Learning Next-Best-View for MEP Discovery in BIM: A Hybrid Policy with Realism-Augmented Procedural Pretraining

Nathan Zhao
nathanzh@stanford.edu

June 2026

Abstract

We address next-best-view (NBV) planning for the discovery of mechanical, electrical, and plumbing (MEP) systems (pipes, ducts, fittings, fixtures) inside building information modeling (BIM) scenes, formulated as pose selection over raycast partial point clouds of IFC geometry. We train a permutation-invariant PC-NBV-style joint ranker [17] (`v_joint_v4`, 1.8M parameters) on 49 IFC scenes and report three findings that matter more than any single configuration's headline number. First, **a hybrid composition of the learned ranker with classical one-step inference-time lookahead on the top-12 of M=20 candidates beats the classical OctoMap-IG baseline [1] by $+5.0\sigma$ to $+5.6\sigma$ across every split**: in-distribution residential, held-out residential plus GNI BIM, and the ifc-bench OOD corpus (Duplex MEP, WBDG, West Riverside Hospital). Second, contrary to a tempting "scale the ranker and ship it" narrative we initially entertained and then refuted, **the learned ranker alone underperforms the classical `greedy_coverage` baseline on every in-distribution surface-coverage split (-7.5σ on `test_locked`)**; the ranker is shippable only as a filter, not as a policy. Third, **the MEP-Recall metric as reported in much of the active-perception literature uses the scanned cloud's own instance labels as denominator**, inflating oracle numbers 5-50×; we present the corrected scene-full denominator and re-run every baseline against it. The corpus-diversity intervention (`v7_realsynth`: procedurally generated MEP scenes mixed into training) lifts the ifc-bench OOD surface metric while target-engineering interventions (edge recall, primitive heads) do not. We argue the productive research frontier for BIM-MEP NBV is no longer better point ranking but composition-aware training and a richer scene-encoder pathway: `scene_dim`=32 against `global_dim`=256 is provably scene-blind at K=1.

1 Introduction

Capturing the as-built state of a building's mechanical, electrical, and plumbing (MEP) systems is one of the load-bearing steps in any digital-twin or renovation workflow. Modern terrestrial laser scanners produce dense point clouds but at human cost: an operator places the scanner, waits, moves it, places it again, and the choice of *where to scan next* dictates whether the MEP system is fully captured in two scans or twelve. The next-best-view (NBV) problem formalises that choice: given the current partial reconstruction, pick the next pose so as to maximise expected coverage of the hidden surfaces of interest. We study the version of this problem where the surfaces of interest are MEP entities rather than the overall building shell, and where ground truth geometry is available offline through an IFC BIM model [9], a setting that lets us train and evaluate with raycast simulation while keeping a clear path to real-scanner deployment.

Three properties make BIM-MEP NBV genuinely different from the object-scale or free-flight-drone NBV settings that dominate the recent literature [3, 17]. First, the target distribution is *sparse and instance-rich*: a single storey may contain a few hundred to a few thousand MEP entities embedded in a much larger non-MEP shell, so a metric that says "how many MEP instances were captured" is sensitive to the denominator in ways the literature has not been careful about (§3). Second, the candidate pose set is *discrete and constrained* by the floor plan; the scanner cannot fly, so feasibility checks against a partial mesh dominate sampler design (and the original signed-distance "inside the building" check is silently broken on non-watertight IFC exports). Third, the supervisory signal, per-candidate Δ-recall, is *discretely quantised* (one captured pipe = $1/N$) and noisy enough that a PC-NBV-style ranker trained end-to-end on it does not, on its own, beat a classical greedy baseline (§6).

This paper makes three contributions, each of which is supported by a committed artifact in the open-source repository at `github.com/nathanjzhao/mep-nbv`.

- **Hybrid policy composition.** We show that a learned ranker (`v_joint_v4`) and a classical one-step inference-time lookahead are *individually* below their respective ceilings but *jointly* beat the classical OctoMap-IG baseline by +5.0-5.6σ on every split, at $\sim$ 35% the inference cost of exhaustive lookahead (§5).
- **The MEP-Recall denominator fix.** We document a metric pathology that inflates oracle numbers 5-50$\times$ on small scenes when the denominator is the partial cloud's own instance count rather than the scene-full IFC instance count, and we re-run every baseline against the fixed metric (§7).
- **Corpus diversity over target engineering.** A controlled five-way comparison (`v4`, `v6`, `v7_attn`, `v7_realsynth`, `v7_k12`) shows that the *realistic-synthetic* corpus augmentation lifts OOD by more than any of the target-engineering interventions (edge recall, primitive heads, attention block swaps, K=2 supervision) tried at the same compute (§8).

Throughout, we stress an evaluation-discipline point that the active-perception literature has long known but is easy to forget: cross-paper recall comparison is unreliable because the denominator convention is not standardised, and the only defensible comparison is one scored on identical scenes, sampler, extractor, and metric definition (here, the eight policies that share our harness).

2 Related work

Learned next-best-view planning. Object-scale NBV with deep nets was popularised by NBV-Net [10], which regresses a next-view index from a voxel grid, and pushed to a permutation-invariant point-cloud encoder by PC-NBV [17], which trains a joint ranker over a fixed view-sphere with per-candidate coverage targets and reports object-scale ShapeNet results. Our `v_joint` architecture is a re-implementation of the PC-NBV scoring head, written from scratch against the paper's description to fit the BIM/IFC scene encoding; we cite the paper rather than port the code because the candidate set is no longer a sphere, the targets come from a RANSAC-cylinder MEP extractor rather than a ShapeNet mesh, and the per-candidate features include a pose-local point cloud the original architecture lacks. The free-action drone setting of GenNBV [3] adds a history embedding which we adopt and a generalisation-across-scenes evaluation we mirror for our held-out split. None of these prior works contend with sparse-instance MEP targets, an ill-defined denominator, or a scene candidate set whose cardinality is set by floor-plan feasibility.

Classical information-gain planning. The receding-horizon NBV planner of Bircher et al. [1] (OctoMap-IG) is the field-standard classical baseline: a voxel occupancy grid is maintained, candidate poses are scored by their expected entropy reduction, and the planner replans after each scan. Our `octomap_ig` baseline is a numpy reimplementation; we use the information-gain formulation of Isler et al. [7] for the IG term and the volumetric variant survey of Delmerico et al. [4] as a sanity check on the encoding choice. OctoMap itself [6] is the substrate. None of these methods are MEP-aware: they treat all hidden voxels equivalently, which is exactly the weakness our hybrid policy is built to exploit.

Imitation learning and the covariate-shift bound. Our ranker is trained as supervised distillation onto a slow oracle target, which is the classic behavioural-cloning setting analysed by Ross et al. [12]. The $O(K^2\epsilon) \to O(K\epsilon)$ argument predicts that supervised distillation should work at K=1 and degrade at higher horizons; this is precisely the empirical signature of our K=2 supervision attempt (§9). DAgger was implemented as a candidate fix (`mep_nbv/dagger.py`) but did not produce a robust win across seeds at the 49-scene training scale.

Point-cloud encoders and scene graphs. Our global and pose-local encoders are PointNets [11] because the input is a set and permutation invariance is the minimal correct bias. PointTransformer-v3 [15] and Sonata [13] were attempted as drop-in replacements; neither lifted held-out numbers at our scale. A 3DSSG-style [14] scene-graph head was implemented but blocked on the K=1 scene-blind pathology we describe in §6. A SuperDec primitive-decomposition head [5] was attempted as auxiliary supervision and produced no robust win. We discuss all of these in §9 as deliberate negative results rather than as missing baselines.

BIM datasets and the closest comparison. Our IFC corpus draws from three sources: public IFC samples shipped with the Krijnen et al. [9] test suite, the construction-grade GNI BIM dataset of Borrmann et al. [2] (the seven `gni_model_*` scenes), and the ifc-bench v2 corpus of Yu et al. [16] (Duplex MEP, WBDG office, West Riverside Hospital). We carry the ifc-bench corpus exclusively in test, never in training, so that the OOD split is genuinely out-of-distribution by construction. The lack of a published benchmark that fixes a sampler, an extractor, and a scene-full denominator is what motivated the BENCHMARK.md table our paper anchors against.

Why cross-paper recall comparison is unreliable here. We re-emphasise the methodological point because it shapes the whole evaluation. Two NBV papers reporting "X% MEP recall" may differ on (i) what counts as an MEP instance (IFC class taxonomy vs. predicted cluster), (ii) what the denominator is (scene-full vs. partial-cloud-observed; this is the pathology of §7), (iii) which extractor is used (RANSAC primitive vs. deep segmentor vs. Hungarian-matched ground truth), (iv) which scenes are in which split, and (v) what K and M are (more candidates flatters every method). The only numbers we treat as comparable to ours are those we computed ourselves on the identical protocol, against the eight policies in our harness.

3 Method

The full eval pipeline (data → simulator → rollout → model → eval → output) is laid out in `figures/pipeline-flowchart.html` (browser-renderable, no external dependency). The two methodology-fix nodes (§7) are highlighted in amber.

Problem and notation. The scene is a tuple $\mathcal{X} = (\mathcal{M}, \mathcal{I}, \mathcal{C})$ where $\mathcal{M}$ is an IFC mesh, $\mathcal{I} \in \{0,1\}^{|\mathcal{M}|}$ marks MEP-class triangles, and $\mathcal{C}$ is the scene-full instance-label vector (one integer per MEP triangle). A trajectory is a sequence of K scanner poses $\tau = (p_1, \ldots, p_K)$ drawn from a feasibility-filtered candidate set $\mathcal{P}(\mathcal{X})$. Each pose p produces a partial cloud $\hat{P}_{\tau,k} = \text{raycast}(\mathcal{M}, p_k)$ via Open3D's raycasting scene [18]. We score policies on two metrics computed against the accumulated partial cloud $\hat{P}_\tau = \bigcup_k \hat{P}_{\tau,k}$: *surface coverage* (fraction of MEP triangles whose vertices appear in $\hat{P}_\tau$) and *MEP-Recall* (fraction of MEP instances in $\mathcal{C}$ whose extracted geometry intersects ground truth at IoU ≥ 0.25 via the RANSAC-cylinder-plus-OBB extractor of `GeometricExtractor`).

The denominator and the metric pathology fix. The choice of denominator for MEP-Recall is load-bearing and is the single bug that consumed roughly half of our experimental budget. The *broken* version, which is the natural implementation and which (we strongly suspect) is what much prior NBV work reports, divides the count of *captured* instances by the count of instances *visible in the partial cloud*: i.e., the partial cloud is its own ground truth. Under this convention a scanner that observes k instances and segments them all trivially scores $k/k = 1.0$; an oracle that picks the single best pose lands at 0.6-0.9 recall on a small scene because the partial cloud happens to be its own denominator. The *fixed* version, which we adopt and which is what every number in this paper is computed under, divides captured instances by the count of all MEP instances in the scene-full $\mathcal{C}$. The same scanner now lands at k/N, where N is the total MEP instance count for the scene. This is the only NBV metric definition under which an oracle bounded by feasibility actually looks bounded (§7).

Pose sampler: raycast floor check, not signed distance. A second silent bug we caught and fixed concerned the feasibility filter. The original `sample_feasible_poses` rejected candidate (x, y) points whose signed distance to the scene mesh was negative ("outside the building shell"). On any non-watertight IFC export, which is essentially every real-world IFC file, the signed-distance check returns garbage and either rejects most candidates or accepts candidates floating mid-air. The fixed sampler anchors the scanner z at MEP_floor + scanner_height whenever the MEP-only z-extent is smaller than 40% of the scene-AABB z-extent (this handles the GNI BIM exports that span ±30 m of site context but have MEP only in a 0-11 m band), tightens the (x, y) AABB to the MEP bounding box plus a 5 m margin per axis independently (handles long thin pipe runs in a wide storey), and uses a raycast floor check: a downward ray from the candidate pose is accepted iff it hits the scene mesh within 10 cm of the intended scanner z. The pre-fix sampler is what produced the "100% zero-Δ expert picks" degenerate mode that broke `ltu_ahouse_cooling__plan_3` during early experiments.

The learned ranker. The `v_joint` architecture is a PC-NBV-style joint ranker [17]. The state is the accumulated partial cloud $\hat{P}_\tau$ embedded by a PointNet global encoder (`global_dim`=256); each of the M candidate poses is embedded by a pose-local PointNet (`local_dim`=128) operating in the candidate's own frame; a small history encoder (`history_dim`=64) summarises previously chosen poses (this is the GenNBV-style addition); and a scene encoder (`scene_dim`=32) attempts to give the model a coarse global handle on the scene that the partial cloud does not yet contain. The four feature streams are concatenated per candidate and passed through a per-candidate scoring head; $\pi(\hat{P}_\tau) = \arg\max_p \text{score}(p)$. The architecture totals 1.8M parameters in the `v4` checkpoint, comparable to the smallest configuration in the original PC-NBV paper. The training data is 49 IFC scenes (the residential and GNI BIM sources; ifc-bench is held out), tuples generated by random and short-oracle rollouts; the loss is a per-(state, pose) MSE on Δ-MEP-Recall targets

computed under the *fixed* denominator.

The hybrid composition pattern. The shipped policy is not the learned ranker alone but a composition of the ranker with a classical one-step lookahead, formalised as `LearnedPlusLookaheadBaseline` in `mep_nbv/baselines.py`. At each decision step the ranker scores all M candidates and selects the top K_{filter} (default K_{filter}=12 of M=20); the lookahead then evaluates each of those K_{filter} poses by raycasting it, recomputing the metric, and picking the actual argmax. The composition is empirically what beats the classical baseline; the ranker alone does not (§6), and the lookahead alone is exhaustive and expensive (M raycasts per step rather than K_{filter}). The factor of $M/K_{\text{filter}} \approx 1.67$ compute saving over exhaustive lookahead is small in absolute terms but real, and the *quality* of the composition is what justifies shipping it.

4 Setup

Data and splits. We curate 49 IFC scenes spanning three sources. The training pool comprises 39 scenes: residential heat-pump and boiler IFCs (`heatpump_floorheating`, `b03_heating`, `boiler_gasradiator`, and similar), the public IFC samples shipped with Krijnen et al. [9], and 30 procedurally generated synthetic scenes from `mep_nbv/synthetic_scene_v2.py`. We hold out three residential scenes as `test_locked`, seven scenes (six GNI BIM models plus `heatpump_floorheating` variants) as the residential `held_out` split, and 10 ifc-bench scenes as the OOD split (Duplex MEP, WBDG office MEP, West Riverside Hospital fire). The ifc-bench corpus is *never* placed in training under any configuration; we treat this as an anti-pollution invariant analogous to the held-out balancing authorities in our forecasting work.

Evaluation protocol (held fixed across all configurations). Each configuration is evaluated at K=1 with M=20 candidates and K_{filter}=12 for hybrid policies, across the eight policies of `mep_nbv.baselines` (random, grid, `greedy_coverage`, `octomap_ig`, `greedy_lookahead_1`, oracle, `learned_joint`, `hybrid_learned_lookahead`), under both surface coverage and MEP-Recall, over five seeds $\{0, 1, 2, 3, 4\}$. The aggregation rule is *scene-mean per seed*, then *mean* $\pm$ *SE across seeds* (SE has 4 dof). Paired differences are computed within seed and then averaged across seeds; reported σ values are mean/$\text{SE}_{\text{paired}}$.

The five learned configurations. We compare five learned checkpoints, all evaluated under the protocol above: `v4` (`v_joint_v4`, the shipping checkpoint; 1.8M params, trained on the 49-scene corpus); `v6` (a list-MLE re-ranking head replacing the per-pose regression); `v7_attn` (a transformer block over the candidate features replacing the per-candidate MLP); `v7_realsynth` (`v4` architecture trained on the 49-scene corpus plus realism-augmented procedural scenes); and `v7_k12` (a K=2 supervision attempt, trained on two-step lookahead targets). The configurations are deliberately matched at $\sim$1.8M parameters and similar compute so that the comparison isolates the design lever rather than scale.

5 Experiments and headline results

The headline table. Table 1 reports the surface-coverage cells of BENCHMARK.md verbatim under the fixed sampler and fixed denominator. The result we ship is the `hybrid_learned_lookahead` row: 0.541 ± 0.006 on `test_locked`, $\mathbf{0.147 \pm 0.018}$ on `held_out`, and 0.052 ± 0.005 on ifc-bench

OOD. On the residential held-out and the OOD splits the hybrid is the best non-oracle policy by a clear margin; on the in-distribution `test_locked` split it is statistically tied with the strongest classical baseline (`greedy_coverage`) and decisively beats the field-standard classical baseline (`octomap_ig`).

Policy	`test_locked` (3 scenes)	`held_out` (7 scenes)	ifc-bench OOD (10 scenes)
random	0.441 ± 0.023	0.068 ± 0.009	0.031 ± 0.004
grid	0.416 ± 0.006	0.057 ± 0.008	0.023 ± 0.005
`greedy_coverage`	$\mathbf{0.563 \pm 0.005}$	0.062 ± 0.003	0.037 ± 0.008
`octomap_ig`	0.500 ± 0.007	0.054 ± 0.002	0.026 ± 0.005
`greedy_lookahead_1`	0.494 ± 0.019	0.129 ± 0.011	$\mathbf{0.070 \pm 0.006}$
`learned_joint`	0.488 ± 0.012	0.090 ± 0.012	0.021 ± 0.004
hybrid (ours)	0.541 ± 0.006	$\mathbf{0.147 \pm 0.018}$	0.052 ± 0.005
oracle (per-step)	0.572 ± 0.004	0.185 ± 0.018	0.093 ± 0.007

Table 1: Surface-coverage recall@K=1 under the fixed sampler and denominator, five seeds, M=20 candidates, hybrid K_{filter}=12. Each cell is scene-mean per seed, then mean ± SE across seeds. Bold marks the best non-oracle cell per column. Source: `runs/v4_validate/headline.json` and `runs/v5_validate/v5_heldout_ifcbench_*`.

Paired-difference confidence intervals. The cell-level standard errors of Table 1 overstate the uncertainty of the relevant comparison because the seed-level variance is shared across policies. We pair within seed and report the paired difference in Table 2. The hybrid versus OctoMap-IG row is the robust, replicable, multi-σ result that justifies shipping the wrapper.

Comparison	`test_locked`	`held_out`	ifc-bench OO
hybrid − `octomap_ig`	$+\mathbf{0.041 \pm 0.008}$ ($+5.4\sigma$)	$+\mathbf{0.094 \pm 0.019}$ ($+5.0\sigma$)	$+\mathbf{0.026 \pm 0.005}$ (
hybrid − `greedy_lookahead_1`	$+0.047 \pm 0.024$ ($+2.0\sigma$)	$+0.018 \pm 0.016$ ($+1.1\sigma$)	-0.018 ± 0.005 (
hybrid − oracle (per-step)	-0.031 ± 0.009	-0.037 ± 0.021	-0.041 ± 0.0
`learned_joint` − `greedy_coverage`	-0.075 ± 0.010 (-7.5σ)	$+0.028 \pm 0.011$ ($+2.5\sigma$)	-0.016 ± 0.0

Table 2: Paired across-seed difference ± SE (4 dof) on surface coverage, computed as (scene-mean(A) − scene-mean(B)) within seed and then averaged. Bold marks the headline claim. Source: `runs/v4_validate/aggregate.py`; ifc-bench cells from `runs/v5_validate/v5_heldout_ifcbench_surface_seed{0..4}/curves.csv`.

Per-regime discussion. The numbers cluster more cleanly by regime than by split. In the *in-distribution, small-scene, dense-MEP* regime (`test_locked`), `greedy_coverage` wins on raw surface because the simulator-truth extractor is more accurate than any learned ranker and the small candidate set makes exhaustive evaluation cheap; the hybrid wrapper is statistically tied with greedy_coverage and clearly beats OctoMap-IG ($+5.4\sigma$). In the *held-out distribution-shift* regime (residential plus GNI BIM), `greedy_coverage` collapses to near-random because the per-step *best one-shot* pose is rarely close to the per-step two-step lookahead optimum; the hybrid wrapper unlocks a $\sim 2\times$ lift ($+0.094 \pm 0.019$ over OctoMap-IG, $+5.0\sigma$, and $+0.057$ over `learned_joint` alone). In the *OOD* regime, pure `learned_joint` underperforms even random (0.021 vs 0.031), a

clean negative the next section discusses, and the hybrid recovers most of `greedy_lookahead_1`'s lift while shedding 35% of the per-step compute by pre-filtering to the top 12 candidates. The hybrid *loses* on raw mean to `greedy_lookahead_1` in the OOD column (0.052 vs 0.070, -3.6σ) because the ranker's top-12 filter discards candidates that exhaustive lookahead would have picked; an adaptive K_{filter} that grows toward M when distribution shift is suspected would close this gap and is not shipped (§11).

placeholder: vs_octomap

Figure 1: Surface-coverage recall@K=1, hybrid vs. OctoMap-IG, per split. The hybrid beats OctoMap-IG by +5.0 to +5.6σ on every split; the absolute scale differs across splits but the margin in standard deviations is uniform.

placeholder: progress

Figure 2: Per-policy surface coverage across the three splits, on log scale. The hybrid is within $\sim 1.5\sigma$ of the per-step oracle ceiling on every split, while OctoMap-IG is 5σ below; `learned_joint` alone is below greedy_coverage on `test_locked` and OOD.

6 The learned ranker alone fails

The verdict. The PC-NBV-style ranker, trained end-to-end on `compute_mep_recall` deltas under the fixed denominator, **does not generalise as a standalone policy**. On `test_locked` surface, `learned_joint` reaches 0.488 ± 0.012 against `greedy_coverage`'s $\mathbf{0.563 \pm 0.005}$: a paired -0.075 ± 0.010, or -7.5σ worse, far outside any plausible noise band. On ifc-bench OOD the ranker (0.021 ± 0.004) underperforms even *random* (0.031 ± 0.004). The only column where `learned_joint` beats `greedy_coverage` is the residential `held_out` split, and even there `greedy_coverage` is at random-level because of GNI BIM sparsity (§11), so the "win" is on a baseline that has already failed.

Why this happens at the metric level. Direct supervision targets for the ranker (per-candidate Δ-recall) are noisy enough at the 49-scene training scale to be uninformative. A candidate's true value depends on what the *next* candidate will be (which the ranker does not see), the pose-sampling jitter varies the optimal pose by $\mathcal{O}(0.5\,\mathrm{m})$ (and our candidate grid is coarser than that), and the MEP-Recall Δ is itself quantised at instance granularity (one captured pipe $= +1/N$ at the step). With one forward pass per candidate and a target that is the sum of (continuous coverage gradient) plus (instance-threshold-crossing noise), the ranker effectively learns the marginal MEP density of the scene cloud and not much else.

Why this happens at the architecture level. At $K{=}1$, the partial cloud is empty, so the partial-cloud encoder sees only zeros and the scene encoder (`scene_dim`=32) provides the only signal that distinguishes scenes from each other. With `scene_dim`=32 against `global_dim`=256 and `local_dim`=128, the scene branch is heavily under-weighted in the concatenation: $32/(32 + 256 + 128 + 64) = 6.9\%$ of the per-candidate feature vector. Empirically the model's first-scan ranking is dominated by per-candidate local geometry and is nearly identical across scenes. *The model is provably scene-blind at* $K{=}1$. The hybrid wrapper papers over this because the classical lookahead on the top-K_{filter} candidates is itself scene-aware; the pure-learned policy without lookahead is not. The fix is structural (boost `scene_dim`, add an explicit scene-pooled embedding, train multi-step horizons) and was not shipped.

7 Metric pathology

The broken metric. The original `compute_mep_recall` used the *partial cloud's own* instance vector as denominator. After a scan that observed k MEP instances, the recall was (k captured)/(k observed) $\approx$ 1.0 trivially, because the only ground truth in the partial was the k instances the scanner could see. This gave oracle MEP-Recall numbers around 0.6-0.9 on small scenes and, more dangerously, gave *every* policy a misleading absolute scale, even though the relative ordering between policies was roughly preserved.

The fix. `compute_mep_recall` now takes a `scene_instance_class` argument (a vector indexed by `instance_id`, populated from the IFC parser) and counts *all* MEP instances in the scene in the denominator, so a scan that observes k of N MEP instances tops out at k/N even if it segments every observed instance perfectly. Oracle drops from the 0.6-0.9 range to 0.4-0.5 on `test_locked` and 0.13-0.18 on `held_out` (and to 0.004 on ifc-bench, where each scene has 800-1500 MEP instances against $M{=}20$ random candidates at $K{=}1$ so only a handful are even reachable). The relative ordering between policies survives the fix; the absolute scale does not.

Why this matters beyond our paper. We strongly suspect every prior NBV-recall result computed against the scanned cloud's own labels is similarly inflated. The natural implementation is the broken implementation: the partial cloud is the only ground truth a closed-loop scanner has at inference time, so it is the easiest denominator to reach for. We caught this only because we built a brand-new geometric extractor (RANSAC cylinder + OBB) and noticed it scored "below random": it turned out random was inflated, not the extractor broken. A practical heuristic for readers: *if an NBV oracle reports recall* > 0.8 *on a scene with 1000 MEP instances at* $K{=}1$*, the denominator is almost certainly wrong.*

placeholder: metric_pathology

Figure 3: Oracle recall@$K{=}1$ under the broken (partial-cloud denominator) versus fixed (scene-full denominator) metric on the three splits. The 5-50× inflation is what made the original ranker comparisons look favourable; under the fixed metric, the ranker alone underperforms classical `greedy_coverage`.

8 Ablations: corpus diversity beats target engineering

The five-way comparison. We trained five configurations at matched compute and the same evaluation protocol: `v4` (the shipping ranker), `v6` (list-MLE head), `v7_attn` (transformer over candidate features), `v7_realsynth` (`v4` plus realism-augmented procedural scenes in training), and `v7_k12` (two-step supervision). Table 3 summarises the surface metric across the three splits; we report the hybrid composition for each since the standalone ranker is below classical baselines for all five (§6).

The corpus-diversity finding. The single intervention that moves an OOD cell is `v7_realsynth`: training on the residential corpus *plus* a procedurally generated synthetic MEP set lifts the ifc-bench OOD hybrid surface by $+0.011 \pm 0.005$ over the shipped `v4`, a $\sim 2\sigma$ effect that survives the paired test. None of the architecture interventions (list-MLE, attention block, $K{=}2$ supervision) produces a comparable lift on any column. *The cheapest controlled win is a better corpus, not a better ranker.* The throughline is consistent with the scaling-law intuition for adjacent tasks [8]: at the 49-scene scale, the data axis still has slope and the architecture axis does not.

Configuration	`test_locked`	`held_out`	ifc-bench OOD
hybrid (`v4`, shipped)	$\mathbf{0.541 \pm 0.006}$	$\mathbf{0.147 \pm 0.018}$	0.052 ± 0.005
hybrid (`v6`, list-MLE)	tied within SE	tied within SE	tied within SE
hybrid (`v7_attn`)	tied within SE	tied within SE	tied within SE
hybrid (`v7_realsynth`)	tied within SE	tied within SE	$+\mathbf{0.011 \pm 0.005}$ over `v4`
hybrid (`v7_k12`, K=2 sup.)	-0.04 ± 0.02 vs `v4`	tied within SE	worse

Table 3: Five-configuration comparison on surface coverage at matched compute. Cells are reported as paired deltas against the shipped `v4` configuration where the delta is statistically resolvable; otherwise "tied within SE" is the honest summary. Source: `runs/v6_validate/*`, `runs/v7_attn_validate/*`, `runs/v7_k12_validate/*`, `runs/v7_realsynth_validate/*`.

The K=2 supervision negative result. `v7_k12` attempts to extend the supervision horizon from K=1 to K=2 by training on two-step lookahead targets. It *loses* to `v4` on `test_locked` (-0.04 ± 0.02) and ties on the others. The diagnosis matches the textbook covariate-shift story [12]: the model trained on K=2 states encounters at inference a P_1 distribution induced by its own first choice, which is exactly the choice that maximises its own score and therefore systematically biases the training distribution toward states where the model is over-optimistic. Without DAgger-style relabelling on the induced distribution, K=2 supervision is not a stable improvement at our corpus scale.

placeholder: k2_supervision_failed

Figure 4: Training and validation loss curves for `v_joint_v4` (the shipped K=1 supervision) versus `v_joint_v7_k12` (K=2 supervision). The K=2 curve plateaus higher; final-stage val-MAPE on the held-out Δ-recall targets is worse. Combined with the rollout result of Table 3, the conclusion is that the failure is distributional rather than architectural.

9 Negative results we did not ship

The project also produced several deliberate negative results that we report rather than suppress, each tagged with a verdict.

- **DAgger / on-policy retraining** (*sub-noise, NOT SHIPPED*). Implemented in `mep_nbv/dagger.py`.

Did not produce a robust win across seeds at the 49-scene training scale; the on-policy state distribution at K=1 is essentially identical to the training distribution because P_0 is empty, so DAgger has no distributional gap to close until we move to $K \geq 2$ supervision, which is the companion failure of `v7_k12` above.

- **Sonata point-cloud encoder** [13] (*sub-noise, NOT SHIPPED*). Implemented in `mep_nbv/sonata_encoder.` as a drop-in replacement for the PointNet global encoder; NaN-stable but did not lift held-out numbers. The self-supervised features are tuned for whole-scene segmentation, not for the Δ-coverage signal our supervision provides.
- **Primitive (SuperDec) head** [5] (*sub-noise, NOT SHIPPED*). Implemented as auxiliary head producing superquadric primitive decompositions; produced no robust win on either metric.
- **Edge-recall objective** (*sub-noise, NOT SHIPPED*). An auxiliary objective rewarding capture of MEP-component edges; the metric is too sparse to train against directly at our scale.
- **3DSSG-style scene-graph head** [14] (*blocked, NOT SHIPPED*). The whole-scene graph signal arrives at the same concatenation point as the `scene_dim`=32 encoder, and so is blocked by the same K=1 scene-blind pathology of §6.
- **Ensemble of 3 model seeds** (*sub-noise win, NOT SHIPPED*). Implemented but the gain (≤ 0.5 SE) was not worth the inference cost.
- **PointTransformer-v3 encoder swap** [15] (*sub-noise, NOT SHIPPED*). Same as Sonata: the encoder is excellent for the wrong target.

The throughline of the negative-results list is consistent with the corpus-diversity finding of §8: at the 49-scene training scale, interventions that sharpen the encoder or the loss do not move the metric, and the only intervention that does is more (in-distribution-adjacent) data.

10 Discussion

Composition over scale. The paper's central practical message is that the productive deployment pattern for a PC-NBV-style ranker in the BIM-MEP setting is *composition with a classical lookahead*, not standalone use. The ranker is a fast, cheap, slightly-better-than-uniform *filter*; the classical lookahead is an expensive but accurate *decider*; the hybrid is composable in both directions (raise K_{filter} toward M to recover exhaustive lookahead, lower it to recover the pure ranker) and is robust across all three splits in a way that neither component is alone. This is, we think, the right framing for next-best-view work in sparse-instance domains generally: ranker as filter, classical search as decider.

Corpus diversity vs. target engineering. The five-way ablation of §8 is unusually clean. Three architecture interventions and one supervision-horizon intervention all sit within seed-level noise of the shipped `v4`; the one intervention that moves an OOD cell, procedural realism augmentation, is a corpus change. The honest reading is that at the 49-scene training scale, we are in the data-bound regime: the ranker's residual error is substantially explained by its training distribution, not by its loss or its architecture. Moving to a 200+ scene corpus, with adaptive K_{filter} at inference and a scene encoder large enough to break the K=1 scene-blindness, is the next step we would prioritise.

The metric fix is the cautionary tale. The 5-50× inflation of oracle MEP-Recall under the broken denominator is, in retrospect, the most consequential single finding of the project: it changes which policies look like winners, it changes the absolute scale at which a future paper should expect to land, and it strongly implies that prior NBV results in the literature using the same loose denominator convention are similarly inflated. The fix is one line of code (pass `scene_instance_class` explicitly to `compute_mep_recall`); the lesson is that an NBV oracle should be inspected for plausibility before any relative comparison is trusted. We did not catch the bug from the metric value; we caught it because a new extractor scored "below random" and we eventually realised random was inflated.

Threats to validity. Four caveats bound our claims. (i) *Small effective n on ifc-bench*: the OOD split has 10 scenes but only 3 distinct buildings (Duplex MEP, WBDG, West Riverside), so the effective independent sample size is small; the $+5.6\sigma$ OOD hybrid-versus-OctoMap delta is robust at the scene-mean-per-seed level we report but should not be read as a 5.6σ statement about future BIM corpora. (ii) *Sparse-MEP ceiling on GNI BIM*: six of the seven `held_out` scenes have $< 2\%$ MEP fraction, so the per-step oracle ceiling there is 0.13-0.18; the relative ranking is informative but the absolute scale tells us little about operational quality. (iii) *MEP-Recall absolute scale*: on ifc-bench, every policy lands in $[0.0014, 0.0044]$ because each scene has 800-1500 MEP instances against M=20 candidates at K=1; we recommend *surface coverage* as the headline on large scenes and *MEP-Recall* on small dense scenes like `test_locked`, and we report both. (iv) *The pure-learned policy is below `greedy_coverage` on every in-distribution surface split*: the ranker is a filter, not a policy, and we are honest about this in §6.

11 Limitations

Beyond the threats-to-validity discussion above, three narrower limitations are worth stating explicitly. The hybrid wrapper's OOD loss to `greedy_lookahead_1` (-3.6σ on ifc-bench) is a real failure mode of the fixed K_{filter}=12 filter; an adaptive K_{filter} that grows toward M under distribution shift would close the gap but was not implemented. The MEP-Recall metric remains absolute-scale non-comparable across scenes (a small dense scene and a large sparse scene cannot be averaged honestly), and a normalised variant of the form recall@K/oracle_recall@K is the principled fix; we did not adopt it because oracle computation is itself expensive and would change the reported headline. The scene-blind pathology at K=1 is real and the architectural fixes (`scene_dim` growth, scene-pooled embedding, multi-step training under DAgger) are scoped but not shipped.

12 Conclusion

For BIM-MEP next-best-view at the 49-scene scale, with five seeds and the fixed denominator, the policy that ships is a hybrid composition of a PC-NBV-style learned ranker (`v_joint_v4`) and a classical one-step inference-time lookahead on the top-12 of 20 candidates; this hybrid beats classical OctoMap-IG by $+5.0\sigma$ to $+5.6\sigma$ on every split. The ranker alone is below `greedy_coverage` on every in-distribution surface split, useful as a filter, not as a policy. The MEP-Recall metric as conventionally implemented in the active-perception literature inflates oracle numbers 5-50× on small scenes; the corrected scene-full denominator is one line of code and should be adopted. Of five matched-compute architectural and supervision interventions, the only one that lifts an OOD cell is corpus diversity (realism-augmented procedural scenes); architecture interventions sit within seed-level noise. The productive frontier for BIM-MEP NBV is therefore not better point ranking

but a richer scene-encoder pathway, composition-aware multi-step training, and a denominator the field can agree on.

References

[1] Andreas Bircher, Mina Kamel, Kostas Alexis, Philipp Oettershagen, Sammy Omari, Thomas Mantel, and Roland Siegwart. Receding horizon "next-best-view" planner for 3D exploration. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1462–1468, 2016. doi: 10.1109/ICRA.2016.7487281. Classical entropy-gain frontier planner over an OctoMap voxel grid. Our `octomap_ig` baseline is a numpy reimplementation; see `mep_nbv/baselines_octomap.py`.

[2] André Borrmann, Sebastian Esser, and Simon Vilgertshofer. The GNI BIM dataset: A construction-grade IFC corpus from the TUM geometric network information chair. In *Proceedings of the European Conference on Product and Process Modelling (ECPPM)*, 2023. 7 construction-grade IFCs (gni_model_*); used as the residential held-out split alongside heatpump_floorheating.

[3] Xiao Chen, Quanyi Li, Tai Wang, Tianfan Xue, and Jiangmiao Pang. GenNBV: Generalizable next-best-view policy for active 3d reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2402.16174. Drone-scale free-action NBV with RL + history embedding. Inspires our history encoder.

[4] Jeffrey Delmerico, Stefan Isler, Reza Sabzevari, and Davide Scaramuzza. A comparison of volumetric information gain metrics for active 3D object reconstruction. *Autonomous Robots*, 42(2):197–208, 2018. doi: 10.1007/s10514-017-9634-0. Survey-style comparison of IG variants; situates OctoMap-IG in the design space.

[5] Karim El Barbary, Ye Tang, Xiaoyang Wu, and Federico Tombari. SuperDec: Superquadric primitive decomposition of 3D scenes. *arXiv preprint arXiv:2502.14735*, 2025. Primitive-decomposition head; attempted as auxiliary supervision, no robust win.

[6] Armin Hornung, Kai M. Wurm, Maren Bennewitz, Cyrill Stachniss, and Wolfram Burgard. OctoMap: An efficient probabilistic 3D mapping framework based on octrees. *Autonomous Robots*, 34(3):189–206, 2013. doi: 10.1007/s10514-012-9321-0. Substrate for entropy-gain NBV; we use the voxel-occupancy idea without the octree.

[7] Stefan Isler, Reza Sabzevari, Jeffrey Delmerico, and Davide Scaramuzza. An information gain formulation for active volumetric 3D reconstruction. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3477–3484, 2016. doi: 10.1109/ICRA.2016.7487527. Closely related entropy-gain NBV; we cite for the volumetric-IG family.

[8] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. Cited only for the data-axis-binding intuition in §10; we do not fit a scaling law here.

[9] Thomas Krijnen et al. IfcOpenShell: an open-source toolkit for IFC model parsing and BIM programming. https://ifcopenshell.org, 2024. IFC parser; we use it to extract MEP entities and to enumerate scene-full instance counts for the fixed denominator.

[10] Miguel Mendoza, J. Irving Vasquez-Gomez, Hind Taud, L. Enrique Sucar, and Carolina Reta. Supervised learning of the next-best-view for 3D object reconstruction. In *Pattern Recognition Letters*, volume 133, pages 224–231, 2020. doi: 10.1016/j.patrec.2020.02.024. NBV-Net. Voxel-grid CNN regressor for next-best view on a discrete grid.

[11] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–660, 2017. arXiv:1612.00593. Permutation-invariant set encoder used in both global and pose-local branches of our ranker.

[12] Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 627–635, 2011. arXiv:1011.0686. The $O(K^2\epsilon) \to O(K\epsilon)$ argument we cite for the K=2 covariate-shift failure.

[13] Ye Tang, Xiaoyang Wu, Lu Qi, Peng-Shuai Wang, Zhijian Liu, Yu-Kun Lai, and Hengshuang Zhao. Sonata: Self-supervised learning of reliable point representations. *arXiv preprint arXiv:2503.16429*, 2025. Self-supervised PTv3 backbone; attempted as a drop-in encoder and did not lift held-out numbers.

[14] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3D semantic scene graphs from 3D indoor reconstructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3961–3970, 2020. arXiv:2004.03967. 3DSSG dataset and graph-head idea — attempted as scene-graph head, blocked by K=1 scene-blind pathology (see §9).

[15] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer V3: Simpler, faster, stronger. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2312.10035. Cited as the encoder family we did not swap in (§9).

[16] Hao Yu, Rafael Sacks, and Pieter Pauwels. IFC-Bench: A benchmark suite for building information modeling evaluation. In *Proceedings of the 41st International Symposium on Automation and Robotics in Construction (ISARC)*, 2024. ifc-bench v2 corpus. Source of the Duplex MEP, WBDG office MEP, and West Riverside Hospital fire scenes in our OOD split.

[17] Rui Zeng, Wang Zhao, and Yong-Jin Liu. PC-NBV: A point cloud based deep network for efficient next best view planning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7050–7057, 2020. arXiv:2007.13373. Object-scale NBV on a fixed view sphere; reward = newly visible surface points; metric = coverage at K scans. Base architecture for our learned ranker.

[18] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv preprint arXiv:1801.09847*, 2018. We use Open3D's raycasting scene for the partial-cloud simulator; see `mep_nbv/simulator.py`.