TRACE2STABLE: LEARNING FROM PHYSICAL SEARCH FOR 3D SCENE REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Image-based 3D reconstruction often produces visually plausible but physically invalid scenes: objects may fall or interpenetrate when simulated. Physical refinement must resolve these violations without replacing the recovered geometry or unnecessarily changing the observed arrangement. Existing per-scene search repeatedly simulates candidate pose corrections to find a stable, collision-free layout. We propose **Trace2Stable**, which amortizes physical search by learning reusable repair hypotheses from optimization traces. A graph-conditioned flow proposes multiple pose corrections, and candidate support relations represent ambiguous contacts. Simulation verifies the proposed repairs; measured outcomes guide residual refinement, while candidate ordering, alternative support hypotheses, and adaptive rollout continuation direct the remaining verification budget. Prior searches thus guide which repairs to test, while physical tests choose a valid arrangement or preserve an already valid input. We compare against the cross-entropy method (CEM), which searches for pose corrections from scratch in each scene. On 32 controlled layouts with matched simulation objects, Trace2Stable increases the proportion of stable outputs without illegal collisions from 21.88 to 59.38 percent while using 46.95 percent fewer candidate-environment simulation steps than this per-scene search baseline.

1 INTRODUCTION

Reconstructing an indoor scene from a single image requires recovering object geometry, poses, and support relations. A convincing image does not establish that the recovered arrangement can withstand gravity and contact: a cup may appear to rest on a table yet fall when simulated, and neighboring furniture may intersect outside the camera's view. For embodied learning and interactive simulation, the scene must remain physically valid while retaining the spatial arrangement conveyed by the observation. Refinement should therefore correct unstable or colliding placements while preserving those already valid. Figure 1 contrasts falling, interpenetration, and missing contact with their corrected arrangements in a shared scene.

One approach is to search for pose corrections through physical simulation. REST3D (Ma et al., 2026) reconstructs objects and a support tree, then uses the cross-entropy method (CEM) to refine local and global placements. Each scene starts a new search: useful corrections do not improve proposals for the next reconstruction. The fixed support tree also commits to one explanation of occluded contacts. Yet the search itself produces useful supervision: it reveals which changes resolve a violation and which arrangements remain plausible under simulation. The central question is how to reuse physical search experience to prioritize repairs in new scenes while verifying them by simulation.

The challenge is to transfer useful alternatives, rather than a single answer. Several corrections may preserve the observed layout, while their average intersects an obstacle or loses support. A different candidate parent can also require a different pose correction. This ambiguity couples hypothesis generation to verification: a proposal is useful if testing it helps resolve the current scene. Verification effort should adapt accordingly. An accepted group can stop early, whereas an unresolved group may benefit from a different candidate, an alternative support relation, or a longer simulation.

We develop **Trace2Stable** around this connection between reusable physical search experience and scene-specific verification. Rather than restarting pose search from an uninformed distribution, a graph-conditioned flow learns preferred corrections from previous optimization traces. Multiple

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.

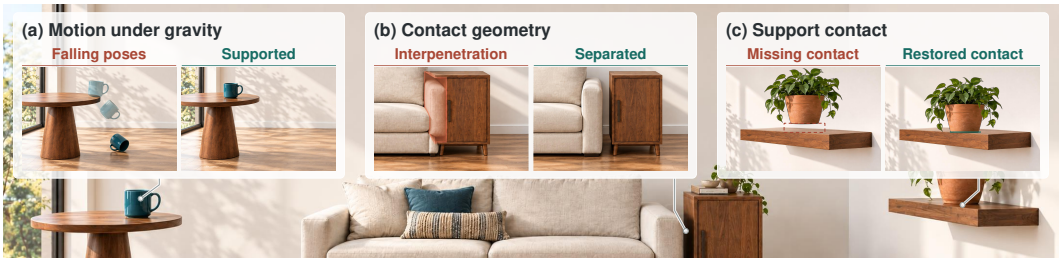


Figure 1: **Three physical failures and their corrected arrangements.** The shared scene locates each case. Paired insets contrast (a) falling mugs with a supported tabletop pose, (b) a cabinet intersecting a sofa with a separated placement, and (c) a floating pot with restored shelf contact. Translucency marks overlap; the dashed marker indicates missing support.

proposals retain alternatives for physical testing and residual refinement. For the controlled furniture benchmark, we combine flow proposals with residual search. Support branching, bank opening, and trajectory continuation direct a finite query budget toward unresolved uncertainties. This hypothesis generation and physical verification pipeline implements *budgeted physical evidence acquisition*: learned proposals guide queries, and simulation checks each candidate against the physical criteria.

This reuse is valuable when it improves the returned scene while reducing the physical work needed to find it. Extensive experiments examine both outcomes, from controlled comparisons to image-derived reconstructions. On 32 controlled layouts with matched simulated objects, the joint stability and safety rate rises from 21.88% to 59.38% while simulation steps fall by 46.95%. ReplicaCAD generator comparisons isolate proposal choice under shared feedback; transfer studies on the Habitat Synthetic Scenes Dataset (HSSD), ScanNet++, ProcTHOR, and iTHOR examine support selection and layout preservation. Candidate-level analysis links these outcomes to the method by measuring whether learned proposals cover valid repairs and whether budgeted physical verification selects one from the candidate bank.

Our contributions are summarized as follows:

- **Amortized physical search.** We learn reusable repair hypotheses from optimization traces while retaining simulator tests to decide physical validity.
- **Contact-aware hypothesis generation.** We model ambiguous support relations and generate diverse pose alternatives conditioned on scene structure.
- **Budgeted physical verification.** We evaluate refinement quality under explicit simulation budgets and analyze the trade-off between candidate discovery and physical validation.

2 Related Work

Observation-anchored 3D reconstruction. Single-view reconstruction combines visual evidence with scene structure: systems use grammars and physical commonsense (Huang et al., 2018; Chen et al., 2019), jointly recover layout, pose, and meshes (Nie et al., 2020), or refine appearance through inverse rendering (Yeh et al., 2022). Generative reconstruction also distills video diffusion priors into feed-forward 3D scenes (Bahmani et al., 2026). REST3D (Ma et al., 2026) adds a support tree and hierarchical CEM. We address the repeated physical search that follows geometry recovery.

Relational and physically plausible scene synthesis. Indoor generators model object arrangements through graph search (Wang et al., 2019), transformers (Wang et al., 2021; Paschalidou et al., 2021), graph-conditioned latent models (Dhamo et al., 2021; Zhai et al., 2023), or diffusion (Tang et al., 2024). Physical variants account for collision, reachability, or stable placement (Yang et al., 2024; Yoneda et al., 2023; Luo et al., 2026). We preserve an observed reconstruction while refining its arrangement under a measured simulation budget. SimuScene (Lee et al., 2026) diagnoses per-scene simulation failures to resample shape and correct gravity-axis geometry. ϕ-Scene (Li et al., 2026a) uses support-ordered assembly, signed-distance penetration

resolution, and rigid-body settling. Trace2Stable learns reusable pose hypotheses and organizes their physical verification while keeping reconstructed geometry fixed.

Generative amortized inference. Latent, score-based, and flow models represent multimodal distributions (Kingma & Welling, 2014; Ho et al., 2020; Song et al., 2021; Lipman et al., 2023; Tong et al., 2024); equivariant and manifold variants respect 3D pose geometry (Fuchs et al., 2020; Satorras et al., 2021; Corso et al., 2023; Chen & Lipman, 2024). Simulation-based inference (SBI) and amortized optimization reuse simulation or search experience (Cranmer et al., 2020; Gloeckler et al., 2024; Amos, 2023; Byravan et al., 2022), while selective prediction allocates test-time computation (Lakshminarayanan et al., 2017; Ovadia et al., 2019; Geifman & El-Yaniv, 2017; Zuo & Zhu, 2026). We combine these views: a learned policy allocates simulator queries whose outcomes, rather than an analytic constraint map (Li et al., 2026b) or learned dynamics surrogate (Sanchez-Gonzalez et al., 2020; Pfaff et al., 2021), determine acceptance.

3 METHOD

3.1 PROBLEM FORMULATION AND FRAMEWORK

Trace2Stable couples hypothesis generation and physical verification: search traces train reusable repair proposals, which simulation tests within a budget. Support exploration, candidate ordering, and rollout continuation direct tests toward unresolved repair candidates.

Inputs and outputs. We fix the meshes supplied by REST3D and infer object poses and support relations. The input comprises geometry, poses, and a support graph:

$$S^0 = (X, T^0, \mathcal{G}^0), \qquad X = \{\boldsymbol{x}_i\}_{i=1}^N, \quad T^0 = \{T_i^0 \in \mathrm{SE}(3)\}_{i=1}^N, \quad \mathcal{G}^0 = (\mathcal{V}, \mathcal{E}^0), \tag{1}$$

where $N = |\mathcal{V}|$ counts objects, $\mathrm{SE}(3)$ is the rigid-pose group, and $\boldsymbol{x}_i$ contains observation features, scaled collision geometry, and mobility. An edge $(j, i) \in \mathcal{E}^0$ means that j supports i; immobile objects provide collision context. We infer support graph $\hat{\mathcal{G}}$ and placement poses $\hat{T}$; independent terminal replay starts from the saved placements and measures their subsequent motion and collision.

For the optimized objects $\mathcal{V}_o$, the joint action has six coordinates per object:

$$\boldsymbol{\xi} = (\Delta \boldsymbol{t}_i, \boldsymbol{\omega}_i)_{i \in \mathcal{V}_o} \in \mathbb{R}^{6|\mathcal{V}_o|}, \tag{2}$$

where $\Delta \boldsymbol{t}_i$ updates translation and $\boldsymbol{\omega}_i$ parameterizes rotation. Local rotations are residual; Global rotations follow the optimizer's absolute convention. Since each correction requires physical testing, we seek final physical validity subject to an explicit budget on online simulation:

$$\min_{q_\theta, \pi_\phi} \ \mathbb{E}[1 - Y_H(\hat{\mathcal{G}}, \hat{T})] \quad \text{subject to} \quad B_{\mathrm{steps}}(\pi_\phi) \le B, \tag{3}$$

where q_θ generates pose candidates, π_ϕ selects simulation actions, and θ, ϕ are their learned parameters; the expectation covers scenes and sampled candidates. B is the online step budget, and B_{steps} counts candidate-environment steps over the executed simulation actions. Y_H is the success indicator at task-specific terminal horizon H, with its physical criteria specified in *Appendix A*. Separate losses train q_θ to model trace-derived repairs and π_ϕ to rank simulator tests using outcomes.

Overall inference. Figure 2 couples support and pose refinement: each object's candidate set $\mathcal{C}_i$ contains the floor and nearby supports. For legal branches $\mathcal{G}_b$ retained by the support posterior, the flow constructs a bank of joint actions conditioned on the chosen support graph:

$$\mathcal{B}_b = \{\boldsymbol{\xi}_{b,m}\}_{m=1}^{M_b}, \qquad \boldsymbol{\xi}_{b,m} \sim q_\theta(\cdot \mid \mathcal{G}_b, X, T^0), \qquad \mathcal{H} = \bigcup_b \{(\mathcal{G}_b, \boldsymbol{\xi}_{b,m})\}_{m=1}^{M_b}, \tag{4}$$

where b indexes graphs, m indexes candidates, and M_b is bank size. Verification tests the finite set $\mathcal{H}$. To reduce action dimension, Local inference fixes the parent while refining its children. Global inference moves roots while preserving child-relative poses. Full-scene verification includes every descendant so that inter-group collisions remain visible.

The furniture route tests 16 Global flow and 16 residual-CEM proposals. Algorithm 1 tracks Local work, budget B_q, query cost $\Delta b(a)$, and measured response z.

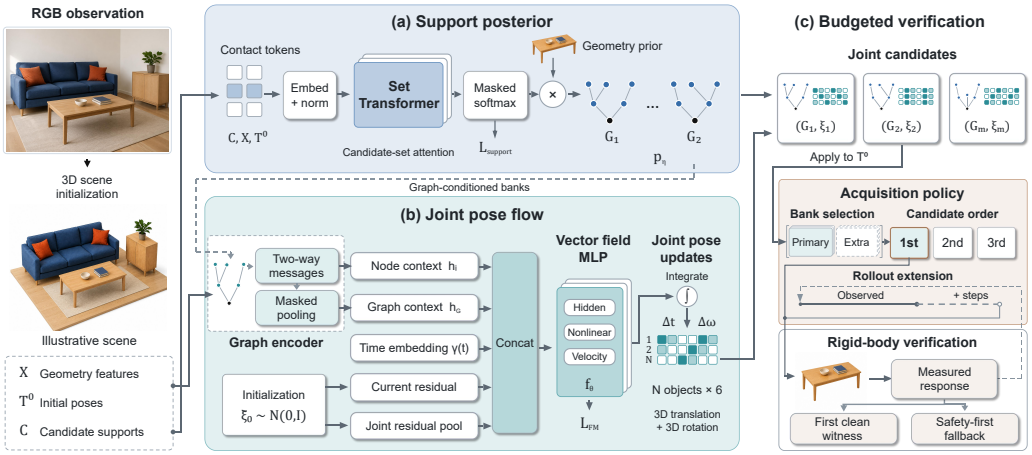


Figure 2: **Search traces yield hypotheses for physical verification.** (a) Support choices and (b) a graph-conditioned flow propose repairs. (c) Budgeted simulator tests order candidates, guide feedback, and continue rollouts when needed.

3.2 Learning Support and Pose Hypotheses

Support posterior. Candidate retrieval keeps multiple contact explanations available until verification resolves support and pose together. A set Transformer without positional encoding maps contact geometry and initial-pose features to logits ℓ_{ij} for $j \in \mathcal{C}_i$. The parent probability combines these logits with a geometry prior over candidate supports, π^{geo}_{ij}:

$$p_\eta(j \mid i, X, T^0) = \frac{\exp(\ell_{ij})\pi^{\text{geo}}_{ij}}{\sum_{k\in\mathcal{C}_i}\exp(\ell_{ik})\pi^{\text{geo}}_{ik}}, \qquad j \in \mathcal{C}_i. \tag{5}$$

Masks restrict normalization to valid parents, with a uniform prior in the evaluated scorers. Support cross-entropy $\mathcal{L}_{\text{support}}$ trains parameters η. For the controlled task's single ambiguous child, each retained parent defines a graph and its conditional pose distribution:

$$q(\mathcal{G}, \boldsymbol{\xi} \mid X, T^0) = p_\eta(\mathcal{G} \mid X, T^0) q_\theta(\boldsymbol{\xi} \mid \mathcal{G}, X, T^0). \tag{6}$$

Appendix I specifies candidate retrieval and the geometry- and image-conditioned support instances.

Empirical pose targets. To transfer search experience, we retain the alternatives explored by CEM rather than just its final answer (Rubinstein, 1999; Ma et al., 2026). Traces record actions, composite costs, collisions, and motion. The condition $c = (\mathcal{G}, T^0, X, \text{phase})$ distinguishes Local child refinement from Global root refinement. The preferred set $\mathcal{S}(c)$ uses clean candidates satisfying both motion and collision criteria, then motion-bounded candidates if none are clean. Local training retains motion-bounded groups; Global training falls back to the lowest-risk candidates when neither set is available. With indicator $\mathbf{1}[\cdot]$, these choices define the empirical target

$$p_w(\boldsymbol{\xi}_j \mid c) = \frac{\mathbf{1}[\boldsymbol{\xi}_j \in \mathcal{S}(c)] w_j}{\sum_k \mathbf{1}[\boldsymbol{\xi}_k \in \mathcal{S}(c)] w_k}. \tag{7}$$

Local targets are sampled uniformly. Let E_{trace} denote the recorded teacher penalty (*Appendix B*). The furniture Global model uses $w_j = \exp[-E_{\text{trace}}(\boldsymbol{\xi}_j)/\tau]$, with temperature τ selected on validation sources under an entropy floor. These weighted targets retain preferred alternatives, including feasible repairs when available, whose average could violate contact.

Graph-conditioned pose flow. Search traces contain distinct acceptable repairs, so flow matching samples pose alternatives for physical verification. Parent-to-child and child-to-parent messages encode how motion changes support and collision context, yielding object states h_i and a masked pooled scene state $h_{\mathcal{G}}$ (Zaheer et al., 2017; Veličković et al., 2018). For a normalized joint action $\tilde{\boldsymbol{\xi}}_t$, a shared action encoder g embeds each object's action; pooling these embeddings

Algorithm 1 Query stage of a configured Trace2Stable task

Require: conditioned hypotheses $\mathcal{H}$, query allowance B_q, task instance κ
Initialize the active bank from $\mathcal{H}$; set $\mathcal{O} = \varnothing$.
Fix bank prefixes and query order using the rule for κ.
while an allowed action remains **do**
  If κ uses feedback, append it once after initial measurements.
  Apply κ's bank-opening and continuation rules when their predicates hold.
  Let a be the next query allowed by κ; stop if $\Delta b(a) > B_q$.
  $z \leftarrow \mathrm{Sim}(a)$; $B_q \leftarrow B_q - \Delta b(a)$; $\mathcal{O} \leftarrow \mathcal{O} \cup \{(a, z)\}$.
  if κ uses witness stopping and z satisfies its online predicate **then**
    return the corresponding graph and saved placement
  end if
end while
return the measured selector for κ on $\mathcal{O}$; failure if $\mathcal{O} = \varnothing$

provides joint action context. With time embedding $\gamma(t)$, a multilayer perceptron (MLP) f_θ predicts the per-object velocity using graph and joint-action context:

$$v_\theta(\tilde{\boldsymbol{\xi}}_t, t \mid c)_i = f_\theta\Big(h_i, h_{\mathcal{G}}, g(\tilde{\boldsymbol{\xi}}_{t,i}), \mathrm{pool}_{j \in \mathcal{V}_o}\, g(\tilde{\boldsymbol{\xi}}_{t,j}), \gamma(t)\Big). \tag{8}$$

Translations and rotations use separate training scales. Let $\tilde{p}_w$ be the normalized target distribution. Conditional flow matching (CFM) (Lipman et al., 2023; Tong et al., 2024) draws $\tilde{\boldsymbol{\xi}}_1$ from this target, $\tilde{\boldsymbol{\xi}}_0$ from a standard Gaussian, and t uniformly between zero and one. Its path and target velocity are

$$\tilde{\boldsymbol{\xi}}_t = [1 - (1 - \sigma_{\min})t]\tilde{\boldsymbol{\xi}}_0 + t\tilde{\boldsymbol{\xi}}_1, \qquad \tilde{\boldsymbol{u}}_t = \tilde{\boldsymbol{\xi}}_1 - (1 - \sigma_{\min})\tilde{\boldsymbol{\xi}}_0, \tag{9}$$

where $\sigma_{\min}$ is the terminal noise scale; training minimizes

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\left[\|v_\theta(\tilde{\boldsymbol{\xi}}_t, t \mid c) - \tilde{\boldsymbol{u}}_t\|^2_{\mathcal{V}_o}\right], \tag{10}$$

where the masked squared norm averages over optimized-object coordinates. Integrating the learned field transports Gaussian samples into joint pose candidates. Local decoding composes rotations with the input pose; Global decoding uses the optimizer's absolute rotation convention. Both add world translation. *Appendix C* provides implementation details of the variable-object representation, flow architecture, and training procedure.

3.3 Budgeted Physical Verification

Candidate priority and measured selection. To avoid redundant tests after a clean candidate, a graph-conditioned model predicts clean feasibility, motion/contact risk, and priority s_ϕ. Classification, risk regression, and within-bank ranking train it separately from the flow (*Appendix Eq. 17*). Measurements determine acceptance: a clean witness satisfies the task's motion and collision criteria. Otherwise, safety-first selection prioritizes measured feasibility and collision before motion risk and learned score. The furniture comparison uses a measured quality selector and one residual-CEM update, allocating its 32 candidates equally to flow proposals and feedback proposals. The update fits a diagonal Gaussian to four measured elites and samples new joint actions rather than composing another transform onto earlier candidates. The final selector compares physically measured candidates from both rounds.

Selecting proposal banks. For each Local group g, a lower confidence bound on estimated full-bank feasibility selects the smallest sufficient prefix of the 16-candidate bank:

$$K_g^* = \min\big(\{16\} \cup \{K \in \{2, 4, 8\} : \mu_K - \beta\sigma_K \geq \tau_{\mathrm{suf}}\}\big), \tag{11}$$

where K is prefix size, μ_K and σ_K are its ensemble mean and standard deviation, and $\beta \geq 0$ penalizes dispersion. If none qualifies, we use 16. Sufficiency means retaining full-bank feasible-candidate availability. Validation selects β and τ_{suf} to retain at least 95% of full-bank-positive development groups. A separate dual-bank policy compares independently trained flow banks using a scale-normalized symmetric set distance D. It opens the second bank when the first has no clean witness and $D > \tau_D$, where validation fixes τ_D. *Appendix C.3* defines both.

Table 1: **Trace2Stable improves returned scenes across three tasks.** Stable+Safe rates compare each route with its task-matched reference; detailed protocols follow.

Task	Criterion ↑	Comparison method	Ref.	Ours
ReplicaCAD refinement	Stable+Safe	Diffusion + shared feedback	71.43	**80.95**
Hierarchical refinement	Stable+Safe	Full-scene CEM	21.88	**59.38**
Support–pose refinement	Stable+Safe	Fixed-graph Trace2Stable	37.50	**50.00**

Continuing an observed trajectory. A short rollout may leave a promising candidate unresolved. To decide which trajectories merit further simulation, we summarize paid observations at step h as drift, speed, support, and penetration in z_h, then estimate

$$v_\psi(z_h, h, n) = \Pr(Y_H = 1 \mid Y_h = 0, z_h, h, n), \tag{12}$$

where ψ denotes the recovery model's learned parameters, n is object count, and Y_h indicates physical success at horizon h; the continuation studies set the terminal horizon to 240 steps. Recovery ranks unresolved trajectories for additional paid steps, whose simulation determines the outcome. This action revisits an observed candidate rather than opening a new bank; *Appendix C.3* specifies its continuation rule and the observations available at each decision.

Full-scene verification and stopping. Composed Local groups can collide, so each Global candidate includes all transported descendants. The fixed-bank study queries five-step dynamics in priority order and accepts its first zero-illegal-collision witness, with motion and H240 success evaluated separately. Other task instances use their stated joint motion/collision witness. If the budget is exhausted, fallback minimizes measured collision and then uses priority to break ties. Within a fully scanned route, early stopping preserves the existence of a short-horizon witness; *Appendix Proposition 2* states the assumptions. Every continuation and candidate evaluation contributes its actual additional steps to the budget; the selected output undergoes independent terminal replay.

4 Experiments

4.1 Experimental Setup

Datasets and baselines. We ask whether learned proposals improve repair, how much simulation they require, and when changing support helps. ReplicaCAD uses 42 training, 20 validation, and 21 test groups. Separate controlled layouts isolate budget allocation and support changes; reconstructed and simulated indoor scenes test transfer. Table 1 gives a within-task overview. Baselines include REST3D, deterministic graph neural networks (GNNs), variational autoencoders (VAEs), diffusion, and learned CEM initialization.

Evaluation metrics. Our main endpoint is **Stable+Safe (S+S)**: every movable object meets the motion criterion and no illegal non-support collision remains. Stable tests translation and rotation drift against 0.1 m and 0.1 rad; ReplicaCAD H240 also checks speeds against 0.1 m/s and 0.1 rad/s. H240 denotes a 240-step replay of the selected placement, separate from online selection. Coll. reports collision percentage; Drift reports translation after placement in meters. Short-verifier columns describe online measurements, not terminal success. Figure badges report object rates rather than whole-scene success. Support tasks additionally test contact or input-pose retention (Stable+Input); their endpoints are compared within each task.

Implementation details. Controlled studies use PhysX through Isaac Gym (Makoviychuk et al., 2021) on NVIDIA L20 GPUs; HSSD and reconstruction replays use PyBullet. Validation selects the 1.78M-parameter furniture flow refit on 62 development groups; test groups are excluded. Paired intervals resample source clusters and retain paired observations.

Table 2: **Learned proposals improve ReplicaCAD success.** H240 results on 21 groups compare complete routes (a) and shared-feedback generators (b) at equal online cost.

	Proposal properties					Short verifier		H240 replay	
Method	Learn	Sample	Iter.	Update	Props. ↓	Stable ↑	Coll. ↓	S+S ↑	Drift ↓
(a) Archived proposal and complete-route evaluation									
REST3D released	×	×	×	✓	32	42.86	14.29	57.14	0.1585
Deterministic GNN	✓	×	×	×	**1**	47.62	23.81	57.14	0.1474
Conditional VAE	✓	✓	×	×	32	47.62	14.29	57.14	0.3174
Learned-CEM	✓	✓	×	✓	32	52.38	**0.00**	61.90	0.0972
Graph-Diffusion	✓	✓	✓	×	32	52.38	**0.00**	61.90	0.1323
Graph-CFM (proposal only)	✓	✓	✓	×	32	52.38	**0.00**	76.19	0.0662
Trace2Stable (ours)	✓	✓	✓	✓	32	**61.90**	**0.00**	**85.71**	**0.0559**
(b) Shared feedback: diffusion 15/21; Trace2Stable 17/21									
Graph-Diffusion + feedback	✓	✓	✓	✓	32	–	–	71.43	**0.0903**
Trace2Stable + shared feedback	✓	✓	✓	✓	32	–	–	**80.95**	0.1031

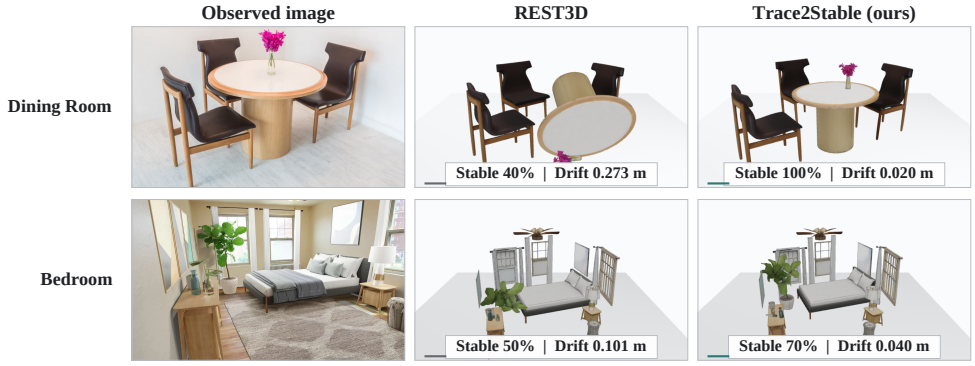


Figure 3: **Measured repairs stabilize released reconstructions.** Paired H120 replays use the same geometry and cameras; badges summarize object stability and drift.

4.2 Reusable Hypotheses and Returned Scene Quality

Image-derived furniture. The nine REST3D-released reconstructions (87 movable objects) test whether physical refinement repairs violations while retaining valid placements. Geometry and cameras remain fixed, so H120 replay measures the effect of the returned poses. Matched replay repairs 13/33 initially unstable objects and preserves 49/54 initially stable objects; five lose stability. Six scene outputs receive corrections, while three remain unchanged. The split captures cases in which correcting one object destabilizes another that already passes replay. The original placement stays in the candidate bank, so measured selection can preserve it when it passes verification. The paired reconstructions in Figure 3 illustrate repairs; Figure 4 contrasts a tabletop collision removal with an unchanged WorldLab placement. Its silhouette intersection over union (IoU) compares returned and released layouts from the same camera, with shared pixels measuring retained projected arrangement.

Matched reconstruction methods. The matched comparisons distinguish learned proposals from subsequent physical search. Table 2(a) lists proposal properties alongside returned quality. The columns Learn, Sample, Iter., and Update denote cross-scene learning, learned multi-hypothesis sampling, iterative neural refinement, and physical feedback updates, respectively. Sample refers to a trained proposal distribution; REST3D instead samples its CEM search distribution. The proposal-only flow improves returned quality over the earlier generators, and measured feedback improves it further without changing the learned flow. Table 2(b) gives both generators 16 proposals and 16 residual-CEM candidates, with identical selection and 1,536 simulation steps. The matched rerun favors flow over diffusion, although the paired source interval spans zero. Fixed selection and budget isolate proposal choice, while the interval reflects variation across the held-out scenes.

Table 3: **Matched refinement raises success at lower step cost.** (a) Complete routes; (b) shared-bank stopping. Wall time covers online inference.

Method	Layouts	Stable ↑	S+S ↑	Coll. ↓	Drift ↓	Rollouts ↓	Steps ↓	Full Q ↓	Wall (s)↓
(a) Scope-matched retrospective correction									
Scratch CEM, expanded actors	32	65.62	21.88	65.62	**0.0167**	80.00	4800.00	16.00	63.68
Trace2Stable (ours), expanded actors	32	**78.12**	**59.38**	**28.12**	0.0181	**39.56**	**2546.25**	16.00	63.66
(b) Same-bank acquisition component analysis									
Trace2Stable with exhaustive verification	32	**75.00**	**59.38**	**18.75**	**0.0105**	–	4093.75	8.00	–
Trace2Stable with learned witness ordering	32	71.88	56.25	**18.75**	0.0114	–	**4066.09**	**2.47**	–

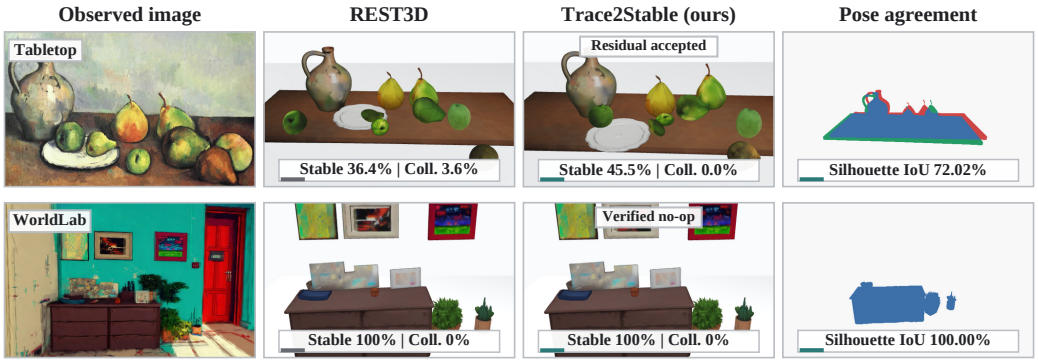


Figure 4: **Verification repairs failures and retains valid inputs.** One frozen policy accepts a tabletop correction and leaves the stable WorldLab placement unchanged. Blue marks shared silhouettes; red and green mark regions unique to the released and repaired layouts, respectively.

Table 4: **Support branching raises observed success on ambiguous layouts.** H240 outcomes cover 48 layouts; brackets give the paired 95% source-bootstrap interval.

Method	Structure	Stable ↑	S+S ↑	Coll. ↓	ΔS+S vs. fixed
Trace2Stable w/o graph posterior	fixed observed graph	66.67	37.50	54.17	reference
Trace2Stable (ours)	joint support–pose	**72.92**	**50.00**	**37.50**	+12.50 [−8.33, 33.33]

4.3 Simulation Budget and Physical Refinement

Steps sum Local and Global simulation, excluding terminal replay; a rollout advances one candidate, and Full Q counts full-scene queries. Table 3(a) shows a higher S+S rate with fewer simulator steps than scratch CEM but similar wall time, so step savings do not yet yield faster end-to-end inference. With the bank fixed in Table 3(b), ordering cuts full-scene queries, while shared Local work limits total-step savings and terminal success falls slightly. H5 checks collision, whereas H240 also tests motion; short-witness acceptance carries that horizon trade-off.

4.4 Resolving Structural Ambiguity

Alternative supports. Support ambiguity couples contact assignment and pose refinement. Across controlled layouts, branching improves joint success and reduces collisions relative to a fixed graph (Table 4). It requires roughly four more rollouts per layout because each branch pays for Local refinement before Global verification. The interval is broad; drift includes failed repairs.

An HSSD replay contrasts retained support with support removal and floor placement (Figure 5). The H0/H16/H240 poses trace the unsupported fall, while markers compare input-relative distances before and after replay. Verified support meets both stability and input retention for all 15 children; the alternatives fail this joint test. Removing support retains the initial pose but lets the object fall; floor placement moves little during replay because it starts far from the input pose. Retaining support instead limits motion and preserves the observed layout.

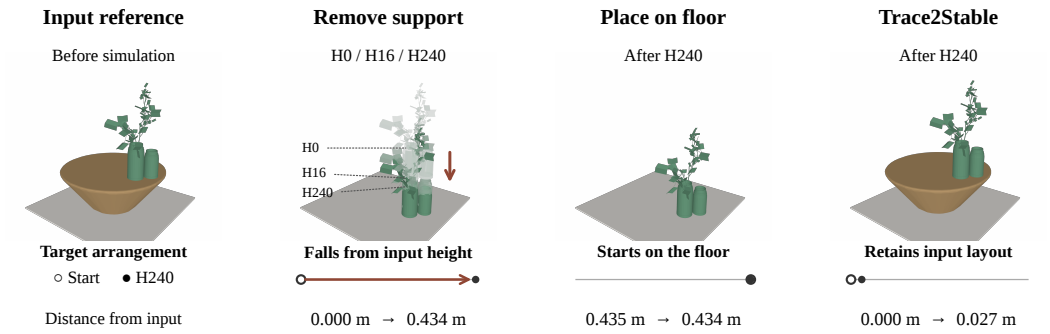


Figure 5: **Preserving support limits motion and pose change.** Shared-camera panels contrast falling, floor placement, and retained contact at H0/H16/H240. Markers show input-relative displacement before and after replay.

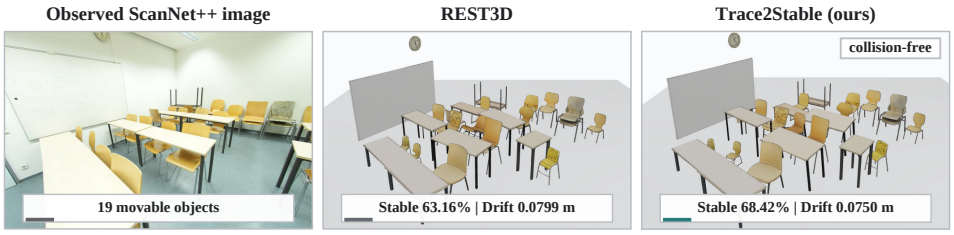


Figure 6: **Measured selection improves stability in a crowded scene.** The 19-object ScanNet++ I reconstruction uses the released camera and H120 replay; refinement stabilizes more objects without introducing detected pair collisions in the returned arrangement.

The two failed controls finish at nearly the same input-relative distance, despite different trajectories. The terminal pose alone therefore cannot distinguish a loss of contact from a changed initial arrangement. The joint test matters because low replay motion can coexist with a changed starting layout, while an initially faithful pose can still fall. The retained-support branch satisfies both requirements, linking the contact decision to a repair that remains close to the reconstruction after replay.

Crowded reconstruction. The ScanNet++ I reconstruction extends the evaluation to a dense scene (Figure 6). Under matched H120 replay, refinement improves object-level stability while keeping the returned layout free of detected pair collisions. The higher stability rate accompanies a modest decline in mean translation drift. The stability metric thresholds each object's translation and rotation, while drift averages translation across the scene. Remaining unstable objects keep this case below the all-object endpoint, separating local repair from whole-scene success. A binary scene score hides this partial repair; object-level stability reveals gains before the whole scene passes. Together, the two cases show why verification must assess both local contact choices and the remaining failures of the complete reconstructed scene.

5 Conclusion

Trace2Stable learns pose alternatives from physical search traces and verifies them with measured dynamics under a fixed budget. Support branches and flow proposals retain alternative contacts and poses for verification. On 32 scope-matched layouts, the S+S rate rises from 21.88% to 59.38% while online simulation steps fall by 46.95% relative to scratch CEM. Across nine released reconstructions, H120 replay repairs 13 of 33 initially unstable objects and preserves 49 of 54 initially stable objects. Candidate-bank analysis separates two remaining bottlenecks: generating a valid repair and selecting it through short physical tests. Future work will investigate how reusable physical experience scales beyond indoor scenes and how verification remains reliable under imperfect contact models and simulator mismatch.

AI USE STATEMENT

Generative AI tools, including large-language-model-based coding assistants, support this work in four areas: writing and language polishing; literature retrieval, discovery, and summarization; research ideation and execution; and drafting sections of the manuscript. Research assistance includes hypothesis refinement, conceptual and mathematical presentation, experimental design, implementation and debugging, data preparation, simulation automation, and analysis of results. AI tools also help prepare scientific figures and tables and check consistency between the manuscript and experimental records. The authors set the research aims, assess methodological choices, and direct manuscript revisions; AI-generated suggestions and drafts inform these decisions. Reported quantitative results come from executed training, simulation, and evaluation programs, including AI-assisted implementations, and are checked against saved outputs. Suggested references are verified against their original sources. The authors retain responsibility for reviewing the accuracy, attribution, and validity of AI-assisted text, code, figures, and analyses, and for the final claims and artifacts submitted with this paper.

ETHICS STATEMENT

This work uses released research datasets, scene assets, and reconstruction examples under their associated terms. The study involves neither recruited human participants nor newly collected personal data. The method is a research prototype for reconstructing physically plausible indoor scenes; its simulator-based acceptance criterion is not a real-world safety certificate. Failures in geometry, support inference, contact parameters, or simulator transfer can still produce collisions or unstable arrangements, so deployment in robotics or other safety-critical settings requires independent real-world validation and appropriate safeguards. Training and physical simulation also consume computational energy. The proposed amortized inference and selective verification reduce simulator rollouts in the reported scratch-search comparisons; training and evaluation costs are reported separately. We follow the ICLR Code of Ethics and emphasize transparent reporting of limitations, source-disjoint evaluation, and unsuccessful variants.

REPRODUCIBILITY STATEMENT

The model, objective, and physical acceptance rules are specified in Sections 3–4. *Appendix A* defines simulator budgets, terminal metrics, data splits, and statistical analysis; *Appendices B and C* specify hierarchical inference, empirical training targets, architecture, losses, and optimization. Complete furniture, graph–pose, and HSSD proposal-transfer protocols appear in *Appendices G, I, and J*. The anonymous manuscript package contains the LaTeX source and figure assets. The experiment descriptions identify evaluation populations, simulated geometry, online budgets, terminal horizons, and model-selection data. Experimental outcomes are summarized in the tables and analyses for controlled allocation, candidate neighborhoods, and transfer. Paired perturbations remain within their source clusters in the reported statistical analysis. The model definitions and protocols specify training and simulator evaluation.

REFERENCES

Brandon Amos. Tutorial on amortized optimization. *Foundations and Trends in Machine Learning*, 16(5):592–732, 2023.

Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David B. Lindell, Zan Gojcic, Sanja Fidler, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra: Generative 3D scene reconstruction via video diffusion model self-distillation. In *International Conference on Learning Representations*, 2026.

Arunkumar Byravan, Leonard Hasenclever, Piotr Trochim, Mehdi Mirza, Alessandro Davide Ialongo, Yuval Tassa, Jost Tobias Springenberg, Abbas Abdolmaleki, Nicolas Heess, Josh Merel, and Martin Riedmiller. Evaluating model-based planning and planner amortization for continuous control. In *International Conference on Learning Representations*, 2022.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *International Conference on Learning Representations*, 2024.

Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3D holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *IEEE/CVF International Conference on Computer Vision*, pp. 8648–8657, 2019.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. DiffDock: Diffusion steps, twists, and turns for molecular docking. In *International Conference on Learning Representations*, 2023.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-scale embodied AI using procedural generation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Helisa Dhamo, Fabian Manhardt, Nassir Navab, and Federico Tombari. Graph-to-3D: End-to-end generation and manipulation of 3D scenes using scene graphs. In *IEEE/CVF International Conference on Computer Vision*, pp. 16352–16361, 2021.

Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. SE(3)-transformers: 3D roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems*, 2020.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Manuel Gloeckler, Michael Deistler, Christian Dietrich Weilbach, Frank Wood, and Jakob H. Macke. All-in-one simulation-based inference. In *International Conference on Machine Learning*, pp. 15735–15766, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, and Song-Chun Zhu. Holistic 3D scene parsing and reconstruction from a single RGB image. In *European Conference on Computer Vision*, pp. 187–203, 2018.

Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X. Chang, and Manolis Savva. Habitat synthetic scenes dataset (HSSD-200): An analysis of 3D scene scale and realism tradeoffs for ObjectGoal navigation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16384–16393, 2024.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Aniruddha Kembhavi, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An interactive 3D environment for visual AI. *arXiv preprint arXiv:1712.05474*, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Inhee Lee, Sangwon Baik, Sungjoo Kim, Hyeonwoo Kim, Hyunsoo Cha, and Hanbyul Joo. SimuScene: Simulation-ready compositional 3D scene reconstruction from a single image. *arXiv preprint arXiv:2606.03994*, 2026.

Haodong Li, Lulu Shao, Haolin Lu, Yu Fu, Yen-Ru Chen, Seemandhar Jain, and Manmohan Chandraker. ϕ-Scene: Physically grounded image-to-3D scene reconstruction. *arXiv preprint arXiv:2606.21596*, 2026a.

Xinpeng Li, Enming Liang, and Minghua Chen. Gauge flow matching: Efficient constrained generative modeling over general convex set and beyond. In *International Conference on Learning Representations*, 2026b.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Zhen Luo, Yixuan Yang, Xudong Xu, Jinkun Hao, Zhaoyang Lyu, Feng Zheng, Jiangmiao Pang, and Yanwei Fu. STABLE: Simulation-ready tabletop layout generation via a semantics-physics dual system. *arXiv preprint arXiv:2605.16137*, 2026.

Xiaoxuan Ma, Jiashun Wang, Nicolas Ugrinovic, Yehonathan Litman, and Kris Kitani. REST3D: Reconstructing physically stable 3D scenes from a single image. *arXiv preprint arXiv:2605.30338*, 2026.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac Gym: High performance GPU-based physics simulation for robot learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3DUnderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 55–64, 2020.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, 2019.

Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. ATISS: Autoregressive transformers for indoor scene synthesis. In *Advances in Neural Information Processing Systems*, 2021.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W. Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021.

Reuven Y. Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability*, 1:127–190, 1999.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter W. Battaglia. Learning to simulate complex physics with graph networks. In *International Conference on Machine Learning*, pp. 8459–8468, 2020.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *International Conference on Machine Learning*, pp. 9323–9332, 2021.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. DiffuScene: Denoising diffusion models for generative indoor scene synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20507–20518, 2024.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Kai Wang, Yu-An Lin, Ben Weissmann, Manolis Savva, Angel X. Chang, and Daniel Ritchie. PlanIT: Planning and instantiating indoor scenes with relation graph and spatial prior networks. *ACM Transactions on Graphics*, 38(4):132:1–132:15, 2019.

Xinpeng Wang, Chandan Yeshwanth, and Matthias Nießner. SceneFormer: Indoor scene generation with transformers. In *International Conference on 3D Vision*, pp. 106–115, 2021.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. PhyScene: Physically interactable 3D scene synthesis for embodied AI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16262–16272, 2024.

Yu-Ying Yeh, Zhengqin Li, Yannick Hold-Geoffroy, Rui Zhu, Zexiang Xu, Miloš Hašan, Kalyan Sunkavalli, and Manmohan Chandraker. PhotoScene: Photorealistic material and lighting transfer for indoor scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18562–18571, 2022.

Takuma Yoneda, Tianchong Jiang, Gregory Shakhnarovich, and Matthew R. Walter. 6-DoF stability field via diffusion models. *arXiv preprint arXiv:2310.17649*, 2023.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems*, 2017.

Guangyao Zhai, Evin Pınar Örnek, Shun-Cheng Wu, Yan Di, Federico Tombari, Nassir Navab, and Benjamin Busam. CommonScenes: Generating commonsense 3D indoor scenes with scene graph diffusion. In *Advances in Neural Information Processing Systems*, 2023.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5745–5753, 2019.

Bowen Zuo and Yinglun Zhu. Strategic scaling of test-time compute: A bandit learning approach. In *International Conference on Learning Representations*, 2026.

Trace2Stable: Learning from Physical Search for 3D Scene Refinement

Appendix

Anonymous authors

SUPPLEMENTARY CONTENTS

Reading map. Three linked parts organize method details, experimental results, and discussion.

Part I: Method (A–D), p. 14 **Part II: Results** (E–J), p. 20 **Part III: Discussion**, p. 32

Section index.

A: Evaluation protocol	14	**F**: Scene complexity	20
B: Hierarchical physical inference	15	**G**: ReplicaCAD proposal learning	22
C: Model architecture and training	16	**H**: ScanNet++ transfer	27
D: Derivations and properties	19	**I**: Support–pose inference	27
E: Qualitative results	20	**J**: HSSD proposal transfer	31

Highlights.

Learning reusable hypotheses. Section C defines the empirical pose distribution and training. Sections G and H examine **candidate coverage, selection, and transfer** on reconstructed layouts.

Allocating physical evidence. Sections F and C.3 separate **opening proposal banks** from **continuing observed trajectories**. Long-horizon replay in Section F measures checkpoint validity.

Inspecting physical behavior. Section E illustrates **repairs, verified unchanged layouts, and residual failures**. Sections I and J connect support choices to pose correction and terminal pose alternatives.

PART I: METHOD DETAILS AND ANALYSIS

Physical objectives, empirical proposal learning, and finite-budget inference.

A EVALUATION PROTOCOL

Task-specific execution. Algorithm 1 instantiates task-specific queries. Furniture measures 16 Global flow and 16 residual-CEM proposals at H48, then applies the validation-fixed quality selector. Hierarchical refinement uses Local prefixes and descendant-inclusive Global CEM. Shared-bank querying follows descending learned priority, stopping at a zero-illegal-collision H5 witness with collision-first fallback. Each support branch pays Local and Global costs. A second bank opens when the primary has no witness and $D > \tau_D$; continuation ranks observed unresolved trajectories by recovery probability. H240 independently evaluates saved placements.

A.1 PHYSICAL BUDGET AND VERIFICATION

A rollout advances one layout for a stated horizon; eight vectorized candidates count as eight rollouts. For scene s, candidate-environment work is

$$B_s^{\text{steps}} = \sum_{e \in \mathcal{E}_s} n_e \Delta h_e, \tag{13}$$

where $\mathcal{E}_s$ contains executed simulation events, n_e is the number of environments advanced, and Δh_e counts actual steps. Continuing H30 to H60 costs 30 steps; restarting to H60 costs 60. Local, Global, and continuation work all contribute. Controlled Local evaluation uses H60 and full-scene acquisition uses H5; furniture verification uses H48 with selective H120

replay. H240 evaluates the frozen output independently. Online budgets exclude this evaluation unless explicitly stated. Table 3(a) reduces physics work by 46.95%, while measured online latency is 63.68 s for CEM and 63.66 s for Trace2Stable. Query counts, physical steps, and elapsed time therefore quantify distinct aspects of efficiency.

Synchronized timing diagnostic. Four fixed eight-object inputs receive two fresh-process requests per method on an L20, with warm filesystem caches, alternating method order, and GPU synchronization. REST3D/Trace2Stable mean times are 34.08/29.72 s overall, 8.70/8.80 s for simulator initialization, 4.46/3.08 s for advancement, and 12.17/8.43 s for geometric collision tests; Trace2Stable proposal generation takes 0.34 s. Operation times are inclusive and nonadditive. This diagnostic identifies geometry and initialization as substantial costs; the 32-layout latencies above retain their original measurement conditions.

A.2 Terminal Physical Metrics

Terminal replay evaluates the selected arrangement without changing the selection. All layouts start from frozen initial root states. Declared parent–child contacts are legal support; child–child and unrelated penetrations remain violations. Each endpoint specifies its physical horizon. In the main tables, Props. counts proposals, Coll. is the collision percentage, and drift is measured in meters. Full Q counts full-scene queries; Stable and S+S are percentages. Table dashes denote unreported quantities or operations absent from a method; they are distinct from measured zeros.

Name	Definition	Use
Stable	Translation drift ≤ 0.1 m and rotation drift ≤ 0.1 rad for every movable object; ReplicaCAD H240 also requires linear/angular speeds ≤ 0.1 m/s and rad/s.	Scene-level motion endpoint; object rates are labeled separately.
Stable+Safe (S+S)	The cohort's motion endpoint and zero illegal non-support collision pairs.	Terminal endpoint under the stated cohort predicate.
Stable+Input	Stable and within the frozen tolerance from the Stage-3 input pose.	HSSD pose-retention endpoint.
Clean	Stable, contact-valid, and without illegal penetration under the stated support protocol.	HSSD candidate analysis.
Contact	Required child–parent contact is retained after replay.	Structural diagnostic.
H5–H120	Short checkpoint used for ordering or continuation.	Online / multi-fidelity evidence.
H240	Four-second independent replay at 60 Hz.	Primary terminal evaluation.
H960	Sixteen-second continuation of the same frozen H240 layout.	Delayed-instability audit.

A.3 Data Splits and Statistical Analysis

To retain the dependence between repeated observations, all candidates and perturbations of one geometry share a source cluster. Paired effects are aggregated within source before averaging across sources. Confidence intervals (CIs) use 20,000 source-cluster bootstrap samples unless a table specifies otherwise; Wilson intervals describe unpaired binary rates (Wilson, 1927). Models and thresholds are selected on the development sources. The repeated-source allocation trials have a time-ordered development role, as detailed in Section G.3.

Allocation uses three trials on the same 21 ReplicaCAD sources and a separate 16-source procedural cohort. Complexity tests apply four perturbations per source, while structural refinement evaluates 48 independently generated layouts with ambiguous support relations.

B Hierarchical Physical Inference

Local inference corrects children relative to their fixed support and resolved context. Global inference moves group roots while preserving descendant transforms. The resulting traces record actions, costs, motion, and contacts. The released Global selector simulates roots and ungrouped movable objects. The scope-matched experiment in Table 3(a) expands both methods to all descendants: two optimized roots, eight simulated objects, and 16 Global candidates on 16 procedural sources with two perturbations each. Under this common scope, the Stable+Safe gain is 37.50 points (95% source-bootstrap interval $[12.50, 59.38]$). Unconditional drift changes from 0.0167 to 0.0181 m, while feasibility improves and physical work decreases.

Local groups follow tree order with per-group candidate caps; Global verification uses its own population. All executed stages contribute to cost. The furniture route uses two populations of 16 H48 queries, totaling 1,536 candidate-environment steps.

Teacher cost and target weights. The furniture teacher uses this negative reward:

$$E_{\text{trace}} = D_t + D_R + 6L_t + L_R + V_{12} + 0.5C_{48} + 0.5C_0, \tag{14}$$

where D_t, D_R sum centroid translation (meters) and geodesic rotation (radians) from placement to H48 across the simulated objects. L_t, L_R sum the same distances from the reconstruction reference to H48. V_{12} sums linear speeds at H12 in meters per second. C_0, C_{48} count intersecting convex-hull pairs at placement and H48 using the geometric collision test, before support-pair filtering. Objects without a valid hull contribute no pair flag. The weighted sum is an optimization penalty with heterogeneous units. Within the preferred set, sampling uses $\mathrm{softmax}[-(E_{\text{trace}} - E_{\min})/0.25]$; subtracting its minimum preserves normalized weights. Independent H240 success supplies the evaluation objective, distinct from this H48 teacher cost.

C Model Architecture and Training

C.1 Variable-Object Graph Representation

Each example pairs a pre-action Local graph with physically evaluated actions. For N scene objects and N_o optimized objects, node features lie in $\mathbb{R}^{N\times 49}$ and each action lies in $\mathbb{R}^{N_o\times 6}$. Directed support edges encode parent–child context. The 49 node channels include normalized half extents, canonical centroid/root offsets, live state in the group-root frame, continuous 6D relative rotation, linear/angular velocity, world height, node roles, prefix progress, graph-size normalization, gravity in the root frame, and missing-state indicators. Unsimulated siblings carry a missing-state indicator. Current-object features use the pre-action state, and observed dynamics enter only after their query.

C.2 Graph-Conditioned Conditional Flow

Let $c = (\mathcal{G}, T^0, X, \text{phase})$ denote one pre-action condition and $\mathcal{S}(c)$ its finite preferred set of stable or low-risk CEM actions. Local training samples this clean-first, motion-fallback set uniformly; the Table 2 Global model uses the energy weights in Eq. 7. An action $\boldsymbol{\xi}_1$ from this target measure is mapped to normalized coordinates $\tilde{\boldsymbol{\xi}}_1$, while $\tilde{\boldsymbol{\xi}}_0$ is standard Gaussian noise. Equations 9 and 10 give the interpolation, target velocity, and training loss for both Local and Global flows. The terminal noise scale controls the path's residual spread; flow time describes interpolation, not simulation. For a training batch, the masked norm divides the sum of squared coordinate errors by $6\sum_{b,i} M_{bi}$, where M_{bi} selects optimized, unpadded objects in example b. Thus each valid coordinate has equal weight; the training loader supplies nonempty optimized-object sets. The variable t is interpolation time in $[0, 1]$, distinct from physical horizon h. The Local model uses directed parent–child messages, 128 hidden channels, three layers, 16-dimensional category/relation embeddings, and eight time frequencies. A shared graph encoding and 16-step midpoint solver transport K Gaussian samples in parallel. Shared-bank querying generates eight root hypotheses from directed graph, root, and pooled scene features; the actor-scope comparison instead uses scratch Global CEM. The furniture model has 192 channels and four layers. Its 41-epoch refit uses 15,872 candidates from 62 development sources; diffusion uses the same data for 86 epochs. Flow and diffusion fitting take 31.67 and 52.17 s, respectively; 21 evaluation sources remain held out.

Actions use three translation and three exponential-coordinate channels. The Local flow decodes $\boldsymbol{\xi} = S\tilde{\boldsymbol{\xi}}$ with a six-dimensional diagonal scale and applies $R_i = R_i^0 \exp([\Delta\boldsymbol{\omega}_i]_\times)$. The full-context Global flow decodes $\boldsymbol{\xi} = \boldsymbol{\mu} + S\tilde{\boldsymbol{\xi}}$, adds translation in world coordinates, and follows the released optimizer's absolute rotation convention $R_i = \exp([\boldsymbol{\omega}_i]_\times)$. Continuous 6D rotation (Zhou et al., 2019) is used as a conditioning feature; generated actions use gravity-frame-aware product tangent coordinates. The Local and Table 2 flows use frozen six-dimensional action scales; the full-context Global flow uses the mean and standard deviation fitted on its training actions. Masks are applied to messages, pooling, and flow loss; Local decoding masks padded objects, while padded full-context roots are ignored downstream. Here S repeats the six translation/rotation scales over optimized objects,

and $\boldsymbol{\mu}$ is the training-action mean for the centered decoder. R_i^0 and R_i denote input and decoded rotations; $[\cdot]_\times$ forms the skew-symmetric cross-product matrix, and exp is the matrix exponential.

Figure 2(b) summarizes this conditional field. "Two-way messages" encode both edge directions; masked pooling forms graph context $h_{\mathcal{G}}$. "Concat" joins node and graph context, current-action embeddings, their joint pool, and time embedding $\gamma(t)$. Gaussian initialization supplies the starting actions. The hidden and nonlinear layers of f_θ predict flow velocity; integration produces six coordinates per optimized object, with padding masked. $\mathcal{L}_{\mathrm{FM}}$ supervises this field. The diagram's $\Delta\omega$ denotes Local rotation residuals; Global decoding uses the absolute convention above.

Feedback proposal operator. The shared-feedback comparison first measures 16 generated actions for 48 steps. A feasibility-first lexicographic order, converted to rank rewards, selects four elites. Coordinatewise elite means and population standard deviations define a diagonal Gaussian:

$$\mu_d^+ = \tfrac{1}{4}\sum_{a\in\mathcal{E}} \xi_{a,d}, \qquad \sigma_d^+ = \sqrt{\tfrac{1}{4}\sum_{a\in\mathcal{E}} (\xi_{a,d} - \mu_d^+)^2} + 10^{-5}, \tag{15}$$

where $\mathcal{E}$ is the four-element elite set and d indexes joint action coordinates. Sampling forms a 32-action pool and evaluates its first 16 entries, with the first entry set to μ^+. The update uses no momentum or elite carryover. Gaussian feedback applies no hard clipping; its spread follows the measured elites. Translation remains relative to input roots and rotation uses absolute exponential coordinates. The final quality selector compares measured candidates across both iterations. Feedback updates the proposal distribution rather than composing another transform onto a candidate.

Action-set distances. To measure how well proposals represent the teacher's alternatives, consider proposal set $\mathcal{Q}$ and nonempty preferred target set $\mathcal{Y}$. Let $d(y, q)$ be the root mean squared difference over valid objects' six action coordinates after division by the frozen action scales. For Local groups, we report the following set distances:

$$\begin{aligned}
\mathrm{BestPair} &= \min_{y\in\mathcal{Y}, q\in\mathcal{Q}} d(y,q),\\
\mathrm{Coverage} &= |\mathcal{Y}|^{-1} \sum_{y\in\mathcal{Y}} \min_{q\in\mathcal{Q}} d(y,q),\\
\mathrm{Precision} &= |\mathcal{Q}|^{-1} \sum_{q\in\mathcal{Q}} \min_{y\in\mathcal{Y}} d(y,q),\\
\mathrm{Diversity} &= \binom{|\mathcal{Q}|}{2}^{-1} \sum_{q<q'} d(q,q'),
\end{aligned} \tag{16}$$

where the last sum ranges over unordered proposal pairs and is zero for a singleton. Distances are dimensionless action-coordinate errors, averaged over units within each source and then over sources. In the 96-source Local development evaluation, Gaussian prior and CFM both use 16 proposals: mean Coverage is 0.8750 versus 0.8388 across three training seeds, a 4.14% reduction. This distance-based Coverage differs from scene-level candidate availability. At equal bank size, this comparison isolates learning; we evaluate deterministic one-candidate prediction separately.

C.3 Proposal Value and Set Sufficiency

Query priority must distinguish likely acceptance from the severity of a possible failure. For candidate a, a graph-conditioned head therefore predicts clean feasibility, component-wise risk, and a scalar ordering score. Binary cross entropy, robust risk regression, and within-bank pairwise ranking jointly supervise these predictions through the following value objective:

$$\mathcal{L}_{\mathrm{value}} = \lambda_c \,\mathrm{BCE}(z_a, q_a) + \frac{\lambda_e}{3}\sum_k \mathrm{SL1}(\log(1+\hat{r}_{a,k}), \log(1+r_{a,k})) + \lambda_r \mathbb{E}_{q_a > q_b}[m - s_a + s_b]_+, \tag{17}$$

where z_a is a clean-feasibility logit, $q_a \in \{0,1\}$ its measured target, and $\hat{r}_{a,k}, r_{a,k}$ are predicted and observed nonnegative risks for pose, residual motion, and contact ($k = 1,2,3$). BCE is class-weighted binary cross entropy on logits; SL1 is smooth L1 loss. The weights $\lambda_c, \lambda_e, \lambda_r$

are nonnegative, m is the ranking margin, and $[x]_+ = \max(x, 0)$. Scores s_a, s_b refer to the priority s_ϕ of two valid candidates in one bank; only pairs with $q_a > q_b$ contribute. The loss averages valid candidates and uses zero ranking loss when no such pair exists. Specifically, the positive BCE weight is the batch negative-to-positive count ratio clipped to $[1, 20]$, with the positive denominator bounded below by one. SL1 uses transition parameter one and averages valid candidates; the three component losses then receive equal weight. The model uses softplus for nonnegative risk predictions and clips observed risks below at zero before $\log(1 + r)$. All valid clean–nonclean ordered pairs within each bank enter the ranking average. The default weights are $(1, 0.5, 0.25)$ and $m = 0.25$. Priority is $s_a = z_a - \hat{r}_{a,1} - \hat{r}_{a,2} - 0.5\hat{r}_{a,3}$.

For characteristic object scale L, pose risk is the maximum of translation drift divided by $0.05L$ and rotation drift divided by ten degrees. Motion risk is the maximum of linear speed divided by $0.05\sqrt{9.81L}$ and angular speed divided by $0.15\sqrt{9.81/L}$. Contact risk counts twice the newly appearing undesired pairs plus persistent undesired pairs. These quantities use the evolving prefix and current objects; fixed support ancestors provide collision context. The dynamic-risk predictor and pose flow use independent training objectives matched to their respective prediction targets.

Local allocation asks how many alternatives a group needs, rather than whether one particular proposal succeeds. For nested prefixes $\mathcal{B}_2 \subset \mathcal{B}_4 \subset \mathcal{B}_8 \subset \mathcal{B}_{16}$, the label

$$F_K = \mathbf{1}\{\exists m \le K : w_m = 1\}, \qquad S_K = \mathbf{1}\{\neg F_{16} \vee F_K\}, \tag{18}$$

indicates that prefix K is sufficient relative to the complete bank. w_m is the measured graph-feasibility label: bounded pose and residual motion with no new or persistent undesired contact. F_K indicates that prefix K contains such a candidate; $\neg$ and $\vee$ denote logical negation and disjunction. Thus an infeasible complete bank yields a sufficient-prefix label without changing its physical success label. Binary cross entropy against S_K defines $\mathcal{L}_{\text{suf}}$. A 64-member source-bootstrap ensemble estimates mean μ_K and standard deviation σ_K for $K \in \{2, 4, 8\}$. The chosen budget is the smallest prefix whose dispersion-penalized statistic meets the calibration threshold, with $K = 16$ as fallback. Since all three predictions precede simulation, the first threshold crossing equals that of their running maximum, even for nonmonotone scores. Here $S_{16} = 1$ and F_K is nondecreasing; S_K labels relative sufficiency rather than absolute feasibility.

Flow, value, and prefix-sufficiency losses are optimized independently. For posterior routing, two equal-size joint-pose banks are compared by the symmetric nearest-neighbor distance shown in Figure 7(a), after applying the frozen translation–rotation action scale.

Figure 7 links three distinct decisions: which bank to test, which trajectory to continue, and when to stop. Purple circles and blue-gray triangles denote two proposal banks; dashed links are nearest-neighbor matches. For physical actions ξ, ζ, define $\rho(\xi, \zeta) = \|S^{-1}(\xi - \zeta)\|_2 / \sqrt{6|\mathcal{V}_o|}$. The scale S is applied once. Directed distance d_{12} averages the minimum ρ from each primary-bank sample to the secondary bank; d_{21} reverses the banks. Their mean D is dimensionless. The second bank opens when $D > \tau_D$ and the primary has no witness. In panel (b), h is the paid horizon and Δh is additional simulation. Observed drift, speed, support, penetration, horizon, and object count condition the Platt-calibrated recovery estimate v_ψ. In panel (c), crosses mark failures, the outline marks the current query, and pale cards are unqueried. The dashed "Not yet accepted" path returns unresolved observations to continuation; its added steps are charged before verification resumes. The inset's $\mathcal{L}_{\text{suf}}$ supervises relative sufficiency for nested Local prefixes $K = 2, 4, 8, 16$, with $K = 16$ as fallback. "Clean witness" names the configured online predicate, including the collision-only H5 predicate in the fixed-bank study. "Safety-first fallback" returns the best measured candidate under that task's rule, rather than certifying terminal success. Figure 2(c) summarizes these alternatives as bank selection, candidate order, and rollout extension.

Training and scale. AdamW fits the Local model on 34 training and eight validation sources; validation selects epoch 44. A separate scale study trains three seeds on 288 procedural sources and evaluates 96 validation sources. Sampling multiple feasible targets reduces coverage distance by 10.9% relative to capacity-matched deterministic regression; six-object scenes follow the same ordering.

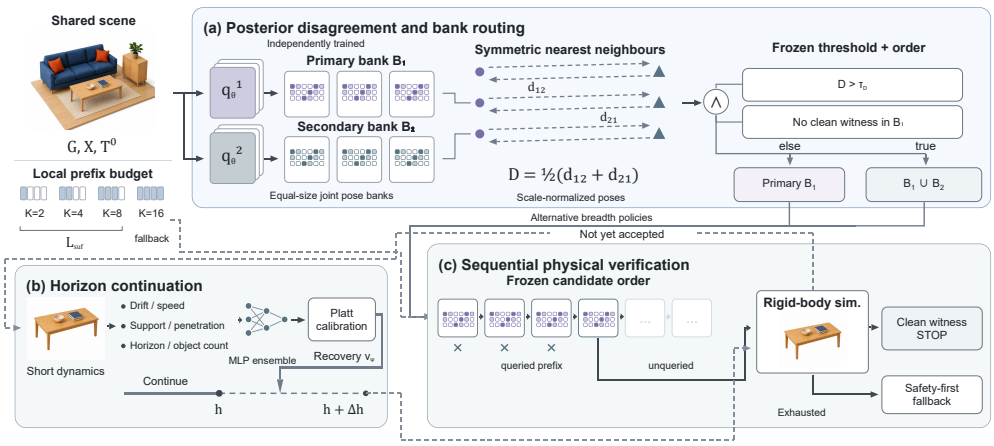


Figure 7: **Physical evidence directs each additional query.** (a) Bank disagreement opens alternatives; (b) recovery scores extend observed trajectories; and (c) a measured witness stops the scan. The inset shows Local-prefix allocation.

D Derivations and Computational Properties

D.1 Finite-Bank Constraint Preservation

Let $c(\boldsymbol{\xi}) \in \mathbb{N}_0$ count illegal pairs under a fixed candidate verifier, and let $\mathcal{B}$ be a fixed finite bank. A larger learned priority s_{ord} induces an earlier query. Table 3(b) stops at the first five-step $c = 0$ witness. If none is found and the budget permits a full scan, it returns the lexicographic minimum of $(c, -s_{\text{ord}}, \text{id})$. The following collision property also holds for stricter witnesses that imply $c = 0$, provided the same fallback is used.

Proposition 1. *Assume the budget permits a full scan when no early witness is found. For any ordering π, the selected candidate $\hat{\boldsymbol{\xi}}$ then satisfies*

$$c(\hat{\boldsymbol{\xi}}) = \min_{\boldsymbol{\xi} \in \mathcal{B}} c(\boldsymbol{\xi}). \tag{19}$$

Proof. If the procedure stops early, $c(\hat{\boldsymbol{\xi}}) = 0$. Non-negativity gives $0 \le \min_{\boldsymbol{\xi} \in \mathcal{B}} c(\boldsymbol{\xi}) \le c(\hat{\boldsymbol{\xi}}) = 0$, proving equality. Otherwise, fallback scans the complete bank and minimizes c, which establishes the same equality in Eq. 19 for the exhaustive route. □

D.2 Witness Preservation and Query Order

Proposition 2 (Route-relative witness preservation)**.** *For a fixed candidate route, witness predicate, and fallback, assume the budget permits a complete scan when no early witness is found. Stopping at the first witness returns a witness if and only if exhaustive verification of that route finds one.*

Proof. An early return satisfies the fixed witness predicate. Otherwise, a complete scan examines every candidate, so any witness on the route is encountered. If none exists, the procedure uses the stated fallback. Under a smaller budget, the argument applies to the observed prefix. □

Proposition 3 (Cost-aware witness ordering)**.** *For independent Bernoulli witness events with probabilities p_i, fixed query costs $c_i > 0$, and stopping at the first witness, ordering by decreasing p_i/c_i minimizes the expected cost of finding a witness or exhausting the bank.*

Proof. Adjacent queries i, j cost $c_i + (1 - p_i)c_j$ in that order and $c_j + (1 - p_j)c_i$ in reverse. The former is no larger exactly when $p_i/c_i \ge p_j/c_j$. Swapping inverted pairs proves the result; suffix cost is unchanged because both orders reach it with the same probability. □

Condition on a fixed scene and decision context c. For witness indicators W_i and query permutation π of a bank of M candidates, the expected total query cost C_π is

$$\mathbb{E}[C_\pi \mid c] = \sum_{k=1}^{M} c_{\pi(k)} \Pr\left(\bigcap_{j<k} \{W_{\pi(j)} = 0\} \mid c \right). \tag{20}$$

Conditional independence factors the joint failure probability into $\prod_{j<k}(1 - p_{\pi(j)})$, yielding the proposition. Independent latent draws under a fixed model and deterministic verifier can satisfy this condition. Shared unobserved physical parameters or coupled feedback can instead make priorities depend on observed history. The proposition supplies a reference objective for query priority; experiments on the instantiated banks measure the resulting cost and terminal quality.

PART II: ADDITIONAL EXPERIMENTAL RESULTS

Proposal quality, physical evidence allocation, and transfer across sources.

E QUALITATIVE RESULTS

Aggregate success rates summarize outcomes but do not show how a repair changes the scene. Figures 8, 9, and 10 separate accepted repairs, unchanged outputs, and complementary repairs. Together with Figure 6, they cover all nine author-released reconstructions once. Figure 11 examines HSSD transfer. Paired renderings share a camera and crop within each row; object-centered crops enlarge the measured layouts without changing poses. The stability badges in the reconstruction grids report the fraction of movable objects that pass, rather than the all-object scene success used in the furniture tables. Pen. denotes the measured penetration rate and drift is in meters. The reconstruction grids below use saved outputs and measured simulator states. Together, the grids cover 87 movable objects and 117 actors across the nine paired scenes. Under the same H120 translation/rotation criterion, 13 of 33 initially unstable objects become stable (39.39% repair), while 49 of 54 initially stable objects remain stable (90.74% preservation). These are paired object counts within nine scenes; pair penetration is reported separately. ScanNet++ II exhibits increased drift alongside reduced penetration in the complementary grid.

In the HSSD grid, the reference and proposal columns show placements before simulation, whereas the identity and Trace2Stable H240 columns show replay endpoints.

Dining Room, Bedroom, and Replica Scan6 receive measured residuals. Figure 9 groups the three unchanged outputs: Bedroom II, WorldLab, and Simpson Room. Their identical paired renderings show the policy's explicit identity choice under the same verification rule.

Tabletop and both ScanNet++ rooms receive measured residuals. The four figures account for six accepted repairs and three no-op decisions across all nine scenes.

The HSSD examples show one scene-level living-room recovery and the two largest positive position-drift reductions among bed anchors. Each row uses one camera across columns; identity and Trace2Stable endpoints are fresh 240-step terminal replays.

F SCENE COMPLEXITY

To test how object interactions affect refinement, we evaluate four new perturbations of 21 held-out sources, producing 84 paired layouts. At H240, Trace2Stable uses 32 proposals versus 64 for CEM and attains 70.24% Stable+Safe versus 57.14%. The prespecified 7–8-object comparison contains six sources and 24 layouts: the gain is 12.50 points, with a source-cluster 95% interval of $[4.17, 20.83]$. Both methods reach 84.38% in the 4–5-object stratum. In the six-object stratum, Trace2Stable improves Stable+Safe by 28.57 percentage points relative to scratch CEM.

Temporal persistence. At H960, 61/84 Trace2Stable and 51/84 CEM outputs pass, a pooled gain of 11.90 points ($[2.38, 21.43]$). The high-complexity contrast is 8.33 points ($[-8.33, 20.83]$). All

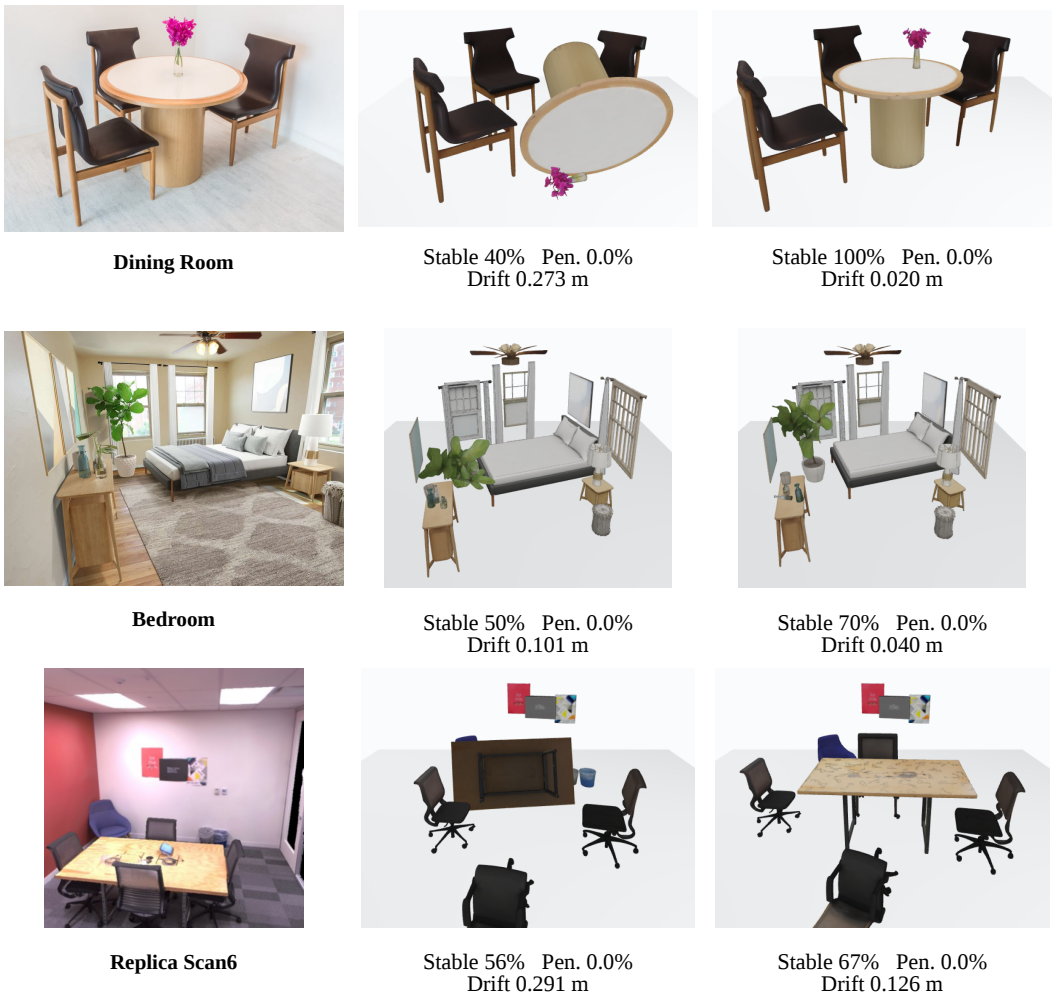


Figure 8: **Accepted repairs stabilize three released scenes.** Dining Room, Bedroom, and Replica Scan6 share H120 replay and report object stability, penetration, and drift.

59 Trace2Stable and 48 CEM H240-successful outputs remain successful at H960. Evaluation measures checkpoint displacement from frozen placement, maximum over objects, together with checkpoint speeds and collision flags. Each trajectory continues from the preceding checkpoint using the same reference. At H480, Trace2Stable/CEM retain 58/46 successes, recover 3/3 earlier failures, regress on 1/2 earlier successes, and leave 22/33 failures unchanged. At H960, these counts are 59/48 retained, 2/3 recovered, zero regressions, and 23/33 unchanged failures. Both additional Trace2Stable successes at H960 cross the rotation threshold; one also crosses the angular-speed threshold. CEM's three additional successes comprise two collision-flag changes and one angular-speed transition. At H480, one Trace2Stable and two CEM previously successful outputs have an undesired collision flag; all three pass again at H960.



Figure 9: **Verification preserves three already valid placements.** Bedroom II, WorldLab, and Simpson Room retain their released layouts, with identical views and H120 metrics.

G Source-Disjoint ReplicaCAD Furniture Amortization

G.1 Sources and Shared-Feedback Comparison

The 83 ReplicaCAD living-room groups have 42/20/21 source-disjoint training/validation/test splits. Each source supplies a clean anchor and three perturbed layouts, yielding 21,248 teacher candidates. Validation selects energy temperature $\tau = 0.25$ under a normalized-entropy floor of 0.70. The main refit uses 15,872 candidates from the 62 development sources. Local support structure is evaluated separately on twelve additional groups containing cushions, tableware, and monitors.

Shared-feedback generator comparison. Table 2(b) fixes residual CEM, inputs, actors, and the validation-selected quality rule. Each scene uses 32 H48 candidates (1,536 steps), followed by independent H240 replay. Across 21 sources, CFM improves success by 9.52 points (95% interval $[-14.29, 33.33]$; 10,000 paired source-bootstrap draws). Terminal collision is 4.76% versus 14.29% for diffusion, with higher unconditional drift for CFM. This rerun returns 17/21

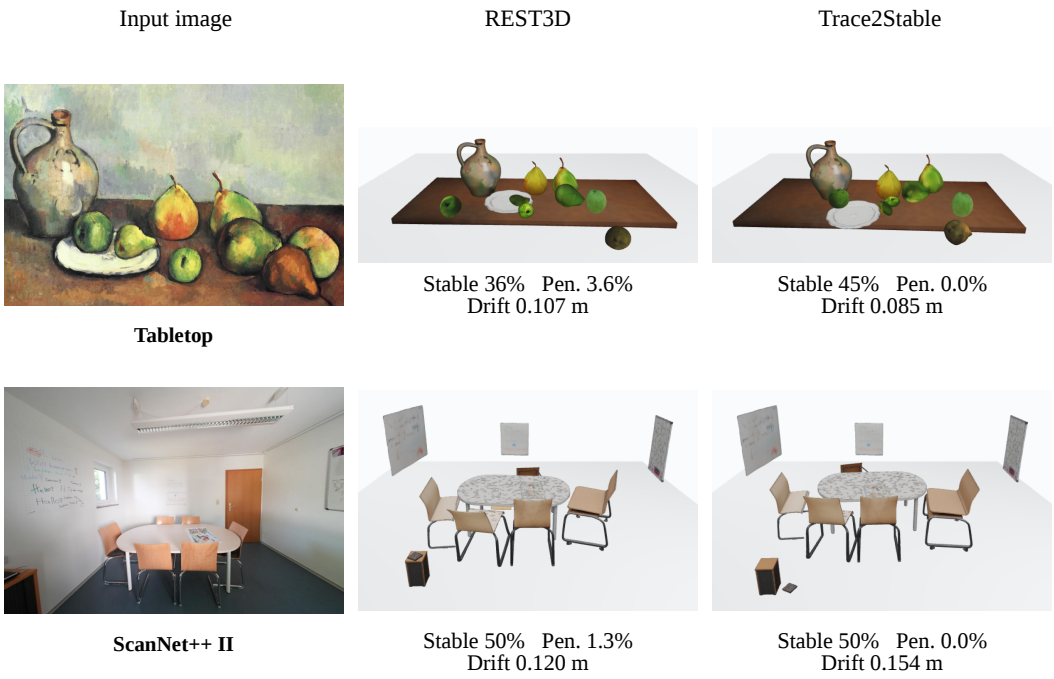


Figure 10: **Additional repairs retain the observed layouts.** Paired H120 views of Tabletop and ScanNet++ II report object stability, penetration, and drift under a shared camera.

CFM successes versus the archived 18/21; the same checkpoint and actor scope yield three different selections, illustrating contact-sensitive physical feedback.

G.2 Candidate Coverage and Terminal Quality

Coverage and selection on one frozen label set. Table 5 evaluates coverage and the frozen online selection using one set of bank H240 labels. For source s, $Y_{s,m}$ labels clean terminal candidates and $\hat{m}_s$ is the selected index. Let $F_s = \mathbf{1}[\exists m : Y_{s,m} = 1]$ indicate candidate availability and N_{src} count sources. Coverage and selection recall are defined by

$$\text{Coverage} = \frac{1}{N_{\text{src}}} \sum_s F_s, \qquad \text{SelectionRecall} = \frac{\sum_s F_s Y_{s,\hat{m}_s}}{\sum_s F_s}, \tag{21}$$

where a clean candidate requires all four motion maxima (translation, rotation, linear speed, and angular speed) at most 0.1 in their respective units and zero illegal collisions. Among 21 sources, diffusion covers 18 and selects a clean candidate on 13 (72.22% selection recall); proposal-only CFM covers 17 and selects 16 (94.12%); the archived flow–feedback bank covers 19 and selects 18 (94.74%). Table 2(b) reports the separate shared-feedback rerun.

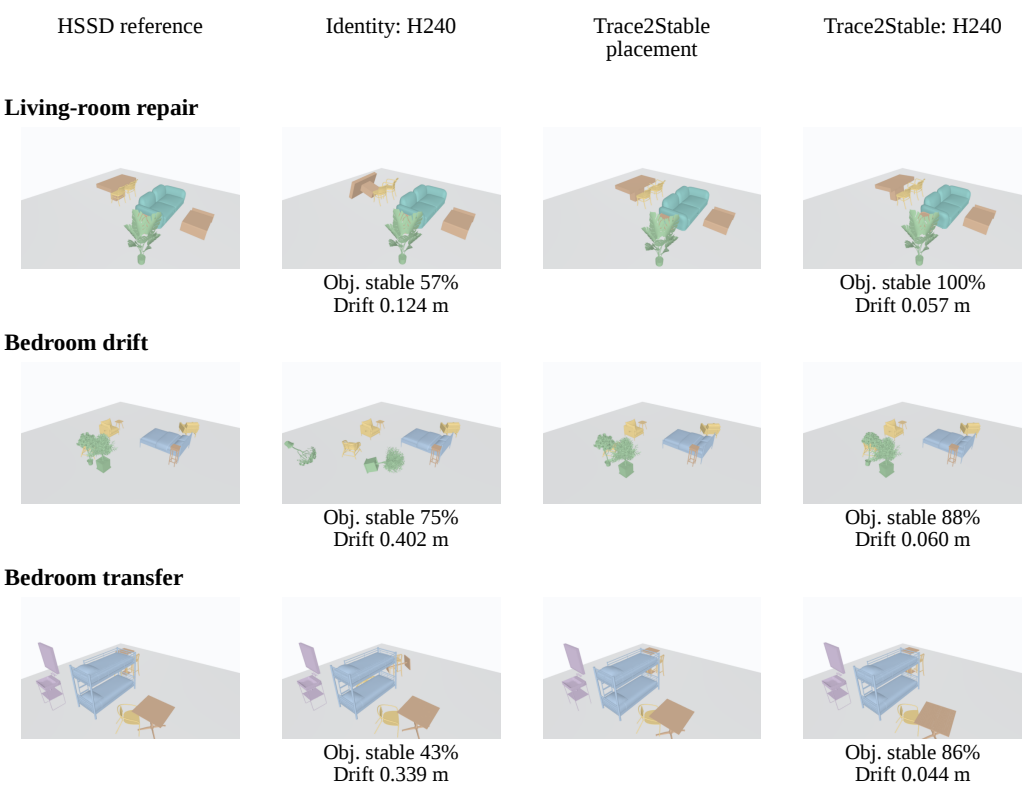


Figure 11: **HSSD transfer repairs selected layouts.** Living-room and bed-anchor rows compare shared-camera placements with H240 replay at $K = 16$.

Table 5: **Candidate coverage and selected quality differ under H240 replay.** Archived winners and complete role-valid banks reveal which methods find and return valid layouts.

			Selected H240 winner			Complete role-valid bank			
Role	Method	K	S+S ↑	Coll. ↓	d_t ↓	Cov. ↑	Prec. ↑	Comp.	Multi
Baseline	REST3D	32	57.14	14.29	0.1585	66.67	49.11	15.71	61.90
Baseline	Deterministic GNN	1	57.14	19.05	0.1474	57.14	**57.14**	0.57	0.00
Baseline	Conditional VAE	32	57.14	23.81	0.3174	66.67	52.83	0.67	0.00
Baseline	Learned-CEM	32	61.90	9.52	0.0972	71.43	42.71	13.67	66.67
Baseline	Graph-Diffusion	32	61.90	**0.00**	0.1323	85.71	44.94	14.38	71.43
Proposal only	Graph-CFM	32	76.19	4.76	0.0662	80.95	44.49	14.24	66.67
Ours final	Trace2Stable	32	**85.71**	4.76	**0.0559**	**90.48**	50.45	16.14	85.71

The left block independently replays each archived Table 2(a) winner for 240 steps. The right block replays every role-valid proposal; coverage is the fraction of 21 source scenes with a clean candidate, precision is the fraction of clean candidates in the bank, and components use 0.02 m/0.05 rad object-mean terminal distance. Generative rows use $K = 32$; Trace2Stable contains 16 flow and 16 residual-CEM proposals. d_t is translation drift in meters; Multi is the percentage of all 21 scenes containing at least two clean components. Component counts describe geometric diversity.

Connected components use single-link adjacency at the stated translation and rotation thresholds to distinguish neighborhoods of terminal feasible poses.

Source breadth at a fixed training budget. Figure 12 varies nested development sets of 6, 16, 31, 47, and 62 sources, with one fixed subset order and training seed. Every setting samples 496 conditions with replacement per epoch for 41 epochs (656 updates), then uses the same 1,536-step inference route on 21 test groups. Success is 12, 15, 14, 14, and 13 out of 21; the 62-minus-6

paired difference is 4.76 points (95% interval $[-9.52, 19.05]$). Mean dynamic drift falls from 0.2608 to 0.0704 m, a paired reduction of 0.1904 m ($[0.0452, 0.3838]$), while mean input edit changes from 0.1261 to 0.1182 m. The endpoint comparison separates lower terminal motion from the non-monotonic binary success curve under fixed training and physics budgets.

Size of terminal-pose components. Across all 21 sources, Trace2Stable's bank contains 339 clean candidates, each in its own component at the primary 0.02 m/0.05 rad threshold. Proposal-only CFM and diffusion likewise have 299 and 302 singleton components. At the predeclared relaxed threshold of 0.04 m/0.1 rad, the respective mean component sizes are 1.009, 1.003, and 1.017, with 99.40%, 99.66%, and 98.32% singleton components. In contrast, VAE samples form 14 components with mean size 25.36 at all three tested radii. These are distance-defined pose neighborhoods; support-graph experiments examine discrete structural alternatives.

Object-scale sensitivity. For terminal layouts A, B with object positions t_i^A, t_i^B, we normalize each object's translation distance by the diagonal ℓ_i of its fixed canonical-mesh bounding box, then average over objects: $d_T(A, B) = |\mathcal{V}_o|^{-1} \sum_i \|t_i^A - t_i^B\|_2 / \ell_i$. The rotational distance remains the mean quaternion geodesic angle. This distance is invariant to expressing the same scene in meters or centimeters. Figure 13 reports the full exploratory grid of five relative translation thresholds and three rotation thresholds on the same 4,053 candidates and 21 sources. Sources with no clean candidate contribute zero components. The seven curves compare REST3D, deterministic GNN, conditional VAE, Learned-CEM, Graph-Diffusion, Graph-CFM, and Trace2Stable. At the coarsest 8%/0.1 rad setting, mean component sizes are 8.07, 5.16, and 10.41 for Trace2Stable, CFM, and diffusion, respectively. Nearby terminal poses merge as the neighborhood expands; coverage and selected outputs remain fixed across the complete threshold grid.

G.3 Bank Allocation and Continuation

Matched ordering and terminal selection. We compare query orders on shared eight-candidate banks for 32 layouts from 16 sources (Figure 14). Each policy stops at the first zero-illegal-collision H5 witness, with a cap of 1, 2, 4, or 8 queries. Fallback ranks queried candidates by collision count, paid-prefix value score, then candidate identifier. This common rule isolates ordering. Independent H240 labels for all 256 candidates enter evaluation after selection. At the eight-query cap, learned and generation ordering both return 18 valid layouts, using 2.47 and 3.25 mean queries. Their paired source-macro success difference is 0.00 points (95% interval $[0.00, 0.00]$); including shared Local work, costs are 4066.09 and 4070.00 candidate-environment steps. Exhaustive selection returns 19 valid layouts at 4093.75 steps. Since 21 banks contain a valid candidate, selection recall is $18/21$ for learned stopping and $19/21$ for exhaustive short-horizon selection. Random ordering averages 56.72% success and 4069.91 steps over 20 fixed permutations per layout, with repetitions clustered by source. Online costs exclude H240 evaluation.

Shared Local work accounts for the narrow range of total costs.

Selection from identical paid observations. We measure the same generation-order prefix of $K \in \{1, 2, 4, 8\}$ candidates before comparing selectors. Both first minimize illegal collision count. The value selector then maximizes its learned score; the physical selector minimizes $r = \max\{d_{\max}/(0.1\,\mathrm{m}), \alpha_{\max}/(0.1\,\mathrm{rad})\}$, where $d_{\max}$ and $\alpha_{\max}$ are H5 displacement maxima across movable objects; candidate identifiers break ties. Both pay $5K$ Global steps plus the shared Local cost and return 11, 13, 15, and 19 terminal successes out of 32. At $K = 8$, the paired success difference is 0.00 points (95% interval $[0.00, 0.00]$); mean unconditional drift is 0.0105 m for value and 0.0123 m for physical-risk selection. This retrospective control isolates selection from identical paid observations, rather than ordering before measurement.

Opening a second bank. Two independently trained flow banks each contain 32 proposals. The policy tests the primary bank, then opens the secondary when no witness is found and the normalized disagreement exceeds the validation-median threshold 0.0310. The representative third trial evaluates two new perturbations of 21 sources. Earlier trials on these sources inform its threshold; the policy is fixed before the third trial's outcomes. On its 42 layouts, scratch search, a single learned bank, full verification, and Trace2Stable reach respectively 50.00%, 66.67%, 69.05%, and 71.43% Stable+Safe, using 64.00, 32.00, 64.00, and 44.95 mean rollouts. Collision rates

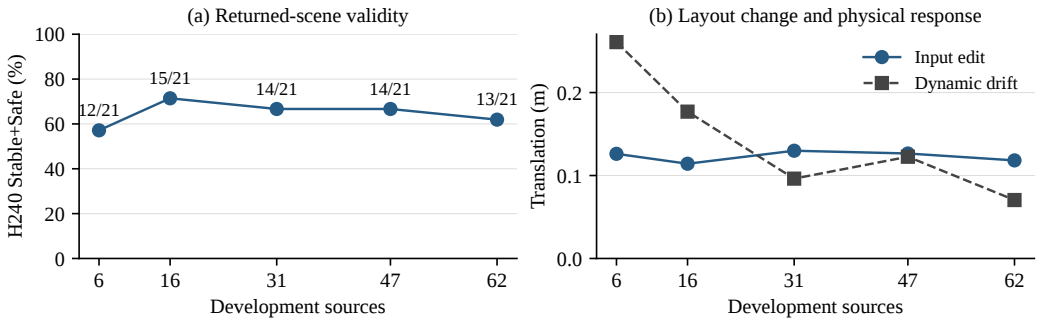


Figure 12: **More trace sources reduce drift without monotonic success gains.** (a) H240 success varies across five nested training sets; (b) input edit and dynamic drift compare the same test inputs and online budget. Lines guide the eye.

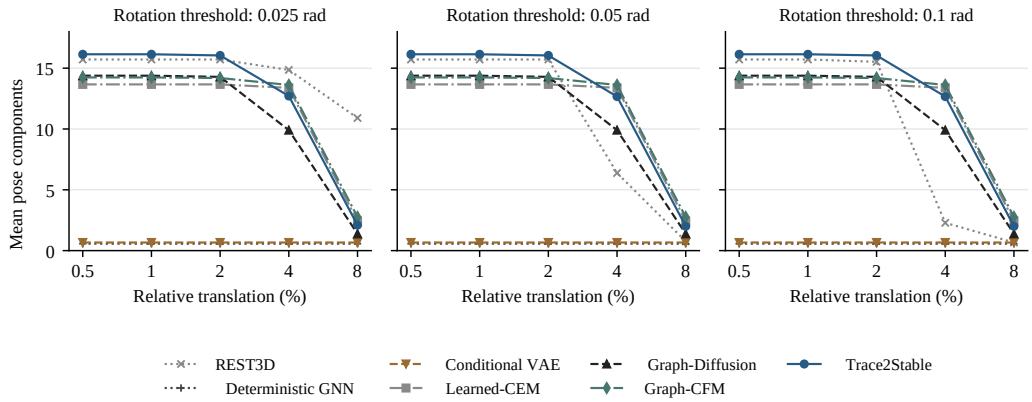


Figure 13: **Terminal-pose neighborhoods merge as tolerance widens.** Panels fix rotation thresholds and vary translation relative to object size; frozen-bank means include every source.

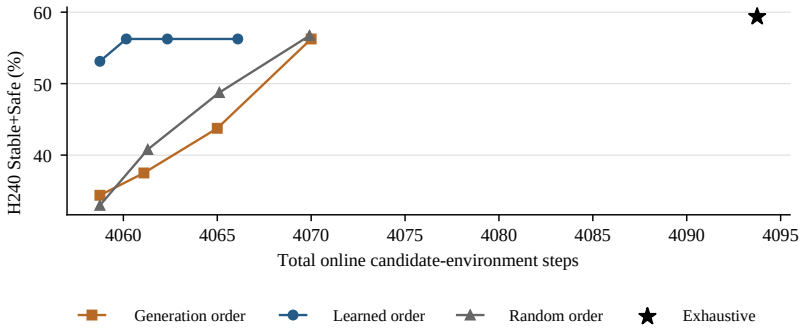


Figure 14: **Early stopping saves few total steps and can lose success.** Four query caps share candidate banks; online cost includes Local work but excludes terminal replay.

are 26.19%, 7.14%, 7.14%, and 4.76%, respectively. The paired Stable+Safe gain over CEM is 21.43 points ($[9.52, 35.71]$), using 29.76% fewer candidate rollouts than full verification. A post-hoc random gate opening the same 17 of 42 banks has expected success 67.63% versus 71.43% for Trace2Stable; its exact randomization tail probability is 0.0987. The terminal-label oracle reaches 73.81%, quantifying the remaining headroom in bank choice.

Recovery-based continuation. To allocate further simulation to unresolved trajectories, a calibrated model predicts H240 recovery from unsafe H30/H60/H120 prefixes. Its inputs are measured drift, speed, support, penetration, horizon, and object count. On 170 unsafe checkpoints from 18

held-out sources, recovery area under the precision–recall curve (AUPRC) is 0.7744 versus 0.4647 for physical-risk ordering. To examine how these predictions affect decisions, the scene-policy evaluation counts each trajectory once and charges only the observed path. With a 6,600-step additional cap on the same 42 banks, learned, physical-risk, and uniform continuation each succeed on 35/42 scenes. Including all H30 prefixes, they use 9,870, 9,450, and 10,320 steps, with mean drift 0.0622, 0.0627, and 0.0635 m. Random ordering over 20 fixed seeds averages 82.38% success and 10,357.5 steps. At the larger 26,460-step cap, all four policies reach 83.33%; learned and risk costs remain unchanged, while uniform and random costs become 10,800 and 10,837.5 steps.

H SCANNET++ MESH-LEVEL TRANSFER

H.1 VALUE TRANSFER AND QUALITATIVE PROPOSAL ANALYSIS

ScanNet++ groups use scan-aligned ShapeNet computer-aided design (CAD) objects, an active anchor–child group, and fixed surrounding collision context. A graph-context value model learns from 1,472 CEM candidates in eight scans; 40-dimensional object features contain geometry, proposed actions, and paid H15/H60 measurements. H240 outcomes provide labels. On 16 new scans and 56 groups, an eight-candidate bank yields 48.21% Stable+Safe versus 46.43% for CEM-energy selection. The source-macro gain is 1.56 points ($[-6.25, 9.38]$); collision decreases by 5.94 points ($[0.00, 13.75]$). The same endpoint at 32 candidates requires four times the online physics.

Transfer across scan cohorts. Sequential 20- and 27-source fits evaluate proposal transfer on 22 new scans and 49 groups. Their pooled source-macro success is 53.79%, compared with 50.00% for scratch CEM and 61.36% for value-ranked CEM. Cohort effects change sign, motivating a separate examination of coverage and selection under source shift.

Qualitative proposal analysis. The following examples use a fixed 27-source ensemble. Each full-route bank contains the observed pose, three flow proposals, and four CEM proposals.

Figure 15 compares the input with scratch-search, learned-only, and full-route H240 layouts. Each row keeps the camera and object identities fixed.

Gray surfaces show ScanNet++ context and colored furniture shows SCANnotate++-aligned ShapeNet CAD. The rows illustrate lower drift, repair of a learned-only failure, and collision removal with residual instability. Badges report Stable+Safe and maximum drift in meters and degrees.

Figure 16 extends the shared-camera comparison to two support-preserving corrections and a cross-source failure, evaluated under the same H240 criteria.

In the final row, every method exceeds the H240 rotation threshold while Trace2Stable reduces translation drift. These six post-evaluation examples include repairs and residual failures.

I JOINT SUPPORT-GRAPH AND POSE COMPONENT EVIDENCE

Protocol. To isolate support inference, we evaluate 12 source-disjoint ReplicaCAD local groups: one desk–monitor group, six sofa–cushion groups, and five tableware groups. Together they contain 47 nested support edges. Parent candidates are generated from object geometry, with the Scene Tree supplying evaluation labels. The two-candidate diagnostic compares an observed parent with its supplied correct alternative; the automatic experiment retrieves parents for every nested child.

Support features and candidate retrieval. The posterior trains on five-candidate sets with floating, penetrating, footprint-offset, undersized, and floor alternatives. Parent categories are disjoint across training and evaluation; 15% of training parents are floors.

To compare contact geometry across object sizes, we use the dimensionless, direction-aware feature

$$\begin{aligned}\tilde{\mathbf{c}}_{cp} = \big[&d/\ell_c,\ \mathrm{clip}_{[-4,4]}(g_y/h_c),\ \min(|g_y|/h_c, 4),\ 1 - o_{xz},\\ &d_{\mathrm{out}}/\ell_c,\ p_y/h_c,\ \mathrm{clip}_{[-4,4]}(\log(A_p/A_c))/4,\ \mathrm{clip}_{[-2,2]}(m_{\mathrm{contain}}/\ell_c)\big],\end{aligned} \tag{22}$$

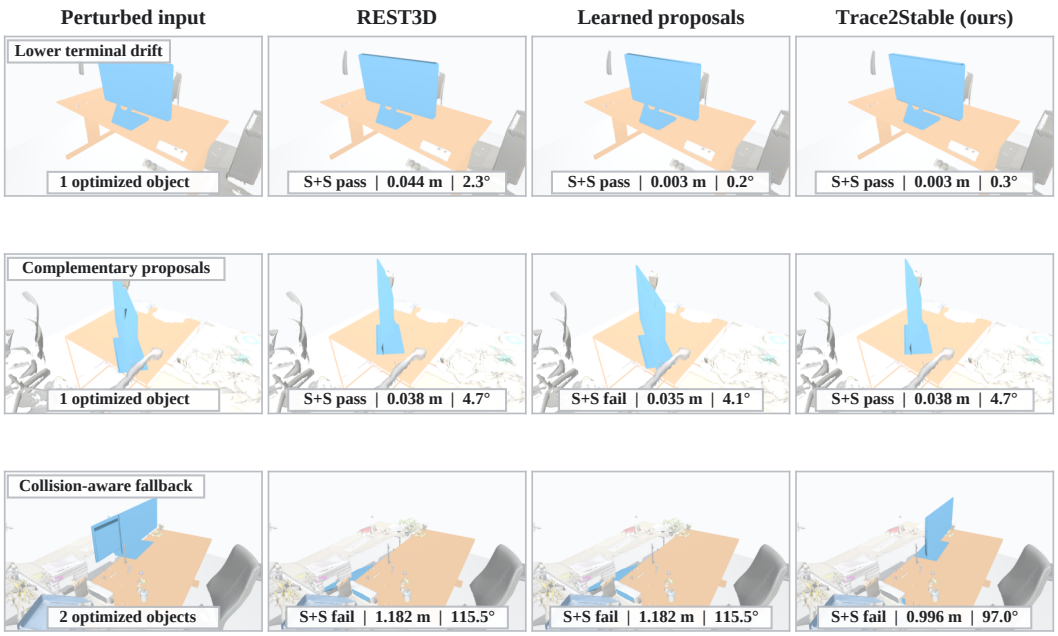


Figure 15: **Scan transfer shows repairs and residual failures.** Shared-camera rows compare input and H240 outputs; badges report Stable+Safe and maximum translation and rotation drift.

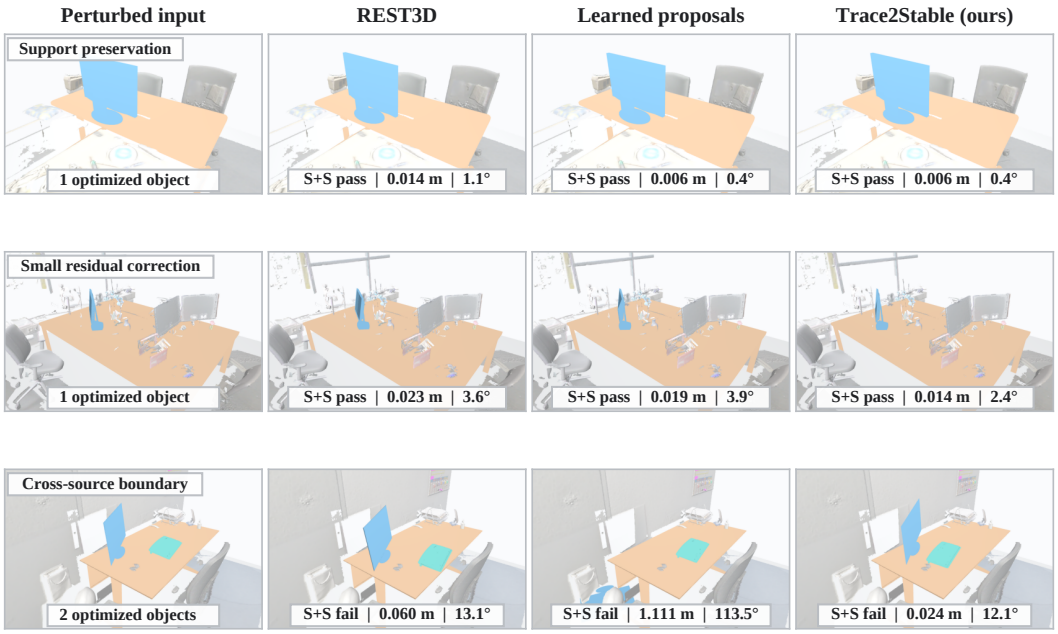


Figure 16: **Support repairs coexist with a cross-source failure.** Two corrected scenes and one remaining failure use matched cameras and H240 physical criteria.

where ℓ_c, h_c, A_c are the box diagonal, height, and horizontal footprint area of child c; A_p is the area of parent p. d is box separation, g_y the signed child-bottom/parent-top gap, o_{xz} their footprint intersection divided by A_c, and d_{out} the child-center distance outside the parent footprint. p_y is vertical overlap capped at h_c; m_{contain} is the minimum signed footprint containment margin. Clipping bounds each stated interval. For floor candidates, the overlap is one, outside distance is zero, and the last two channels are one. Floor height anchors the signed gap; all features are dimensionless.

Candidate probabilities are temperature-calibrated on validation examples. Risk thresholds and candidate cardinality are fixed before physical evaluation.

Automatic parent proposals. For child axis-aligned bounding box (AABB) B_i, retrieval uses unordered world bounds $\{B_j\}$ to select the floor and $K-1$ nearby parent candidates:

$$\mathcal{C}_i^K = \{\text{floor}\} \cup \operatorname*{TopK}_{j\neq i}^{K-1}[-a_{ij}], \qquad a_{ij} = d_{ij}/\ell_i + 0.35 d_{ij}^{xz}/\ell_i + 0.20[y_j - y_i]_+/h_i + 0.15 d_{ij}^{\text{out}}/\ell_i, \tag{23}$$

where y_i, y_j are box-center heights, h_i is child height, ℓ_i its diagonal, d_{ij} is box separation, and d_{ij}^{xz} is horizontal center distance. d_{ij}^{out} is the outside distance defined above. TopK returns the $K-1$ highest negative costs. The broad score tolerates interpenetrating sofa–cushion AABBs; learned reasoning performs the fine ranking. Candidate retrieval uses world-space bounds, with a fixed tie-breaking order. On all 47 frozen ReplicaCAD edges, Recall@3 is 85.11%, Recall@5 is 100%, the mean true-parent rank is 1.45, and the maximum is four.

Permutation-equivariant candidate context. An individually plausible support may become less likely when a better alternative is present. We use set attention to compare parents jointly, writing $\tilde{\boldsymbol{c}}_{ij}$ for the feature in Eq. 22 with child i and candidate parent j. Embedding each token and applying a Transformer with no positional encoding gives

$$\boldsymbol{h}_{ij}^0 = \text{LN}(\text{SiLU}(W_e \tilde{\boldsymbol{c}}_{ij} + b_e)), \quad \{\boldsymbol{h}_{ij}\}_{j\in\mathcal{C}_i} = \text{Tr}_\eta(\{\boldsymbol{h}_{ij}^0\}_{j\in\mathcal{C}_i}), \quad \ell_{ij} = g_\eta(\boldsymbol{h}_{ij}), \tag{24}$$

where W_e, b_e are embedding weights and bias; LN is layer normalization, SiLU the sigmoid linear unit, and Tr_η the set Transformer. The head g_η applies LN, a hidden linear layer, SiLU, and a scalar linear output. Figure 2(a) depicts these operations. Valid-parent logits use a uniform prior and masked softmax (Eq. 5); geometric costs determine set membership. Masked parent cross entropy $\mathcal{L}_{\text{support}}$ trains the probabilities on the retrieved candidate set.

The fixed-graph route retains $\mathcal{G}^0$. In the controlled structural task, a designated child has two supplied parents and all other edges are fixed. Parent probabilities thus index two legal trees. Frozen uncertainty routing retains one or both trees. Each retained branch receives a pose bank of equal size, followed by physical evaluation of its proposed placements.

The two-layer, four-head, width-64 Transformer is permutation equivariant. On 47 held-out ReplicaCAD support edges, contextual ranking attains 95.74% Top-1 versus 72.34% for geometry and 27.66% for independent edge scoring. Five-candidate retrieval covers every true parent; retrieval coverage and ranking accuracy are therefore evaluated separately.

Image-conditioned mesh-based structural confirmation. ProcTHOR-10K meshes and nested receptacles (Deitke et al., 2022) run in AI2-THOR (Kolve et al., 2017). Of 200/80/96 source-disjoint train/validation/test houses, 151/65/71 expose a target and at least two nearby receptacles in one red–green–blue (RGB), depth, and instance-mask view, selected without terminal outcomes. Both scorers and the image heuristic use these simulator-rendered observations. Ten category-free projected-contact and depth relations condition a two-layer, four-head set Transformer without positional encoding. For the ProcTHOR comparison, image-only and image–geometry models share training examples; validation negative log-likelihood (NLL) selects checkpoints and temperature.

The 225-house evaluation combines 71/80/74 eligible houses from disjoint intervals $[96, 192)$, $[192, 288)$, and $[288, 384)$. The last 74 houses and targets also appear in the earlier category-transfer cohort $[288, 480)$; pooling counts each house once. Parameters and temperature stay fixed. Of 288 requested houses, 61 lack an eligible child and two fail visibility requirements. The category-transfer cohort has 152 eligible houses, including 29 training-unseen ordered category pairs. A separate study selects 62 of 320 eligible houses in the 424-house window $[576, 1000)$: 25 unseen-child, five unseen-parent, and 32 seen-category controls. These subsets retain separate evaluation roles.

Table 6 evaluates the learned posterior against input-matched image and geometry baselines. ECE denotes expected calibration error, and Brier reports squared probability error. Physical acceptance evaluates the top-ranked parent, using two-second ProcTHOR replay and eight-second (H160) iTHOR replay. The learned posterior has higher observed ProcTHOR parent accuracy and physical acceptance (Table 6). Allowing a second parent query after a failed first test changes the evaluation to budgeted selection. With a two-query cap on 118 iTHOR rooms, both it and the image heuristic reach 88.14% joint validity, using 1.12 versus 1.14 mean queries. For these 118 rooms, continuing failed H40 branches to H160 reduces mean steps from 178.98 to 74.24 relative to direct H160 verification. Across all 391 candidate branches, the 149 H40 witnesses remain valid at H160.

Table 6: **Learned image support ranking has the highest observed accuracy.** ProcTHOR and iTHOR results compare matched image and geometry baselines under physical replay.

Role	Method	Top-1 ↑	Recall@2 ↑	NLL ↓	Brier ↓	ECE ↓	Phys. accept ↑
ProcTHOR: 225 evaluable houses							
Baseline	Geometry nearest	76.44	94.22	0.6179	0.3666	0.1150	71.56
Baseline	Image heuristic	97.78	99.56	0.0710	0.0342	0.0279	89.78
Ablation	Image + geometry posterior	98.22	**100.00**	**0.0556**	0.0201	0.0134	90.22
Trace2Stable final	Learned image posterior	**99.56**	**100.00**	0.0815	**0.0120**	**0.0093**	**91.11**
iTHOR: unchanged model on 118 hand-designed rooms							
I / Baseline	Geometry nearest	50.00	71.19	1.0972	0.6146	0.1338	61.02
I / Baseline	Image heuristic	96.61	**100.00**	**0.1172**	0.0582	0.0629	86.44
I / Trace2Stable	Learned image posterior	**98.31**	**100.00**	0.1767	**0.0387**	**0.0277**	**88.14**

Observation robustness. On 76 HSSD objects from 31 scenes, a geometry-anchored candidate union gives 84.21% Stable+Input versus 78.95% for the geometry anchor. Frozen mask, depth, and occlusion perturbations lower this to 82.37% under severe joint corruption. These trials reuse fixed physical candidate trajectories and measure observation-based routing. Anchor-first stopping saves 28.22% of queries with the same measured outputs over 2,736 realizations.

Joint support and pose inference. The 48-source cohort has 24 thin-support and 24 coplanar-support layouts. Each graph bank uses eight full-scene candidates, with Local work included in the costs reported alongside Table 4. H5 PhysX selection precedes H240 replay. Two further 16-source replications yield a pooled 11.25-point gain ($[-6.25, 28.75]$); the last reverses sign.

Equal-cost support allocation. To separate structural allocation from additional simulation, we compare five frozen routes on 16 historical controlled layouts, eight per support topology (Table 7). Each layout supplies two legal graph hypotheses. A single graph receives 16 candidates per group; a two-graph route receives eight per graph. Both spend 1,920 Local and 960 Global steps, including all descendant actors. A frozen Local flow supplies proposals, H60 measurements select a placement, and independent H240 replay evaluates the returned scene. Geometric ranking uses vertical gap and footprint overlap; the random route selects one graph with a fixed seed. Uniform allocation opens both graphs, while learned routing chooses a single graph or splits the same budget.

Table 7: **Graph routing reduces collisions but increases drift at equal cost.** Five routes share 16 layouts and 2,880 online steps; H240 Coll. is a scene percentage and Drift is in meters.

Route	Steps ↓	Coll. ↓	Drift ↓
Observed graph	2,880.0	75.00	0.2346
Geometric ranking	2,880.0	75.00	0.2346
Seeded random graph	2,880.0	56.25	0.5236
Uniform two-graph allocation	2,880.0	56.25	0.6185
Learned graph routing	2,880.0	50.00	0.6499

Learning and random graph selection each return one valid layout; the other routes return zero. The learned-minus-observed difference is 6.25 percentage points, with a paired source-bootstrap interval of $[0.00, 18.75]$. Geometric ranking selects the observed graph on every layout. Learned routing reduces terminal collision from 75.00% to 50.00% relative to that graph, while mean maximum-object drift rises from 0.2346 to 0.6499 m. Across all five routes, at most one layout satisfies the joint terminal criterion. This comparison separates collision removal from motion stability when both stages share a fixed physical budget.

On these same layouts, adding learned query ordering to flow and support routing retains 1/16 H240 successes at 2,880 steps. Splitting the Global allowance equally between initial and feedback candidates yields 0/16 at 2,861.25 mean steps, with a paired success difference of -6.25 points (95% source-bootstrap interval $[-18.75, 0.00]$). This component comparison uses the same Local outputs, H60 measurements, and H240 terminal criterion.

J CROSS-SOURCE HSSD PROPOSAL TRANSFER

Source transfer and terminal pose components. HSSD (Khanna et al., 2024) supplies 20 validation and 28 test scenes; the anchor-local floor rule retains 18/26. Flow proposals undergo H60 PyBullet selection and H240 replay. Two independent banks provide 368 candidate trajectories from 46 scenes. At 0.02 m/0.05 rad object-mean distance, clean poses form multiple components in 21/46 scenes (45.65%, 95% interval $[32.61, 60.87]$), indicating distinct terminal pose neighborhoods.

J.1 SUPPORT CLOSURE

Retrieval searches within 1.5 m for up to 12 parents. The selected branches in Figure 5 pass Stable+Input on all 15 held-out children; the two interventions fail this joint criterion. These interventions isolate how the selected support affects the final placement.

PART III: DISCUSSION

From candidate availability to returned quality. Physical refinement needs both a valid candidate and a verifier that returns it. Under the archived complete-bank H240 labels in Table 5, Trace2Stable covers 19 of 21 scenes and selects a clean candidate in 18. Two scenes have no clean candidate; one has an available repair that remains unselected. This separates proposal coverage from query and selection limits using the same terminal labels.

Budget and decision granularity. A physical query has a cost determined by the number of candidate environments and the steps actually advanced. Hierarchical refinement adds another dimension: opening a support branch can repeat Local work before a Global candidate is tested. The fixed-bank study therefore reports both Global queries and total steps. Its query reduction accompanies a smaller change in total work and a different terminal success rate. Likewise, support routing trades additional rollouts for access to alternative structures. These measurements identify where computation changes and make the quality–cost trade-off interpretable at the level of the complete inference route. Wall time remains a separate measurement because initialization and parallel execution also contribute.

Repair and preservation. The input arrangement supplies a geometric reference as well as a starting point for physical testing. Input-to-placement edit measures how much the method changes that arrangement; placement-to-terminal drift measures its subsequent response to dynamics. The two quantities serve different purposes. The nine released reconstructions illustrate both accepted residuals and verified identity choices. Their object-level repair and preservation counts make these behaviors visible alongside penetration measurements. An unchanged placement preserves the reconstructed layout, while acceptance is governed by the stated online test. Rendering agreement with that layout is consequently an input-preservation measure, with the camera and reference specified.

Transfer and physical interpretation. Source transfer tests learned proposals or support decisions on new inputs. The ScanNet++ analysis separates candidate valuation from pose generation, while HSSD replay examines terminal pose alternatives and support retention. Geometry, collision proxies, mobility, friction, and motion thresholds define the acceptance event in each setting. Input-relative pose changes and terminal motion then describe how a selected correction preserves the arrangement and responds to the evaluated dynamics. These complementary measurements link the proposed hypotheses to their observed physical behavior. For physical deployment, sensed contact and motion provide the corresponding check on simulator-based acceptance. This interpretation connects reusable search experience to a concrete, measured decision about the current scene.