ACTIVECONTACT: FACTORIZED PHYSICAL HYPOTHESES FOR ACTIVE IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Different physical worlds can produce similar motion, so one push is often not enough to know what is happening. We present ACTIVECONTACT, which treats physical understanding as an active identification problem. The model keeps several possible physical worlds, then chooses the next probe that is expected to separate them best. The discrete world contains four directional contact factors and one global friction factor. Local factors take the states *native*, *rounded*, or *compliant*; the friction factor is *isotropic*, *x-grain*, or *y-grain*. This gives $3^5 = 243$ combinations, together with continuous parameters for mass, friction, restitution, and inertia. Expected information gain (EIG) scores each unused probe by how much it is expected to reduce uncertainty. With six probes, Active reaches **87.5%** exact-world accuracy on ID tasks, **84.6%** on unseen combinations, and **79.9%** under geometry shift, compared with 37.0%, 29.4%, and 35.7% for Random. On composition-OOD, five Active probes already reach **77.3%**, above Random (29.4%) and Fixed (32.0%) with six probes. We also test two real-data interfaces. Physics101 is an external passive-video benchmark. Separately, we collect 72 tabletop push videos of a cube, sphere, and box; 69 are tracked successfully. In leave-one-repeat-out evaluation, EIG identifies the held-out object with 77.8% accuracy after four probes, compared with 53.6% for Random and 22.2% for Fixed. These results show that factorized physical hypotheses and information-seeking probes are useful when the correct explanation is not known in advance.

1 INTRODUCTION

Objects that look similar can behave differently. A light object with high friction can move like a heavier object with lower friction, and the same shape can respond differently when contact happens at a different local region. This means that appearance, or even one short interaction, may not uniquely reveal the physics.

We ask a simple question: *when several explanations still fit the observations, what should we try next?* Our goal is not to create the largest motion. It is to choose a probe that is expected to remove the most uncertainty.

Physical-world inference in our setting poses two distinct challenges: **representation** and **identification**. The representation challenge is combinatorial: local contact behavior can differ across signed directions while global friction and continuous parameters vary independently. Treating every joint configuration as one class wastes this reusable structure. The identification challenge is sequential: after one push, several candidate worlds may still explain the same motion, so the system must decide what to try next.

We address representation with a **compositional physical world**. The discrete structure is

$$h = (h_{+x}, h_{-x}, h_{+y}, h_{-y}, h_f), \tag{1}$$

where the first four factors describe signed local-contact modes and h_f describes the global friction pattern. Each factor has three states, giving $3^5 = 243$ joint worlds. Because the factors are predicted separately, familiar local states can be recombined into joint worlds that were never labeled during training; this is the central motivation for our composition-OOD split.

We address identification **actively**. The model maintains a posterior over the compositional world and continuous physical parameters. If plausible worlds respond similarly to one probe but differently to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

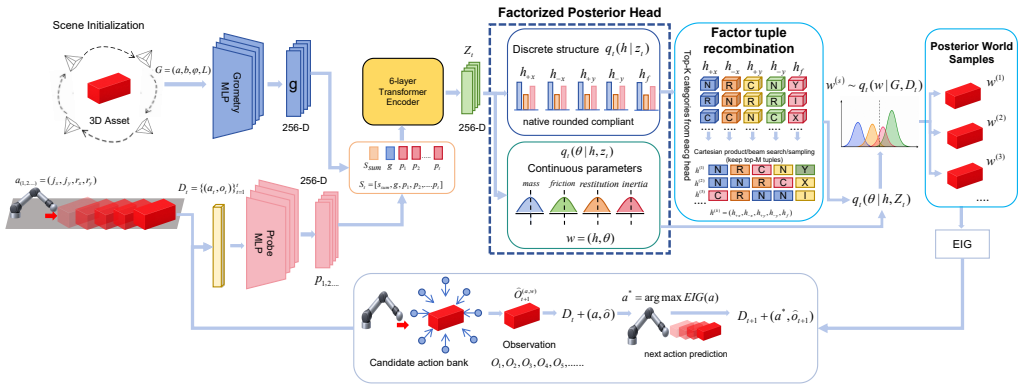

Figure 1: **Overview of ACTIVECONTACT around two challenges. Representation:** a physical world is modeled compositionally as four signed local-contact factors, one global friction factor, and continuous parameters, with a compact directional geometry proxy. **Identification:** the model maintains multiple plausible worlds and actively chooses the next probe by EIG to separate them.

another, the latter is more informative. Expected information gain (EIG) scores each unused probe by its expected posterior-entropy reduction, so interaction is chosen to discriminate the remaining hypotheses rather than merely to produce large motion. The anisotropic Gaussian footprint is a compact directional geometry interface for this controlled planar setting, not a claim of a general reconstructed-3DGS physics representation.

Figure 1 summarizes the resulting loop: composition makes the hypothesis space reusable, while EIG decides which interaction is most useful for identifying the remaining physical world.

We evaluate the method in a controlled simulation benchmark and with real data. The simulation benchmark separates three cases: new samples from the training distribution (ID), unseen combinations of familiar factors (composition-OOD), and unseen geometry (geometry-OOD). For real data, we keep two protocols separate. Physics101 (Wu et al., 2016) is an external passive-video benchmark. We also collect our own tabletop data with a cube, sphere, and weighted box, recording four push directions, two push strengths, and three repeats per object. This gives 72 videos and lets us test whether the same information-seeking idea is useful on recorded physical responses.

Our contributions follow this decomposition: **(1)** a **compositional 243-world contact representation** whose factors have explicit local meanings and can be recombined into unseen tuples; **(2)** an **amortized information-seeking rule** that selects counterfactual probes by approximate expected uncertainty reduction; and **(3)** controlled evaluation on ID, composition-OOD, geometry shift, external passive video, and an offline real-contact pilot. The anisotropic Gaussian footprint is used as a directional geometry interface in the planar benchmark; we do not claim a general reconstructed-3DGS collision model or superiority over mesh/point-cloud geometry.

2 RELATED WORK

Gaussian geometry and physical models. 3D Gaussian Splatting (3DGS) represents scenes with explicit anisotropic Gaussians (Kerbl et al., 2023). Later work connects Gaussian primitives to geometry, dynamics, and physical simulation (Xie et al., 2024; Li et al., 2025; Abou-Chakra et al., 2025; Zhao et al., 2025; Wang et al., 2026; Huang et al., 2024; Guédon & Lepetit, 2024; Luiten et al., 2024; Wu et al., 2024). We focus on a different question: when several physical explanations remain possible, which interaction should be used next?

Compositional physical reasoning. Structured and object-centric models learn reusable parts rather than one monolithic state (Kipf et al., 2020; Kossen et al., 2020; Locatello et al., 2020; Goyal et al., 2021). Work on systematic generalization shows why this matters when familiar pieces appear in a new combination (Ahmed et al., 2021; Wiedemer et al., 2024; Schug et al., 2024). We apply the same idea to local contact rules.

Active identification and real video. System identification estimates hidden dynamics from observations (Ljung, 1999). Bayesian experimental design and active learning additionally choose measurements expected to be informative (Lindley, 1956; MacKay, 1992; Chaloner & Verdinelli, 1995; Houlsby et al., 2011; Kirsch et al., 2019). We use EIG to choose physical probes. Physics101 and related work show that physical properties can also be learned from real video (Wu et al., 2015; 2016; Bear et al., 2021; Chen et al., 2022); in our paper, Physics101 tests passive real-video transfer, while our self-collected contact set tests offline action selection on real responses.

3 METHOD

3.1 PROBLEM FORMULATION

A probe is a planar impulse

$$a_t = (J_x, J_y, r_x, r_y), \tag{2}$$

where (J_x, J_y) is the impulse and (r_x, r_y) is the contact offset. In the synthetic benchmark, each probe returns a short pose trajectory

$$o_t = \{(x_k, y_k, \sin\phi_k, \cos\phi_k)\}_{k=1}^{19}. \tag{3}$$

The interaction history is

$$\mathcal{D}_t = \{(a_\tau, o_\tau)\}_{\tau=1}^{t}. \tag{4}$$

The hidden world is $\boldsymbol{w} = (h, \theta)$, with discrete contact structure h and continuous parameters θ. In our controlled benchmark, the hidden physical world consists of four directional contact factors, one global friction factor, and continuous physical parameters. Each discrete factor has three states, giving $3^5 = 243$ compositional physical worlds.

Given the observed geometry G and the interaction history $\mathcal{D}_t$, the model maintains the posterior

$$q_t(\boldsymbol{w}) = q(\boldsymbol{w} \mid G, \mathcal{D}_t). \tag{5}$$

At each step, the model evaluates the unused candidate probes and chooses the unused probe with the highest expected information gain. In the controlled benchmark, all actions come from a predefined low-energy action bank, so $\mathcal{A}_{\text{safe}}$ is simply that bank; there is no separate safety planner.

3.2 FACTORIZED PHYSICAL REPRESENTATION

To address the representation challenge, we describe each physical world by an observed directional geometry G, a factorized discrete structure h, and continuous physical parameters θ.

For the controlled planar benchmark, the object geometry is represented by a compact anisotropic footprint

$$G = (a, b, \phi_0, L), \tag{6}$$

where a and b are the semi-major and semi-minor axes, ϕ_0 is the initial orientation, and L is the arena half-extent. Let Σ denote the corresponding planar covariance; for a unit direction n, the support radius is

$$r(n) = \sqrt{n^\top \Sigma n}. \tag{7}$$

This preserves the directional geometry relevant to lateral pushing while keeping the benchmark controlled. Since interaction is restricted to the tabletop plane, we consider the four signed contact directions $\pm x$ and $\pm y$; the z direction is not part of the action space.

The discrete physical structure is factorized into four directional contact factors and one global friction factor:

$$h = (h_{+x}, h_{-x}, h_{+y}, h_{-y}, h_f), \qquad h_j \in \{0, 1, 2\}. \tag{8}$$

Each factor has three states, giving $3^5 = 243$ possible joint worlds. The key design is to predict these five factors separately rather than treating the 243 configurations as unrelated classes, so local contact states can be reused in combinations that were not observed during training.

Table 1: **Meaning of the three discrete states.** The four signed contact factors share the same local modes.

Factor	State 0	State 1	State 2
$h_{\pm x}, h_{\pm y}$	native	rounded	compliant
h_f	isotropic	x-grain	y-grain

The *native* mode preserves the anisotropic Gaussian support, whereas the *rounded* mode replaces it with an area-matched isotropic support. This native-versus-rounded pair provides a controlled intervention for isolating the effect of directional Gaussian geometry while keeping the object footprint area comparable. The *compliant* mode weakens the local tangential and bounce response. The friction factor is either isotropic or stronger along one of the two planar axes.

In addition to the discrete structure, each world has continuous physical parameters

$$\theta = (m, \mu, e, I), \tag{9}$$

where m is the mass, μ is the friction magnitude, e is the restitution coefficient, and I is the inertia scale.

This representation separates observable directional geometry from reusable discrete contact mechanisms and continuous physical properties, while also enabling a controlled anisotropic-versus-isotropic geometry comparison. It provides the hypothesis space used by the posterior network in Sec. 3.3.

3.3 Factorized Amortized Posterior

For each observed interaction $(a_\tau, o_\tau) \in \mathcal{D}_t$, we concatenate the 4-D action with the flattened 19×4 pose trajectory to form an 80-D input. A probe encoder maps this input to a 256-D token $\mathbf{p}_\tau$, while a separate geometry encoder maps the 4-D geometry descriptor G to a 256-D token $\mathbf{g}$. The geometry token, the observed probe tokens $\{\mathbf{p}_\tau\}_{\tau=1}^{t}$, and a learned summary token are processed by a 6-layer, 8-head Transformer (Vaswani et al., 2017). The Transformer output at the summary-token position is denoted by $z_t \in \mathbb{R}^{256}$ and provides a fused representation of the geometry G and the accumulated interaction history $\mathcal{D}_t$.

The summary output feeds five 3-way classification heads and a continuous parameter head. The posterior is

$$q_t(h, \theta \mid G, \mathcal{D}_t) = \left[\prod_{j=1}^{5} q_t(h_j \mid z_t)\right] q_t(\theta \mid h, z_t). \tag{10}$$

The five discrete heads predict the factorized structure $h = (h_{+x}, h_{-x}, h_{+y}, h_{-y}, h_f)$, allowing previously learned local states to be recombined into unseen joint worlds. The continuous head predicts $\theta = (m, \mu, e, I)$ conditioned on the discrete structure.

Training uses factor-wise cross-entropy for h and Gaussian negative log-likelihood for the normalized continuous parameters. The probe encoder, geometry encoder, Transformer, and posterior heads are optimized jointly during offline posterior training. The number of visible probes is randomly varied during training so that the same network can update the posterior throughout sequential interaction. A flat 243-way head is used only as an ablation.

3.4 Active Identification with Counterfactual EIG

The synthetic action bank contains 144 predefined probes, formed by 12 planar directions, six impulse magnitudes, and two contact offsets. Previously executed probes are removed from the candidate set.

For each unused candidate action, we sample eight worlds from the current posterior, simulate their counterfactual responses, and pass the resulting hypothetical interactions through the posterior network. The candidate is scored by

$$\mathrm{EIG}_t(a) \approx H[q_t(\boldsymbol{w})] - \frac{1}{8}\sum_{s=1}^{8} H[q_{t+1}^{(s)}(\boldsymbol{w})] - \ 0.004\,\frac{\|J(a)\|_2^2}{6.5^2}. \tag{11}$$

Table 2: **Three evaluation splits.** "Seen" means sampled from the same factor/geometry family used for training.

Split	Factor combination	Geometry
ID	seen 80% tuple pool	train range
Composition-OOD	held-out 20% joint tuples	train range
Geometry-OOD	seen 80% pool	shifted aspect/size

Here, $q_{t+1}^{(s)}$ is the hypothetical posterior after the s-th counterfactual response. The first two terms estimate the expected uncertainty reduction, while the final term mildly penalizes large impulses.

The hybrid entropy combines the five discrete-factor entropies with the uncertainty of the four continuous parameters; implementation details are provided in the appendix.

Accordingly, we select the unused probe that maximizes the score in Eq. (11). Unlike motion-based exploration, EIG favors probes for which the currently plausible physical worlds predict different outcomes. After executing the selected probe, its observed response is appended to $\mathcal{D}_t$, Eq. (10) is recomputed, and the process repeats until the probe budget is exhausted.

4 Experiments

4.1 Experimental setup

Synthetic benchmark. All synthetic experiments use a custom PyTorch planar rigid-body contact simulator with the compact anisotropic Gaussian proxy in Sec. 3.2. Each probe runs for 90 steps at 60 Hz, with every fifth state and the final state recorded as a 19-frame trajectory. Continuous parameters are sampled from $m \in [0.45, 2.20]$, $\mu \in [0.03, 0.55]$, $e \in [0.15, 0.92]$, and $I \in [0.55, 1.80]$, with Gaussian observation noise of standard deviation 0.006. We generate 262,144 training tasks and 16,384 validation/test tasks per split, each training task containing eight randomly selected probes. The posterior network is trained offline using factor-wise cross-entropy for h and Gaussian negative log likelihood for normalized (m, μ, e, I), and is then reused for sequential belief updates and EIG-based probe selection. And additional rollout-training visualizations are provided in Appendix. Formal Active/Random/Fixed evaluation uses 128 fresh tasks per split, three checkpoints, and a six-probe budget; full optimization details are given in the supplementary material.Active selects the unused probe with the highest EIG, Random samples an unused probe uniformly without replacement, and Fixed follows the same predefined probe order for every task.

The composition split is deterministic and keeps every individual factor state in training; only the full tuple is held out. For ID and composition-OOD, the semi-major axis is 0.105–0.175 and the aspect ratio is 0.36–0.76. Geometry-OOD changes these to 0.090–0.190 and 0.20–0.34, while keeping the same factor pool. This makes the object more elongated and moves the geometry outside the training family.

Metrics. *Exact* is one only when all five factors are correct. *Factor* is the mean accuracy over the five factors. *Param* is the mean absolute error after min–max normalizing (m, μ, e, I) by the declared parameter ranges; Table 3 reports this value times 100. *Roll* is the RMSE between predicted and true 19×4 pose trajectories over 12 held-out evaluation actions, with x, y normalized by arena size.

Real data. Physics101 (Wu et al., 2016) is an external passive-video benchmark with 1,189 processed clips from 100 objects (70/11/19 object-disjoint train/validation/test). It does *not* validate active probing; it asks whether the learned physical representation transfers to real visual observations. Separately, we collect our own active-contact dataset with a cube, sphere, and weighted box. Each object has four directions, two push strengths, and three repeats, giving 72 videos. The real protocol is described in detail in Section 4.5 and the appendix.

4.2 Main synthetic results

Figure 3 shows how the benefit of Active grows with the probe budget. At one or two probes, the strategies are often close. By three probes, Active starts to separate clearly. On composition-OOD,

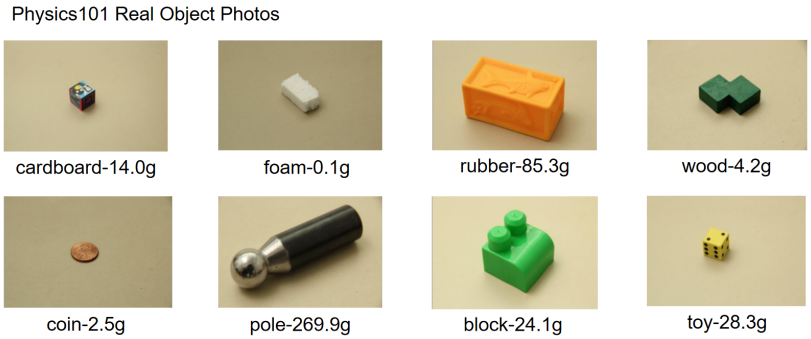

Figure 2: **Representative objects from the external Physics101 benchmark.** We show real object photos from the processed real-video transfer set. The benchmark spans very different materials, masses, and shapes, so transfer is tested on genuinely diverse real objects rather than on one narrow family.

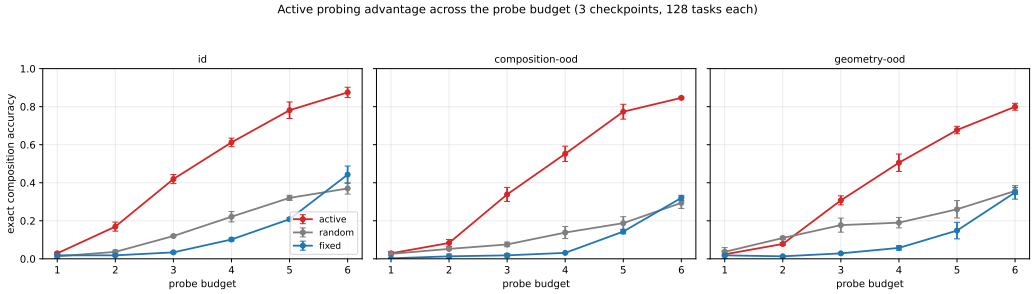

Figure 3: **Exact-world accuracy versus probe budget.** Active becomes clearly better after several probes. On composition-OOD, five Active probes already beat Random and Fixed using all six probes.

Active reaches **77.3%** with five probes, already higher than Random (29.4%) and Fixed (32.0%) with six probes. Across the 36 split–budget–baseline comparisons in the paired bootstrap analysis, Active is significantly better in 26, indistinguishable in 10, and never significantly worse. Table 3 gives the final six-probe results: Active reaches **87.5%** on ID, **84.6%** on composition-OOD, and **79.9%** on geometry-OOD.

The task-level bootstrap gives the same picture. For composition-OOD, the exact-accuracy gain over Random is 0.570 with a 95% interval of [0.477, 0.656]. The result is therefore not caused by only a few easy tasks.

4.3 Why factorization matters

Composition-OOD contains factor values the model has seen before, but combines them into full tuples that were never labeled during training. A flat 243-way classifier treats each tuple as a separate class, so it cannot assign the correct unseen class. The factorized model instead learns the five pieces separately and can recombine them.

Table 4 also shows an important trade-off. The flat classifier is better on ID and geometry-OOD in this eight-observation structure-only test because it can memorize the seen joint classes more directly. Factorization gives up some of that memorization in exchange for systematic generalization: on composition-OOD the flat model drops to 0%, while the factorized model remains at 36.1% before active test-time probing. With Active probing, the final six-probe composition-OOD exact accuracy reaches 84.6%.

The composition-order test in the appendix reaches the same qualitative conclusion when several factors change together.

Table 3: **Formal six-probe evaluation, mean ± std over three checkpoints.** Exact and Factor are accuracies (%); Param is normalized parameter MAE (%, lower is better); Roll is predictive rollout RMSE (lower is better).

Split	Strategy	Exact ↑	Factor ↑	Param ↓	Roll ↓
ID	**Active**	**87.5±2.7**	**97.45±0.63**	**0.691±0.009**	**0.223±0.006**
	Fixed	44.3±4.5	85.89±0.89	0.978±0.048	0.274±0.006
	Random	37.0±3.0	80.73±1.02	1.406±0.095	0.312±0.004
Composition-OOD	**Active**	**84.6±0.5**	**96.46±0.24**	**0.722±0.024**	**0.214±0.002**
	Fixed	32.0±1.4	81.35±0.65	1.106±0.035	0.297±0.007
	Random	29.4±3.0	78.70±1.19	1.299±0.047	0.305±0.005
Geometry-OOD	**Active**	**79.9±1.8**	**95.21±0.36**	**1.079±0.067**	**0.311±0.008**
	Fixed	34.9±3.5	81.98±0.74	1.578±0.085	0.355±0.005
	Random	35.7±1.8	79.32±0.24	1.770±0.110	0.366±0.004

Table 4: **Factorized vs. flat structure inference using the eight stored training-style observations.** Factorized numbers are mean ± std over three seeds; flat 243-way uses seed 8101. This is a structure ablation, separate from the six-probe active evaluation.

Method	ID Exact ↑	Comp.-OOD Exact ↑	Geom.-OOD Exact ↑
Flat 243-way	**58.37**	0.00	**55.42**
Factorized	44.05±0.62	**36.13±0.62**	38.66±0.90

4.4 PHYSICS101: EXTERNAL REAL-VIDEO TRANSFER

Physics101 is not an active-probing experiment. We use it only to test whether the learned physical representation transfers to real RGB video for passive mass inference. Figure 2 makes this setting more concrete: the benchmark contains lightweight foam pieces, small cardboard and toy objects, and heavier wood, rubber, and metal-like objects. In other words, the transfer test is visually and physically diverse, making the real-video result more meaningful. During a re-audit, we found that the old low-label sampler selected at least one object per family, so the nominal 5% and 10% settings accidentally used the same 15 training objects. We therefore reran the experiment using a true global object draw. Figure 4 shows the corrected result. With 4, 7, 18, 35, and 70 training objects, the pretrained MAE is 0.2041, 0.1241, 0.1008, 0.0816, and 0.0778, compared with 0.2079, 0.1324, 0.0997, 0.0837, and 0.0808 from scratch. The point estimates are usually better with pretraining, but the difference is smaller than the three-seed variation at every setting. We therefore report pretraining as competitive, not as a statistically confirmed improvement.

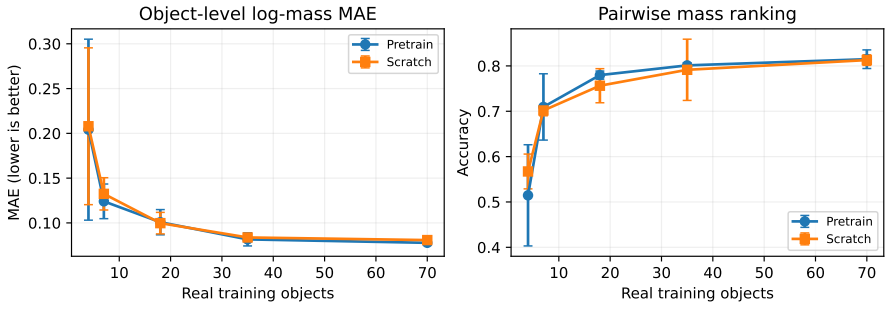

Figure 4: **Corrected Physics101 low-label sweep.** The x-axis shows the actual number of real training objects. Error bars are standard deviations over three seeds.

4.5 OUR REAL-WORLD DATA AND EVALUATION

We collect a real-contact dataset using a red cube, sphere, and weighted box, with measured masses of 20.2, 18.7, and 24.8 g and static-friction coefficients of 0.73, 0.82, and 0.86. Four push directions ($\pm x, \pm y$), two categorical strengths, and three repeats per object yield 72 videos. Weak/strong pushes

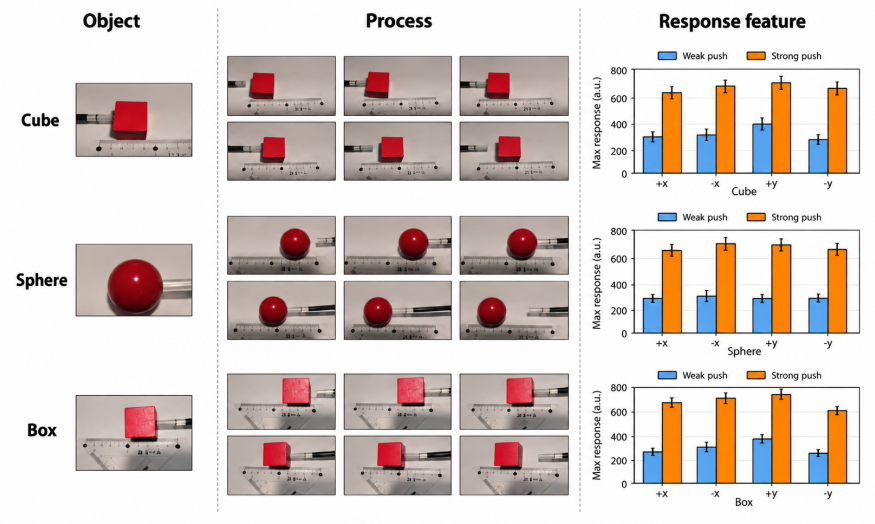

Figure 5: **Our self-collected real-contact dataset.** Each row shows one object, a representative 2×3 push sequence, and the tracked response feature. Blue/orange bars denote weak/strong push conditions.

Table 5: **Real-world evidence summary.** Physics101 is external; captured geometry and real contact use our own data.

Interface	Metric	Ours	Baseline	Delta
Physics101, 70 objs.	log-mass MAE ↓	**0.0778**	0.0808 Scratch	−0.0030
Captured geom., Comp.-OOD	Exact ↑	**38.7%**	12.6% Random	+26.1 pp
Captured geom., Geom.-OOD	Exact ↑	**39.5%**	16.5% Random	+23.0 pp
Captured box, Comp.-OOD	Exact ↑	**62.9%**	17.8% Random	+45.1 pp
Real contact, 4 probes	Object ID acc. ↑	**77.8%**	53.6% Random	+24.2 pp
Real contact tracking	Valid tracks ↑	**95.8%**	–	–

correspond to nominal 2/5 cm travel and are hand-executed at the marked center contact point, so the labels are not force-calibrated.

Videos are recorded at 30 fps. We track the largest red component with an HSV mask and use centroid displacement along the commanded axis as the response feature. Three tracking failures leave 69 valid trajectories.

We evaluate leave-one-repeat-out object identification over the three objects. For each held-out repeat, the other two define an action-conditioned Gaussian response model for each object. Starting from a uniform prior, Active selects unused actions by EIG, Random samples without replacement, and Fixed follows a preset order. EIG uses each hypothesis mean response as a deterministic sigma point and does not reuse the synthetic neural posterior. After four probes, Active reaches 77.8% accuracy, versus 53.6% for Random and 22.2% for Fixed, showing that information-seeking action selection remains effective on measured responses.

We also replace only the simulator geometry with proxies from the same objects. Across four checkpoints and three seeds, Active reaches 38.7% exact accuracy on composition-OOD and 39.5% on geometry-OOD, versus 12.6% and 16.5% for Random. This is a geometry-domain test rather than a full reconstructed-3DGS collision model.

4.6 Ablation Summary

The ablations separate the two main ideas. Factorization is what allows unseen combinations to be represented (Table 4); EIG is what improves the rate at which the correct world is found as more probes are allowed (Figure 3). The corrected Physics101 study is a transfer check rather than evidence

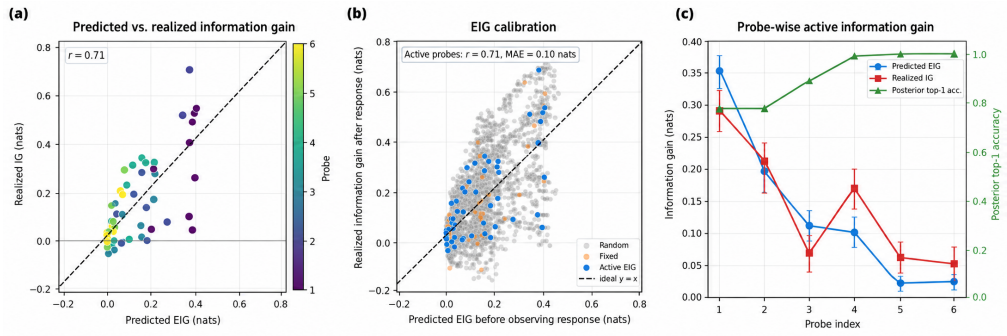

Figure 6: Comparison between predicted information gain and realized information gain.

for active probing. The appendix also keeps a separate anisotropic-vs-isotropic contact diagnostic from the earlier 3DV protocol.

5 Discussion

What the active policy is doing. Factorization and active probing solve different problems. Factorization makes a new combination representable. EIG then decides which action is most useful for telling the remaining candidates apart. A large motion is not automatically informative; the useful action is the one that changes the belief.

EIG as an effective action-selection score. Figure 6 compares predicted information gain with the entropy reduction realized after observing the response. For Active probes, the two are positively correlated ($r = 0.71$), with a mean absolute calibration error of 0.10 nats. Both predicted and realized information gain generally decrease as evidence accumulates, while posterior top-1 accuracy increases. This behavior is consistent with the intended role of EIG: early probes are more informative when multiple hypotheses remain plausible, whereas later probes add less once the posterior has concentrated. We therefore use EIG as an action-selection criterion, while final performance is evaluated by identification accuracy and probe efficiency.

Interpretation of the real-world experiments. The real protocols answer different questions. Physics101 tests passive RGB transfer. The captured-geometry study tests how sensitive the synthetic pipeline is to geometry derived from real objects. The 72-video contact set tests whether information-seeking action selection still helps on recorded physical responses. The three results should therefore be read together, but not merged into one claim.

Role of Gaussian geometry. Directional Gaussian support gives a compact way to describe how local shape changes with contact direction, which is why we use it in the controlled benchmark. The current main ablations prove the value of factorization and EIG, not that Gaussians are the only possible geometry representation. The appendix contains a separate anisotropic-vs-isotropic contact diagnostic, but a matched point-cloud or mesh encoder comparison remains future work.

6 Conclusion

We presented ActiveContact, an active physical-identification framework that keeps several possible worlds instead of committing after one observation. The factorized representation lets familiar local rules be recombined, and EIG chooses probes that best separate the remaining hypotheses. The method performs strongly on ID, composition-OOD, and geometry-OOD simulation tests. Physics101 provides an external real-video transfer check, while our own captured geometry and 72-video contact set provide controlled real-world evidence for geometry shift and information-seeking action selection.

Reproducibility statement

The supplementary material records the exact factor semantics, simulator ranges, observation interface, network architecture, training settings, EIG approximation, metrics, split definitions, Physics101 preprocessing, captured-geometry tests, and our real-contact capture and offline EIG protocol. The asset package includes CSV/JSON summaries, code snapshots, figures, and provenance notes. We keep the current ICLR experiments separate from the earlier 3DV diagnostics. Re-running the active selector on the same machine is bit-identical; across two GPU architectures, near-tied EIG choices can change a small number of actions, shifting the composition-OOD six-probe result from 0.8464 to 0.8490 without changing the conclusion.

AI use statement

Generative AI tools were used for language editing, manuscript organization, LaTeX formatting, and figure-layout drafting. They were not used as a source of experimental evidence. The authors remain responsible for checking the protocols, numbers, citations, equations, claims, and final wording.

Ethics statement

This work studies low-energy interaction with inanimate tabletop objects and uses no human-subject or personally identifying data. A deployed robot would still need calibrated force limits, collision monitoring, and task-specific safety checks.

References

Jad Abou-Chakra, Krishan Rana, Feras Dayoub, and Niko Suenderhauf. Physically embodied gaussian splatting: A visually learnt and physically grounded 3d representation for robotics. In *Conference on Robot Learning*, 2025.

Faruk Ahmed, Yoshua Bengio, Harm van Seijen, and Aaron Courville. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations*, 2021.

Daniel Bear, Chaofei Fan, Damian Mrowca, Yunzhu Li, Stephen Alter, Aran Nayebi, Jeremy Schwartz, Li Fei-Fei, Jiajun Wu, Joshua Tenenbaum, et al. Physion: Evaluating physical prediction from vision in humans and machines. In *NeurIPS*, 2021.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

Zhenfang Chen, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B. Tenenbaum, and Chuang Gan. Comphy: Compositional physical reasoning of objects and events from videos. *arXiv preprint arXiv:2205.01089*, 2022.

Arnaud Doucet, Nando de Freitas, and Neil Gordon (eds.). *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, and Bernhard Sch"olkopf. Recurrent independent mechanisms. In *International Conference on Learning Representations*, 2021.

Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *CVPR*, 2024.

Neil Houlsby, Ferenc Huszar, Zoubin Ghahramani, and Mate Lengyel. Bayesian active learning for classification and preference learning. In *NIPS Workshop on Bayesian Optimization, Experimental Design and Bandits*, 2011.

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH*, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Thomas Kipf, Elise van der Pol, and Max Welling. Contrastive learning of structured world models. In *International Conference on Learning Representations*, 2020.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *NeurIPS*, 2019.

Jannik Kossen, Karl Stelzner, Marcel Hussing, Claas Voelcker, and Kristian Kersting. Structured object-aware physics prediction for video modeling and planning. In *International Conference on Learning Representations*, 2020.

Wenxuan Li, Hang Zhao, Zhiyuan Yu, Yu Du, Qin Zou, Ruizhen Hu, and Kai Xu. Pin-wm: Learning physics-informed world models for non-prehensile manipulation. In *Robotics: Science and Systems*, 2025.

Dennis V. Lindley. On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Lennart Ljung. *System Identification: Theory for the User*. Prentice Hall, 2 edition, 1999.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *NeurIPS*, 2020.

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *3DV*, 2024.

David J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

Simon Schug, Seijin Kobayashi, Yassir Akram, Maciej Wolczyk, Alexandra M. Proca, Johannes von Oswald, Razvan Pascanu, Joao Sacramento, and Angelika Steger. Discovering modular solutions that generalize compositionally. In *International Conference on Learning Representations*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Meizhong Wang, Wanxin Jin, Kun Cao, Lihua Xie, and Yiguang Hong. Contactgaussian-wm: Learning physics-grounded world model from videos. *arXiv preprint arXiv:2602.11021*, 2026.

Thaddaeus Wiedemer, Jack Brady, Alexander Panfilov, Attila Juhos, Matthias Bethge, and Wieland Brendel. Provable compositional generalization for object-centric learning. In *International Conference on Learning Representations*, 2024.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *CVPR*, 2024.

Jiajun Wu, Ilker Yildirim, Joseph J. Lim, William T. Freeman, and Joshua B. Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *NeurIPS*, 2015.

Jiajun Wu, Joseph J. Lim, Hongyi Zhang, Joshua B. Tenenbaum, and William T. Freeman. Physics 101: Learning physical object properties from unlabeled videos. In *BMVC*, 2016.

Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *CVPR*, pp. 4389–4398, 2024.

Haoyu Zhao, Hao Wang, Xingyue Zhao, Hao Fei, Hongqiu Wang, Chengjiang Long, and Hua Zou. Physsplat: Efficient physics simulation for 3d scenes via mllm-guided gaussian splatting. In *ICCV*, pp. 5242–5252, 2025.

A Supplementary Details

A.1 Scope of each protocol

The simulation benchmark, Physics101, captured-geometry test, and real-contact dataset answer different questions. The 243-world simulation benchmark tests the full discrete/continuous belief and active probe policy. Physics101 tests passive real-video transfer. Our captured-geometry test replaces only the simulator geometry with proxies made from real captured objects. Our real-contact test uses recorded physical pushes and evaluates action selection offline. Earlier 3DV results use a different hypothesis space and are kept separate. We do not claim a universal physical world model or closed-loop robot probing.

A.2 Factor states and simulation details

The four local factors use the same three modes. For state 0 (*native*), contact uses the anisotropic support radius and tangent term. State 1 (*rounded*) replaces the support radius with the area-matched isotropic radius and removes the tangent term. State 2 (*compliant*) uses a softened support rule, scales the tangent term to 0.30 of native, and scales restitution to 0.45. The friction factor uses multipliers $(1, 1)$ for isotropic, $(1.55, 0.62)$ for x-grain, and $(0.62, 1.55)$ for y-grain.

The simulator uses $\Delta t = 1/60$ s for 90 steps and records every fifth step plus the final state, producing 19 frames. The continuous prior ranges are

$$m \in [0.45, 2.20], \quad \mu \in [0.03, 0.55], \quad e \in [0.15, 0.92], \quad I \in [0.55, 1.80].$$

Training samples use the interior 4–96% of each range. For ID and composition-OOD, $a \in [0.105, 0.175]$ and aspect $b/a \in [0.36, 0.76]$; geometry-OOD uses $a \in [0.090, 0.190]$ and $b/a \in [0.20, 0.34]$. Initial angle is sampled from $[-0.75, 0.75]$ rad and arena half-extent from $[0.55, 0.73]$ in all splits. Observation noise has standard deviation 0.006.

A.3 Network architecture and training

For one probe, the input is 4 action values plus 19×4 observation values, for 80 scalars. The probe MLP is $80 \to 256 \to 256$ with GELU and LayerNorm. The geometry MLP is $4 \to 256 \to 256$. A learned summary token, one geometry token, and the available probe tokens are processed by a 6-layer Transformer with width 256, 8 attention heads, feed-forward width 1024, and dropout 0.05. The factorized structure head outputs 5×3 logits. The continuous head outputs four normalized means and four standard deviations, with factor-conditioned offsets.

Training uses AdamW for 20,000 steps, batch size 2,048, learning rate 3×10^{-4}, weight decay 10^{-4}, cosine decay to 5% of the initial learning rate, gradient clipping at 5.0, and bfloat16 autocast. Random observation-length prefixes are sampled during training, and 0.002 Gaussian augmentation noise is added to observations.

Qualitative simulator-training check Figure 7 shows how the learned rollout model improves during training. Each column is one representative contact rollout from the supplied simulator assets. The top row (GT rollout) is the ground-truth simulator trajectory, and the lower rows are model predictions at epochs 0, 20, 60, and 120. In each frame, the red cuboid is the main 3DGS object body, the orange side face only provides a stronger 3D depth cue, the yellow and blue blobs mark two different local Gaussian support regions on the object, the small red dots indicate the external contact probe, and the dark grid is the simulator ground plane used to visualize translation and rotation. Importantly, the yellow and blue regions are only visualization markers for local Gaussian-supported regions; they should not be interpreted as fixed labels for mass, friction, or elasticity. At epoch 0, the model has not yet learned the contact-conditioned dynamics, so the predicted pose, rotation, and final position differ clearly from the simulator rollout. By epoch 20, the model starts to capture the rough motion direction, but the contact-driven rotation and displacement are still unstable. By epochs 60 and 120, the predicted translation, rotation, and pose become much closer to the simulator ground truth. This makes the figure a useful qualitative sanity check: it shows that the learned dynamics model is indeed approaching the simulator behavior that later supports active identification

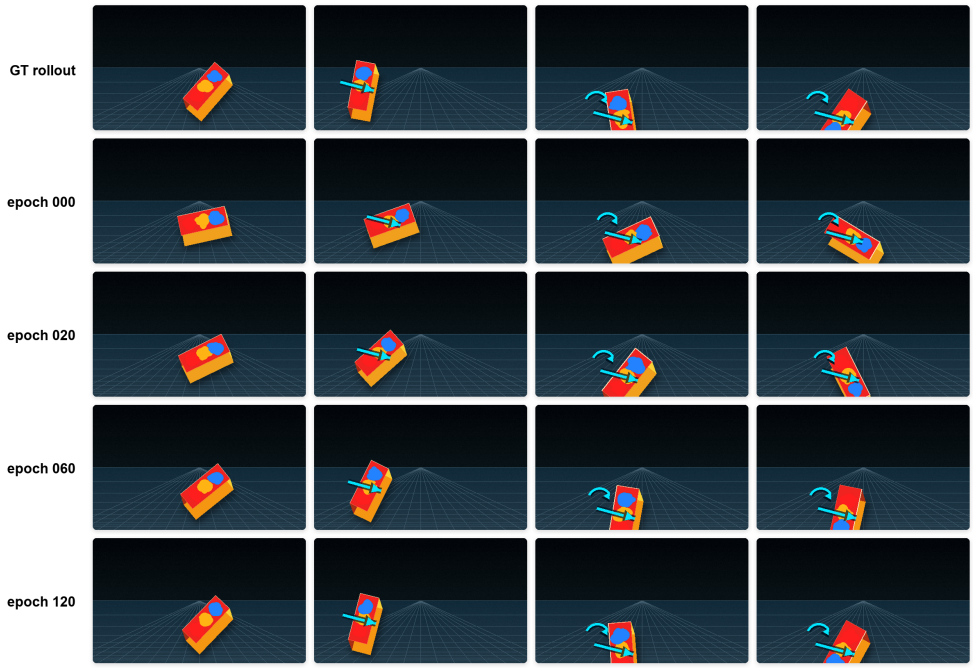

Figure 7: **Training intermediate visualization: predictions approach simulator rollout.** Each column corresponds to one representative simulator rollout case. The top row shows the ground-truth (GT) rollout generated by the simulator, while the lower rows present the model predictions at training epochs 0, 20, 60, and 120. The red cuboid denotes the main object body, and the orange side face provides a 3D depth cue. The yellow and blue blobs indicate two local Gaussian support regions on the object. The cyan arrows visualize the contact action direction and the associated motion tendency under interaction, while the dark grid represents the simulator ground plane. As training proceeds, the predicted translation, rotation, and overall pose become increasingly consistent with the ground-truth simulator rollout, showing that the learned dynamics model gradually captures the contact-conditioned object behavior.

A.4 EIG APPROXIMATION, ACTION BANK, AND METRICS

The synthetic action bank has 144 actions: 12 planar directions, impulse magnitudes $\{0.8, 1.4, 2.2, 3.4, 4.8, 6.5\}$, and offsets $r_y \in \{-0.045, +0.045\}$. Previously used actions are masked and cannot be selected again. For each candidate action we draw 8 posterior worlds, simulate one noisy fantasy observation per world, append that observation virtually, and re-run the posterior network. The entropy score used in implementation is the sum of factor entropies plus the log standard deviations of the continuous head (constants omitted). We subtract an energy term $0.004\|J\|_2^2/6.5^2$. Thus Eq. (11) is computed by full action enumeration with an eight-sample Monte Carlo approximation.

Exact accuracy requires all five discrete factors to be correct. Factor accuracy averages correctness over the five factors. Normalized parameter MAE first maps each of (m, μ, e, I) to $[0, 1]$ using the ranges above and then averages absolute error; the main table multiplies this value by 100. Rollout RMSE uses 12 evaluation actions spread across the action bank and compares predicted with true 19-frame $(x, y, \sin\phi, \cos\phi)$ trajectories; x, y are normalized by arena half-extent.

A.5 REAL-CONTACT PROCESSING AND OFFLINE EIG

The real videos are recorded at 30 fps with a fixed camera; the dataset contains both landscape and portrait 1080p recordings depending on push direction. The camera model is not stored in the metadata. A ruler is visible as a scale cue and the pose is reset between trials. Weak/strong are nominal 2 cm/5 cm hand-push conditions rather than calibrated force levels.

Tracking uses an HSV red mask, keeps the largest red connected component, and follows its centroid. The scalar response used for the reported offline policy is the absolute displacement along the commanded axis. Three tracks fail because of mask/occlusion issues and are excluded.

For each held-out repeat, the other two repeats fit a Gaussian likelihood for each of the three object identities and each of the eight actions. A standard-deviation floor of 25 pixels avoids overconfident two-sample fits. Active EIG starts from a uniform prior over object identity and evaluates every unused action. For each candidate action, the mean response under each object hypothesis is used as a deterministic sigma point; posterior entropy after those three possible responses is averaged under the current object probabilities. The action with the largest expected entropy reduction is selected. Random averages 100 action orders; Fixed uses a preset order. The reported 77.8% after four probes is therefore held-out *object-identification* accuracy over 3 objects $\times$ 3 held-out repeats, not 243-world accuracy.

A.6 Matched finite-particle SMC baseline

The matched-lite SMC baseline uses 486 stratified particles, about two continuous particles per discrete world, and receives the same six selected actions as seed 8101. Table 6 shows the result. At this particle budget, the amortized posterior is much stronger. This should not be read as a general claim that SMC is weak; larger particle sets and better proposals may improve it (Doucet et al., 2001).

Table 6: **Matched-lite SMC reference, seed 8101.** Exact accuracy and normalized parameter MAE after six probes.

Split	Method	Strategy	Exact ↑	Param MAE ↓	Roll RMSE ↓
ID	ActiveContact	Active	0.8438	0.0070	0.2295
ID	SMC-486	Active	0.0234	0.1635	0.4754
Comp.-OOD	ActiveContact	Active	0.8438	0.0074	0.2164
Comp.-OOD	SMC-486	Active	0.0156	0.1542	0.4678
Geom.-OOD	ActiveContact	Active	0.7891	0.0102	0.3038
Geom.-OOD	SMC-486	Active	0.0078	0.1509	0.4855

A.7 Composition-order stress test

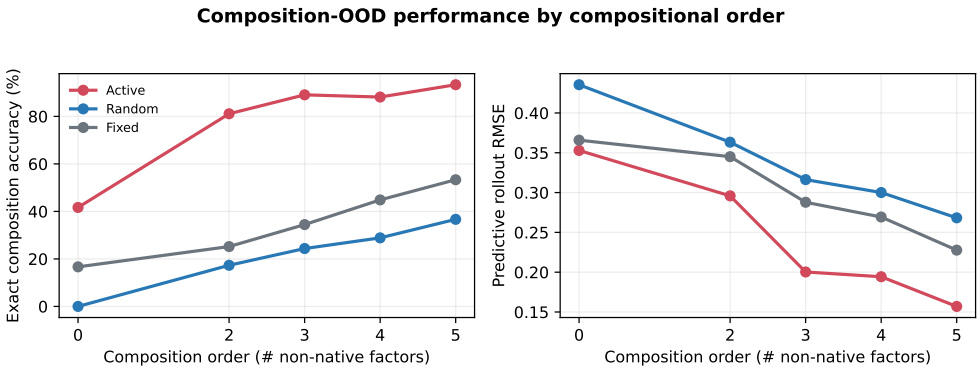

Figure 8: Composition-OOD performance grouped by the number of non-native factors. Active stays ahead across the observed orders. We do not assume that a larger order is always harder.

A.8 Corrected Physics101 low-label sweep

The old sampler spent the requested fraction separately inside each object family and always selected at least one object per family. Because the families are small, the nominal 5% and 10% settings both selected 15 of 70 objects. We corrected this by sampling globally over training objects. With only three seeds, the pretrain-minus-scratch difference is not statistically resolved at any fraction, so the main paper does not claim a significant gain.

Table 7: **Corrected global-object low-label sweep.** Mean ± std over three seeds.

Requested	Objects	Init.	MAE ↓	Pairwise ↑
5%	4	Scratch	0.2079 ± 0.0876	0.5673 ± 0.0383
5%	4	Pretrain	0.2041 ± 0.1010	0.5146 ± 0.1116
10%	7	Scratch	0.1324 ± 0.0180	0.7018 ± 0.0117
10%	7	Pretrain	0.1241 ± 0.0193	0.7096 ± 0.0731
25%	18	Scratch	0.0997 ± 0.0120	0.7563 ± 0.0376
25%	18	Pretrain	0.1008 ± 0.0141	0.7797 ± 0.0089
50%	35	Scratch	0.0837 ± 0.0027	0.7914 ± 0.0675
50%	35	Pretrain	0.0816 ± 0.0073	0.8012 ± 0.0058
100%	70	Scratch	0.0808 ± 0.0032	0.8129 ± 0.0117
100%	70	Pretrain	0.0778 ± 0.0029	0.8148 ± 0.0205

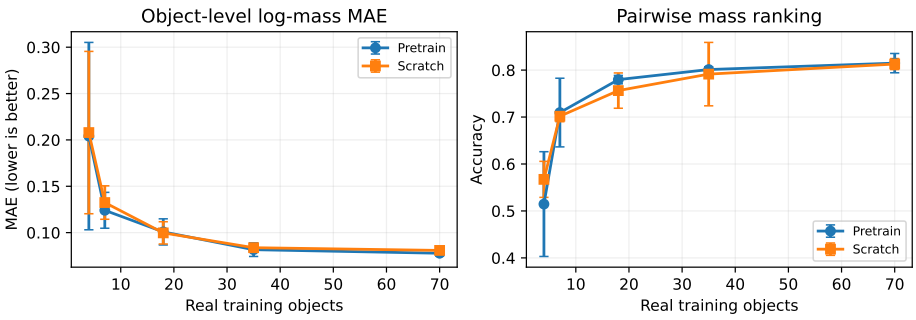

Figure 9: Corrected global-object low-label sweep used in the main paper.

A.9 Captured-object geometry test

We build simple geometry proxies from the three objects collected for the real-contact experiment: cube, sphere, and weighted box. These proxies replace only the simulator geometry field. Across four checkpoints and three seeds each, the mixed captured-object task gives 0.387 exact accuracy on composition-OOD and 0.395 on geometry-OOD under Active. The box transfers much better than the cube and sphere, so the real geometry distribution is still a major source of error.

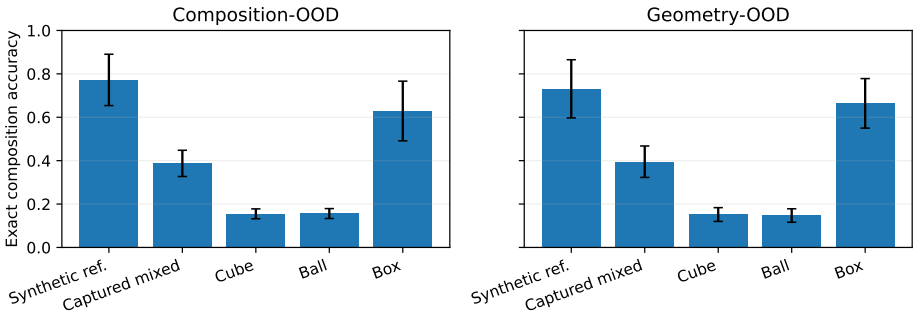

Figure 10: Captured-object geometry substitution. The box transfers better than the near-isotropic cube and sphere; the synthetic reference is shown for comparison.

A.10 Our controlled real-contact dataset

We collect three physical objects: cube, sphere, and weighted box. The action library contains four lateral directions and two strengths. Each action is repeated three times, giving 72 videos. The fixed camera records the motion. Automatic red-object tracking produces 69 valid tracks. We measure mass directly and static friction from the incline-board critical angle.

Table 8: **Measured properties of our real-contact objects.**

Object	Mass (g)	Critical angle	Measured μ_s	Contact surface
Cube	20.2	36.1°	0.73	A4 paper
Sphere	18.7	39.4°	0.82	White coated paper
Weighted box	24.8	40.7°	0.86	Cardstock

For offline policy evaluation, one repeat is held out. The policy reveals one recorded action response at a time. Active chooses the next action with EIG, Random samples without replacement, and Fixed follows a fixed order. This tests action selection on real recorded responses, not closed-loop robot control.

A.11 Legacy anisotropic-contact diagnostic

This diagnostic comes from the earlier 3DV protocol and is included only to show why anisotropic Gaussian geometry can matter for contact. It is not part of the current 243-world quantitative benchmark.

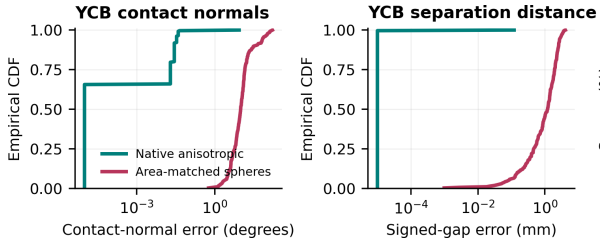

Figure 11: Legacy YCB-style geometry diagnostic. Native anisotropic Gaussian support gives lower contact-normal and signed-gap errors than an area-matched sphere proxy.

A.12 Limitations

The current world family is deliberately controlled: only lateral pushes are used, the real active set has three objects, and the hand-executed weak/strong pushes are not force-calibrated. Some worlds will remain hard to separate whenever all available probes produce similar responses. Larger object-disjoint capture, calibrated actuation, and tighter integration with reconstructed Gaussian geometry are natural next steps.