUST3R	0.109	0.143	0.297	0.347	0.536	0.586
Fast3R	0.115	0.136	0.318	0.352	0.574	0.599
CUT3R	0.125	0.168	0.334	0.397	0.582	0.633
MASt3R	0.165	0.202	0.377	0.431	0.594	0.646
DUST3R	0.068	0.079	0.200	0.222	0.410	0.438
Times worst	10	0	10	0	10	0

Grounding. Tables XXV and XXVI report regular and in-context grounding for the same 12 VLM configurations.

TABLE XXV: T5 regular VQA: native grounding metrics per view and phase (Center: Center-in-GT), pooled over all eight planned question types, including the unscored spatial types, so they are not the inputs of the published $\mathcal{J}$, which uses the five attribute types. **Bold:** best; **shaded:** each model’s worst phase or view.

Model	Acc@0.5 ↑				GIoU ↑				Center ↑			
	Clu.	Int.	Clean	Ego	Clu.	Int.	Clean	Ego	Clu.	Int.	Clean	Ego
Cosmos Reason2 8B	0.489	0.425	0.486	0.632	0.188	0.077	0.113	0.416	0.541	0.476	0.516	0.664
Cosmos Reason2 2B	0.203	0.173	0.172	0.356	−0.393	−0.504	−0.547	−0.181	0.254	0.211	0.197	0.388
Qwen3-VL 8B	0.480	0.383	0.477	0.659	0.273	0.103	0.167	0.450	0.524	0.423	0.499	0.681
Qwen3-VL 2B	0.297	0.288	0.308	0.594	−0.187	−0.225	−0.249	0.315	0.318	0.304	0.318	0.609
InternVL 3.5 14B	0.420	0.307	0.389	0.502	0.252	0.069	0.150	0.252	0.459	0.349	0.424	0.531
InternVL 3.5 1B	0.362	0.296	0.357	0.319	0.149	0.010	0.041	0.011	0.433	0.368	0.405	0.429
PaliGemma 2 10B	0.355	0.275	0.320	0.377	0.068	−0.124	−0.071	0.004	0.392	0.307	0.338	0.398
PaliGemma 2 3B	0.276	0.230	0.270	0.393	0.057	−0.038	−0.033	0.076	0.324	0.270	0.294	0.414
DeepSeek-VL2 4.5B	0.355	0.222	0.295	0.469	0.219	0.041	0.108	0.224	0.420	0.277	0.346	0.541
DeepSeek-VL2 1.0B	0.090	0.054	0.069	0.113	−0.053	−0.108	−0.090	−0.242	0.214	0.137	0.157	0.198
Florence 2 0.77B	0.138	0.129	0.139	0.125	−0.362	−0.371	−0.394	−0.435	0.168	0.157	0.165	0.151
Florence 2 0.23B	0.131	0.118	0.126	0.120	−0.314	−0.368	−0.380	−0.399	0.169	0.147	0.152	0.147
Times worst	0	10	1	1	0	7	2	3	0	10	1	2

TABLE XXVI: T6 in-context VQA: native grounding metrics per view and phase (Center: Center-in-GT), pooled over all planned question types, including the unscored spatial type, as in Table XXV. **Bold:** best; **shaded:** each model’s worst phase or view.

Model	Acc@0.5 ↑				GIoU ↑				Center ↑			
	Clu.	Int.	Clean	Ego	Clu.	Int.	Clean	Ego	Clu.	Int.	Clean	Ego
Cosmos Reason2 8B	0.082	0.060	0.068	0.275	−0.265	−0.418	−0.412	−0.127	0.268	0.206	0.208	0.326
Cosmos Reason2 2B	0.042	0.032	0.035	0.157	−0.473	−0.565	−0.601	−0.305	0.117	0.086	0.082	0.213
Qwen3-VL 8B	0.200	0.103	0.171	0.276	−0.112	−0.276	−0.302	−0.090	0.275	0.180	0.230	0.348
Qwen3-VL 2B	0.118	0.108	0.118	0.274	−0.210	−0.342	−0.396	−0.026	0.175	0.146	0.149	0.336
InternVL 3.5 14B	0.162	0.099	0.124	0.060	−0.073	−0.219	−0.240	−0.192	0.273	0.187	0.212	0.233
InternVL 3.5 1B	0.042	0.022	0.029	0.010	−0.240	−0.370	−0.402	−0.238	0.143	0.079	0.093	0.099
PaliGemma 2 10B	0.067	0.050	0.052	0.068	−0.485	−0.552	−0.582	−0.565	0.087	0.066	0.065	0.085
PaliGemma 2 3B	0.090	0.069	0.082	0.059	−0.472	−0.652	−0.589	−0.663	0.109	0.076	0.094	0.071
DeepSeek-VL2 4.5B	0.036	0.021	0.018	0.065	−0.532	−0.624	−0.640	−0.471	0.072	0.048	0.045	0.144
DeepSeek-VL2 1.0B	0.027	0.019	0.020	0.002	−0.172	−0.230	−0.248	−0.212	0.140	0.093	0.096	0.091
Florence 2 0.77B	0.038	0.026	0.030	0.020	−0.663	−0.746	−0.712	−0.795	0.062	0.034	0.040	0.027
Florence 2 0.23B	0.043	0.044	0.044	0.015	−0.684	−0.740	−0.737	−0.891	0.064	0.051	0.051	0.017
Times worst	0	5	1	6	0	1	8	3	0	5	3	4

How many models are worse during Interaction?

Nearly all trackers and VQA models; half of depth

How large is the drop?

Typically 0.01–0.05; up to 0.125 for one model

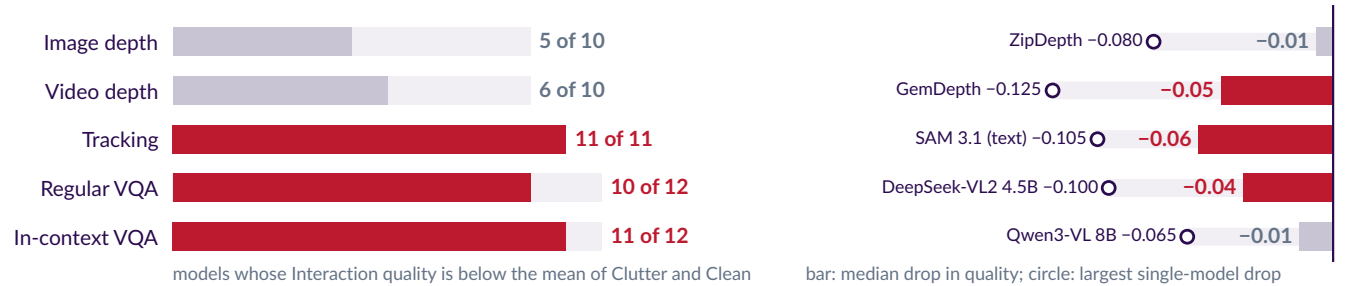


Fig. 16: Quality during Interaction. (a) Models whose Interaction $\mathcal{J}$ is below the mean of their Clutter and Clean $\mathcal{J}$. (b) Median drop $\mathcal{J}_{\text{Int}} - \frac{1}{2}(\mathcal{J}_{\text{Clu}} + \mathcal{J}_{\text{Cln}})$ per task (bar) and the largest single-model drop (circle, labelled). T4 excludes Interaction and is omitted.

calls and 1,213 egocentric frames in five calls. Timing covers synchronized adapter processing and excludes dataset I/O, weight loading, scoring, and persistence.

Cost metrics. For calls $k = 1, \dots, N_c$ with wall time t_k over n_k units (frame, pair, or query), the mean processing

cost, tail cost, throughput, and the ratio between matched configurations A and B are

$$\bar{c} = \frac{\sum_k t_k}{\sum_k n_k}, \quad c_{95} = P_{95}(t_k/n_k), \quad (54)$$

$$\text{Rate} = 1/\bar{c}, \quad r_{A/B} = \bar{c}_A/\bar{c}_B. \quad (55)$$

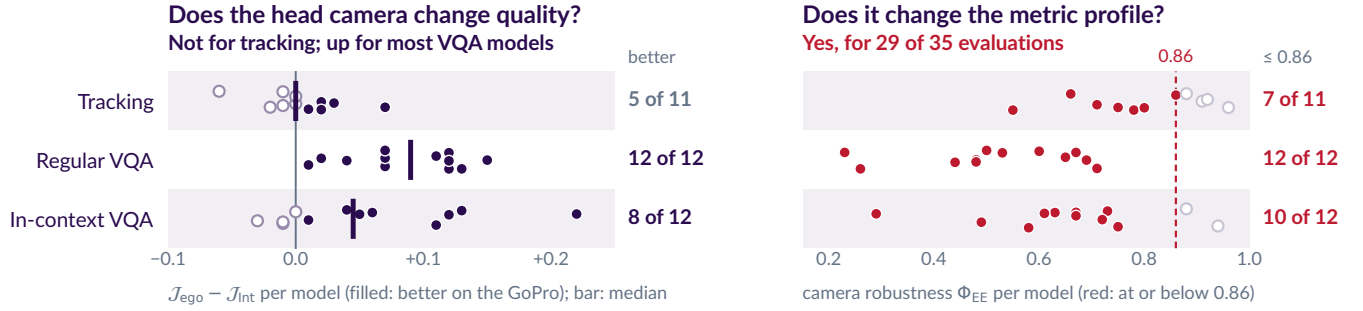


Fig. 17: Egocentric (GoPro, 1920×1080) versus exocentric (D435, 640×480) evaluation during Interaction for T3, T5, and T6; each dot is one model. (a) Quality change $J_{\text{ego}} - J_{\text{int}}$; filled dots are better on the egocentric view and the bar marks the median (0.00 for tracking, +0.09 for regular VQA). (b) Camera robustness Φ_{EE} (Eq. (51)); red dots are at or below 0.86.

How much does a reference image cost each VLM?

All 12 lose 49–93% of their quality; the order is largely kept (Spearman $\rho = 0.87$)

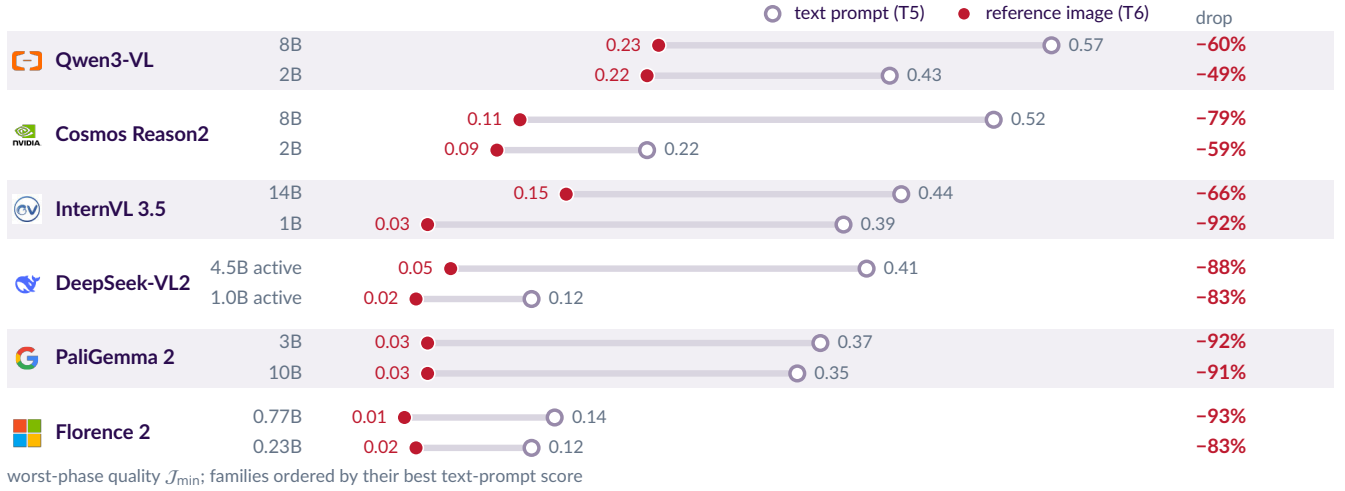


Fig. 18: What a reference image costs each VLM. Each row joins a configuration’s worst-phase quality J_{min} with a text description (T5, hollow) and with a reference image (T6, filled); the right column gives the relative drop Δ_{T6} (Eq. (53)). Rows are grouped by model family, ordered by the family’s best T5 quality; logos identify the developer, and DeepSeek sizes denote active parameters.

For T2 and T3, $n_k > 1$, so $\bar{c}$ is amortized per frame and quantiles describe clip calls. Memory is the peak CUDA reservation for T1–T4 and the process footprint summed over GPUs for T5–T6.

Requirements by task. Figure 19 summarizes all 64 profiles; the hardware tables list every measurement. Median cost is 83, 166, and 129 ms/frame for image depth, video depth, and tracking, 452 ms/pair for relative pose, and 501 and 592 ms/query for regular and in-context VQA. Median memory is 4.2, 15.8, 2.8, and 3.5 GiB for T1–T4 and about 38 GiB for VQA, whose largest model, DeepSeek-VL2, needs two GPUs (85.8 GiB in total). Within a task, cost spans $25\text{--}531\times$ between the fastest and slowest configuration, so model choice matters more than task choice.

Near-best quality is cheaper. In three tasks a near-best model costs a fraction of the best one: UniDepth V2 is $22.7\times$ cheaper than Metric3D V2 for 0.02 lower J_{min} , SAM 2 is $4.7\times$ cheaper than SAM 2 Long for 0.01, and Qwen3-VL 2B is $2.2\times$ cheaper than Qwen3-VL 8B in-context for 0.01. In

video depth, relative pose, and regular VQA the best model (VGGT, DA3, Qwen3-VL 8B) has no cheaper near-tie. The most robust configuration is cheaper than the highest- J_{min} one in T1, T3, T5, and T6, at lower quality.

No consistent cost–quality pattern. Spearman ρ between mean cost and J_{min} is $-0.10, -0.57, +0.48, -0.63, +0.62$, and $+0.41$ for T1–T6, and between mean cost and Φ it is $+0.26, +0.56, +0.24, +0.20, -0.72$, and -0.40 . Only grounding shows a clear pattern: larger, slower VLMs ground better but are less phase-robust (ρ of size with J_{min} : $+0.71$ and $+0.66$; with Φ : -0.56 and -0.46 for T5 and T6). In image depth, parameter count and GFLOPs are uncorrelated with J_{min} ($\rho = +0.02, -0.18$): Lotus-2 spends 283,499 GFLOPs for 0.35, DA3 Metric-L 902 GFLOPs for 0.70.

Protocol overhead and deployment. Egocentric tracking costs a median $1.55\times$ the MOS mean ($1.11\text{--}2.81\times$ across 10 trackers; Eq. (55)), and in-context VQA $1.17\times$ the regular mean across 12 VLMs; the median P95/mean ratio is 1.46 (T5) and 1.60 (T6). DA3 Metric-L (3.1 GiB, 39 ms), Cutie (0.8 GiB,

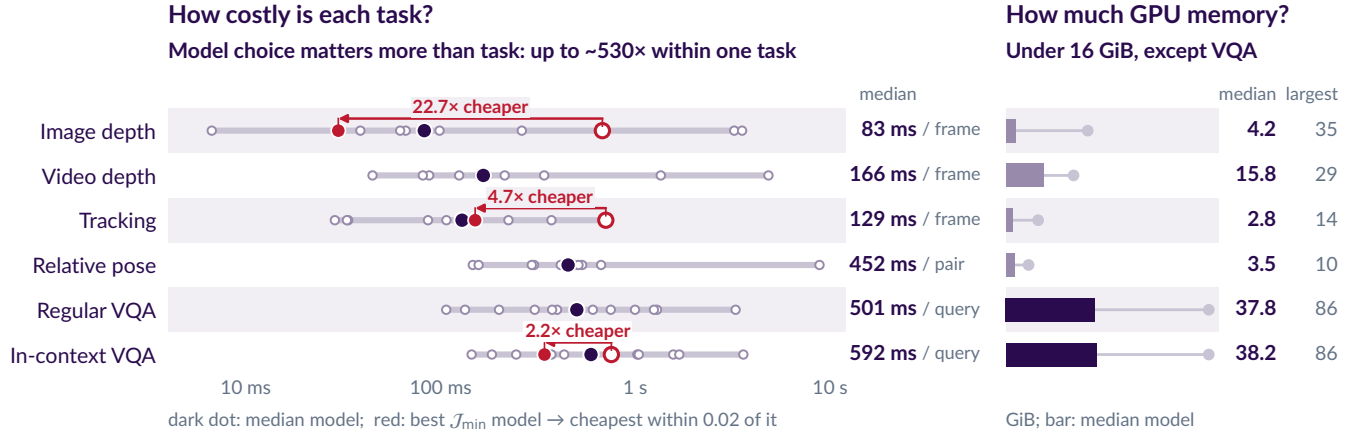


Fig. 19: Cost and memory per task, for the 64 profiled configurations. (a) Mean processing cost (Eq. (54)) of every configuration on a log scale; the dark dot is the median, also given on the right. Red arrows run from the highest- $\mathcal{J}_{\min}$ model (ring) to the cheapest model within 0.02 of it (filled), labelled with the cost ratio r (Eq. (55)); tasks without an arrow have no cheaper near-tie. (b) Median memory per task (bar) and the largest model (grey); memory is peak CUDA reservation for T1–T4 and summed process memory for T5–T6.

Depth profiles. Tables XXVII and XXVIII report image-depth and video-depth processing costs.

TABLE XXVII: T1 image-depth hardware measurements. Peak A/R is peak CUDA memory allocated/reserved; parameters are exact loaded-module counts. Bold: best per column.

Configuration	Protocol	N	Mean	P50	P95	P99	Rate	Peak A/R	Params	GFLOPs
				Latency (ms/unit)			(unit/s)	(GiB)	(M)	/unit
DA-V2 Large	Easy	1000	98.9	98.2	111.3	115.3	10.11	1.67/2.16	335.3	1806.1
DA3 Metric-L	Easy	1000	38.8	37.8	46.1	56.5	25.76	2.04/3.14	334.2	902.2
Depth Pro	Easy	1000	261.5	262.6	277.4	279.1	3.82	3.69/5.92	952.0	19247.4
FE2E	Easy	1000	3218.9	3222.6	3247.5	3262.2	0.31	26.52/31.85	12511.6	195375.4
HyDen	Easy	1000	66.3	66.3	71.3	74.3	15.08	1.81/2.41	374.4	1436.8
Lotus-2	Easy	1000	3532.6	3540.7	3560.6	3569.6	0.28	33.51/34.62	17587.8	283498.9
Metric3D V2	Easy	1000	677.7	684.2	690.9	699.3	1.48	6.32/6.92	1377.7	13469.8
MoGe-2 ViT-L	Easy	1000	62.1	61.7	65.1	68.0	16.10	2.86/4.12	330.9	3162.1
UniDepth V2	Easy	1000	29.9	29.3	33.9	34.8	33.43	2.58/4.26	353.8	1393.5
ZipDepth	Easy	1000	6.7	6.6	6.7	7.2	150.31	0.09/0.13	6.1	7.9

TABLE XXVIII: T2 video-depth hardware measurements, reported per processed frame. Peak A/R is peak CUDA memory allocated/reserved. Bold: best per column.

Configuration	Protocol	N	Mean	P50	P95	P99	Rate	Peak A/R	Params	GFLOPs
				Latency (ms/unit)			(unit/s)	(GiB)	(M)	/unit
ChronoDepth	Easy	1000	340.6	341.4	342.1	342.2	2.94	9.69/15.26	2254.4	18356.5
DA3 (video)	Easy	1000	44.8	43.7	46.9	70.0	22.33	7.83/16.41	410.9	2769.9
DepthCrafter	Easy	1000	1351.5	1353.0	1359.7	1360.7	0.74	11.05/17.74	2254.4	76924.3
DVD v1.1	Easy	1000	87.5	87.3	88.5	88.9	11.43	7.39/9.66	2245.8	6338.8
GemDepth [†]	Easy	1000	158.6	159.3	160.2	160.3	6.31	20.62/28.63	710.8	—
MonST3R	Easy	1000	213.6	210.3	229.4	242.0	4.68	2.57/2.79	571.2	1856.8
RollingDepth	Easy	1000	4822.7	4819.9	4853.5	4871.8	0.21	10.51/16.52	1290.0	495071.2
VGGT	Easy	1000	173.4	173.0	190.0	195.7	5.77	10.04/20.22	1256.5	7880.3
Video-DA	Easy	1000	81.2	79.5	92.4	92.5	12.32	8.43/9.33	384.4	3099.4
ViGeo	Easy	1000	124.7	124.7	126.1	127.0	8.02	10.61/14.81	1255.4	12448.7

34 ms), and XMem (1.4 GiB, 29 ms) fit an 8 GiB envelope at >10 fps with $\mathcal{J}_{\min} = 0.70, 0.93$, and 0.89 . Together, the highest- $\mathcal{J}_{\min}$ image-depth, tracking, and pose models (Metric3D V2, SAM 2 Long, DA3) reserve 19.6 GiB, within a desktop GPU but beyond most edge accelerators; this sum is not a measured concurrent run, and other hardware will differ.

E. TOOLKIT AND EXTENSIBILITY

The **R**PX toolkit is open-source and distributed on PyPI as `rpx-benchmark`¹; installation and usage are documented online². Reference adapters for the evaluated models are packaged as task-specific container images on Docker Hub (`irvlutd/rpx`), so each result can be rerun in its recorded environment.

¹<https://pypi.org/project/rpx-benchmark/>

²<https://irvlutd.github.io/RPX/toolkit-docs/>

Tracking and pose profiles. Tables XXIX and XXX report tracking by camera protocol and relative pose per image pair.

TABLE XXIX: T3 tracking hardware measurements for MOS and egocentric protocols, reported per processed frame. Bold: best per column within each protocol.

Configuration	Protocol	N	Mean	P50	P95	P99	Rate	Peak A/R	Params	GFLOPs
			Latency (ms/unit)				(unit/s)	(GiB)	(M)	/unit
Cutie	Ego	1213	73.4	70.2	78.9	78.9	13.63	0.81/2.03	35.0	582.3
Cutie	MOS	1000	33.6	34.4	35.2	35.2	29.79	0.57/0.82	35.0	420.2
DAM4SAM [†]	Ego	1213	563.0	563.8	573.7	573.7	1.78	3.44/3.87	—	16904.5
DAM4SAM [†]	MOS	1000	371.3	372.5	378.2	378.2	2.69	3.45/3.87	—	16932.4
EdgeTAM	Ego	1213	124.7	122.6	152.2	152.2	8.02	1.26/1.66	13.9	770.1
EdgeTAM	MOS	1000	86.3	90.6	112.2	112.2	11.58	1.21/1.46	13.9	770.9
GSAM2	Ego	1213	202.2	189.2	240.3	240.3	4.95	4.72/5.04	396.7	4903.7
GSAM2	MOS	1000	107.3	105.6	138.0	138.0	9.32	4.65/5.21	396.7	4903.9
MiTS	Ego	1213	93.0	90.9	97.0	97.0	10.75	1.92/2.46	32.7	439.2
MiTS	MOS	1000	33.1	33.1	33.9	33.9	30.17	0.93/1.12	32.7	168.7
SAM 2	Ego	1213	216.8	187.7	324.9	324.9	4.61	3.12/3.56	224.4	6101.4
SAM 2	MOS	1000	150.7	147.1	187.5	187.5	6.64	3.07/3.43	224.4	6110.6
SAM 2++	Ego	1213	242.0	239.5	259.6	259.6	4.13	2.48/2.80	101.2	5299.8
SAM 2++	MOS	1000	153.1	153.0	154.1	154.1	6.53	2.39/2.65	101.2	5308.9
SAM 2 Long	Ego	1213	1054.5	1051.7	1067.3	1067.3	0.95	6.01/6.66	224.4	25580.3
SAM 2 Long	MOS	1000	705.7	703.5	729.7	729.7	1.42	2.68/3.05	224.4	25609.5
SAM 3.1 (box)	Ego	1213	247.3	247.6	252.9	252.9	4.04	15.74/17.05	873.2	12182.7
SAM 3.1 (box)	MOS	1000	223.3	223.8	225.5	225.5	4.48	12.58/13.66	873.2	12198.2
XMem	Ego	1213	80.2	78.1	83.9	83.9	12.47	1.32/1.88	62.2	1575.1
XMem	MOS	1000	28.7	28.1	33.2	33.2	34.85	1.04/1.39	62.2	1145.0

TABLE XXX: T4 relative-pose hardware measurements, reported per image pair. Bold: best per column.

Configuration	Protocol	N	Mean	P50	P95	P99	Rate	Peak A/R	Params	GFLOPs
			Latency (ms/unit)				(unit/s)	(GiB)	(M)	/unit
CUT3R	Easy	1000	533.3	554.6	648.4	679.9	1.87	3.42/3.72	793.3	3211.2
DA3	Easy	1000	146.5	141.1	169.0	194.1	6.82	7.11/9.60	1355.7	6186.2
DUS3R	Easy	1000	489.5	490.7	533.5	624.1	2.04	2.58/2.80	571.2	3713.6
Fast3R	Easy	1000	511.1	480.6	712.2	1029.5	1.96	4.24/4.58	647.6	2957.9
MASt3R	Easy	1000	8819.2	8776.9	9219.8	9353.5	0.11	3.09/3.34	688.6	4094.2
MonST3R	Easy	1000	667.4	576.5	1176.9	1517.8	1.50	2.58/2.80	571.2	3713.6
MUS3R	Easy	1000	302.1	299.6	341.4	398.6	3.31	2.00/2.02	423.4	2030.6
Pi3X [†]	Easy	1000	156.9	153.5	176.0	181.9	6.37	5.90/6.79	992.5	—
Reloc3R	Easy	1000	294.2	291.2	331.5	410.7	3.40	2.48/2.50	426.5	1505.1
VGGT	Easy	1000	414.2	420.4	495.1	519.3	2.41	7.00/9.62	1256.5	5533.6

E.1 Adding models

A model subclasses `BenchmarkModel`, declares its task, loads weights once in `setup()`, and maps a batch of `Sample` objects to typed predictions in `predict(batch)` (e.g. `DepthPrediction`). Relative-depth models set `depth_output_kind="relative"` so that the runner applies scale-and-shift alignment before scoring. The runner validates predictions and reports native metrics, $\mathcal{J}_p$, $\mathcal{J}_{\min}$, and Φ .

E.2 Adding tasks

A new task registers a `TaskSpec` (primary metric, required modalities, run callable) with `register_task`, and each of its metrics registers a `MetricSpec` with direction and desirability window. Because $\mathcal{J}$ and Φ consume only these specifications and per-cell scores, the MANOVA and desirability pipeline applies to the new task unchanged.

Grounding profiles. Table XXXI reports T5 and T6 over 1,000 queries per configuration, including GPU count and aggregate process memory.

TABLE XXXI: T5 and T6 hardware measurements over 1,000 queries per model. Peak process memory is the sum across participating GPUs; model size is the declared roster size. Bold: best per column within each task.

Model	GPUs	N	Mean	P50	P95	P99	Rate	Peak mem.	Size
			Latency (ms/query)				(query/s)	(GiB)	(B)
<i>T5 regular VQA</i>									
Cosmos Reason2 2B	1	1000	396.6	323.1	739.1	758.8	2.52	41.75	2.44
Cosmos Reason2 8B	1	1000	996.9	940.0	1540.7	1671.8	1.00	41.66	8.00
DeepSeek-VL2 4.5B	2	1000	374.1	338.3	539.4	582.0	2.67	85.79	4.50
DeepSeek-VL2 1.0B	1	1000	199.8	179.5	319.6	338.2	5.00	43.10	1.00
Florence 2 0.23B	1	1000	107.0	96.8	151.4	206.0	9.34	1.13	0.23
Florence 2 0.77B	1	1000	133.7	126.0	185.8	229.2	7.48	2.32	0.77
InternVL 3.5 14B	1	1000	3272.0	3350.3	3656.1	3812.7	0.31	33.85	15.10
InternVL 3.5 1B	1	1000	1292.1	1288.6	1455.7	1573.0	0.77	4.72	1.10
PaliGemma 2 10B	1	1000	1246.4	1162.6	1757.6	2980.1	0.80	20.08	10.00
PaliGemma 2 3B	1	1000	605.4	530.7	896.7	1200.8	1.65	6.83	3.00
Qwen3-VL 2B	1	1000	304.7	286.7	588.8	622.2	3.28	41.75	2.00
Qwen3-VL 8B	1	1000	748.3	670.7	1215.4	1290.1	1.34	41.70	8.00
<i>T6 in-context VQA</i>									
Cosmos Reason2 2B	1	1000	431.4	367.7	699.4	761.6	2.32	41.79	2.44
Cosmos Reason2 8B	1	1000	1023.9	947.0	1518.6	1585.8	0.98	41.76	8.00
DeepSeek-VL2 4.5B	2	1000	371.1	354.1	534.6	621.6	2.69	85.79	4.50
DeepSeek-VL2 1.0B	1	1000	244.7	224.7	378.7	400.5	4.09	43.14	1.00
Florence 2 0.23B	1	1000	144.5	130.5	250.0	313.0	6.92	1.13	0.23
Florence 2 0.77B	1	1000	183.6	175.1	289.3	349.6	5.45	2.33	0.77
InternVL 3.5 14B	1	1000	3586.4	3468.2	4329.2	4420.5	0.28	34.63	15.10
InternVL 3.5 1B	1	1000	1566.8	1578.1	1778.4	1813.5	0.64	5.06	1.10
PaliGemma 2 10B	1	1000	1681.8	1536.5	3784.2	3877.3	0.59	20.08	10.00
PaliGemma 2 3B	1	1000	1041.9	921.8	2097.5	2200.0	0.96	6.83	3.00
Qwen3-VL 2B	1	1000	341.9	299.1	602.4	638.6	2.93	41.83	2.00
Qwen3-VL 8B	1	1000	753.5	693.5	1254.3	1303.6	1.33	41.87	8.00