DEPTH ANY SEEN: WHICH SURFACES AND HOW FAR?

Anonymous authors
Paper under double-blind review

ABSTRACT

When several surfaces are visible along a ray, recovering visible 3D structure from one image requires jointly estimating their presence and metric depth. Depth Any Seen represents these surfaces as image-conditioned multi-Bernoulli depth sets, whose components each contribute one depth or remain absent. Its auxiliary-free Exact Multi-Bernoulli objective (ExactMB) learns depth and presence by marginalizing one-to-one assignments to complete, distinct targets. Our analysis shows that matching expected count can leave component–surface assignment unresolved. We extend real and synthetic layered-depth benchmarks to evaluate depth accuracy, recovered support, and overprediction. Compared to depth stacking, ExactMB reduces overprediction by a relative 88.2% on LD-Real and 80.5% on MD-3K while retaining most ordinal accuracy, with comparable conditional metric-depth error on LD-Syn. Further ablation studies show that ordered assignment improves depth-accurate recall and precision over marginalization, whereas the count-regularized configuration achieves higher deeper-rank precision than ordered assignment at lower recall. Our code will be publicly released.

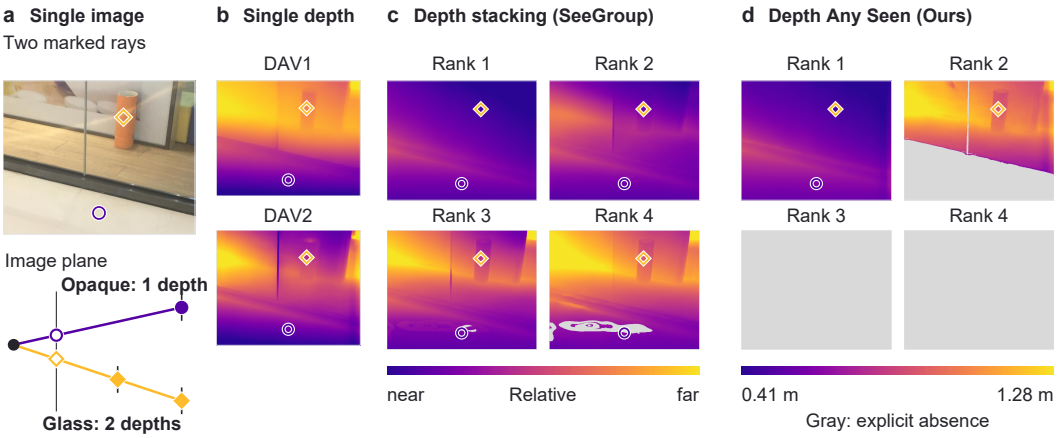


Figure 1: **Recovering visible 3D geometry requires two answers: which surfaces are present, and how far away are they?** (a) A single RGB image: purple circles and amber diamonds mark opaque and through-glass rays. (b) Depth Anything V1/V2 (Yang et al., 2024a;b) compress layered visibility into one depth per ray. (c) SeeGroup's released evaluation decoder (Wen & Deng, 2026) retains up to four ranked relative depths per ray. (d) Depth Any Seen predicts ray-adaptive metric depth sets with explicit absence for unused components. Gray denotes no retained depth.

1 INTRODUCTION

A single image can reveal more surfaces than a single depth map can represent. Through glass, the surface and the scene behind it can both be visible along the same ray. **Recovering everything seen requires learning which surfaces are present and where they lie.** As shown in Figure 1(d), their depths must share a metric scale to capture both variation within layers and separation between them.

Layered depth images, support masks, and adaptive layers represent multiple surfaces (Shade et al., 1998; Dhamo et al., 2019). LayeredDepth supplies large-scale synthetic layered-depth data (Wen

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.

et al., 2025), while SeeGroup learns unordered components with variable retained counts (Wen & Deng, 2026). Recovery also requires deciding which predictions represent visible surfaces. Count alone is insufficient: extra predictions can offset missing surfaces. Depth error on retained predictions cannot reveal omissions. This motivates joint depth–presence learning and support-aware evaluation.

Depth Any Seen learns these decisions jointly through an **image-conditioned random finite set**. Its elements are coexisting surfaces with uncertain presence and depth, not alternative estimates of one depth. We leverage multi-Bernoulli set prediction (Hess et al., 2022): components independently emit one depth or remain absent. Our Exact Multi-Bernoulli likelihood (ExactMB) marginalizes one-to-one assignments to complete, distinct targets, coupling presence and depth.

Components receive presence credit for explaining observed depths. These credits sum to the target count, linking geometric explanation to presence supervision. Our analysis distinguishes expected-count correction from the allocation of presence credit among components. **Matching expected count can therefore leave assignment unresolved.** We consequently test how count-regularized configurations and assignment supervision shape visible-surface recovery beyond ExactMB.

Our joint depth–presence benchmarks combine synthetic metric depth, real-data ordinal relations and membership, and fixed-ray set recovery to assess where surfaces lie and which are recovered.

This evaluation shows that auxiliary-free ExactMB reduces annotation-defined LD-Real overprediction from 98.4% to 11.6% compared with depth stacking, retaining most ordinal accuracy on LD-Real and MD-3K with comparable conditional metric-depth error on LD-Syn. Additional supervision changes this balance: the count-regularized configuration improves aggregate count accuracy and deeper-rank precision, while ordered assignment improves depth-accurate recall and precision over marginalization, with a higher mean MD-3K overprediction rate.

Contributions. 1) Joint depth and membership. ExactMB specializes the multi-Bernoulli model to dense visible-depth sets with exact assignment marginalization and no auxiliary losses. **2) Count and assignment supervision.** We show that matching expected count need not resolve surface assignment, and that count-regularized configurations and ordered assignment favor different precision–recall trade-offs. **3) Joint depth and presence benchmarking.** We extend existing multilayer-depth benchmarks to jointly evaluate depth accuracy and explicit surface presence. Results show lower overprediction than depth stacking while retaining most ordinal accuracy, and expose localization–support trade-offs across emission and assignment design choices.

2 Related Work

Visible-depth and amodal prediction. LayeredDepth presents a large-scale synthetic layered-depth dataset (Wen et al., 2025), while SeeGroup learns unordered components with intensity and coverage losses (Wen & Deng, 2026). MDA permits transparent multilayers through independent weights (Bian et al., 2026). With some frozen monocular models, Xu et al. (2026) elicit unordered ordinal depth pairs using RGB and Laplacian Visual Prompting (LVP) in two-layer transparent scenes. Amodal methods extend beyond visible surfaces: LaRI learns stopping for amodal intersections (Li et al., 2026), World Tracing reconstructs visible and occluded intersections (Zhang et al., 2026), and TRELLIS and SAM 3D generate complete object geometry (Xiang et al., 2025; 2026; Chen et al., 2026b). Our objective instead couples visible-set membership with metric depth to recover coexisting surfaces seen along each ray, including through transparent foreground objects.

Feed-forward geometry. MVSNet (Yao et al., 2018) learns depth from calibrated views. DUSt3R (Wang et al., 2024) regresses pairwise pointmaps, and MASt3R (Leroy et al., 2024) adds local matching features. VGGT (Wang et al., 2025) and π^3 (Wang et al., 2026b) predict cameras and dense geometry, while MapAnything (Keetha et al., 2026) unifies metric reconstruction. Our complementary goal is coexisting visible depths and their membership on a shared metric scale from one image.

Layered geometry and rendering. Learned layers support view synthesis (Tulsiani et al., 2018), masked reconstruction (Shin et al., 2019), and object-wise decomposition (Dhamo et al., 2019), while layered inpainting adds occluded samples (Shih et al., 2020). For insertion and shading, Engel et al. (2024) infer depth, color, and opacity intervals from semitransparent volume renderings.

Transparent-scene depth and reconstruction. ClearGrasp (Sajjan et al., 2020) and TransCG (Fang et al., 2022) advance RGB-D completion, and local implicit functions jointly predict termination

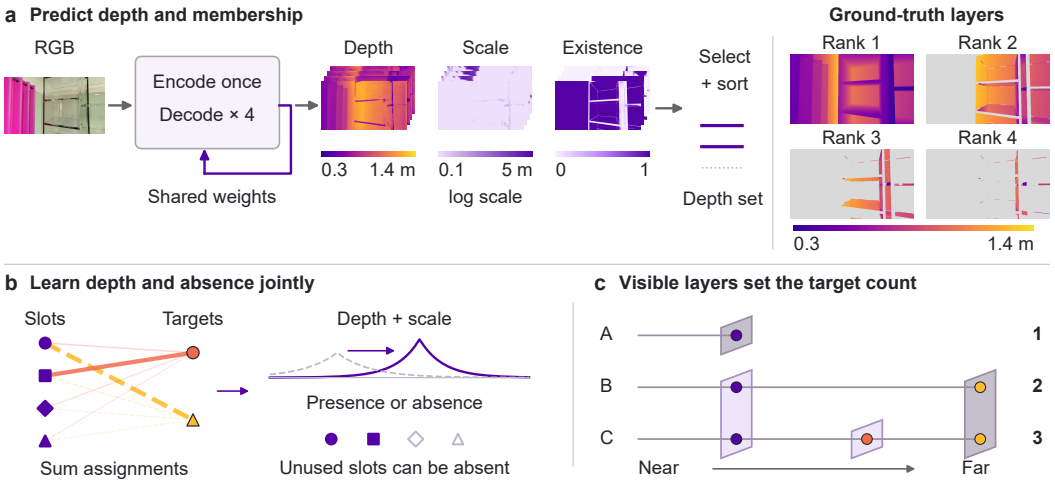


Figure 2: **One likelihood learns depth and absence.** (a) Our default feed-forward model uses recurrent decoding to explicitly predict depth, scale, and existence probability. Selection and sorting form the output set. (b) ExactMB sums over one-to-one assignments to explain observed returns and score unused components as absent. Ordered assignment retains optional emission. (c) Annotated visibility determines the target return count. Scale describes localization conditional on presence.

probability and position (Zhu et al., 2021). Depth4ToM (Costanzino et al., 2023), MODEST (Liu et al., 2025), and SeeClear (Wang et al., 2026a) recover selected surface depth. For multilayer recovery, ASGrasp (Shi et al., 2024) reconstructs two layers from RGB and active stereo for grasping, while DepthFocus (Min et al., 2026) selects layers through distance queries on stereo observations. Using only one monocular image, we jointly model presence and metric depth across visible layers.

3 Depth Any Seen via Ray-Adaptive Depth Sets

Problem statement and assumptions. Given one image, we seek a set of visible depths per viewing ray. Each layer records an intersected object surface's optical-axis depth in meters. The target $\mathcal{T} = \{t_1, \ldots, t_M\}$ contains exactly M depths, with K the maximum supported cardinality ($0 \leq M \leq K$). We assume that objects through which farther layers are visible are transparent and locally modeled as finite-thickness glass. Geometry hidden behind opaque surfaces is excluded. Ray-wise targets leave cross-pixel surface identities unspecified. Qualitative out-of-distribution examples explore irregular transparent objects beyond this approximation.

3.1 Continuous depth, discrete membership

Figure 2(a) shows the prediction pipeline. The image is encoded once, and shared-weight recurrent decoder passes predict K triples (d_j, β_j, e_j): depth center, positive localization scale, and existence logit. A Bernoulli variable $R_j \sim \text{Bernoulli}(q_j)$ determines whether component j contributes a return. It is absent when $R_j = 0$ and emits one depth when $R_j = 1$, with an untruncated Laplace density

$$q_j = \sigma(e_j), \qquad \ell_j(t) = \frac{1}{2\beta_j} \exp\left(-\frac{|t - d_j|}{\beta_j}\right). \tag{1}$$

Here β_j measures conditional localization spread, with a positive floor preventing unbounded likelihood. Densities are normalized on $\mathbb{R}$, although targets and decoded depths are positive. Components are independent given the image and network outputs, defining a multi-Bernoulli random set X. Continuous draws are distinct almost surely, so $|\mathsf{X}| = \sum_j R_j$.

Optional emission lets count vary within capacity K. Component identities are not depth ranks: assignments link components to targets, while selection, sorting, and filtering define decoded ranks.

3.2 EXACTMB: ONE LIKELIHOOD FOR DEPTH AND ABSENCE

To score targets without prescribing component ranks, each one-to-one assignment, or injection, $\phi : \{1, \dots, M\} \hookrightarrow \{1, \dots, K\}$ matches each observed depth to a distinct predicted depth component, with whole-set density

$$f(\mathcal{T}, \phi) = \underbrace{\prod_{i=1}^{M} q_{\phi(i)} \ell_{\phi(i)}(t_i)}_{\substack{\text{explain observed} \\ \text{returns}}} \underbrace{\prod_{j \notin \operatorname{im} \phi} (1 - q_j)}_{\substack{\text{leave unused} \\ \text{components absent}}}. \tag{2}$$

As shown in Figure 2(b), each assignment explains the targets and leaves unused components absent. ExactMB sums their densities to evaluate the multi-Bernoulli likelihood:

$$\boxed{f(\mathcal{T}) = \sum_{\phi \in \mathrm{Inj}(M,K)} f(\mathcal{T}, \phi), \qquad \mathcal{L}(\mathcal{T}) = -\log f(\mathcal{T}).} \tag{3}$$

The sum is invariant to target enumeration. Each explanation gives distinct targets distinct owners, while allowing coincident component centers. Absence factors score unused components, yielding $f(\varnothing) = \prod_j (1 - q_j)$ for the empty set. Absence denotes visible-set nonmembership, not physical nonexistence. ExactMB thus jointly scores localization and membership without auxiliary penalties. Appendix A.1.3 defines the finite-set measure and proves normalization and log-score propriety.

Training and inference. We train the reference model by averaging the ExactMB loss in Equation (3) over image rays. At inference, we threshold the predicted existence maps, then sort and de-duplicate the retained depths following SeeGroup's released evaluation code (Wen & Deng, 2026).

3.3 HOW ASSIGNMENT SUPERVISES DEPTH AND PRESENCE

Marginalized assignments determine how each component learns. Normalized explanation weights give the probability P_{ji} that component j explains target i. Posterior occupancy ρ_j sums these responsibilities over targets:

$$\omega_\phi = \frac{f(\mathcal{T}, \phi)}{f(\mathcal{T})}, \qquad P_{ji} = \sum_{\phi : \phi(i) = j} \omega_\phi, \qquad \rho_j = \sum_i P_{ji}. \tag{4}$$

Here ρ_j is the probability that component j explains any target. Whereas q_j is predicted from the image, posterior occupancy also uses observed depths and competing components. Differentiating the finite marginal likelihood gives the presence gradient

$$\frac{\partial \mathcal{L}}{\partial e_j} = q_j - \rho_j, \tag{5}$$

and the geometric gradients

$$\frac{\partial \mathcal{L}}{\partial d_j} = \sum_{i=1}^{M} P_{ji} \frac{\operatorname{sign}(d_j - t_i)}{\beta_j}, \qquad \frac{\partial \mathcal{L}}{\partial \beta_j} = \sum_{i=1}^{M} P_{ji} \left(\frac{1}{\beta_j} - \frac{|d_j - t_i|}{\beta_j^2} \right). \tag{6}$$

The same assignment responsibilities supervise depth and presence: each component earns presence credit by plausibly explaining an observed surface. Output derivatives use an absolute-value subgradient at zero residual. Raw-head gradients also differentiate through Softplus and scale clipping.

As shown in Figure 2(c), presence credits sum to the annotated target count, $\sum_j \rho_j = M$. Summing the presence gradients therefore gives

$$\sum_j \frac{\partial \mathcal{L}}{\partial e_j} = \sum_j q_j - M = \mathbb{E}|\mathsf{X}| - |\mathcal{T}|. \tag{7}$$

The aggregate signal leaves component–surface assignment unresolved. Each assignment selects an occupied subset and pairs it with targets. At $M = 1$, within-subset correspondence is unique,

leaving only selection ambiguous. At $M = K$, the subset is fixed and every $\rho_j = 1$ despite uncertain correspondence. Intermediate counts involve both decisions.

ExactMB scores observed cardinality jointly with geometry conditional on count:

$$\mathcal{L}(\mathcal{T}) = \underbrace{-\log \Pr(|\mathsf{X}| = M)}_{\text{observed cardinality}} + \underbrace{\left[-\log f(\mathcal{T} \mid |\mathsf{X}| = M)\right]}_{\text{geometry at that cardinality}}. \tag{8}$$

The factors share parameters: relative presence probabilities can affect conditional geometry. In contrast, for complete distinct targets with $0 < M < K$, finite logits, fixed centers, and positive fixed scales, a common logit shift preserves posterior assignment probabilities. Appendix A.3 proves that its unique minimum of the set loss matches expected count to M.

4 Evaluating Visible Depth Layers

Count and assignment must also be distinguished in evaluation: correct counts can hide missed surfaces and extra predictions, while correct ordering can coexist with incorrect presence. We distinguish whether a surface is recovered from how accurately its depth is estimated. We extend existing benchmarks to measure depth accuracy, recovered support, and overprediction together.

4.1 Benchmarks and annotation conventions

LayeredDepth (LD) (Wen et al., 2025) provides 300 real validation images (LD-Real) with sparse ordinal queries and 500 synthetic images (LD-Syn) with metric-depth channels L1/L3/L5/L7. Headline LD-Real evaluation uses valid and all-absent quadruplets that can span rays. **MultiDepth-3K (MD-3K)** complements these with 3,161 material-labeled point pairs (Xu et al., 2026). We align its up-to-two-layer annotations with LD: L1 denotes foreground and L3 optional background. For presence evaluation, L1 is always valid, L3 only at transparent points, and unused slots are absent. Dense synthetic labels test metric depth and count, while sparse real labels test presence and ordering.

4.2 Joint depth–presence evaluation

Ordinal accuracy and presence. LD-Real ACC requires all requested depths in a valid quadruplet to be present in annotated strict near-to-far order. Conversely, OverPred is the fraction of all-absent quadruplets with any requested prediction. On MD-3K, pattern ACC requires both points' validity patterns to match their targets. Recall and OverPred measure L3 emission at transparent and opaque points, respectively. Headline ordinal ACC requires correct foreground and background ordering on transparent–transparent (TT) and transparent–opaque (TO) pairs. Background order uses L3 at transparent points and L1 at opaque points, without filling missing L3. Joint ACC requires both pattern and ordinal correctness across all material-pair types.

Metric depth and count. On LD-Syn, decoded ranks 1–4 are paired with corresponding target channels. AbsRel averages relative depth error where target and prediction are both valid. Because omitting difficult surfaces can lower this conditional error, we also report target-support recall: the fraction of target-valid pixels with a valid prediction. Depth-accurate recall additionally requires $\max(\hat{d}/d, d/\hat{d}) < 1.25$, while depth-accurate precision divides the same accurate matches by all prediction-valid pixels. Count ACC/MAE pool occupied target rays and retain supplied ties.

Shared protocol and reporting. Primary ablations use native selection followed by common geometric filters, without alignment or clipping. Summary AbsRel equally weights four conditional rank errors per seed. We report means and sample SD across seeds. Rank-wise scores complement fixed-ray set-recovery diagnostics. Rates and AbsRel are percentages, while count MAE is in layers.

5 Experiments

We first compare ExactMB with depth stacking and released models, then examine count and assignment choices. Primary ablations and a separate matched-recipe study use four seeds per configuration, reporting means±sample SD and seed-paired differences.

Table 1: **Primary cardinality and assignment ablations.** Separately trained variants (means±sample SD). All denotes depth stacking and Marginalized denotes ExactMB. Count reg.: explicit count regularization. Ordered: compacted annotation-channel assignment. Rows (iii–v) differ only in recorded assignment per seed. Section 4 defines ACC/OverPred. AbsRel (%) averages conditional errors over four ranks. MAE is in layers. Markers match Figure 3(a).

			LD-Real		MD-3K			LD-Syn		
	Emission	Assignment	Ordinal ACC↑	OverPred↓	Recall↑	OverPred↓	Pattern ACC↑	AbsRel↓	Count ACC↑	MAE↓
(i)	■ All	Ordered	$67.4_{\pm 0.7}$	$98.4_{\pm 0.3}$	$100.0_{\pm 0.0}$	$98.5_{\pm 0.2}$	$0.03_{\pm 0.04}$	$18.3_{\pm 0.1}$	$3.7_{\pm 0.2}$	$2.244_{\pm 0.031}$
(ii)	● Count reg.	Ordered	$58.9_{\pm 0.9}$	$7.4_{\pm 0.6}$	$79.1_{\pm 1.0}$	$8.6_{\pm 0.5}$	$65.9_{\pm 1.2}$	$17.2_{\pm 0.1}$	$95.0_{\pm 0.0}$	$0.067_{\pm 0.001}$
(iii)	▲ Presence	MAP	$60.7_{\pm 1.2}$	$12.1_{\pm 1.1}$	$81.7_{\pm 0.7}$	$20.9_{\pm 1.4}$	$47.7_{\pm 2.7}$	$18.7_{\pm 0.3}$	$88.4_{\pm 0.2}$	$0.183_{\pm 0.002}$
(iv)	◆ Presence	Marginalized	$61.7_{\pm 1.8}$	$11.6_{\pm 1.3}$	$81.5_{\pm 2.9}$	$19.2_{\pm 3.8}$	$46.3_{\pm 4.0}$	$19.2_{\pm 0.2}$	$88.2_{\pm 0.5}$	$0.194_{\pm 0.007}$
(v)	▼ Presence	Ordered	$63.2_{\pm 1.9}$	$10.1_{\pm 1.2}$	$82.6_{\pm 3.9}$	$21.7_{\pm 1.2}$	$52.4_{\pm 3.1}$	$17.6_{\pm 0.2}$	$89.7_{\pm 0.1}$	$0.154_{\pm 0.002}$

Table 2: **Released-model transfer references.** Single estimates from unretrained models for different tasks. Scale denotes native output. † AbsRel (%) averages four conditional rank errors after per-image/layer GT least-squares calibration and clipping; primary errors are unaligned.

		LD-Real		MD-3K			LD-Syn		
Released model	Scale	Ordinal ACC↑	OverPred↓	Recall↑	OverPred↓	Pattern ACC↑	GT-cal. AbsRel† ↓	Count ACC↑	MAE↓
SeeGroup (Wen & Deng, 2026)	Relative	74.2	100.0	100.0	100.0	0.0	15.8	2.4	2.756
WT r69l (Zhang et al., 2026)	Relative	58.2	100.0	100.0	100.0	0.0	24.5	2.4	2.789
LaRI scenes (Li et al., 2026)	Relative	52.0	100.0	100.0	100.0	0.0	29.5	2.4	2.789
LaRI objects (Li et al., 2026)	Relative	12.9	50.2	79.1	80.5	38.1	37.7	33.1	0.750
WT r75b (Zhang et al., 2026)	Metric	17.0	100.0	100.0	100.0	0.0	46.1	2.4	2.789

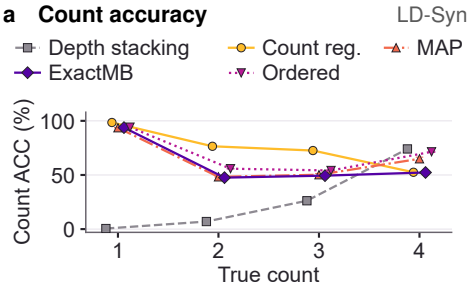
Main ablation setup. Table 1 compares: **(i) Depth stacking** regresses supplied channels with SeeGroup's decoder, selecting all K slots without absence supervision. It differs from native SeeGroup's many-to-one maximum-density coverage. **(ii) Explicit count regularization** adds categorical count supervision, averages valid-entry depth loss, and selects the argmax-count prefix. **(iii) MAP assignment** uses the highest-scoring injective Bernoulli–Laplace explanation. **(iv) ExactMB** marginalizes all injective explanations without fixed assignment constraints. **(v) Ordered assignment** fixes matches in compacted annotation-channel order, marking matched components present and others absent. Rows (i–ii) compare complete constant-gate configurations differing in count supervision, depth-loss reduction, and selection. Rows (iii–v) share presence-gated recurrence and the remaining recorded training settings, differing only in their assignment criterion within each seed.

Implementation. All variants use 14,800 synthetic training images, $K = 4$, terminal-800k checkpoints, and zero geometric auxiliary weights. Real benchmarks assess synthetic-to-real transfer. They share metric Depth Anything V2-L initialization (Yang et al., 2024b) and SeeGroup's decoder (Wen & Deng, 2026). Presence-based variants select $q_j > .01$. All variants then sort finite depths above .02 m and retain centers more than .02 m beyond the last retained depth. These filters can reduce even the depth-stacking count. Further details appear in Appendix B.1.

5.1 Joint depth–presence modeling reduces overprediction

Auxiliary-free ExactMB substantially reduces overprediction. As shown in Table 1, ExactMB lowers LD-Real OverPred from 98.4% to 11.6% and MD-3K OverPred from 98.5% to 19.2% compared with depth stacking. It retains most LD-Real ordinal accuracy (61.7% versus 67.4%), with LD-Syn conditional AbsRel of 19.2% versus 18.3%. Joint depth–presence learning thus suppresses false emissions with similar conditional depth accuracy, without auxiliary count or geometric losses.

Visible-set membership matters even with learned stopping. The released-model results in Table 2 show why optional emission need not recover visible membership. LaRI objects allows absence via its ray-stop mask, yet LD-Real and MD-3K OverPred are 50.2% and 80.5%, versus ExactMB's 11.6% and 19.2%. MD recall is comparable (79.1% versus 81.5%). These transfers complement the shared-training comparison: released models target different tasks and retain native adapters.



b Paired effects — Separate recipe

Metric	PPP → MAP	MAP → ExactMB
MD pattern ACC gain (pp)	+14.39 ± 2.84	+1.96 ± 1.37
MD OverPred reduction (pp)	-0.32 ± 1.59	+5.31 ± 1.49
LD-Syn count MAE reduction (layers)	+0.1211 ± 0.0030	-0.0069 ± 0.0030

Figure 3: **Count and presence metrics favor different modeling choices.** (a) Primary-study LD-Syn Count ACC by true count. (b) Separate-recipe paired effects. Positive values indicate improvement. ExactMB denotes marginalization. pp denotes percentage points. Means±sample SD.

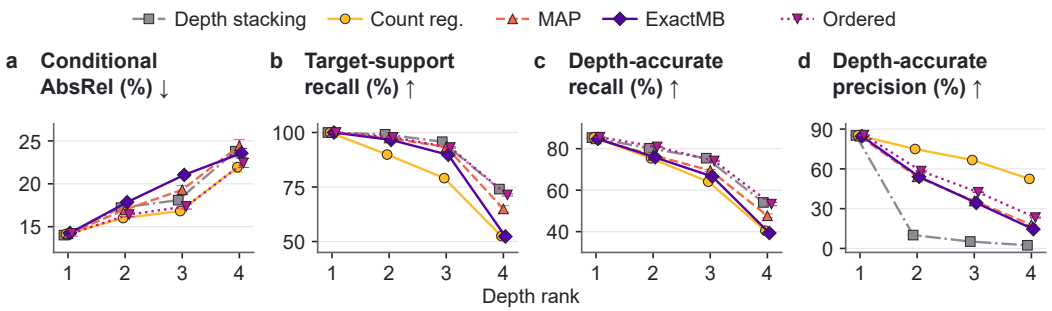


Figure 4: **Evaluating depth quality requires both localization and recovered support.** The same five LD-Syn ablations, means±sample SD. (a) AbsRel on jointly valid target–prediction pixels. (b) Target-support recall. (c) Depth-accurate recall. (d) Depth-accurate precision. Section 4 defines these metrics. Decoded ranks 1–4 are paired with supplied target channels L1/L3/L5/L7. Principal outputs are unaligned and unclipped. Full rank-wise scores and denominators are in Appendix C.2.

Count regularization and ordered assignment further shape recovery. The count-regularized configuration exceeds ExactMB on MD pattern ACC (65.9% versus 46.3%) and synthetic count ACC (95.0% versus 88.2%). One-return rays dominate the latter: an always-one predictor achieves 87.49% micro count ACC and .2112-layer MAE on occupied LD-Syn rays. Figure 3(a) stratifies the results by cardinality. Count regularization leads on one- through three-return rays, while ordered assignment reaches 71.4% on four-return rays versus ExactMB's 52.3%. Stratified scores therefore test multilayer counting beyond performance on the dominant one-return case.

Assignment marginalization shapes recovery trade-offs. Figure 3(b) compares a Poisson point-process (PPP) reference, MAP, and ExactMB under a separate recipe. MAP→ExactMB lowers MD OverPred by 5.31 ± 1.49 points and raises pattern ACC by 1.96 ± 1.37, while count MAE worsens by $.0069 \pm .0030$ layers. The primary-study pattern-ACC change is -1.35 ± 6.69 points: marginalization's benefits depend on the recipe and evaluation metric.

5.2 Depth Accuracy Must Be Read With Recovered Support

Geometric evaluation distinguishes selective accuracy from recovery. Which surfaces remain accurately recovered at these lower emission rates? Figure 4 compares four ranks across the same 20 checkpoints and 500 LD-Syn images. Relative to depth stacking, ExactMB raises second-rank depth-accurate precision from 10.07% to 54.04%, while recall falls from 80.20% to 75.78%. The count-regularized configuration lowers conditional errors relative to marginalization, but second- and third-rank target-support recall falls by $6.71 \pm .26$ and $10.68 \pm .28$ points. Lower conditional error can therefore accompany selective omission rather than more complete recovery.

Ordered assignment improves deeper geometry and support. Relative to ExactMB, ordered assignment lowers conditional AbsRel and increases target-support recall at ranks 2–4 in every seed. Fourth-rank depth-accurate recall and precision improve by $13.95 \pm .58$ and $8.76 \pm .59$ points.

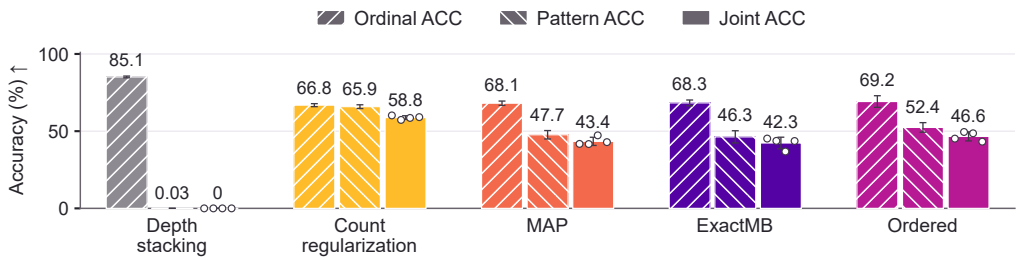


Figure 5: **Correct ordering does not guarantee correct visible-surface membership.** All 3,161 MD-3K pairs; principal operating point. Joint ACC requires correct presence and ordering against the annotations. Bars/whiskers: four-seed means/sample SD; circles: per-seed Joint ACC.

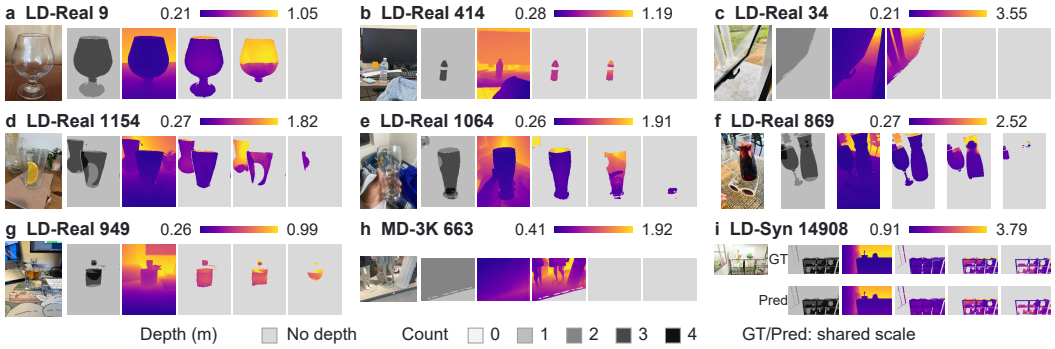


Figure 6: **Visible depth layers with spatially varying support.** Selected auxiliary-regularized ExactMB predictions. Columns: RGB, count, front-to-back depth ranks 1–4. (a–h) Real predictions. (i) Synthetic GT/prediction on a shared metric scale, without alignment. Gray: absent or invalid depth. The supplementary **video demo** provides additional qualitative cases and clearer views.

These joint gains indicate more accurate recovery, not just more emitted depths. In contrast, count regularization attains higher deeper-rank precision than ordered assignment at lower recall, though it also exceeds marginalization in fourth-rank recall and precision. Ordered assignment's geometric gains accompany an increase in MD OverPred from 19.23% to 21.65%. These configurations yield different precision–recall operating points for deeper-layer recovery.

Correct ordering can coexist with incorrect membership. Figure 5 reports ordinal results on all 3,161 MD pairs. Depth stacking obtains 85.13% ordinal ACC but 0% Joint ACC. ExactMB, ordered assignment, and the count-regularized configuration attain 42.34%, 46.62%, and 58.79% Joint ACC, respectively. As detailed in Appendix C.3.2, ordered assignment's gain over ExactMB reflects better presence-pattern recovery despite lower ordering accuracy conditional on correct presence. Joint ACC measures annotated presence and ordering together, not metric-set recovery.

5.3 Multilayer 3D Reconstruction and Applications

ExactMB's retained depth sets support layered 3D reconstruction. Figure 6 shows ExactMB depth and spatial support, including out-of-distribution MD-3K and LD-Real examples after synthetic-only task training. Unprojecting only retained depths carries ray-adaptive membership into 3D: foreground and background surfaces can coexist without turning unused components into additional mesh layers. Figure 7 shows the resulting RGB-textured meshes rendered with inserted lights and assigned thin-glass materials. These pinhole reconstructions offer a starting point for calibrated refractive ray models (Agrawal et al., 2012). Beyond rendering, Figure 8 shows candidate robot trajectories screened against selected first-layer surfaces of the reconstructed partial scene. Red and green indicate rejected and locally feasible simulated rollouts. Extending this local geometric test to closed-loop control and whole-scene safety verification is a promising direction for future work.

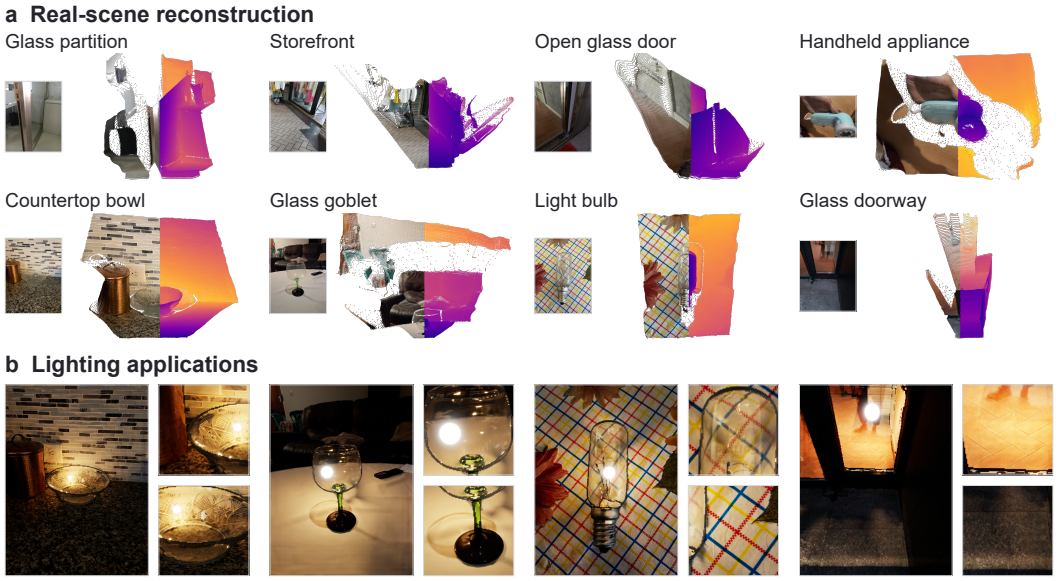


Figure 7: **Layered 3D reconstruction for relighting.** (a) Textured and depth-colored unprojections (purple near, yellow far, per-scene scales). (b) Inserted-light re-rendering and detail crops.

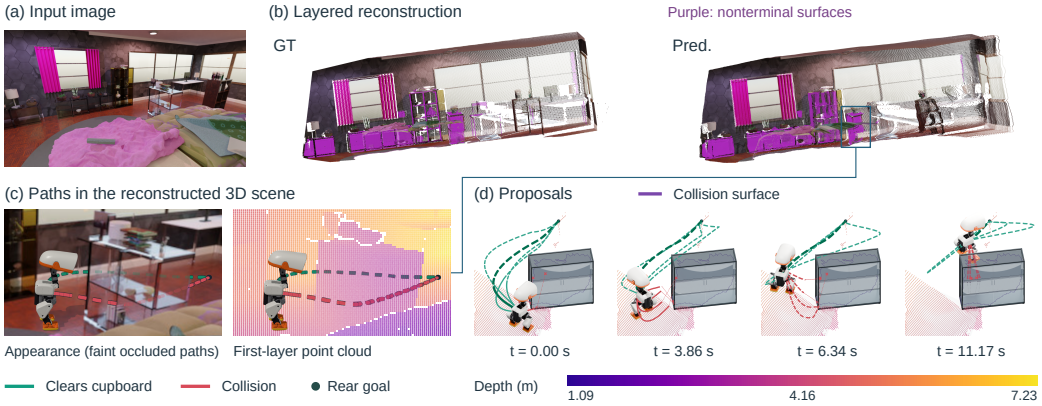


Figure 8: **Local route screening with reconstructed surfaces.** (a) Monocular input. (b) Layered reconstructions. (c) MicroDuck routes (Pollen Robotics, 2026) clear or intersect selected first-layer facets (green/red). (d) Rejected (red) and feasible (green) rollouts.

6 Conclusion

Depth Any Seen learns which visible surfaces are present and where they lie on a shared metric scale through ExactMB's joint depth–presence likelihood. Our analysis distinguishes expected-count correction from component–surface assignment. Experiments show that conditional depth accuracy and correct ordering can conceal incomplete or excessive predictions, motivating joint evaluation of localization, recovered support, and overprediction. The central goal is to recover the right surfaces at the right depths, rather than merely produce additional depth maps.

Limitations and future work. Our synthetic-only task training leaves room for stronger real-world generalization through mixed synthetic–real supervision. We believe that extending our image-based layered 3D prediction to video and streaming reconstruction would offer promising next steps toward temporally consistent recovery (Hu et al., 2025; Zhuo et al., 2026; Chen et al., 2026a). We hope this work sheds light on joint depth–presence learning for recovering visible 3D structure.

AI USE STATEMENT

Generative AI assistants (Codex) supported manuscript preparation, grammar checking, language refinement, consistency checks, and code for analysis and figure layout. No reported experimental results or plotted values were synthesized by generative AI assistants.

ETHICS STATEMENT

For our quantitative benchmarking, we adapt existing datasets to evaluate visible multilayer depth and presence, without collecting new data. Our method aims to improve the reliability of multilayer 3D representations and support safer downstream tasks, including robot navigation. Future work should validate real-world deployment safety and downstream control.

REPRODUCIBILITY STATEMENT

The appendix details the formulation, implementation, evaluation protocols, and study populations. Primary comparisons use multi-seed experiments, with single-seed analyses identified separately. Training histories and experiment records are archived in Weights & Biases (W&B). We will publicly release all our code and model checkpoints with the camera-ready paper to support reproducibility.

REFERENCES

Amit Agrawal, Srikumar Ramalingam, Yuichi Taguchi, and Visesh Chari. A theory of multi-layer flat refractive geometry. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3346–3353, 2012. doi: 10.1109/CVPR.2012.6248073.

Siyuan Bian, Congrong Xu, and Jun Gao. Modeling depth ambiguity: A mixture-density representation for flying-point-free depth estimation. *arXiv preprint arXiv:2606.02552*, 2026.

Lin-Zhuo Chen, Jian Gao, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, et al. Geometric context transformer for streaming 3d reconstruction. In *European Conference on Computer Vision*, 2026a.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7220–7232, 2026b.

Alex Costanzino, Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Learning depth estimation for transparent and mirror surfaces. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9210–9221, 2023. doi: 10.1109/ICCV51070.2023.00848.

Helisa Dhamo, Nassir Navab, and Federico Tombari. Object-driven multi-layer scene decomposition from a single image. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5368–5377, 2019. doi: 10.1109/ICCV.2019.00547.

Dominik Engel, Sebastian Hartwig, and Timo Ropinski. Monocular depth decomposition of semi-transparent volume renderings. *IEEE Transactions on Visualization and Computer Graphics*, 30(7):3981–3994, July 2024. ISSN 1077-2626. doi: 10.1109/TVCG.2023.3245305. URL https://doi.org/10.1109/TVCG.2023.3245305.

Hongjie Fang, Hao-Shu Fang, Sheng Xu, and Cewu Lu. Transcg: A large-scale real-world dataset for transparent object depth completion and a grasping baseline. *IEEE Robotics and Automation Letters*, 7(3):7383–7390, 2022. doi: 10.1109/LRA.2022.3183256.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Georg Hess, Christoffer Petersson, and Lennart Svensson. Object detection as probabilistic set prediction. In *European Conference on Computer Vision*, pp. 550–566. Springer, 2022.

Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2005–2015. IEEE, 2025.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. In *2026 International Conference on 3D Vision (3DV)*, pp. 499–509. IEEE, 2026.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Rui Li, Biao Zhang, Zhenyu Li, Federico Tombari, and Peter Wonka. LaRI: Layered ray intersections for single-view 3D geometric reasoning. In *Proceedings of the 43rd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2026.

Jiangyuan Liu, Hongxuan Ma, Yuxin Guo, Yuhao Zhao, Chi Zhang, Wei Sui, and Wei Zou. Monocular depth estimation and segmentation for transparent object with iterative semantic and geometric fusion. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11162–11168, 2025. doi: 10.1109/ICRA55743.2025.11128401.

Junhong Min, Jimin Kim, Minwook Kim, Cheol-Hui Min, Youngpil Jeon, and Minyong Choi. Depthfocus: Controllable depth estimation for see-through scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12595–12605, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Pollen Robotics. Microduck Sandbox. Hugging Face Space, 2026. URL https://huggingface.co/spaces/pollen-robotics/microduck-simulator. Robot model and simulator assets, revision e81974b.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Shreeyak Sajjan, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, and Shuran Song. Clear grasp: 3d shape estimation of transparent objects for manipulation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3634–3642, 2020. doi: 10.1109/ICRA40945.2020.9197518.

Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski. Layered depth images. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '98, pp. 231–242, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 0897919998. doi: 10.1145/280814.280882. URL https://doi.org/10.1145/280814.280882.

Jun Shi, Yong A, Yixiang Jin, Dingzhe Li, Haoyu Niu, Zhezhu Jin, and He Wang. Asgrasp: Generalizable transparent object reconstruction and 6-dof grasp detection from rgb-d active stereo camera. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5441–5447, 2024. doi: 10.1109/ICRA57147.2024.10611152.

Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8025–8035, 2020. doi: 10.1109/CVPR42600.2020.00805.

Daeyun Shin, Zhile Ren, Erik Sudderth, and Charless Fowlkes. 3d scene reconstruction with multi-layer depth and epipolar transformers. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2172–2182, 2019. doi: 10.1109/ICCV.2019.00226.

Shubham Tulsiani, Richard Tucker, and Noah Snavely. Layer-structured 3d scene inference via view synthesis. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII*, pp. 311–327, Berlin, Heidelberg, 2018. Springer-Verlag. ISBN 978-3-030-01233-5. doi: 10.1007/978-3-030-01234-2_19. URL https://doi.org/10.1007/978-3-030-01234-2_19.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024.

Xiaoying Wang, Yumeng He, Jingkai Shi, Jiayin Lu, Yin Yang, Ying Jiang, and Chenfanfu Jiang. Seeclear: Reliable transparent object depth estimation via generative opacification. In *European Conference on Computer Vision*, pp. 75–93. Springer, 2026a.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, volume 2026, pp. 10481–10497, 2026b.

Hongyu Wen and Jia Deng. SeeGroup: Multi-layer depth estimation of transparent surfaces via self-determined grouping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Hongyu Wen, Yiming Zuo, Venkat Subramanian, Patrick Chen, and Jia Deng. Seeing and seeing through the glass: Real and synthetic data for multi-layer depth estimation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6715–6725. IEEE, 2025.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21469–21480. IEEE, 2025.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, et al. Native and compact structured latents for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14419–14429, 2026.

Xiaohao Xu, Feng Xue, Xiang Li, Haowei Li, Shusheng Yang, Tianyi Zhang, Matthew Johnson-Roberson, and Xiaonan Huang. One scene, two depths: Probing geometric ambiguity in monocular foundation models. In *European Conference on Computer Vision*. Springer, 2026.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10371–10381. IEEE, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in neural information processing systems*, 37:21875–21911, 2024b.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *European conference on computer vision*, pp. 785–801. Springer, 2018.

Hao Zhang, Mohamed El Banani, Jen-Hao Cheng, Paul Zhang, Yi Hua, Ben Mildenhall, Christoph Lassner, Narendra Ahuja, and Gengshan Yang. World tracing: Generative pixel-aligned geometry beyond the visible. *arXiv preprint arXiv:2606.13652*, 2026.

Luyang Zhu, Arsalan Mousavian, Yu Xiang, Hammad Mazhar, Jozef van Eenbergen, Shoubhik Debnath, and Dieter Fox. Rgb-d local implicit function for depth completion of transparent objects. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4647–4656, 2021. doi: 10.1109/CVPR46437.2021.00462.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming visual geometry transformer. In *International Conference on Learning Representations*, volume 2026, pp. 88055–88072, 2026.

APPENDIX CONTENTS

The appendix connects probabilistic analysis, evaluation protocols, and recovery behavior. Appendices A.2–A.3 explain how count and assignment interact through selection, correspondence, and specialization. The evaluation definitions and study provenance in Appendices B.2–B.3 support the primary ablations and geometric recovery results in Appendices C.1–C.2. Appendix C.7.1 relates geometric operating points to retained support, while Appendices D.1–D.2 trace count, activation, uncertainty, and geometry during training. Appendix figures, tables, and equations use an S prefix.

A Model and probabilistic foundations **15**

A.1 Distributional and objective details . . . 15

A.2 Selecting components versus assigning surfaces . . . 16

A.3 Mass correction and specialization are different directions . . . 17

B Implementation, evaluation, and reproducibility **18**

B.1 Reference implementation and training recipe . . . 18

B.2 Evaluation definitions and adapters . . . 20

B.3 Ablation factors and configuration provenance . . . 22

B.4 Annotation semantics and the training interface . . . 23

C Additional experimental results **23**

C.1 Primary ablations and released-model transfer . . . 23

C.2 Primary four-seed geometric recovery . . . 24

C.3 Error structure across cardinalities and relations . . . 25

C.4 Point-process and assignment controls . . . 27

C.5 Matched geometric-loss interventions . . . 28

C.6 Encoder capacity and decoder design . . . 30

C.7 Archived decoder operating points . . . 31

C.8 Additional real-world qualitative results . . . 33

D Learning dynamics and geometric recovery **48**

D.1 Count, activation, and held-out learning . . . 48

D.2 Fixed-ray uncertainty and geometric recovery . . . 49

D.3 Cardinality weighting and retained support . . . 53

E Output and membership conventions **54**

A Model and Probabilistic Foundations

A.1 Distributional and Objective Details

A.1.1 What the Coverage Objective Encourages

For a nonempty target, the maximum-density coverage terms associated with SeeGroup (Wen & Deng, 2026) take the form

$$\mathcal{L}_{\text{target}} = -\sum_{i=1}^{M} \log \max_{j\in[K]} \ell_j(t_i), \tag{S1}$$

$$\mathcal{L}_{\text{cover}} = -\sum_{j=1}^{K} \log \max_{i\in[M]} \ell_j(t_i), \tag{S2}$$

$$\mathcal{L}_{\text{bi}} = \lambda_t \mathcal{L}_{\text{target}} + \lambda_c \mathcal{L}_{\text{cover}}, \quad \lambda_t, \lambda_c > 0. \tag{S3}$$

The target-side term permits targets to share components; the component-side term encourages every component to explain a target. With distinct targets, $1 \le M \le K$, freely optimized ray-wise centers and scales, and $\beta_j \ge \beta_{\min} > 0$, each density is bounded by $(2\beta_{\min})^{-1}$. Both losses reach their lower bounds by covering every target and centering every component on a target at the scale floor. Positive weights require every global minimum to attain both bounds, so surplus components duplicate target depths. Shared parameters, regularization, finite training, and filtering can alter this idealized behavior. Our depth-stacking baseline instead uses ordered depth regression.

A.1.2 A Poisson Intensity Defines a Different Set Model

Assignment constraints also distinguish ExactMB from Poisson set models. Intensity specifies only a first-order moment; a Poisson point-process assumption with finite total intensity gives the normalized density

$$f_{\text{PPP}}(\mathcal{T}) = \exp(-\Lambda) \prod_{t\in\mathcal{T}} v(t), \qquad \Lambda = \int v(t)\ \mathrm{d}t. \tag{S4}$$

For $v(t) = \sum_j \lambda_j \ell_j(t)$, $\lambda_j \ge 0$, expansion sums over all target-to-component assignments, allowing each component to explain multiple targets. This law has unbounded count and empty-event probability $\exp(-\Lambda)$. ExactMB bounds count through optional singletons and injective assignments.

A.1.3 Reference Measure, Normalization, and Propriety

For ExactMB, let $\mathfrak{F}_{\le K}(\mathcal{Y})$ denote simple finite subsets of $\mathcal{Y} = \mathbb{R}$ of cardinality at most K. This domain accommodates untruncated densities despite positive targets and decoded depths. For symmetric nonnegative measurable or integrable g, define

$$\int g(\mathcal{T})\, \delta\mathcal{T} = g(\varnothing) + \sum_{M=1}^{K} \frac{1}{M!} \int_{\mathcal{Y}^M} g(\{t_1, \ldots, t_M\})\ \mathrm{d}t_{1:M}. \tag{S5}$$

The factorial removes repeated set enumerations. Under the fixed Lebesgue measure in meters, M-element densities have units m^{-M}. Holding existing components fixed, appending a component with $q = 0$ leaves the law unchanged; sigmoid logits approach this boundary as $e \to -\infty$.

Proposition S1 (Multi-Bernoulli normalization)**.** *Under* (S5)*, density* (3) *satisfies* $\int f(\mathcal{T})\, \delta\mathcal{T} = 1$.

Proof. Fix M and substitute (2) into the integral. Each ℓ_j integrates to one. Every size-M component subset admits $M!$ bijections from the integration variables, whose multiplicity is canceled by $1/M!$. The total mass at this cardinality is

$$\sum_{\substack{S\subseteq[K]\\|S|=M}} \prod_{j\in S} q_j \prod_{j\notin S} (1 - q_j). \tag{S6}$$

Summing over $M = 0{:}K$ enumerates each subset once and gives $\prod_j [(1 - q_j) + q_j] = 1$. □

Normalization gives the complete-set law the standard logarithmic-score identity (Gneiting & Raftery, 2007):

Theorem S1 (Finite-set logarithmic-score propriety)**.** *Let $f^\star$ be a normalized finite-set density with finite entropy and let f be normalized on the same space and reference measure. Whenever cross-entropy is finite,*

$$\mathbb{E}_{\mathcal{T}\sim f^\star}[-\log f(\mathcal{T})] = H(f^\star) + \mathrm{KL}(f^\star \| f). \tag{S7}$$

The expected score is uniquely minimized, up to equality almost everywhere, by $f = f^\star$. Within a restricted model family, any attained minimizer is a KL projection.

Proof. Add and subtract $-\int f^\star \log f^\star \, \delta\mathcal{T}$. The remainder is $\mathrm{KL}(f^\star \| f) \geq 0$. If f vanishes on positive $f^\star$ mass, cross-entropy is infinite. Under corresponding input-integrability conditions, the conditional identity can also be averaged over images and camera rays. □

Propriety characterizes the induced density, not component identifiability, calibration, or optimization. Auxiliary penalties, unequal existence weights, temperature changes, and target-dependent loss rescaling can modify this normalized score and require separate analysis.

A.1.4 Count and Conditional Location Density

Normalization also yields the count–geometry factorization. Continuous draws are distinct almost surely, so $N = |\mathsf{X}| = \sum_j R_j$. For $S \subseteq [K]$ of size M, define

$$\begin{aligned} w_q(S) &= \prod_{j\in S} q_j \prod_{j\notin S}(1-q_j), \\ G_S(\mathcal{T}) &= \sum_{\substack{\phi:[M]\hookrightarrow[K] \\ \mathrm{im}(\phi)=S}} \prod_{i=1}^{M} \ell_{\phi(i)}(t_i). \end{aligned} \tag{S8}$$

Grouping by occupied subset gives $f(\mathcal{T}) = \sum_{|S|=M} w_q(S) G_S(\mathcal{T})$. The $M!$ terms in each G_S cancel the cardinality-slice factor $1/M!$ on integration:

$$\begin{aligned} \int_{|\mathcal{T}|=M} G_S(\mathcal{T})\, \delta\mathcal{T} &= 1, \\ p_q(M) = \mathbb{P}(N=M) &= \sum_{|S|=M} w_q(S). \end{aligned} \tag{S9}$$

For $p_q(M) > 0$, the conditional geometric density is

$$f(\mathcal{T} \mid N = M) = \sum_{|S|=M} \frac{w_q(S)}{p_q(M)} G_S(\mathcal{T}), \qquad |\mathcal{T}| = M. \tag{S10}$$

The weights sum to one, giving (8). For $0 < M < K$, relative presence weights can affect conditional geometry through the subset mixture. At $M = K$, only $S = [K]$ remains, eliminating dependence on q; at $M = 0$, the conditional density is one. Conditioning is undefined if $p_q(M) = 0$.

A.2 Selecting Components Versus Assigning Surfaces

The subset mixture separates selection from correspondence. For a complete distinct target, $\Phi \sim \omega$ selects $S = \mathrm{im}(\Phi)$ with $|S| = M$. At $M = 1$, within-subset correspondence is unique, leaving only selection ambiguity. At $M = K$, the subset is fixed despite $K!$ potentially plausible correspondences. Hence, for finite outputs under positive Laplace densities,

$$M = K: \quad \rho_j = 1, \qquad \frac{\partial \mathcal{L}}{\partial e_j} = q_j - 1. \tag{S11}$$

For $M = 0$, occupancy is zero and the gradient is q_j. At identical outputs, MAP and ExactMB thus share full-capacity existence gradients; training can differ through geometry, recurrence, shared parameters, and other cardinalities. Complete four-target rays with four slots therefore have posterior presence targets of one despite uncertain depth-to-component correspondence.

A.3 Mass Correction and Specialization Are Different Directions

To distinguish count correction from component selection, fix locations and scales, and write $R_j(\Phi) = \mathbf{1}[j \in \mathrm{im}(\Phi)]$. Each assignment has log density

$$\log f(\mathcal{T}, \phi) = \sum_j e_j R_j(\phi) - \sum_j \mathrm{softplus}(e_j) + \sum_i \log \ell_{\phi(i)}(t_i).$$

Every injection satisfies $\sum_j R_j = M$. A common finite-logit shift $e_j \mapsto e_j + a$ adds $Ma - \sum_j [\mathrm{softplus}(e_j + a) - \mathrm{softplus}(e_j)]$ to every log weight and preserves the entire ownership posterior. For the shifted loss,

$$\frac{\mathrm{d}\mathcal{L}(a)}{\mathrm{d}a} = \sum_j \sigma(e_j + a) - M,$$

$$\frac{\mathrm{d}^2\mathcal{L}(a)}{\mathrm{d}a^2} = \sum_j \sigma(e_j + a)[1 - \sigma(e_j + a)] > 0.$$

For $0 < M < K$, the derivative increases from $-M$ to $K - M$, yielding a unique finite optimum matching expected count without changing ownership. Fixed centers and positive scales suffice; symmetry is unnecessary. Shared-network updates need not follow this output-space direction.

Relative logit changes, however, can alter selection. The log-partition covariance identity (Wainwright & Jordan, 2008) gives $\partial \rho_j / \partial e_k = \mathrm{Cov}_\omega(R_j, R_k)$, hence

$$\nabla^2_{\mathbf{e}} \mathcal{L} = \mathrm{diag}\{q_j(1 - q_j)\} - \mathrm{Cov}_\omega(\mathbf{R}). \tag{S12}$$

Proposition S2 (Fractional symmetry is a saddle with surplus capacity)**.** *Suppose all slots have the same density value at each target, $0 < M < K$, and $q_j = p = M/K$. The existence gradient vanishes. With $v = p(1-p)$, the Hessian eigenvalues are v along $\mathbf{1}/\sqrt{K}$ and $-v/(K-1)$ along each of the $K - 1$ orthogonal contrast directions between component logits.*

Proof. The posterior over size-M subsets is uniform, with $\rho_j = p$, $\mathrm{Var}(R_j) = v$, and off-diagonal covariance $-v/(K-1)$. The Hessian has zero diagonal and off-diagonal entries $v/(K-1)$. Its row sum is v, and its action on a zero-sum vector is multiplication by $-v/(K-1)$. □

At empty or full capacity, deterministic occupancy instead gives strict convexity in finite existence logits at fixed localization. The network parameterization and optimizer preconditioning determine how optimization in parameter space follows this output-space geometry.

Expected count, count probability, and gates. For decoding, distinguish expected count from the pre-thresholding, pre-merging distribution:

$$p_q(n) = [z^n] \prod_{j=1}^{K} (1 - q_j + q_j z). \tag{S13}$$

Four .25 gates have mean count one, count-one probability $27/64$, and variance $3/4$. A strict cutoff below .25 retains all four; otherwise none survive. Merging can remove coincident depths. For one target and $K > 1$ identical component densities, the tied-gate likelihood maximum at $q = 1/K$ is the independent-logit saddle above. Matching expected count therefore guarantees neither concentration of the pre-decoding count distribution nor correctness of the count after decoding.

B Implementation, Evaluation, and Reproducibility

We describe auxiliary-free ExactMB, evaluation, and annotation semantics, distinguishing primary ablations from point-process, geometric-loss, architecture, and fixed-ray studies.

B.1 Reference Implementation and Training Recipe

B.1.1 From RGB and annotations to slot parameters

As summarized in Table S1, the image is encoded once, and four recurrent passes predict absolute depth centers, Laplace scales, and existence logits:

$$d_j = 20\,\mathrm{softplus}(u_j), \qquad \beta_j = \mathrm{clip}_{[.1,10]}(.1 + \mathrm{softplus}(v_j)), \qquad q_j = \sigma(e_j). \tag{S14}$$

Depth is camera optical-axis Z in meters. The factor 20 scales rather than bounds the output; centers need not increase across passes. The loader converts millimeter PNGs to meters, zeros nonfinite, nonpositive, or above-80-m entries, and compacts valid L1/L3/L5/L7 channels in supplied order without sorting depths or deduplicating ties. RGB and depth/mask tensors share an aspect-preserving resize to cover a 518×518 crop with 14-pixel-compatible dimensions, followed by a shared random crop and horizontal flip with probability .5. RGB uses ImageNet normalization. Training and principal evaluation fit no target-dependent affine scale or shift.

B.1.2 Presence-aligned recurrent update

At encoder scale ℓ and token n, let $\mathbf{F}_{j-1,\ell n}$ be the previous feature state, $\mathbf{C}_j$ decoded features, and $\mathcal{R}_\ell$ the reverse-DPT projection. With $g = \sigma(a)$, a initialized to $\mathrm{logit}(.05)$, $\eta_{\mathrm{cap}} = .25$, and $\epsilon = 10^{-6}$, the update is

$$\mathbf{U}_{j,\ell n} = g\ \mathrm{sg}(q_{j,n})\mathcal{R}_\ell(\mathbf{C}_j)_n, \tag{S15}$$

$$\gamma_{j,\ell n} = \min\left\{1, \frac{\eta_{\mathrm{cap}} \max\{\|\mathbf{F}_{j-1,\ell n}\|_2, \epsilon\}}{\max\{\|\mathbf{U}_{j,\ell n}\|_2, \epsilon\}}\right\}, \tag{S16}$$

$$\mathbf{F}_{j,\ell n} = \mathbf{F}_{j-1,\ell n} - \gamma_{j,\ell n}\mathbf{U}_{j,\ell n},$$

$$\|\mathbf{F}_{j,\ell n} - \mathbf{F}_{j-1,\ell n}\|_2 \le \eta_{\mathrm{cap}} \max\{\|\mathbf{F}_{j-1,\ell n}\|_2, \epsilon\}. \tag{S17}$$

Detaching the presence gate blocks direct gradients from later-slot losses to earlier existence probabilities through feature removal, while indirect shared-feature interactions remain. The shared existence head uses a $128 \to 32$ 3×3 convolution, rectifier, and $32 \to 1$ projection. All K passes execute.

B.1.3 Training loss and the annotation interface

Auxiliary-free ExactMB uses linear-depth Laplace densities with unit outer weight, temperature, and existence weight, and no geometric auxiliary terms. Training averages over all BHW rays, treating all-false masks as $M = 0$ and supplied ties as separate valid entries. Element masks do not distinguish empty targets from unannotated rays. Section B.4 relates these annotations to the likelihood's assumptions of complete, distinct ray-wise depth targets.

B.1.4 Training schedule and sampling

Table S1 specifies the training data, optimizer, and four-seed schedule. Equal-weight cardinality buckets are visited round-robin without replacement, skipping exhausted buckets. This interleaving does not equalize full-epoch bucket mass or per-ray exposure. Primary results use terminal/latest checkpoints at 800k updates. Section B.3 details the ablation factors.

B.1.5 Native selection and geometric decoding

Selection is variant-specific. Depth stacking selects all K slots. Count regularization predicts categorical count probabilities $r_m(x)$ for $m = 0, \ldots, K$ and selects a prefix of length $\arg\max_m r_m(x)$. Bernoulli-slot variants select components with finite $q_j > \tau$, whereas the first two variants ignore this cutoff. A common geometric decoder then keeps finite depths $d_j > d_{\min}$ and sorts them. An empty candidate list yields an empty prediction. Otherwise, it retains the shallowest candidate and

Table S1: **Auxiliary-free ExactMB configuration.** The configuration connects each implementation choice to its operational role. Variant-specific changes are stated with the corresponding comparisons. Section B.3 describes the available historical configuration records and their limits.

Component	Reference choice	Operational meaning
Representation	$K = 4$ unordered Bernoulli–Laplace components	Capacity bounds the emitted set; all four decoder passes execute.
Encoder / decoder	DINOv2 ViT-L (Oquab et al., 2024); metric Depth Anything V2 initialization; shared recurrent DPT (Ranftl et al., 2021)	One encoder, shared depth/scale/existence heads, absolute rather than cumulative depths.
Parameterization	$d = 20\,\mathrm{softplus}(u)$; $q = \sigma(e)$; $\beta = \mathrm{clip}_{[.1,10]}(.1 + \mathrm{softplus}(v))$	Camera-Z meters. The factor 20 is not a depth clamp.
Recurrence	Detached existence gate; learned removal strength initialized .05; update cap .25	Bound (S17) is relative to each feature token's norm.
Targets / mask	Compacted provided L1/L3/L5/L7 channels; element mask from positive valid depth	No numeric target sorting or deduplication; all-false masks contribute $M = 0$ in the historical all-ray mean.
Complete criterion	ExactMB only; outer weight 1, temperature 1, existence weight 1; no alignment	All geometric auxiliary weights are zero. Padded enumeration includes the $(K - M)!$ correction.
Data ordering	14,800 training images; equal-weight four-bucket round robin without replacement	Interleaves remaining buckets without equalizing epoch mass or per-ray cardinality exposure.
Preprocessing	Aspect-preserving resize, 518×518 crop, $p = .5$ horizontal flip, ImageNet RGB normalization	Shared image/depth/mask geometry; metric units are retained.
Optimization	AdamW; 10^{-5} peak LR; encoder multiplier .1; moments $(.9, .999)$; decay .01; gradient-norm clipping .5	Batch 2, one recorded process; four training seeds $7/42/61/123$.
Schedule / endpoint	Warmup 5,000; hold 127,200; cosine decay to .02 of peak at 800,000	Terminal/latest checkpoint; no best-validation alias substitution in principal comparisons.
Principal decoder	Strict $q > .01$, finite $d > .02$ m, ascending sort, .02 m last-retained gap	No relative gap, scale/shift alignment, or ground-truth-dependent selection.

each later candidate more than Δ_{merge} beyond the last retained depth. Rejected candidates leave this reference unchanged: with a .02-m gap, $1.000, 1.015, 1.030$ retains $1.000, 1.030$ m.

The principal thresholds are $\tau = .01$, $d_{\min} = .02$ m, and $\Delta_{\mathrm{merge}} = .02$ m, with strict comparisons and no relative-gap term, target cardinality, alignment, or test-set threshold optimization. Accepted depths form a valid prefix for four-rank MD/LD-Syn scoring. Sorting defines front-to-back evaluator ranks, while target–component assignments are determined by the training objective.

B.1.6 Batched ExactMB implementation and checks

Predictions, targets, and element masks have shape $B \times K \times H \times W$. For target-column validity m_i, the broadcast pair costs are

$$C_{ji} = \begin{cases} |d_j - t_i|/\beta_j + \log(2\beta_j) - \mathrm{logsigmoid}(e_j), & m_i = 1, \\ -\mathrm{logsigmoid}(-e_j), & m_i = 0. \end{cases}$$

Finite placeholders replace invalid targets before residual computation. Stable log-sigmoid terms avoid logarithms of rounded probabilities. For each cached permutation p, sum $S_p = \sum_j C_{j,p(j)}$. The exact per-ray loss is

$$-\mathrm{logsumexp}_p(-S_p) + \log\Gamma(K - M + 1), \qquad M = \sum_i m_i.$$

The correction removes $(K - M)!$ permutations of indistinguishable null columns. The $K = 4$ implementation differentiates through 24 padded permutations at arithmetic cost $O(BHW\,K!K)$.

Table S2: **Complementary annotations for visible-layer evaluation.** Each benchmark defines absence through its annotations. Together, they assess ordering, presence, and metric depth.

Benchmark	Available annotation	Reported measures	Not established by this view alone
LD-Real 300 real images	Sparse valid/invalid tuples, requested ranks, and near-to-far relations	ACC on valid tuples; OverPred on invalid tuples; same/mixed-rank breakdowns	Dense metric accuracy, physical surface count on every ray, or a literal transparent/opaque material partition.
MD-3K 3,161 real pairs	Transparent mask at query points; foreground/background order; TT/TO/OO and Same/Reverse strata	Recall, OverPred, pattern ACC, ordinal ACC, and Joint ACC on their declared point/pair populations	A census of all physical interfaces or dense real metric depth.
LD-Syn 500 synthetic images	Four dense metric channels and validity masks, in camera-Z meters	Count ACC, Count MAE, under/over-count, and conditional same-rank AbsRel with support	Unconditional real-image metric generalization or calibration of existence probabilities.

Empty targets reduce to $-\sum_j \log(1-q_j)$, while full targets need no correction. Checks compare direct injections, padded sums, target permutations, finite-difference gradients, and strict decoding boundaries. For diagnostics, occupancy excludes null columns, and subtracting $\log((K-M)!)$ from padded-permutation entropy gives the entropy over distinct injections.

B.2 Evaluation Definitions and Adapters

The benchmark annotations in Table S2 give ACC different meanings: ordinal-tuple, presence-pattern, or count accuracy. LD OverPred complements invalid-tuple rejection, while MD OverPred complements opaque-point rejection; this conversion preserves sample SD and reverses paired-effect signs. Each metric refers to its annotation population, which provides neither dense real multilayer metric ground truth nor a complete enumeration of physical interfaces at every pixel.

B.2.1 Selection and diagnostic contexts

Table S3: **Decoding rules and populations across evaluation settings.** Each setting specifies its own thresholds and aggregation rules. Diagnostic settings leave the principal decoder unchanged.

Context	Selection and geometry	Population
Principal terminal suite	Native selector; Bernoulli cutoff .01; .02 m floor and last-retained gap	Full splits; first four accepted ranks on MD/Syn; no alignment
Primary cutoff sweep	Same 12 Bernoulli-slot checkpoints; six cutoffs; fixed geometry	Evaluates operating points without treating them as replicates or test-selected optima
Training / Syn histories	Raw expected mass or $q > .5$ activation; no principal geometric filter	Minibatch exposure / fixed 64-image cohort; image-conditional validation means
Fixed-ray geometry	$q > .01$, depth $> .02$ m, gap $> .02$ m; injective tolerance	4,513 fixed supplied-distinct-target rays from 20 images; one auxiliary-regularized run
Qualitative displays	$q > .5$ with depth sorting; pinhole geometry	Selected examples and checkpoints; illustrative geometry and graphics
Released models	Dataset-specific masks, order and scale below	Single released estimates; native-task comparisons remain separate

Table S3 separates principal and diagnostic protocols. "Before fallback" denotes post-decoder validity without shallower-depth substitution, as used for principal presence scores. Appendix C.7.1 presents the 20-checkpoint geometric comparison, with cutoff curves for 12 Bernoulli-slot checkpoints alongside the native fixed-count and count-regularized exports.

B.2.2 LD-REAL: RECOVERY AND REJECTION OF QUERIED RANKS

The 300-image validation split of the LayeredDepth real benchmark supplies sparse valid/invalid pairs, triplets, and quadruplets (Wen et al., 2025). For a tuple of pixel/rank requests with decoded depths $\hat{d}_1, \ldots, \hat{d}_k$, correct recovery requires all requested ranks and their annotated strict near-to-far order. Correct rejection of an annotated-invalid tuple requires every requested rank to be absent:

$$\text{correct}_{\text{real}} = \mathbf{1}[\bigwedge_r (\hat{d}_r \neq \perp) \wedge \hat{d}_1 < \cdots < \hat{d}_k], \tag{S18}$$

$$\text{correct}_{\text{fake}} = \mathbf{1}[\bigwedge_r (\hat{d}_r = \perp)]. \tag{S19}$$

The main-paper ACC is micro-averaged over valid quadruplets. OverPred is the fraction of invalid quadruplets with any requested prediction, so the two metrics have separate denominators. Because tuples can span rays, tuple arity, requested rank, and per-ray cardinality are distinct quantities. Same-rank tuples request one rank index at every point, whereas mixed-rank tuples request different indices. A separate last-visible diagnostic checks the strict ordering of the deepest retained values on mixed valid tuples. Missing predictions count as failures, and invalid tuples have no last-visible target.

B.2.3 MD-3K: PATTERNS AND MATERIAL-CONDITIONED GEOMETRY

MD-3K provides 3,161 real point pairs with foreground/background relations and material labels (Xu et al., 2026). At point p of pair n, we map transparency z_{np} to target validity $\mathbf{v}^*(z) = (1, z, 0, 0)$ in L1/L3/L5/L7 order. For post-decoder bits $\hat{\mathbf{v}}_{np}$,

$$E_n = \bigwedge_{p=1}^{2} [\hat{\mathbf{v}}_{np} = \mathbf{v}^*(z_{np})], \qquad \text{ACC}_{\text{MD}} = N^{-1} \sum_n \mathbf{1}[E_n]. \tag{S20}$$

Pattern ACC checks all eight pair-level bits. Point-level measures distinguish missed backgrounds from extra emissions. With $\hat{z}_{np}$ denoting decoded L3 validity,

$$\text{Recall} = \frac{\sum_{n,p} z_{np} \hat{z}_{np}}{\sum_{n,p} z_{np}}, \qquad \text{OverPred} = \frac{\sum_{n,p} (1 - z_{np}) \hat{z}_{np}}{\sum_{n,p} (1 - z_{np})}. \tag{S21}$$

The frozen split contains 1,485/1,350/326 TT/TO/OO pairs and 4,320 transparent/2,002 opaque requests. This validity convention does not imply that a ray contains at most two physical surfaces.

Foreground ordering uses L1; background ordering uses L3 at transparent points and L1 at opaque points, without replacing missing transparent-point L3 with L1. Let O_n require both orderings to be correct. Headline multilayer ordinal ACC pools TT and TO, while Figure 5 uses all pairs. The all-pair evaluation retains six pairs—three OO/Reverse and three TO/Reverse—whose foreground and background orderings cannot both hold under primary front-to-back decoding. Joint correctness combines presence E and ordering O on a common evaluation population S:

$$\Pr(E \cap O \mid S) = \Pr(E \mid S) \Pr(O \mid E, S). \tag{S22}$$

If no pair satisfies E, conditional accuracy is undefined and joint frequency is zero. Pattern ACC thus evaluates pair-level presence, Recall/OverPred individual L3 requests, and Joint ACC presence with both orderings. Same/Reverse denotes agreeing/flipped foreground and background orders. Released ordinal scores may use shallower fallback, unlike Joint ACC.

B.2.4 LD-SYN: CARDINALITY AND METRIC LOCALIZATION

The 500 held-out images provide four dense metric channels and validity masks (Wen et al., 2025). Target and decoded masks are compact prefixes on this split, so four-bit pattern accuracy equals count accuracy. The principal population is occupied rays $\Omega_+ = \{x : c_x \geq 1\}$, with

$$\text{ACC}_c = \text{mean}_{x:c_x=c}\, \mathbf{1}[\hat{c}_x = c], \qquad \text{Macro} = \tfrac{1}{4} \sum_{c=1}^{4} \text{ACC}_c, \tag{S23}$$

$$\text{MAE} = \text{mean}_{x \in \Omega_+} |\hat{c}_x - c_x|. \tag{S24}$$

Occupied-micro ACC and under/over-count rates pool Ω_+. Explicitly labeled all-pixel fields also include empty masks. Target counts retain supplied ties without distance-based deduplication. For localization, own-slot error compares each decoded rank with its target channel where both are valid. Let $\mathcal{S}_3$ be this intersection for L3:

$$\mathrm{AR}_3^{\mathrm{own}} = \frac{100}{|\mathcal{S}_3|} \sum_{x \in \mathcal{S}_3} \frac{|\hat{d}_{3x} - d_{3x}|}{d_{3x}}. \tag{S25}$$

Errors and counts are pooled before division. Abstaining on difficult rays changes the evaluated population, so conditional error must be read with support. Section C.2 gives all primary rank-wise denominators, support recall, and depth-accurate recall/precision. The complementary injective-set diagnostic uses a distinct cohort and an absolute geometric tolerance.

B.2.5 Replication and retrospective selection

Training seed is the replication unit. Absolute results give means and sample SD; paired comparisons report within-seed changes, SD, and directional agreement. Pixel/image variation is not training-seed replication. Four seed pairs cannot attain $p < .125$ in a two-sided exact sign-flip test, so we emphasize effect sizes and metric trade-offs. Primary MD pattern ACC and accompanying Recall/OverPred, LD tuple outcomes, and synthetic count/geometry were organized retrospectively, without preregistration or score-blind selection. Uncertainty summaries therefore include each cohort's selection context.

B.2.6 Released-system transfer and adapter boundaries

Table 2 evaluates released models without retraining. SeeGroup (Wen & Deng, 2026) decodes with a .01 validity-score cutoff and .02 minimum depth and gap. World Tracing (Zhang et al., 2026) maps its first four intersections to L1/L3/L5/L7 without reordering or merging. LD uses finite positive camera-Z alone, whereas MD requires the mask and finite positive Z. Each image uses one 20-step sample with seed $42 + i$. Releases r69l/r75b differ in resolution and XYZ normalization.

LaRI (Li et al., 2026) keeps native intersection order without sorting or merging. The scene release uses finite positive Z without a mask checkpoint, whereas the separately trained object release applies its ray-stop mask. This compares distinct checkpoints. Scale conventions differ: SeeGroup/LaRI use relative scale, WT r69l returns relative XYZ, and r75b restores metric units with fixed normalization statistics. Ordinal annotations assess ordering, while metric accuracy requires depth ground truth.

All five released LD-Syn evaluations use the same 500 images and 460,008,237 occupied rays as the primary results, retaining supplied ties in the target counts. MD uses the eight-bit events and fixed point denominators defined above. These scores measure transfer to the supplied visible targets. Differences in native masks, ordering, scale, training, and stochastic generation prevent attributing the results solely to architecture or ranking each model on its native task.

Released-model AbsRel uses a different calibration protocol from the primary comparison. For all five models, the evaluator fits affine scale and shift per image and layer on valid GT–prediction pairs with $0 < d, \hat{d} < 10^4$, retains native values when fewer than 16 fitting pixels exist, and clips predictions to $[.001, 30]$ m. We average four conditional same-rank AbsRel scores without shallower-layer inheritance. SeeGroup's calibrated pass covers all 500 images and preserves decoded validity masks. These GT-calibrated estimates assess conditional within-layer localization after alignment, leaving native scale and inter-layer geometry to separate evaluation. Support remains important: LaRI objects retains only $3.63\%/1.64\%$ of third-/fourth-rank target support, so its deeper errors describe a small selected population rather than recovery of all annotated visible layers.

B.3 Ablation Factors and Configuration Provenance

The five primary variants in Table 1 ask how to determine emitted cardinality and assign observed depths to components. All share the recorded data, encoder/recurrent trunk, capacity, optimization schedule, and four-seed terminal-800k protocol, with geometric auxiliary terms disabled.

Count regularization versus fixed cardinality tests a complete global-stopping configuration: categorical predictor, loss, and selector. Count regularization averages depth costs over valid entries; other variants average per-ray sums. For assignment, MAP, ExactMB, and ordered assignment share model

and training fields within each seed, differing only in assignment criterion. Cross-group comparisons also change recurrence gating and selection rather than isolating an additional loss.

A separate four-seed study compares PPP, MAP, and ExactMB. Its recorded synthetic-data sampler and scheduler differ from the primary cardinality-interleaved sampler and warmup–hold–cosine schedule; these snapshots are incomplete launch records. Within this cohort, PPP changes the point-process law, while MAP and ExactMB differ in assignment criterion. The PPP control uses $\lambda_j = \mathrm{softplus}(e_j)$, with decoder score $\sigma(e_j)$ equal to its nonzero-count probability before the numerical rate floor. An intensity component may generate multiple returns. ExactMB minus MAP changes MD pattern accuracy by -1.35 ± 6.69 points in the primary study and $+1.96 \pm 1.37$ in the mechanism study (paired mean and sample SD). Interpretation is therefore recipe-specific.

Endpoint records are fuller than training histories. Checkpoint and evaluator hashes cover all 20 primary and 12 point-process-study endpoints, but synthetic geometry records lack a historical evaluator-source hash. The 20 primary exports describe resumed states without a source commit; the mechanism-study catalog preserves selected evaluator-resolved fields. These records support the comparisons but do not identify the source version throughout every historical training segment.

B.4 Annotation Semantics and the Training Interface

The supplied annotations do not certify the completeness or distinctness assumed by the set likelihood. Compaction retains supplied order and quantized repetitions, and all-false element masks contribute $M = 0$ without a missingness flag. ExactMB sums assignments over these entries, whereas ordered assignment fixes their compacted-channel correspondence. Normalization and propriety hold for the continuous simple-set law, whose assumptions need not be satisfied by these annotations.

In the 20-image 432×768 audit, 202 of 6,635,520 rays at one image's right boundary have all-false masks of unresolved empty-versus-missing status. Of the remaining rays, 11,604 (.175%) have exact valid-depth ties and 66,567 (1.003%) have a pair within 2 cm. The 2 cm prediction filter leaves targets unchanged, so pruning correct predictions of close targets can undercount.

The full-resolution audit tests depth validity (positive and no greater than 80 m) across 460,800,000 pixels in 500 synthetic images. No invalid channel precedes a valid one. Cardinalities one through four occupy 402,453,131 / 29,100,688 / 17,298,739 / 11,155,679 pixels, with 791,763 empty masks. Adjacent valid depths tie in 1,720,016 pixels and descend in 37,774. Thus, compact prefixes equate pattern and count accuracy without certifying sorted, distinct targets or physically empty rays.

C Additional Experimental Results

Count agreement, depth accuracy, and surface recovery can favor different choices. We expand primary four-seed terminal-800k comparisons with paired effects, rank-wise geometry, and error breakdowns, then examine separate point-process, geometric-loss, and architecture studies. Operating-point analyses and real-world visualizations connect decoding choices to retained support.

C.1 Primary Ablations and Released-Model Transfer

The primary variants—depth stacking (fixed count), explicit count regularization, MAP assignment, marginalized assignment (ExactMB), and ordered assignment—share a prediction family while varying how many depths are emitted and how predictions are assigned to targets.

C.1.1 Paired Effects of Explicit Count Regularization

Table 1 reports all five variants without geometric auxiliaries at the principal operating point. Fixed-count and count-regularized configurations differ in objective, depth-loss reduction, and selector. The three presence-based variants instead share gated recurrence and differ only in recorded assignment criterion per seed, as specified in Appendix B.3. We first examine the count-regularized contrast.

Table S4 shows that count regularization reduces false emission and improves aggregate count accuracy, but loses valid returns, especially at full cardinality. These effects compare complete

Table S4: **The count-regularized configuration reduces false emission at the cost of valid returns.** Paired count-regularized minus fixed-cardinality differences, reported as mean±sample SD over seeds $7/42/61/123$, with native selectors and common geometric filters. Rate differences are percentage points. Count MAE is in layers. Every paired seed shares the displayed direction, including losses in LD valid-query accuracy, transparent-point recall, and four-return count accuracy.

Benchmark	Outcome	Signed change Count reg. − fixed
LD-Real	Ordinal ACC ↑	$-8.5_{\pm 1.3}$
LD-Real	Tuple-level OverPred ↓	$-91.0_{\pm 0.8}$
MD-3K	Pattern ACC ↑	$+65.8_{\pm 1.2}$
MD-3K	Transparent-point Recall ↑	$-20.9_{\pm 1.0}$
MD-3K	Opaque-point OverPred ↓	$-89.9_{\pm 0.5}$
LD-Syn	Occupied-micro count ACC ↑	$+91.32_{\pm 0.18}$
LD-Syn	Cardinality-macro count ACC ↑	$+48.07_{\pm 0.28}$
LD-Syn	Count MAE ↓	$-2.177_{\pm 0.032}$
LD-Syn	Four-return count ACC ↑	$-21.55_{\pm 0.80}$

configurations, not an isolated count loss. Converting rejection to overprediction reverses signs without changing SD; four paired seeds give an exact two-sided sign-flip resolution of .125.

Released-model transfer. Table 2 complements controlled ablations with released models, without retraining. Appendix B.2.6 specifies native-output adapters and GT-calibrated depth evaluation.

C.2 Primary Four-Seed Geometric Recovery

Count agreement alone does not establish which surfaces are recovered. Table S5 therefore reports per-rank geometry and support for the same primary terminal-800k checkpoints and four seeds as the main comparison. Each checkpoint is evaluated on all 500 LD-Syn validation images with the prescribed decoder. Ranks 1–4 correspond to benchmark channels L1/L3/L5/L7. We pool pixels within each seed and report the mean and sample SD across seeds $7/42/61/123$. The SD therefore measures training-seed variation, excluding image-sampling uncertainty.

To separate localization from retained support, let G_r and P_r contain pixels with valid annotated and decoded depth at rank r, respectively, and let $I_r = G_r \cap P_r$. Conditional AbsRel averages relative absolute error over I_r. RMS is the square root of mean squared error on I_r. Conditional δ_1 is the fraction of I_r satisfying $\max(\hat{d}_r/d_r, d_r/\hat{d}_r) < 1.25$. Target-support recall is $|I_r|/|G_r|$. With C_r the subset satisfying this depth criterion, depth-accurate recall and precision are $|C_r|/|G_r|$ and $|C_r|/|P_r|$. We recover $|C_r|$ from archived conditional δ_1 and intersection counts, verifying integrality at every checkpoint and rank. These scores assess depth and presence per rank. Appendix D.2.1 tests recovery of all ray returns via absolute-tolerance injective matching on a separate cohort.

Ordered assignment improves depth-accurate recall and precision over marginalization at every rank in all four paired seeds. At rank four, the gains are 13.95 ± 0.58 and 8.76 ± 0.59 percentage points, respectively. Fixed count retains high target coverage with low deeper-rank precision. Compared with ordered assignment, the count-regularized model attains higher precision at ranks 2–4 with lower recall. Thus, localization, coverage, and precision favor different configurations.

Metric and provenance scope. We analyze archived same-rank optional-depth outputs without last-visible fallback, alignment, or clipping, excluding legacy per-layer metrics that clip to $[0.001, 30]$m. The 20 core metric hashes match the ledger, and all 80 same-rank records match archived summaries. The current evaluator confirms these definitions, but historical synthetic records lack a source hash identifying its version. This reanalysis uses existing aggregate outputs without new inference.

Table S5: **Depth error, support, and depth-and-presence recovery for the primary five systems.** Mean±sample SD over four training seeds. All entries except RMS (meters) are percentages. AbsRel, RMS, and δ_1 are conditional on both same-rank values being valid. Support and depth-accurate recall use the fixed GT denominators shown below. Depth-accurate precision uses each system's predicted support. Different conditional supports preclude interpreting AbsRel alone as complete recovery.

Method	AbsRel ↓	RMS (m) ↓	δ_1 ↑	Support ↑	Depth rec. ↑	Depth prec. ↑
Rank 1 (L1)	$\lvert G_r\rvert = 460\,008\,237$ pixels					
Fixed count	$14.02_{\pm 0.19}$	$0.240_{\pm 0.001}$	$85.34_{\pm 0.16}$	$100.00_{\pm 0.00}$	$85.34_{\pm 0.16}$	$85.20_{\pm 0.16}$
Count reg.	$14.14_{\pm 0.07}$	$0.235_{\pm 0.001}$	$85.41_{\pm 0.12}$	$99.99_{\pm 0.00}$	$85.40_{\pm 0.12}$	$85.35_{\pm 0.12}$
MAP	$14.17_{\pm 0.11}$	$0.268_{\pm 0.002}$	$84.71_{\pm 0.10}$	$100.00_{\pm 0.00}$	$84.71_{\pm 0.10}$	$84.61_{\pm 0.10}$
ExactMB	$14.26_{\pm 0.10}$	$0.263_{\pm 0.004}$	$84.72_{\pm 0.16}$	$99.99_{\pm 0.01}$	$84.71_{\pm 0.16}$	$84.62_{\pm 0.16}$
Ordered	$14.22_{\pm 0.32}$	$0.237_{\pm 0.003}$	$85.55_{\pm 0.14}$	$100.00_{\pm 0.00}$	$85.55_{\pm 0.14}$	$85.45_{\pm 0.14}$
Rank 2 (L3)	$\lvert G_r\rvert = 57\,555\,106$ pixels					
Fixed count	$17.22_{\pm 0.21}$	$0.550_{\pm 0.018}$	$80.89_{\pm 0.57}$	$99.15_{\pm 0.04}$	$80.20_{\pm 0.59}$	$10.07_{\pm 0.08}$
Count reg.	$16.00_{\pm 0.11}$	$0.461_{\pm 0.003}$	$83.84_{\pm 0.30}$	$89.89_{\pm 0.28}$	$75.36_{\pm 0.35}$	$74.96_{\pm 0.35}$
MAP	$16.99_{\pm 0.15}$	$0.602_{\pm 0.007}$	$78.77_{\pm 0.46}$	$97.46_{\pm 0.06}$	$76.78_{\pm 0.42}$	$53.95_{\pm 0.49}$
ExactMB	$17.86_{\pm 0.17}$	$0.608_{\pm 0.019}$	$78.46_{\pm 0.44}$	$96.59_{\pm 0.14}$	$75.78_{\pm 0.49}$	$54.04_{\pm 1.24}$
Ordered	$16.41_{\pm 0.25}$	$0.491_{\pm 0.002}$	$82.69_{\pm 0.22}$	$97.77_{\pm 0.18}$	$80.85_{\pm 0.25}$	$58.33_{\pm 0.52}$
Rank 3 (L5)	$\lvert G_r\rvert = 28\,454\,418$ pixels					
Fixed count	$18.09_{\pm 0.16}$	$0.606_{\pm 0.012}$	$78.75_{\pm 0.10}$	$95.76_{\pm 0.09}$	$75.41_{\pm 0.09}$	$5.21_{\pm 0.06}$
Count reg.	$16.82_{\pm 0.19}$	$0.517_{\pm 0.008}$	$81.04_{\pm 0.17}$	$79.08_{\pm 0.10}$	$64.08_{\pm 0.07}$	$66.67_{\pm 0.57}$
MAP	$19.30_{\pm 0.51}$	$0.716_{\pm 0.009}$	$74.47_{\pm 0.40}$	$93.34_{\pm 0.19}$	$69.51_{\pm 0.49}$	$35.73_{\pm 0.12}$
ExactMB	$21.04_{\pm 0.26}$	$0.735_{\pm 0.022}$	$74.19_{\pm 0.09}$	$89.76_{\pm 0.37}$	$66.60_{\pm 0.29}$	$34.35_{\pm 0.64}$
Ordered	$17.36_{\pm 0.30}$	$0.594_{\pm 0.009}$	$79.57_{\pm 0.15}$	$93.11_{\pm 0.29}$	$74.09_{\pm 0.32}$	$42.39_{\pm 0.24}$
Rank 4 (L7)	$\lvert G_r\rvert = 11\,155\,679$ pixels					
Fixed count	$23.80_{\pm 0.40}$	$0.895_{\pm 0.013}$	$73.05_{\pm 0.84}$	$74.05_{\pm 0.75}$	$54.10_{\pm 0.46}$	$2.40_{\pm 0.12}$
Count reg.	$21.93_{\pm 0.59}$	$0.621_{\pm 0.026}$	$77.08_{\pm 0.42}$	$52.51_{\pm 1.06}$	$40.47_{\pm 0.73}$	$52.46_{\pm 0.62}$
MAP	$24.42_{\pm 0.76}$	$0.930_{\pm 0.062}$	$73.38_{\pm 1.16}$	$64.94_{\pm 1.53}$	$47.65_{\pm 0.95}$	$17.84_{\pm 0.24}$
ExactMB	$23.57_{\pm 0.55}$	$0.994_{\pm 0.021}$	$75.22_{\pm 0.84}$	$52.30_{\pm 0.95}$	$39.34_{\pm 0.49}$	$14.51_{\pm 0.25}$
Ordered	$22.43_{\pm 0.53}$	$0.843_{\pm 0.028}$	$74.65_{\pm 0.89}$	$71.39_{\pm 0.50}$	$53.29_{\pm 0.83}$	$23.27_{\pm 0.35}$

C.3 Error Structure Across Cardinalities and Relations

To explain aggregate rankings, we stratify by cardinality and material pattern, then distinguish presence from ordering failures within each experiment population.

C.3.1 Cardinality and Material-Conditioned Reversals

Figure S1 stratifies the five primary variants by true occupied count and MD material pattern. Count regularization leads at $c = 1$, 2, and 3, with the best cardinality macro (75.0%), occupied-micro exactness (95.0%), and MAE (0.067 layers). At $c = 4$, however, its 52.5% exactness trails ordered assignment's 71.4%, reversing their lower-cardinality ranking.

Cardinality and material pattern reverse ablation rankings. Relative to the count-regularized model, ordered assignment changes exact count accuracy by $-4.30/-20.70/-18.43/+18.88$ points for $c = 1/2/3/4$. Its aggregate changes are -6.14 points macro, -5.31 occupied-micro, and 0.087 more MAE. Rankings also depend on material pattern: relative to ExactMB, ordered assignment improves MD pattern ACC by 9.29 ± 6.95 points on TT and 4.56 ± 4.94 on TO, but changes OO by -1.92 ± 4.78 and raises point-level OverPred by 2.42 ± 4.46. Compulsory four-slot emission likewise reaches 74.1% at $c = 4$, despite a 3.7% occupied-micro score and 2.244-layer MAE overall.

C.3.2 Joint Existence and Ordering Failures

Correct cardinality does not ensure correct ordering. At 800k and $\tau = .5$, Figure S2 separates existence and ordering failures on valid LD tuples in the separate ViT-L recurrent geometry study described in Appendix D.3. Figure 5 applies the same decomposition to all MD-3K pairs across the five primary four-seed ablation configurations at the principal operating point.

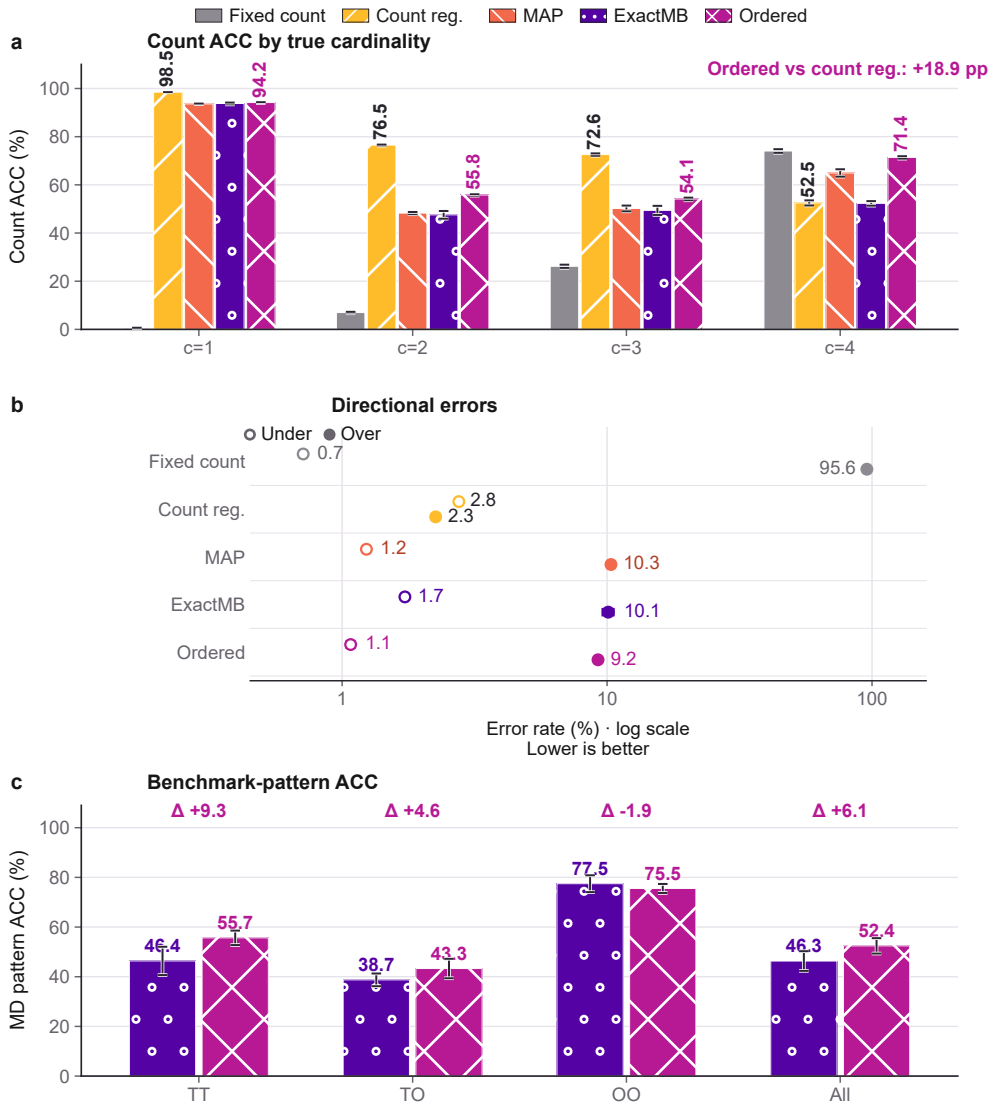


Figure S1: **Conditioning reveals that the aggregate ranking is regime-dependent. (a)** Grouped bars stratify exact decoded count by true occupied count c. Whiskers are sample SD, and direct values identify the count-regularized and ordered-assignment variants, which reverse at $c = 4$. **(b)** Open and filled points show under- and over-counting, respectively, on one logarithmic axis. Printed values retain native percentages and whiskers show sample SD. All intervals are strictly positive, with no numerical offsets or truncation. The log view resolves small error rates alongside the fixed-cardinality variant's 95.6% over-counting. **(c)** Paired bars compare marginalized (ExactMB) and ordered assignment on TT, TO, OO, and overall MD-3K benchmark-pattern exactness. Here, Δ is ordered minus marginalized assignment. Each stratum tests the full decoded validity pattern at both query points, before ordinal fallback. All displayed variants use the same four primary training seeds as the controlled rows in Table 1. The fixed-cardinality variant's high $c = 4$ value reflects compulsory four-slot emission, so it cannot demonstrate successful stopping.

The source of failure changes with both rank composition and arity. From pairs to quadruplets, same-rank LD existence failures remain near 8–10%, while correct-existence/wrong-order outcomes rise from 13.72% to 31.87%. Mixed-rank existence failures rise from 23.97% to 27.15%. Across all

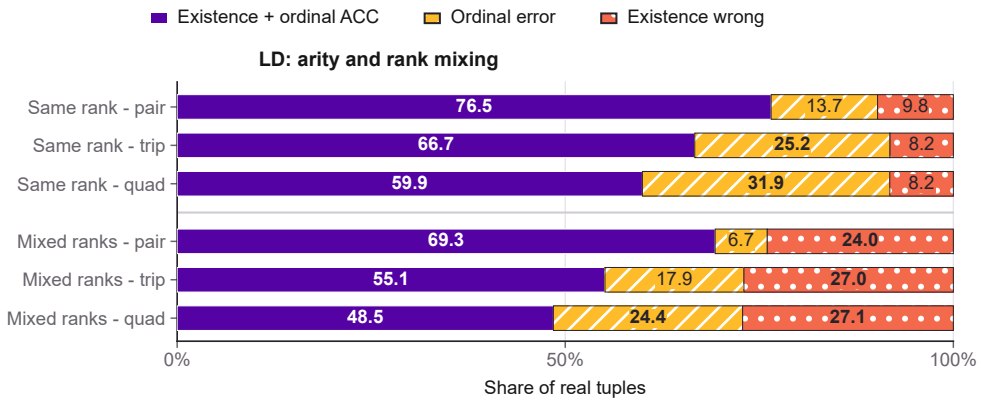


Figure S2: **Ordering accuracy alone can hide incorrect presence patterns.** In the geometry study at latest 800k and τ_{diag}=0.5, each valid LD tuple has exactly one outcome: existence and order both correct, existence correct but order wrong, or existence wrong. Seeds are averaged within recipe family, then the 16 designed families receive equal weight. Mixed-rank tuples have more existence failures. Quadruplets have a larger ordering-failure share than pairs even when queried ranks exist.

recipe families and six thresholds, seed-averaged mixed-rank existence accuracy trails same-rank accuracy at every arity. Quadruplets have larger ordering-failure shares than pairs, consistent with the additional ordering constraints that must hold simultaneously.

In the primary MD study, ordered assignment exceeds ExactMB by 4.28 Joint ACC points and 6.11 pattern ACC points despite lower $\Pr(O = 1 \mid E = 1)$ (88.9% versus 91.4%): better presence-pattern recovery outweighs lower conditional ordering accuracy. These summaries use all 3,161 pairs, whereas headline ordinal ACC is evaluated only on the TT+TO subset.

C.4 Point-Process and Assignment Controls

C.4.1 Point-Process Law and Assignment Reduction

To examine point-process and assignment effects, we compare PPP, MAP, and ExactMB using a shared $K = 4$ model and four seeds at 800k updates. This separate study uses a different recorded sampler and schedule. Appendix B.3 describes its incomplete historical training records. Figure S3 contrasts PPP's global void event and repeated intensity ownership with MAP's injective Bernoulli singleton/null costs and ExactMB's marginalization of these explanations.

Changing the point-process model improves pattern and count recovery. PPP→MAP raises MD pattern ACC by 14.39 ± 2.84 points and synthetic occupied-micro count ACC by 8.79 ± 0.15, and lowers MAE by 0.121 ± 0.003 layers across all four seeds. MD point-level OverPred changes by $+0.32 \pm 1.59$ points, positive in three seeds. PPP already supplies absence evidence, so this contrast combines changes in point-process law, multiplicity, and assignment.

MAP→ExactMB changes only the recorded assignment criterion. Posterior summation lowers LD tuple-level OverPred by 1.56 ± 0.47 points and MD point-level OverPred by 5.31 ± 1.49, and raises MD pattern ACC by 1.96 ± 1.37 across all four seeds. Count MAE worsens by 0.0069 ± 0.0030 layers in every seed. Synthetic occupied-micro exactness changes by -0.014 ± 0.214 points, negative in only one of four seeds. This near-zero mean does not establish equivalence. Marginalization thus reduces false emission at a small count-MAE cost. Appendix B.3 reports the opposite pattern-accuracy change and greater seed variation in the primary four-seed comparison.

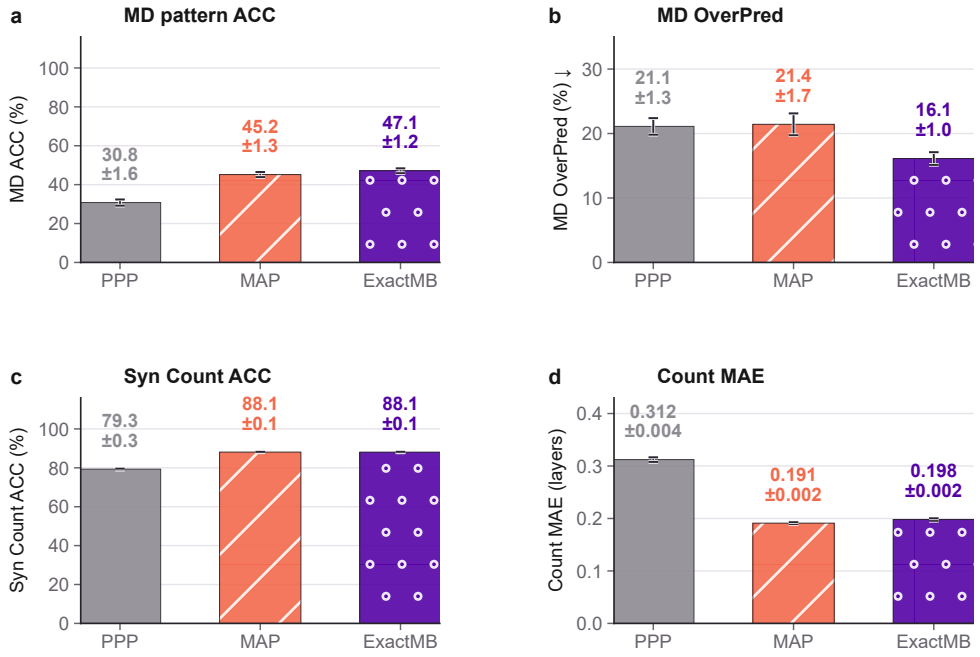


Figure S3: **MAP improves pattern and count accuracy over PPP. Marginalization reduces false emission.** (a) MD pattern ACC. (b) MD point-level OverPred. (c) LD-Syn occupied-micro count ACC. (d) LD-Syn count MAE. PPP has a global void event and permits repeated intensity ownership. MAP assignment uses injective Bernoulli–Laplace costs with null factors. ExactMB sums those explanations. PPP uses numerical rate and log-intensity floors. Normalization describes the underlying law. PPP→MAP assignment is a bundled point-process contrast. MAP→marginalized assignment changes only the criterion class among the available configuration fields. Bars report terminal/latest 800k means over seeds $7/42/61/123$. Whiskers and $\pm$ readouts give sample SD. Higher exactness and lower OverPred and MAE indicate better performance.

C.5 Matched Geometric-Loss Interventions

With ExactMB fixed, Table S6 tests geometric supervision across seven configurations and 26 terminal checkpoints at 800k (epoch 108), sharing a four-slot recurrent decoder and $\tau = .01$. G0 includes geometric regularization, so contrasts concern schedules and added terms. This shared-seed study is separate from the primary ablations; links to historical training code are incomplete.

Table S6: **Seven-arm geometric-loss study design.** All arms use ExactMB. G0 is the reference. G1 changes the final weight and ramp of sorted-layer gradient matching (SortedGrad). G5 and G6 share only seeds 42 and 123, so their difference cannot establish a replicated schedule interaction.

Arm	Intervention	Reference	Seeds
G0	SortedGrad $\lambda = .02$; linear ramp completes at $.15T$	—	7/42/61/123
G1	SortedGrad $\lambda = .05$; ramp completes at $.05T$	G0	7/42/61/123
G2	Depth-channel weights $[.5, 1, 1.5, 2]$ replace $[1, 1, 1, 1]$	G1	7/42/61/123
G3	Depth diversity: margin .01, $\lambda : 0 \to .03$	G1	7/42/61/123
G4	Intensity supervision: $\gamma = .1$, $\lambda = .2$	G1	7/42/61/123
G5	AssignGrad: detached assignment-aligned gradient term, $\lambda = .05$	G0	7/42/123
G6	AssignGrad: detached assignment-aligned gradient term, $\lambda = .05$	G1	42/61/123

Table S7: **Paired effects of the geometric-loss interventions.** Mean±sample SD over the intersecting training seeds in Table S6. Positive denotes improvement: accuracy differences, OverPred reduction, or count-MAE reduction. All rate effects are percentage points. The last column is in layers. A dagger marks an effect with the same nonzero direction in every paired seed. LD last-visible ordering pools mixed valid tuples. MD pattern ACC requires the complete post-decoder pair of validity patterns.

Paired improvement ↑							
	LD-Real			**MD-3K**		**LD-Syn**	
Contrast	**Ordinal ACC**	**OverPred reduction**	**Last-visible ACC**	**Pattern ACC**	**OverPred reduction**	**Count ACC**	**Count MAE reduction**
G1–G0	$+1.18^{\dagger}_{\pm 0.80}$	$+0.48^{\dagger}_{\pm 0.36}$	$+4.34^{\dagger}_{\pm 1.50}$	$+0.69_{\pm 2.53}$	$-0.17_{\pm 0.96}$	$-0.11_{\pm 0.16}$	$-0.0002_{\pm 0.0023}$
G2–G1	$+0.24_{\pm 1.79}$	$+0.04_{\pm 0.99}$	$-0.07_{\pm 2.04}$	$+0.49_{\pm 2.27}$	$+0.65_{\pm 1.39}$	$+0.07_{\pm 0.15}$	$-0.0004_{\pm 0.0021}$
G3–G1	$-0.10_{\pm 1.63}$	$+0.06_{\pm 1.52}$	$-0.12_{\pm 1.87}$	$+0.37_{\pm 3.62}$	$-0.70_{\pm 1.20}$	$+0.09_{\pm 0.16}$	$+0.0009_{\pm 0.0030}$
G4–G1	$-0.02_{\pm 1.95}$	$-1.84_{\pm 1.95}$	$-0.49_{\pm 2.52}$	$-3.01_{\pm 4.84}$	$+1.81_{\pm 3.57}$	$-0.11_{\pm 0.59}$	$-0.0040_{\pm 0.0092}$
G5–G0	$+0.90_{\pm 1.90}$	$+0.30_{\pm 2.38}$	$+2.83^{\dagger}_{\pm 1.50}$	$-1.28_{\pm 1.28}$	$+0.15_{\pm 0.38}$	$-0.12^{\dagger}_{\pm 0.07}$	$-0.0022^{\dagger}_{\pm 0.0020}$
G6–G1	$-0.05_{\pm 0.74}$	$+0.38_{\pm 0.89}$	$-0.85_{\pm 1.26}$	$-0.47_{\pm 1.77}$	$+0.87_{\pm 0.88}$	$-0.05_{\pm 0.08}$	$-0.0010^{\dagger}_{\pm 0.0005}$

Relational gains and set decisions respond differently. Table S7 shows that the stronger, faster-ramping SortedGrad schedule improves LD valid-query accuracy and last-visible ordering while reducing LD OverPred in all four paired seeds. Effects on MD pattern ACC, MD OverPred, and synthetic count accuracy vary across seeds; mean MD OverPred rises and count accuracy falls. Depth weighting and diversity likewise show mixed directions. Intensity supervision reduces mean MD OverPred at the cost of lower MD pattern exactness and higher LD OverPred.

AssignGrad shows a related trade-off: under weak SortedGrad, it improves last-visible ordering while worsening exact count and MAE in all three seeds. Under strong SortedGrad, its mean ordering effect reverses and MAE again worsens in every seed. Matched-seed comparisons could separate schedule effects from the different three-seed populations. These results motivate evaluating geometric supervision through both relational accuracy and full visible-set recovery.

C.6 Encoder Capacity and Decoder Design

Encoder capacity. Holding the recurrent decoder and ExactMB recipe fixed, Figure S4 examines encoder capacity. Across ViT-S/B/L, LD ACC rises from 53.60% to 57.44% and 63.62%, while synthetic count ACC increases from 81.50% to 85.48% and 88.15%. MD pattern accuracy improves mainly from Small to Base (28.52% to 46.69%), with Large at 46.24%. Encoder scaling therefore improves ordinal and count recovery more consistently than exact presence patterns.

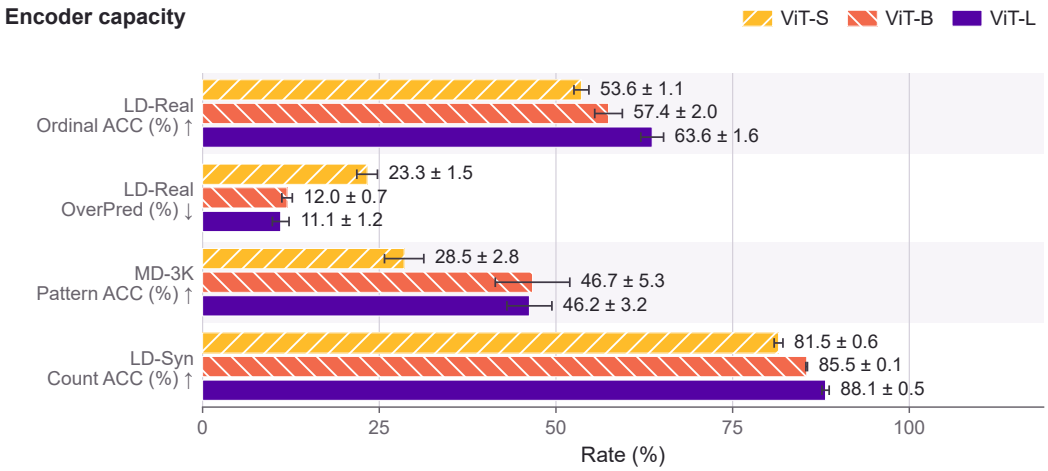


Figure S4: **Encoder scaling improves ordinal and count recovery.** ViT-S/B/L share a recurrent $K = 4$ decoder, ExactMB, and SortedGrad weight .05. Bars and error bars report means and sample SD across four matched seeds ($7/42/61/123$), at terminal-800k checkpoints with $\tau = .01$.

Decoder cost and recovery. Figure S5 examines decoder sharing through cost and recovery in a separate ordered-assignment study. Independent heads improve ordinal and presence-pattern accuracy over shared parallel FiLM, but selectively. LD-Syn occupied-micro count ACC rises by $.38 \pm .13$ points, while four-return count ACC falls by $4.88 \pm .39$ points and LD last-visible ACC falls by 3.87 ± 1.30 points (paired mean$\pm$sample SD). All four seeds share these directions. Aggregate accuracy can therefore improve while high-cardinality recovery and last-visible ordering deteriorate.

These selective gains also increase storage cost: independent heads use four times as many non-encoder parameters and raise peak memory from 1.94 to 2.50 GiB, despite similar operations and throughput. Shared recurrence retains compact storage but has lower profiled throughput. Profiles use randomly initialized models and exclude training, preprocessing, and postprocessing.

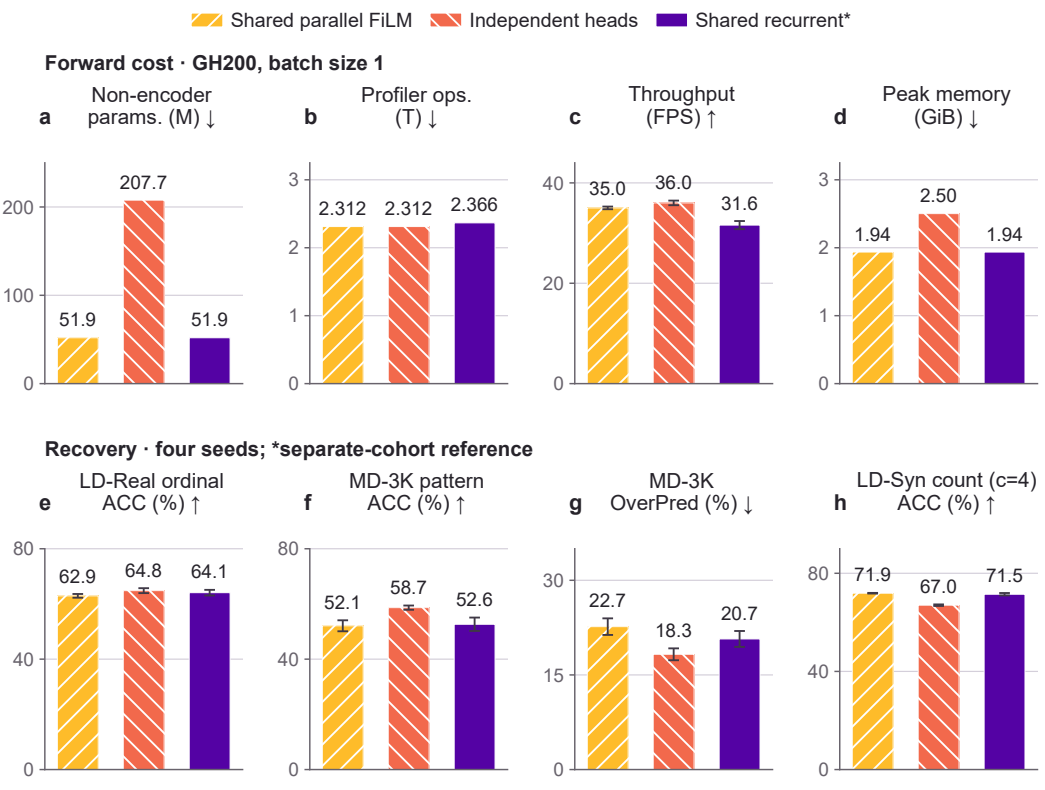


Figure S5: **Decoder design trades storage for selective recovery gains.** Top: GH200 120 GB forward profiles (518×518, batch size 1, TF32), with five warmups and 20 timed forwards per round. Parameters exclude the 304.37M encoder. Operations use the archived MAC convention. Bottom: ViT-L, $K = 4$, ordered assignment, SortedGrad .02, terminal-800k checkpoints, and $\tau = .01$. Parallel heads differ in weight sharing and parameter count. Shared recurrence (*) is a separate-cohort reference. Error bars show sample SD across three timing rounds for FPS and four seeds for recovery.

C.7 Archived Decoder Operating Points

To relate recovery to decoding, Figure S6 shows archived operating points from separate inference passes, not a fixed-prediction threshold sweep. It covers 12 primary MAP, ExactMB, and ordered-assignment checkpoints at $\tau \in \{.01, .1, .3, .5, .7, .9\}$. These terminal 800k checkpoints share finite-depth, depth-floor, and last-retained-gap filters. The principal threshold remains the predeclared .01, without test-set retuning or assumed shared calibration. Count-regularized and fixed-cardinality selectors ignore τ, so export variation cannot reflect threshold sensitivity.

Across same-checkpoint passes, ExactMB's .5 cutoff lowers LD ACC/OverPred by $4.90/6.07$ points, raises synthetic occupied-micro count ACC by 5.82 points, and lowers count MAE by $.117$ layers relative to .01. Count agreement and false emission thus improve while ordinal recovery falls. The geometric precision–recall analysis below examines which accurate depths remain. Since cutoffs index separate exports, calibration and geometric dominance require separate tests.

C.7.1 Geometric Precision–Recall Across Archived Operating Points

The rank-wise geometry archive contains 120 synthetic evaluations of all 20 primary terminal checkpoints, indexed by six cutoffs. The 12 Bernoulli-slot checkpoints trace operating-point curves, while the eight fixed-count and count-regularized checkpoints contribute their native exports. Each evaluation covers all 500 LD-Syn images. For each rank, we recover the integer number of depth-accurate predictions from the archived conditional δ_1 score and valid-intersection count. Following

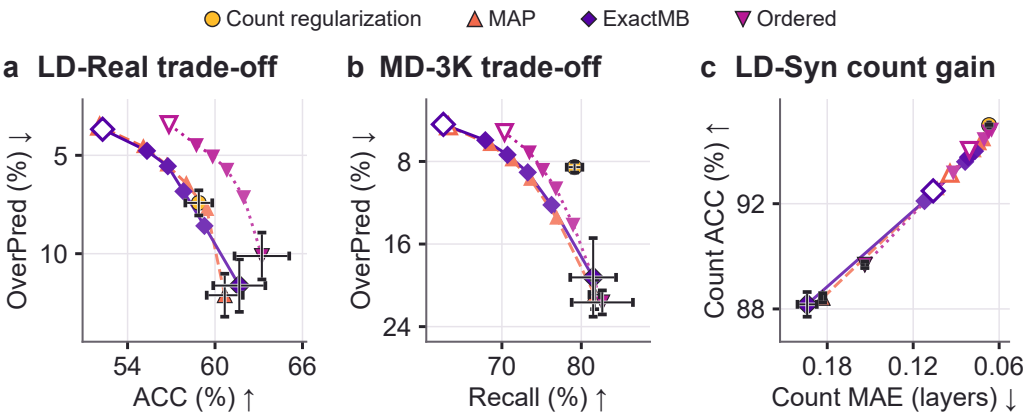


Figure S6: **Archived recovery operating points.** (a) LD-Real ACC versus OverPred. (b) MD-3K recall versus OverPred. (c) LD-Syn occupied-micro count ACC versus MAE. Curves connect separate inference passes at $\tau = .01, .1, .3, .5, .7, .9$. Filled/hollow endpoints mark $.01/.9$. Means and SD are shown at .01. Upper right is preferred. Fixed-count output lies outside the displayed ranges. The count-regularized model retains its native count-based selection.

Section C.2, dividing by GT-valid and prediction-valid counts gives recall and precision, respectively. At the native cutoff, all 80 rank records agree with the original primary-geometry report.

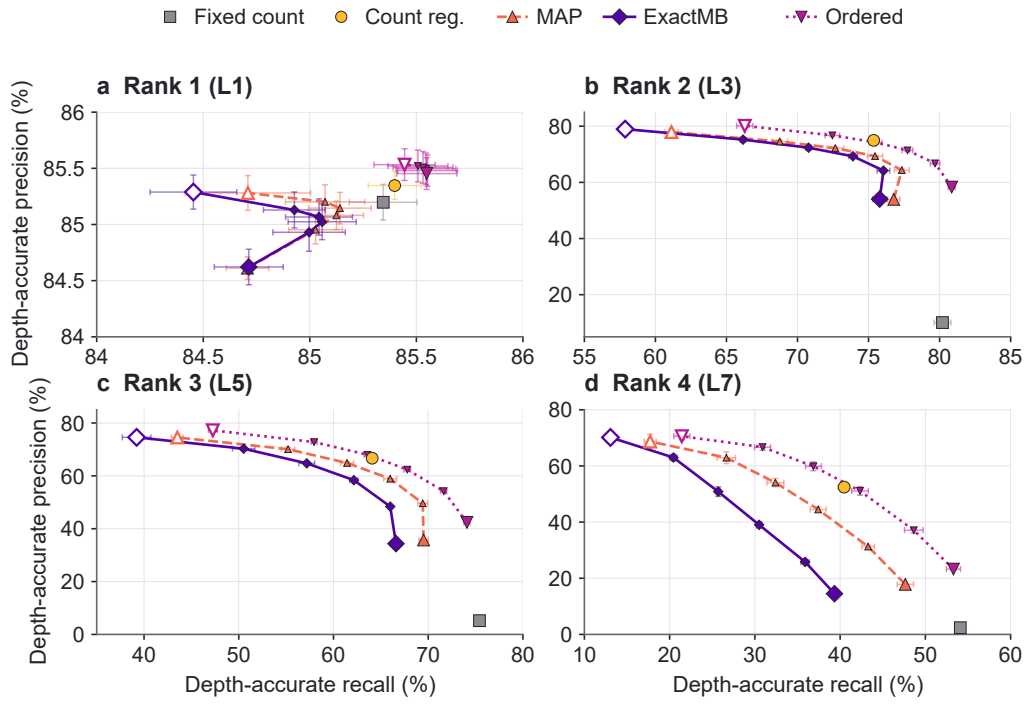


Figure S7: **Rank-wise geometric operating points for the primary models.** Depth-accurate recall and precision use the strict relative-error criterion $\max(\hat{d}/d, d/\hat{d}) < 1.25$, without alignment or optional-depth clipping. Curves connect archived passes at $\tau = .01, .1, .3, .5, .7, .9$. Large filled/hollow markers denote $.01/.9$, with small markers at intermediate cutoffs. Fixed-count and count-regularized variants use native .01 exports. Points and whiskers give four-seed means and sample SD. Rank-wise scores leave injective whole-set recovery unassessed. Axis ranges vary.

Table S8: **Geometry at a shared, existing cutoff.** MAP, ExactMB, and ordered assignment use $\tau = .5$. The count-regularized model retains its native selector and .01 export. Entries are depth-accurate recall/precision (%, mean±sample SD). A common cutoff does not ensure equal recall, equal precision, or matched probability calibration across methods.

		Rank 2 (L3)		Rank 3 (L5)		Rank 4 (L7)	
Method	τ	**Recall↑**	**Precision↑**	**Recall↑**	**Precision↑**	**Recall↑**	**Precision↑**
Count reg.	—	$75.36_{\pm 0.35}$	$74.96_{\pm 0.35}$	$64.08_{\pm 0.07}$	$66.67_{\pm 0.57}$	$40.47_{\pm 0.73}$	$52.46_{\pm 0.62}$
MAP	.5	$72.66_{\pm 0.59}$	$72.20_{\pm 0.29}$	$61.43_{\pm 0.69}$	$64.87_{\pm 0.57}$	$32.41_{\pm 0.99}$	$54.06_{\pm 1.46}$
ExactMB	.5	$70.78_{\pm 0.20}$	$72.35_{\pm 0.59}$	$57.18_{\pm 0.83}$	$64.73_{\pm 0.68}$	$25.70_{\pm 0.20}$	$50.85_{\pm 1.71}$
Ordered	.5	$75.54_{\pm 0.42}$	$74.23_{\pm 0.46}$	$63.58_{\pm 0.35}$	$67.99_{\pm 0.59}$	$36.87_{\pm 0.93}$	$59.88_{\pm 1.55}$

Higher precision can come at the cost of recall. Across ExactMB's archived .01 and .5 passes in Figure S7, fourth-rank precision rises from 14.51% to 50.85% while recall falls from 39.34% to 25.70%: paired changes are $+36.34 \pm 1.72$ and $-13.64 \pm .66$ points. Table S8 shows that ordered assignment at .5 approaches the count-regularized model's rank-2/3 operating points. At rank four, it trades lower recall (36.87% versus 40.47%) for higher precision (59.88% versus 52.46%). The changing precision gaps emphasize the role of the chosen operating point in method comparisons.

Operating-point and metric scope. These are comparisons between archived prediction passes. Despite its τ-independent selector, the count-regularized model varies across exports by up to 2.34 recall and 2.21 precision points. We therefore use its native export while the source of this variation remains unresolved. Curves retain annotation ties and same-rank pairing, with no calibrated or independently validated thresholds. Their aggregate records assess rank-wise recovery but lack the joint per-ray residuals needed for injective matching, whole-set correctness, or geometric-tolerance sweeps. Section D.2.1 examines these on separate fixed rays with absolute tolerances.

C.8 Additional Real-World Qualitative Results

Qualitative out-of-distribution multilayer depth comparisons. Figures S8–S21 visualize retained depth on fourteen real-world images, comparing the supplementary video's auxiliary-regularized ExactMB checkpoint with released SeeGroup, LaRI, and World Tracing models. World Tracing uses the r69l scene and r75b object releases, matching Table 2. These scenes are out of distribution for our synthetic-only task training. ExactMB's deeper layers show spatially selective support, while several released-model outputs repeat foreground or background structure across ranks.

Additional captured scenes. The eight additional photographs in Figures S14–S21 span tabletop glassware, decorative enclosures, flower arrangements, and larger interiors. Glassware and vase examples show support differences around transparent objects; bathroom, restaurant, and cafe views extend the comparison to cluttered interiors with localized transparency. Together, they illustrate how depth and retained support vary across ranks and scene content.

Display protocol. We sort valid decoded depths front to back per pixel. Each method and image uses a logarithmic color scale shared across retained ranks. Colors indicate relative depth, not comparable distances across methods. Gray denotes no retained depth, and dashes mark ranks beyond decoded capacity. ExactMB retains finite depths above 10^{-4} m with $q > .5$, without gap filtering. SeeGroup uses its released evaluation decoder, while LaRI's object model retains native stopping. LaRI and World Tracing object models receive full-frame images without automatic background removal.

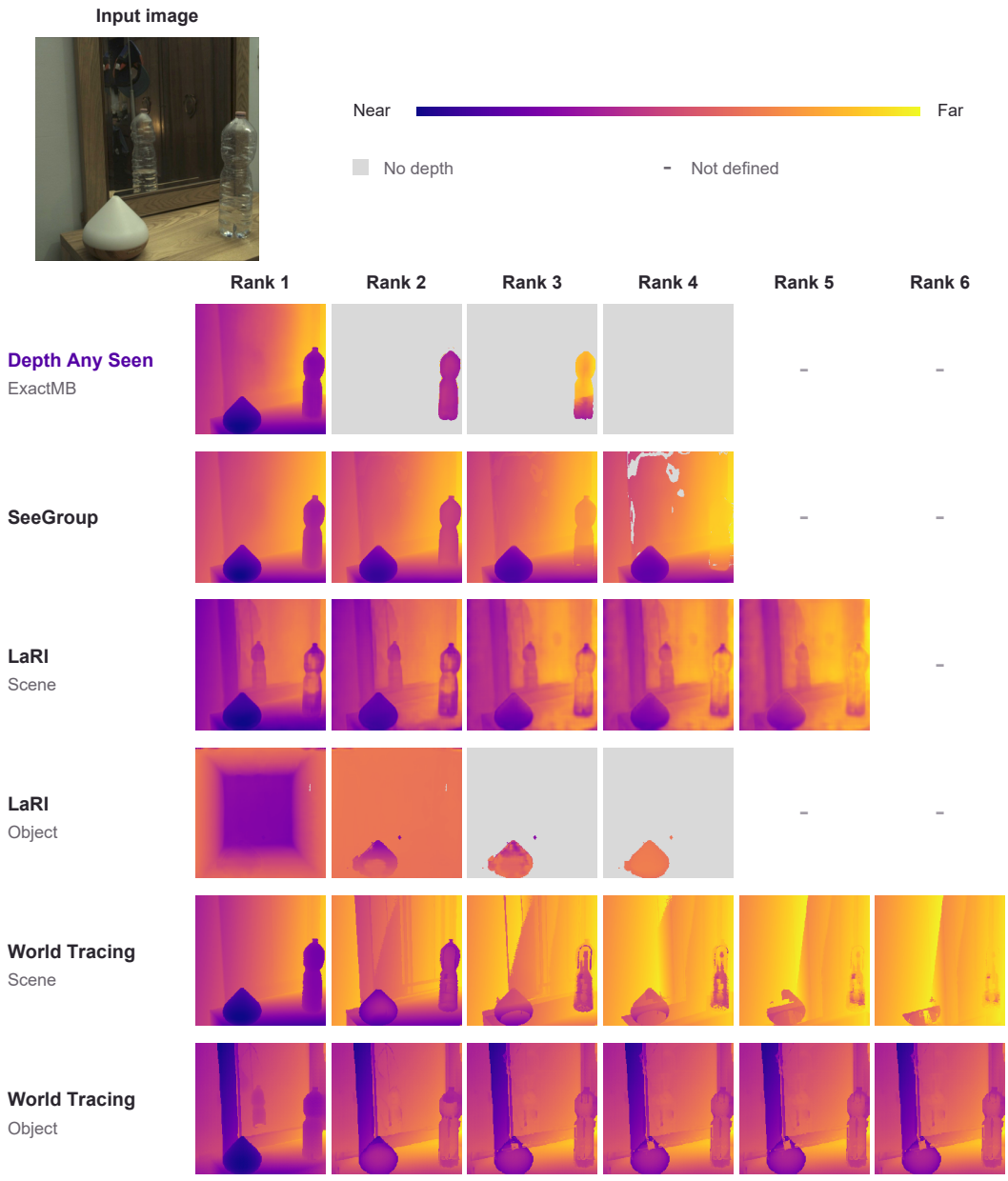


Figure S8: **Qualitative out-of-distribution multilayer depth comparisons.** Mirror and bottle.

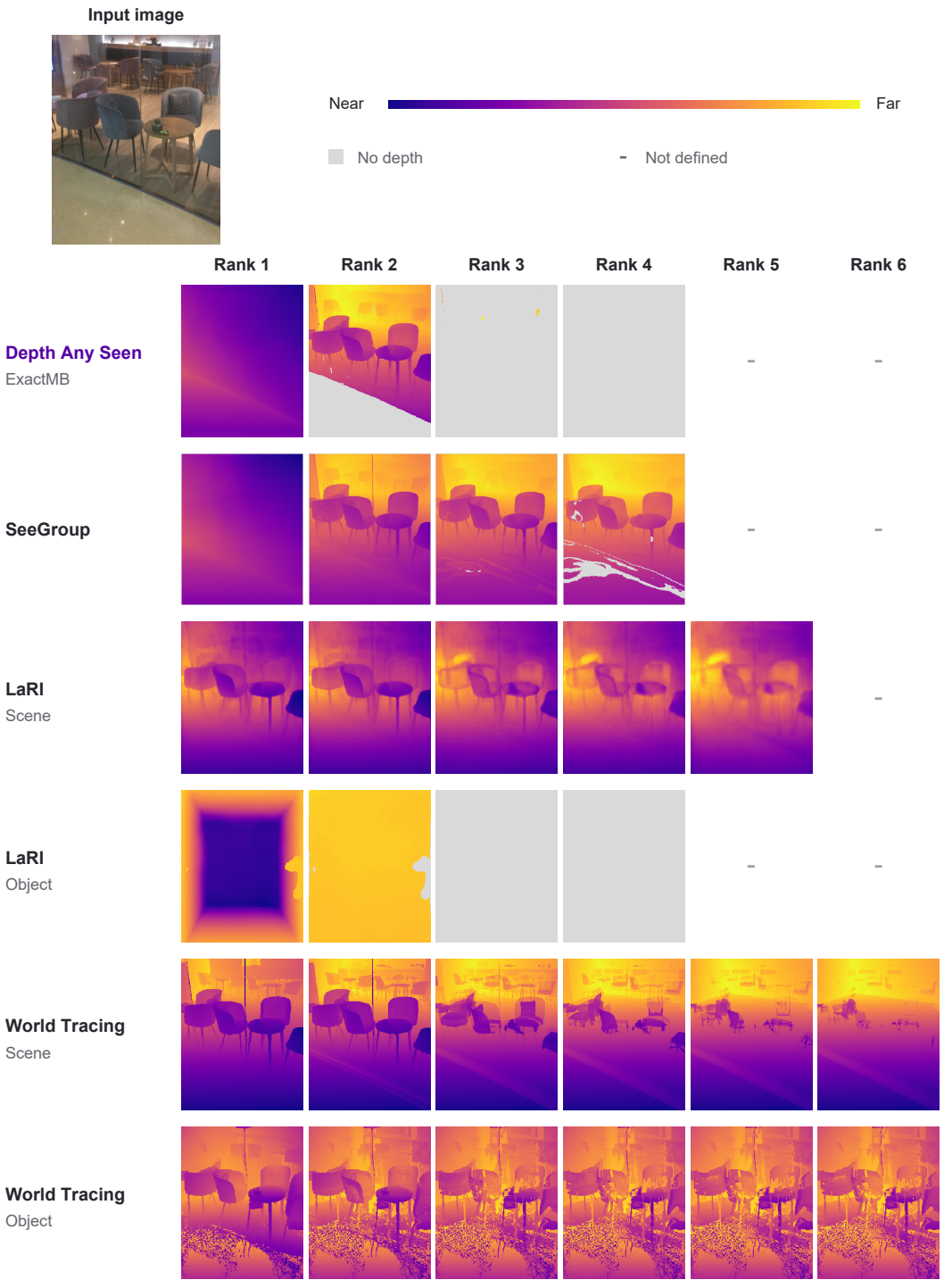


Figure S9: **Qualitative out-of-distribution multilayer depth comparisons.** Chairs and tables viewed through a glass partition, with seating at different depths (MD-3K 743).

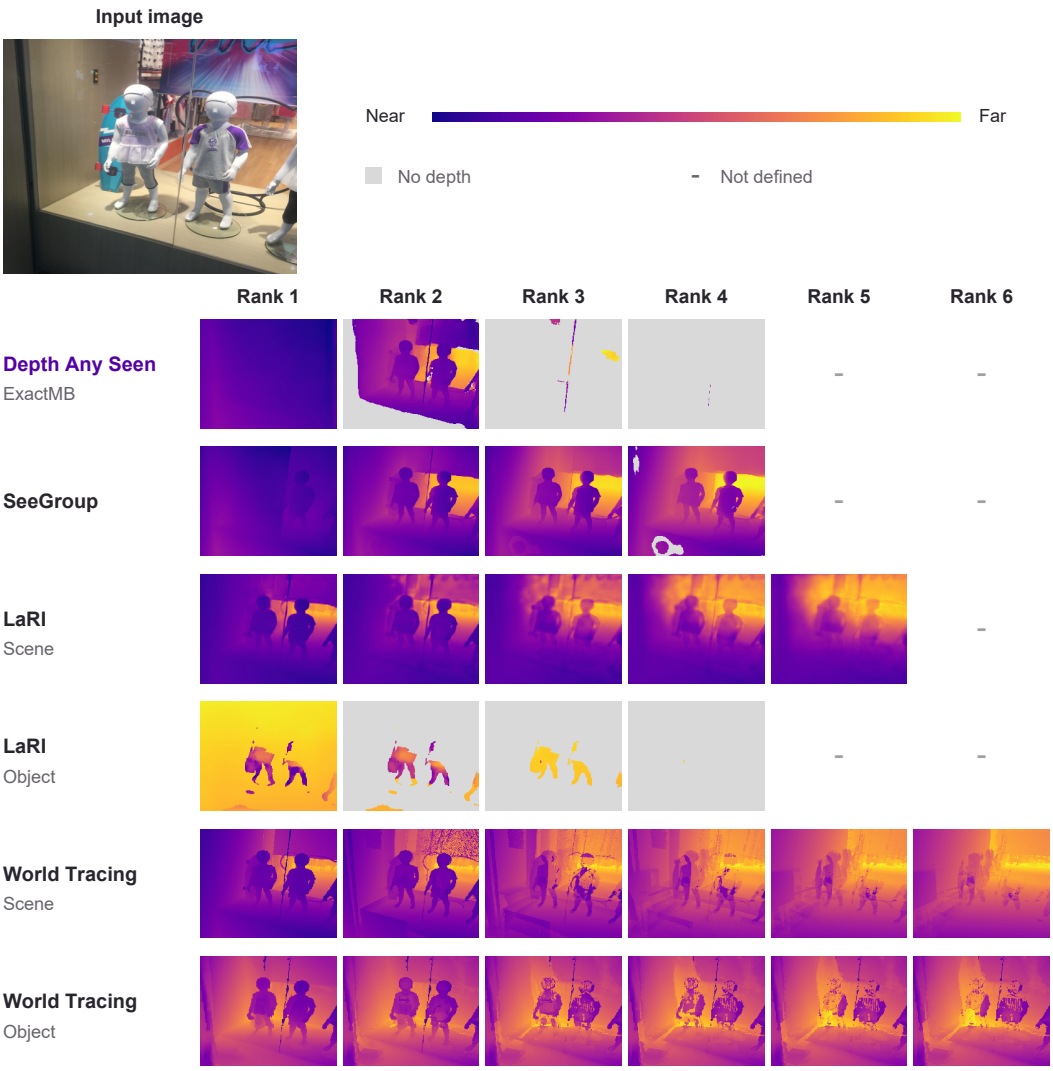


Figure S10: **Qualitative out-of-distribution multilayer depth comparisons.** A window display with two astronaut figures behind glass and a colorful backdrop (MD-3K 540).

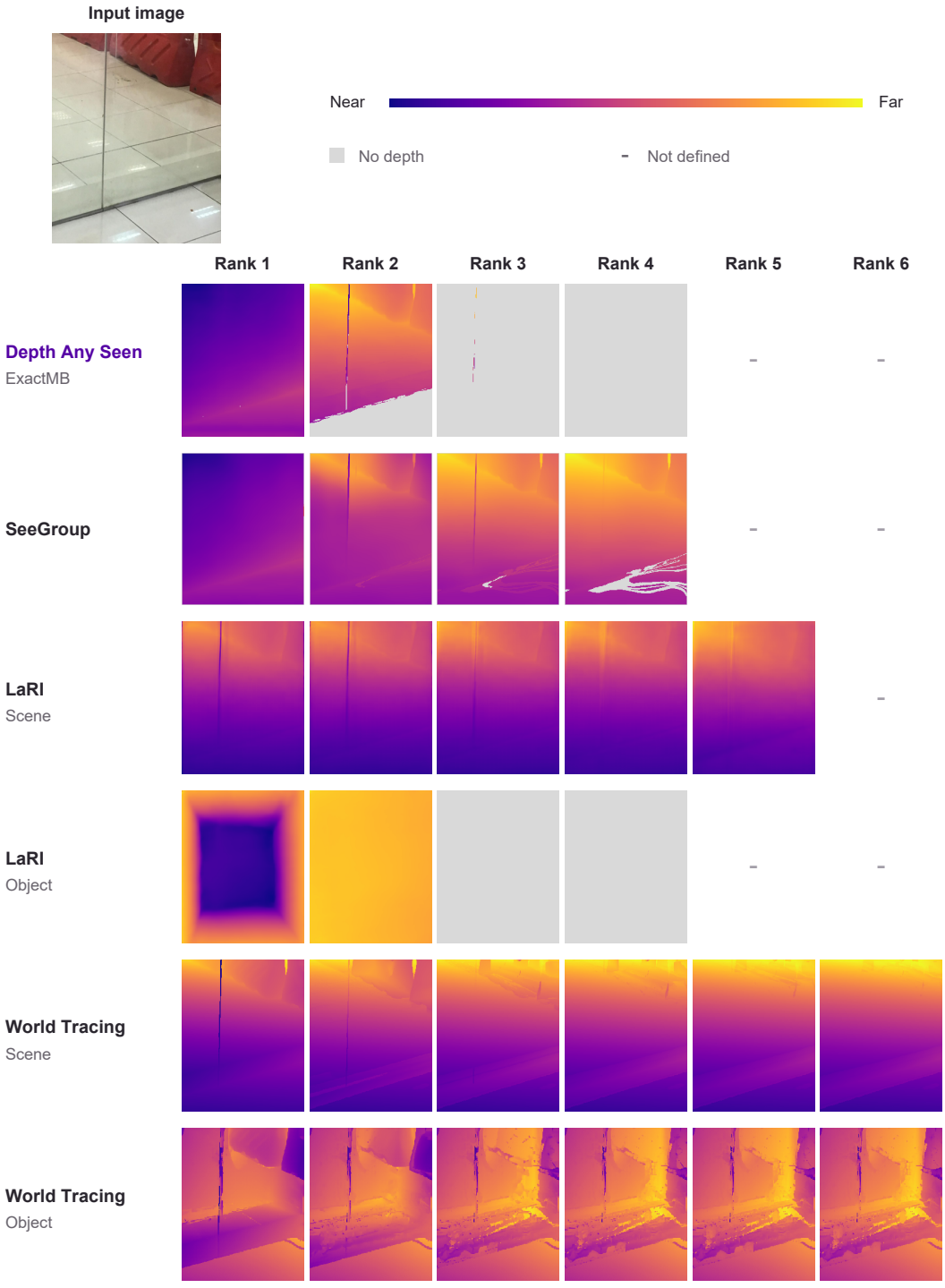


Figure S11: **Qualitative out-of-distribution multilayer depth comparisons.** A glass entrance above a tiled floor, with a door seam and red seating behind the glass (MD-3K 3134).

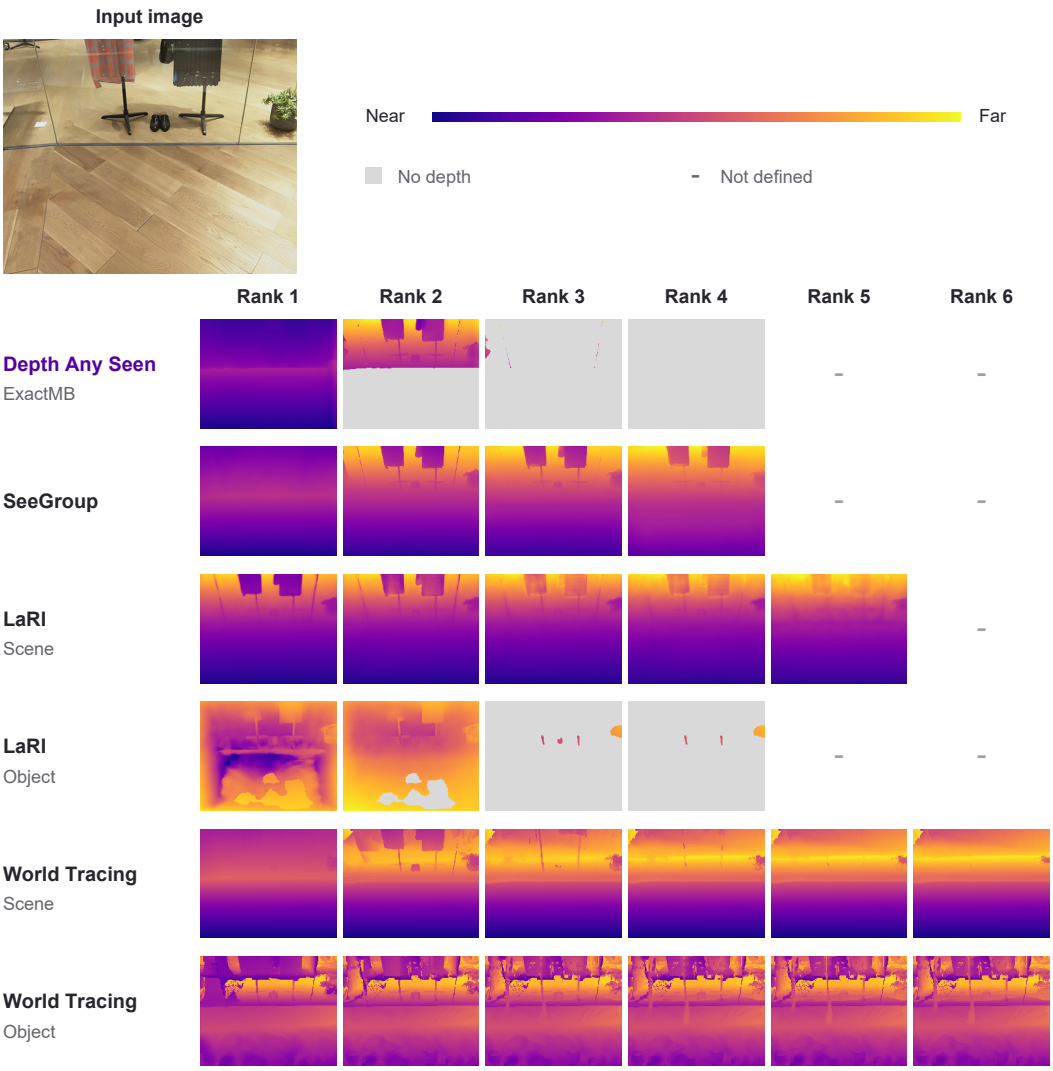


Figure S12: **Qualitative out-of-distribution multilayer depth comparisons.** A clothing display behind glass, with hanging garments above a wooden floor (MD-3K 2814).

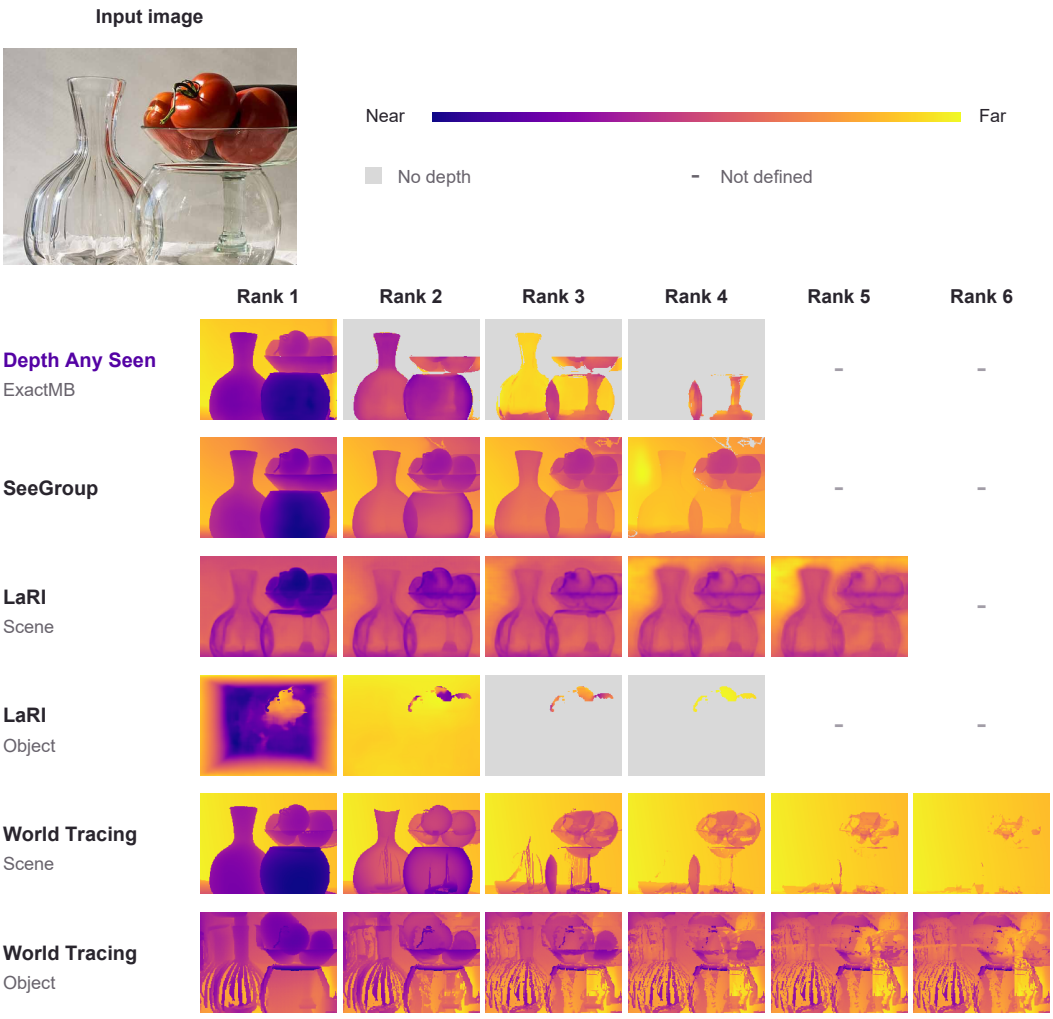


Figure S13: **Qualitative out-of-distribution multilayer depth comparisons.** Clear glass vessels beside a raised fruit bowl, with overlapping transparent surfaces (LD-Real 179).

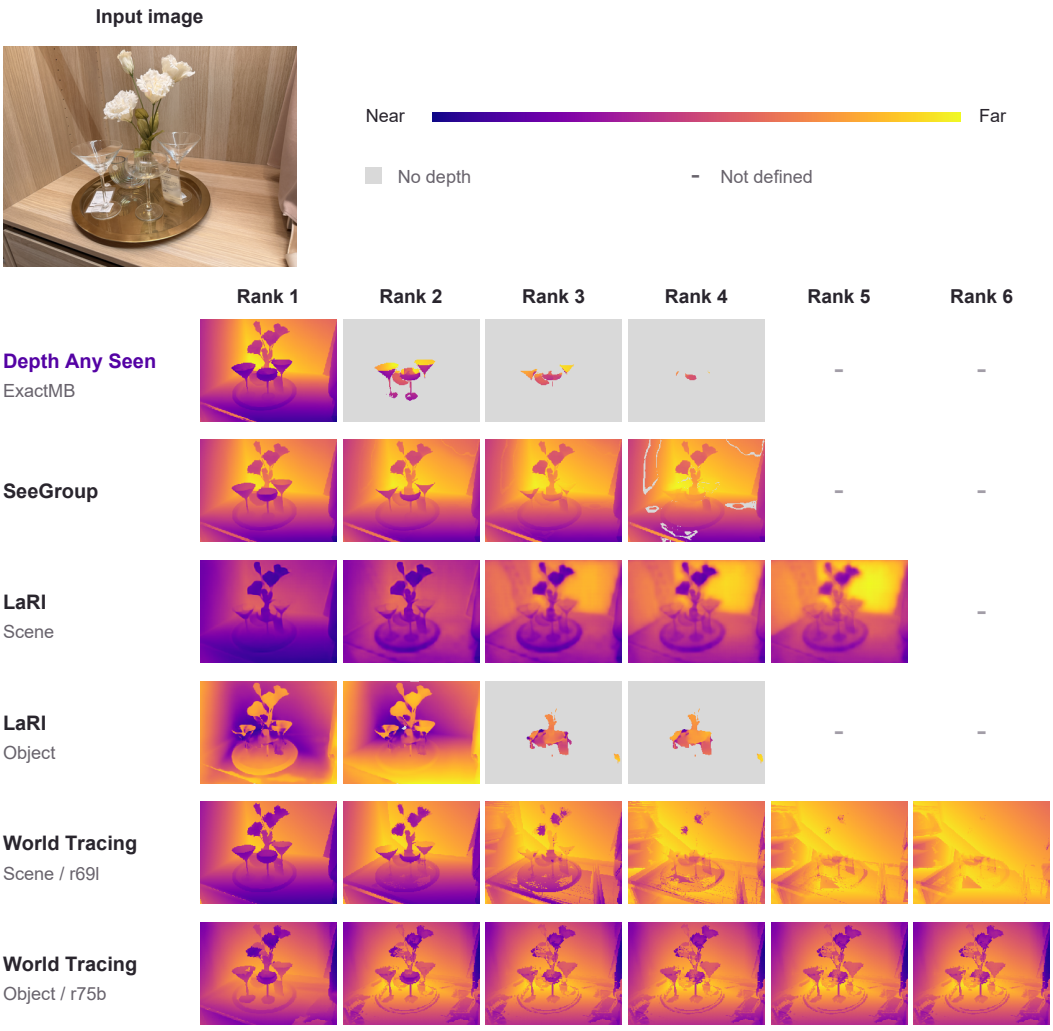


Figure S14: **Qualitative out-of-distribution multilayer depth comparisons.** Captured glassware arrangement with white flowers and stemware on a round tray.

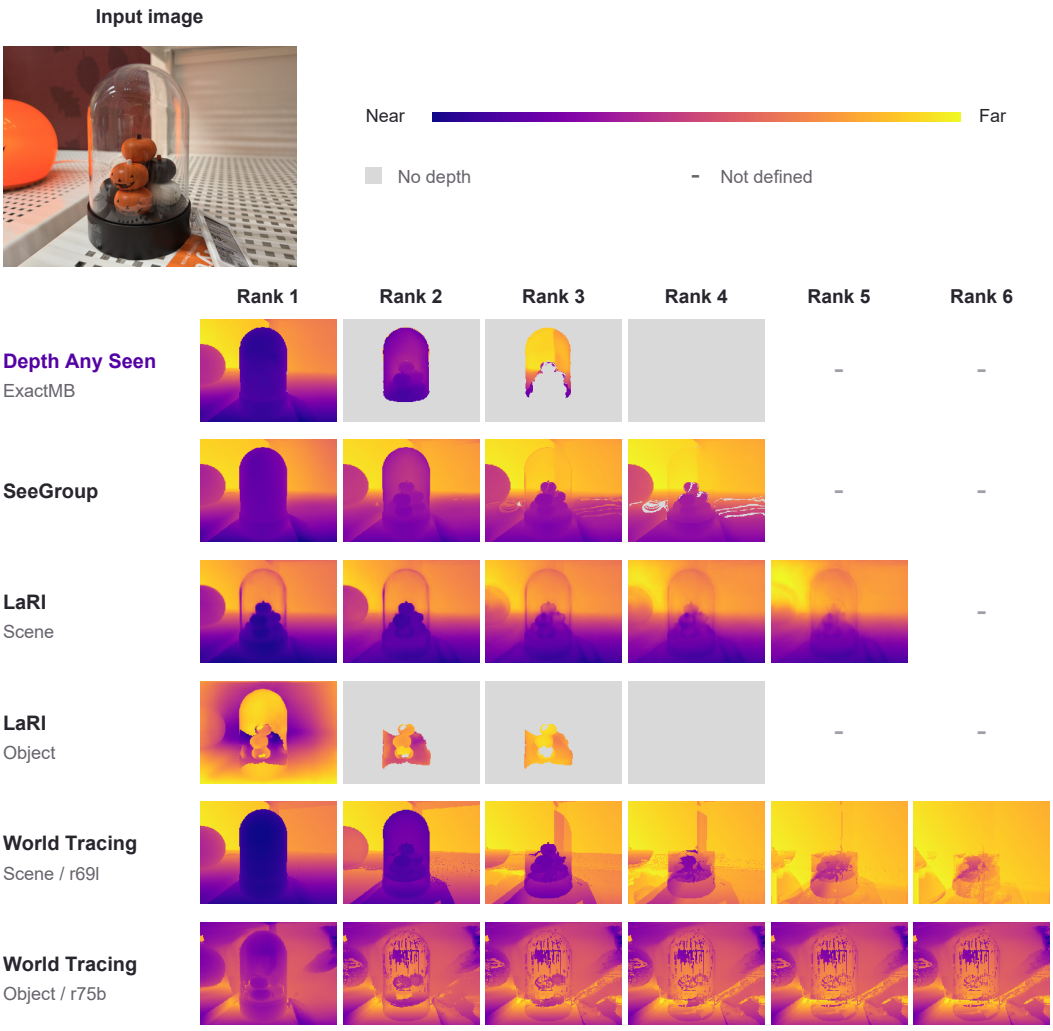


Figure S15: **Qualitative out-of-distribution multilayer depth comparisons.** Captured display of stacked pumpkin decorations enclosed by a transparent glass dome.

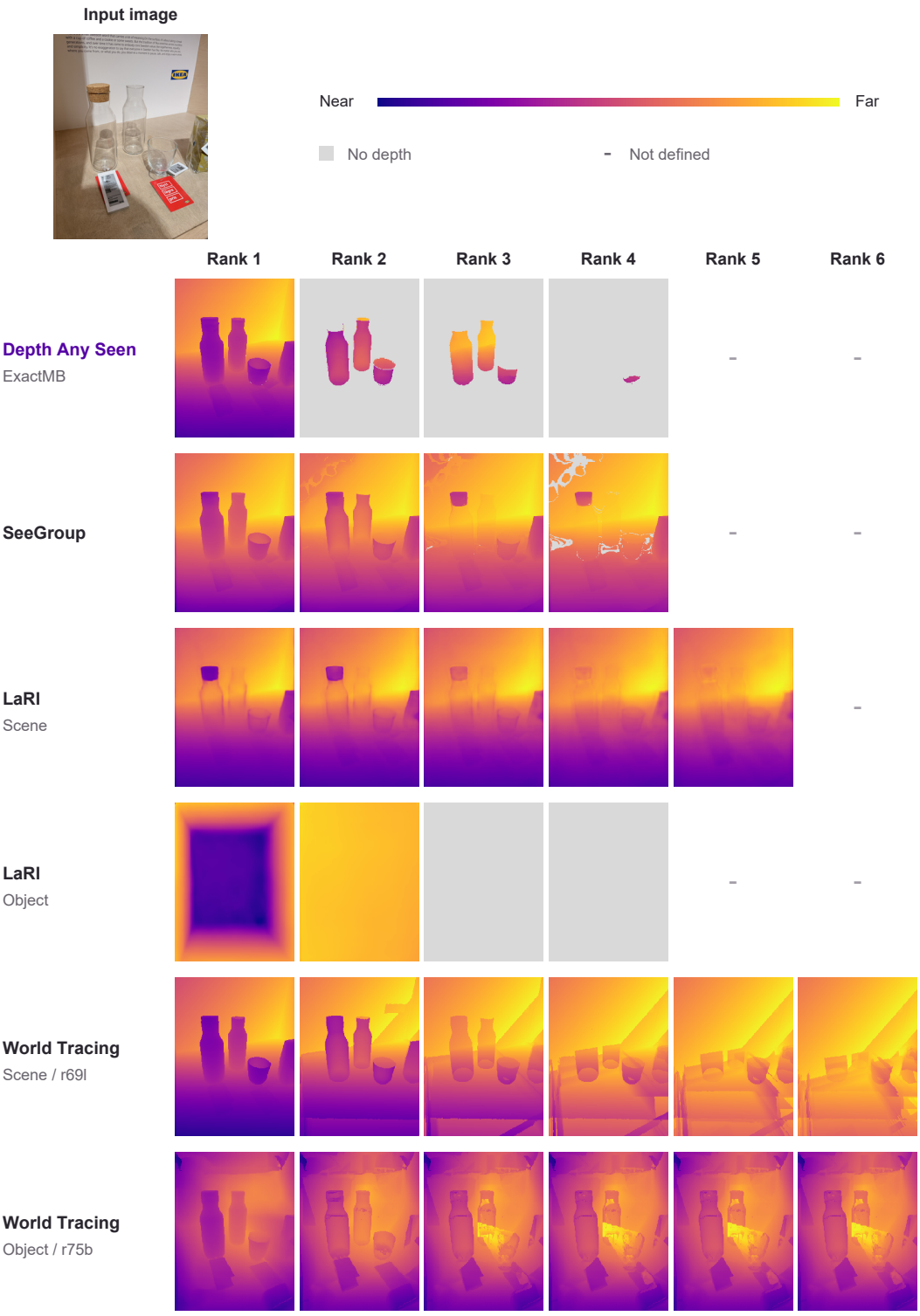


Figure S16: **Qualitative out-of-distribution multilayer depth comparisons.** Captured tabletop display with clear bottles and a tumbler against an opaque background.

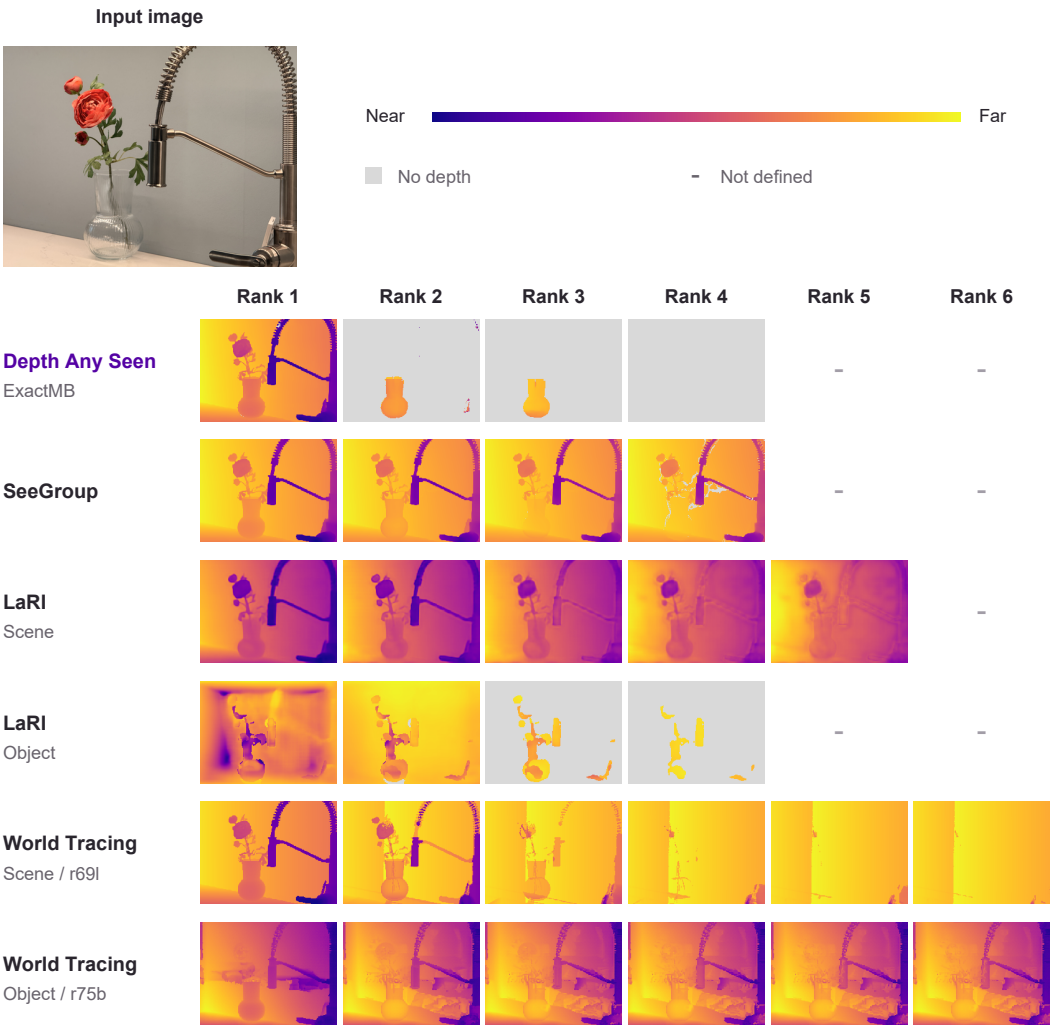


Figure S17: **Qualitative out-of-distribution multilayer depth comparisons.** Captured kitchen scene with red flowers in a ribbed glass vase beside a metal faucet.

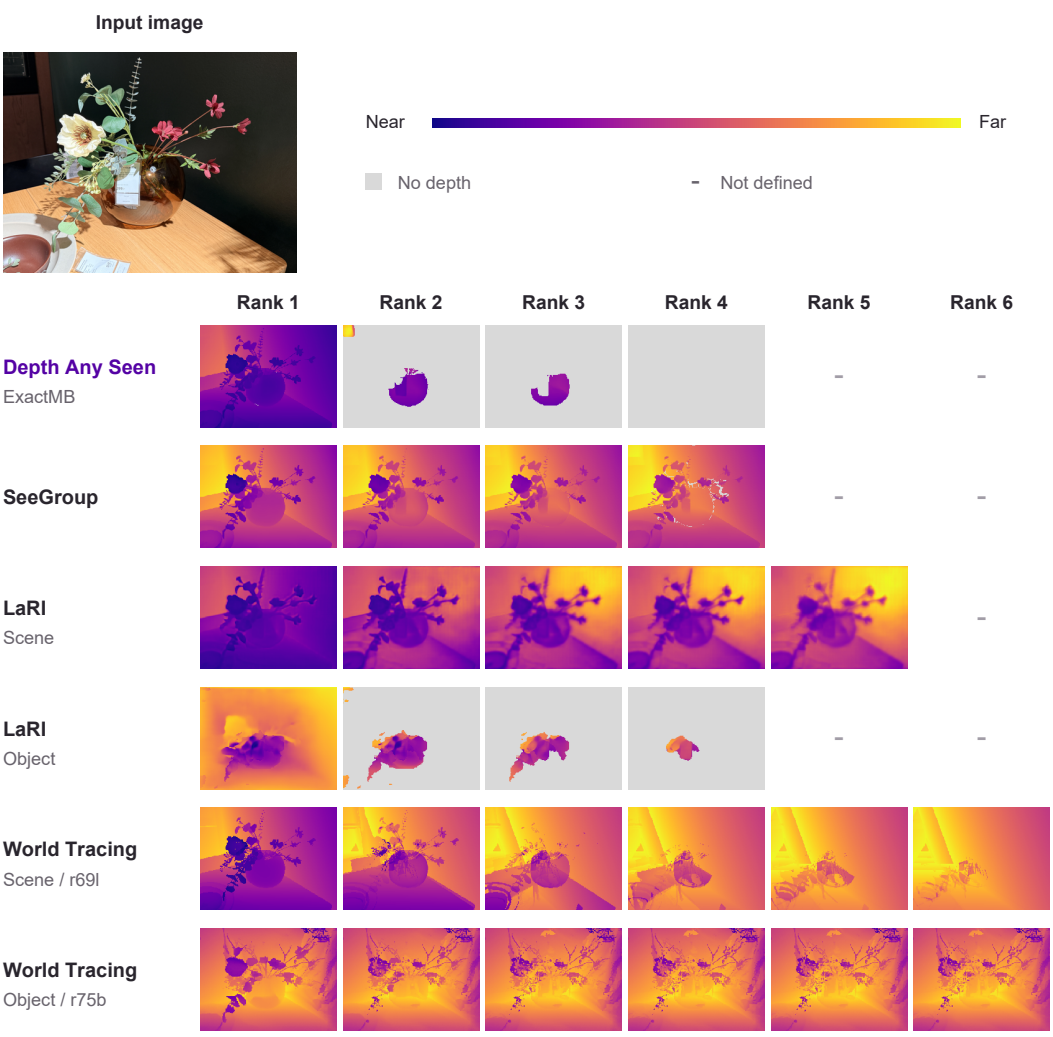


Figure S18: **Qualitative out-of-distribution multilayer depth comparisons.** Captured flower arrangement in a rounded amber glass vase on a wooden tabletop.

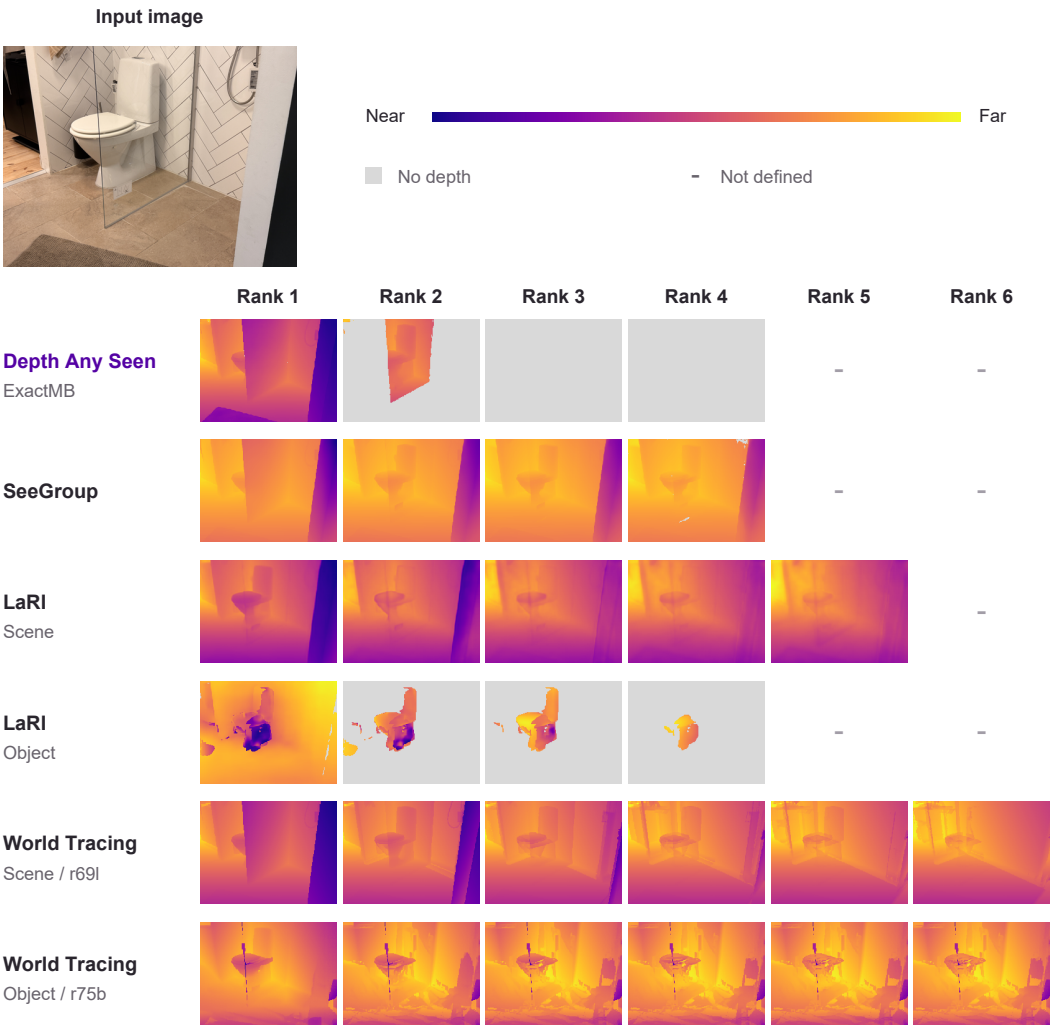


Figure S19: **Qualitative out-of-distribution multilayer depth comparisons.** Captured bathroom interior with a glass shower screen in front of a toilet and tiled wall.

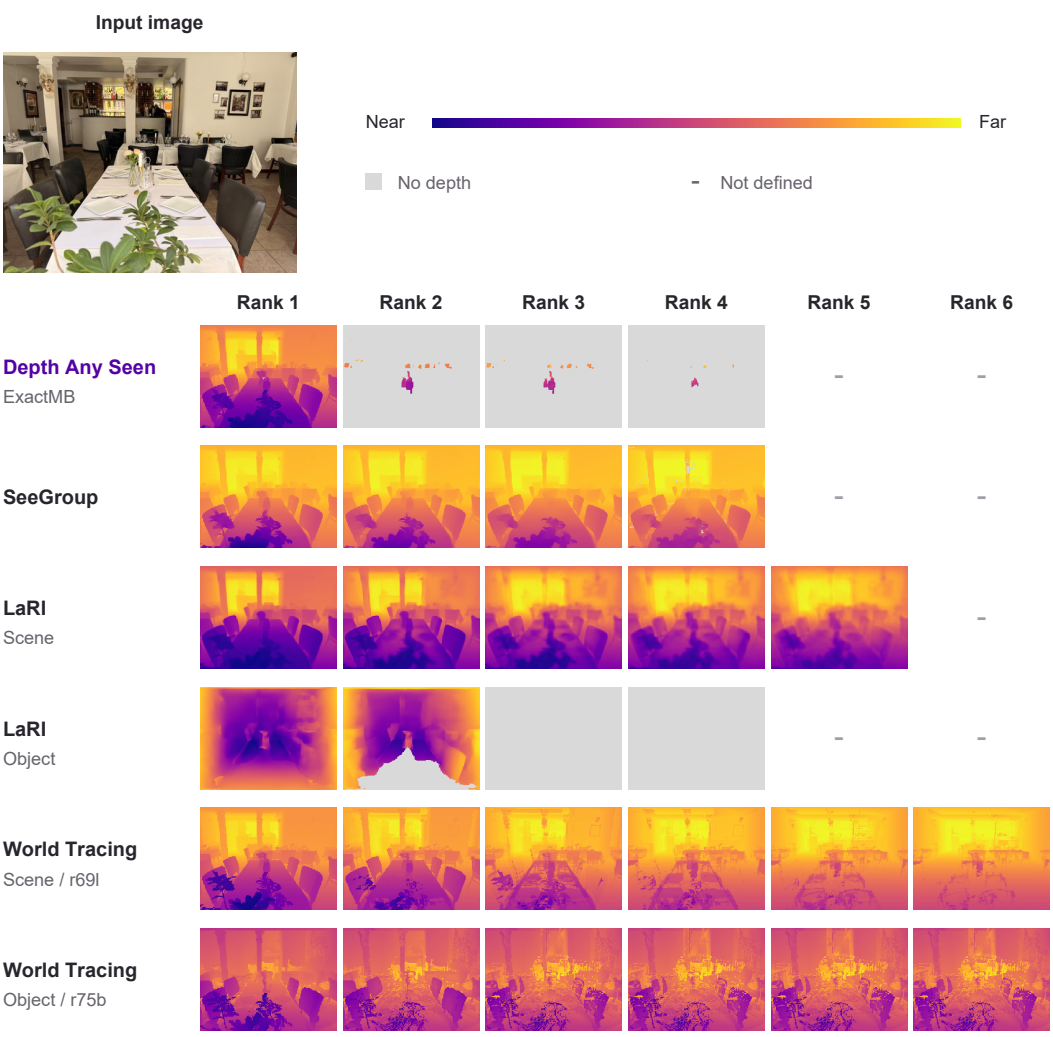


Figure S20: **Qualitative out-of-distribution multilayer depth comparisons.** Captured restaurant interior with glassware along a dining table and furnishings behind it.

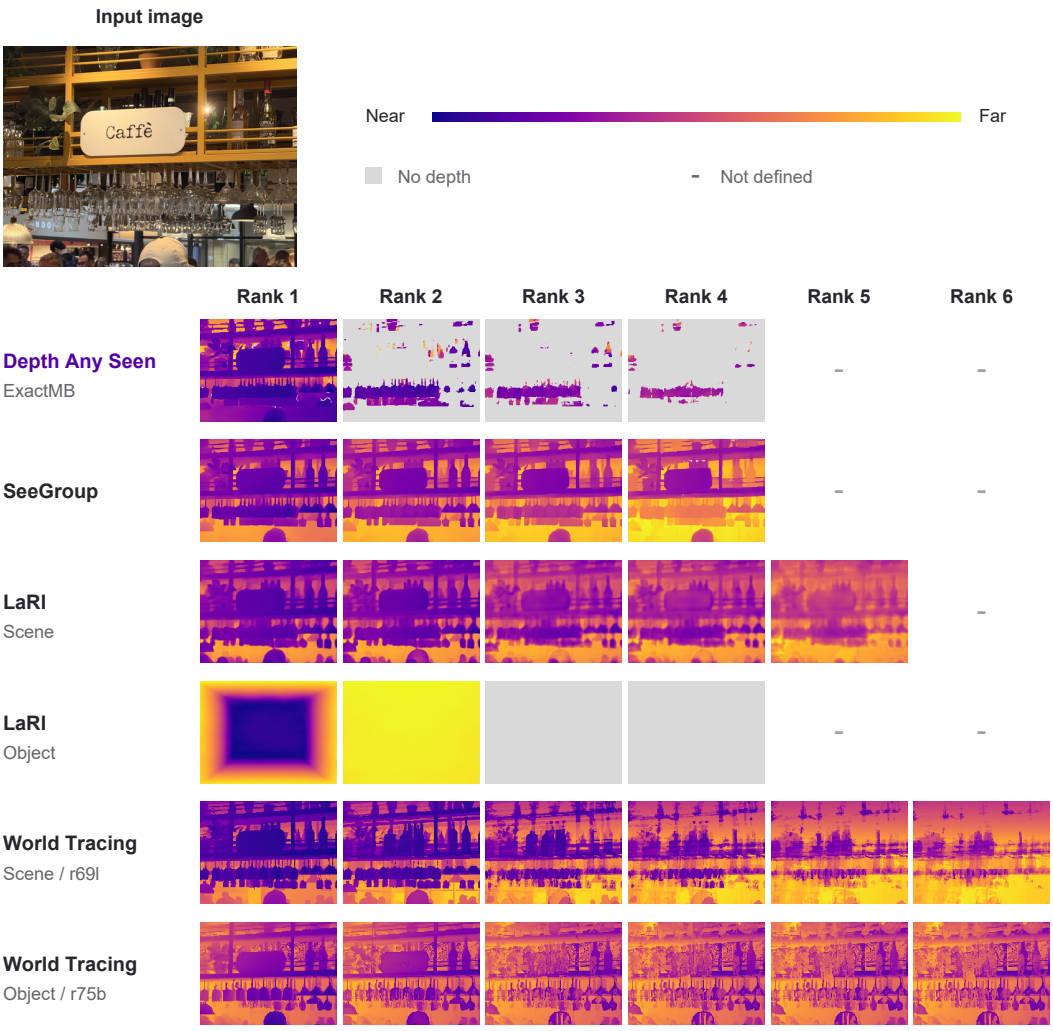


Figure S21: **Qualitative out-of-distribution multilayer depth comparisons.** Captured cafe interior with hanging glassware, bottles, and shelves at different depths.

D Learning Dynamics and Geometric Recovery

The aggregate gradient in Equation (7) links ExactMB to expected count, but emitted cardinality and geometric recovery also depend on individual presence scores and candidate depths. The primary four-seed histories track expected-count fitting and discrete activation. A separate auxiliary-regularized trajectory follows uncertainty and geometry on fixed rays, distinguishing missing candidate depths from losses during decoding. A third study examines cardinality weighting and retained support. The latter two studies provide descriptive evidence beyond the primary replications.

D.1 Count, Activation, and Held-Out Learning

We first compare the primary MAP, ExactMB, and ordered-assignment histories, using seeds $7/42/61/123$ through update 800,000. Training diagnostics average minibatches, while synthetic validation uses 64 fixed images and weights contributing images equally within each target cardinality. We measure raw activation at $q > .5$, compared with $.01$ in the principal decoder. Numerical summaries use unsmoothed records without selecting maxima, with sample SD across training seeds.

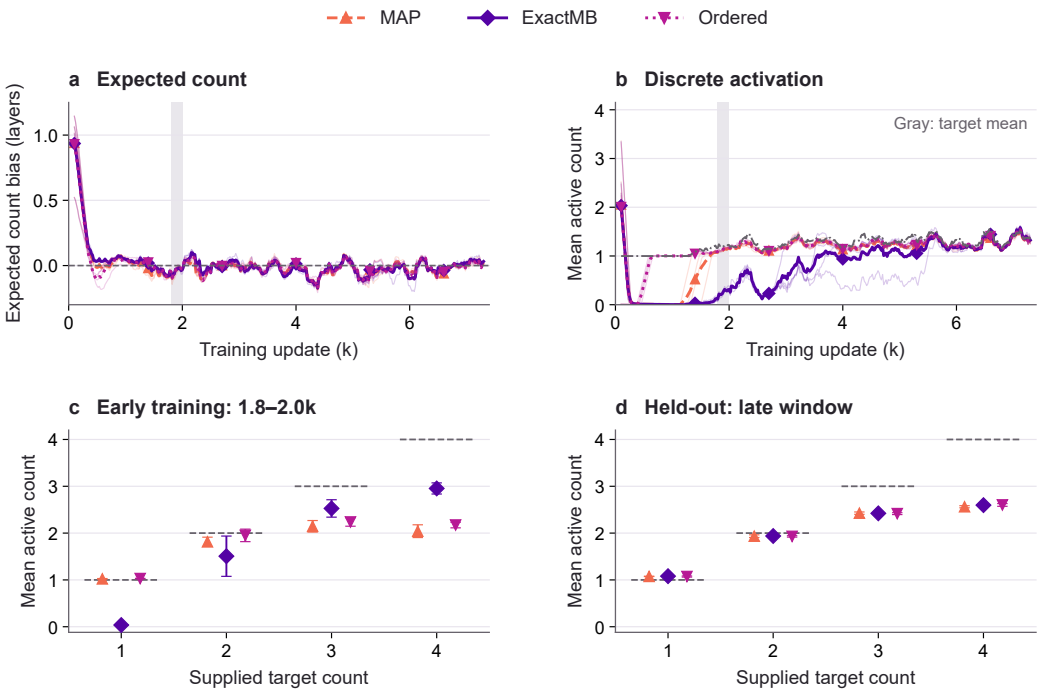


Figure S22: **Expected count and discrete activation follow different trajectories.** Learning histories for primary MAP, ExactMB, and ordered assignment under the protocol in Section D.1. (a) Minibatch expected-count bias. (b) Minibatch thresholded count at $q > .5$. Thin curves show individual seeds and bold curves their means, smoothed over a centered 200-update window for display. Shading marks the raw 1.8–2.0k summary window. (c) Conditional counts in that window, weighted by cardinality-specific pixel exposure. (d) Held-out conditional counts, averaging seven checkpoints from 762.2k to 800k on 64 fixed images. Symbols and whiskers show seed mean $\pm$ sample SD. Dashed marks denote target counts. Diagnostic and principal gates are .5 and .01.

Expected count and discrete activation. The expected count is $\bar{c} = \sum_j q_j$, whereas the thresholded count is $\hat{c}_{.5} = \sum_j \mathbf{1}[q_j > .5]$. The summed gradient in (7) supplies an expected-count signal without determining the occupied subset. For example, four probabilities of .25 have expected count one but activate none at .5. Even expected-count agreement in aggregate can hide canceling ray-wise errors. Figure S22 compares expected and thresholded counts, both overall and by target cardinality.

This distinction is visible early in training. Over raw updates 1.8–2.0k, the target mean is $1.200 \pm .027$ layers. MAP, ExactMB, and ordered assignment predict expected means $1.167 \pm .012$, $1.175 \pm .020$, and $1.174 \pm .021$, but thresholded means $1.126 \pm .012$, $.274 \pm .047$, and $1.145 \pm .022$. ExactMB's activation delay is cardinality-selective: one-layer pixels supply 88.7% of early exposure, and ExactMB activates $.037 \pm .010$ components at $c = 1$, versus MAP's $1.023 \pm .007$ and ordered assignment's $1.029 \pm .005$. At $c = 4$, ExactMB instead activates $2.954 \pm .119$, versus $2.041 \pm .136$ and $2.172 \pm .060$, and is higher in every seed. These activation differences occur despite similar expected counts. Early same-ray diagnostics and neural subset/correspondence controls would help test the mechanism behind this delay and any relation to the symmetric saddle.

Held-out high-cardinality deficit. By termination, the largest activation deficit occurs at high cardinality. Over the seven-checkpoint terminal window, held-out counts at $q > .5$ average 1.07–1.08, 1.93–1.94, 2.42–2.43, and 2.56–2.60 layers for $c = 1/2/3/4$. The $c = 4$ stratum occurs in 53 of 64 images but only 1.42% of pixels, so all-pixel averages dilute a roughly 1.4-layer conditional deficit. To examine these learning signals, we compare costs near 125k and termination: training uses complete 7,400-update cycles and excludes the single-layer-dominated final partial cycle, while validation averages three baseline checkpoints and seven checkpoints in the terminal window.

We probe all three variants with a temperature-one injection posterior and occupied-existence cost $C_{\mathrm{occ}} = -\sum_j \rho_j \log q_j$. Only ExactMB uses this posterior in training; MAP and ordered assignment use their own assignment rules. In all four seeds, each variant reduces this diagnostic cost on training batches but increases it on held-out data. For ExactMB, the paired changes are $-.0365 \pm .0037$ and $+.0681 \pm .0352$ nats/pixel. Interpretation must account for differences in cropping, model mode, composition, averaging, and posterior weights. Training is pixel-exposure-weighted, while validation weights contributing images equally within each cardinality. Matched training–validation analysis would help relate these contextual differences to the held-out high-cardinality deficit.

D.2 Fixed-Ray Uncertainty and Geometric Recovery

Count histories alone do not show whether the proposed depths explain the target surfaces. We therefore track assignment uncertainty and geometric recovery in an ExactMB run (seed 6) with an assignment-aligned gradient auxiliary of weight .05. Native 432×768 predictions cover 20 evenly spaced validation images at nine epochs $22/40/61/82/100/121/142/160/181$ (162.8k–1,339.4k updates). This separate auxiliary-regularized trajectory complements the primary four-seed study with a view of late-stage refinement. A fixed random seed samples up to 64 rays per cardinality and image from first-stage target masks. Sampled coordinates and supplied target depths remain fixed.

Excluding 64 rays with no valid targets and 185 occupied rays with exact ties leaves 4,513 positive, finite, distinct-target rays. The $M = 1/2/3/4$ supports are $1{,}280/1{,}216/1{,}152/865$ rays from $20/19/19/18$ images. Tie exclusion removes 184 of 1,049 sampled four-target rays (17.54%). One three-target image retains only one ray. We first average within each image/cardinality group, then equally across contributing images, with variability reported across images. Without a complete-ray flag, recovery is evaluated against supplied targets of unknown completeness.

Assignment uncertainty. For each supplied distinct target $\mathcal{T}$ of size M, we evaluate the injection posterior ω_ϕ from Equation (4). Let Φ denote the random assignment and $S = \mathrm{im}(\Phi)$ its occupied subset. Using (S8), the subset posterior is

$$\pi_S = \sum_{\phi:\mathrm{im}(\phi)=S} \omega_\phi = \frac{w_q(S) G_S(\mathcal{T})}{\sum_{S'} w_q(S') G_{S'}(\mathcal{T})}. \tag{S26}$$

Geometry reweights each subset's prior. Within a fixed subset, presence factors cancel, leaving correspondence dependent on localization densities. Shannon entropy with natural logarithms separates uncertainty about which components are occupied from uncertainty about their target correspondence:

$$H(\Phi \mid \mathcal{T}) = H(S \mid \mathcal{T}) + \mathbb{E}_{S|\mathcal{T}} H(\Phi \mid S, \mathcal{T}). \tag{S27}$$

The subset and correspondence terms have maxima $\log \binom{K}{M}$ and $\log M!$, respectively. We normalize each term per ray by its own maximum, assigning zero to a singleton event space, then average equally across contributing images within each cardinality. Missing strata are excluded. The separately normalized terms need not sum to the normalized total assignment entropy.

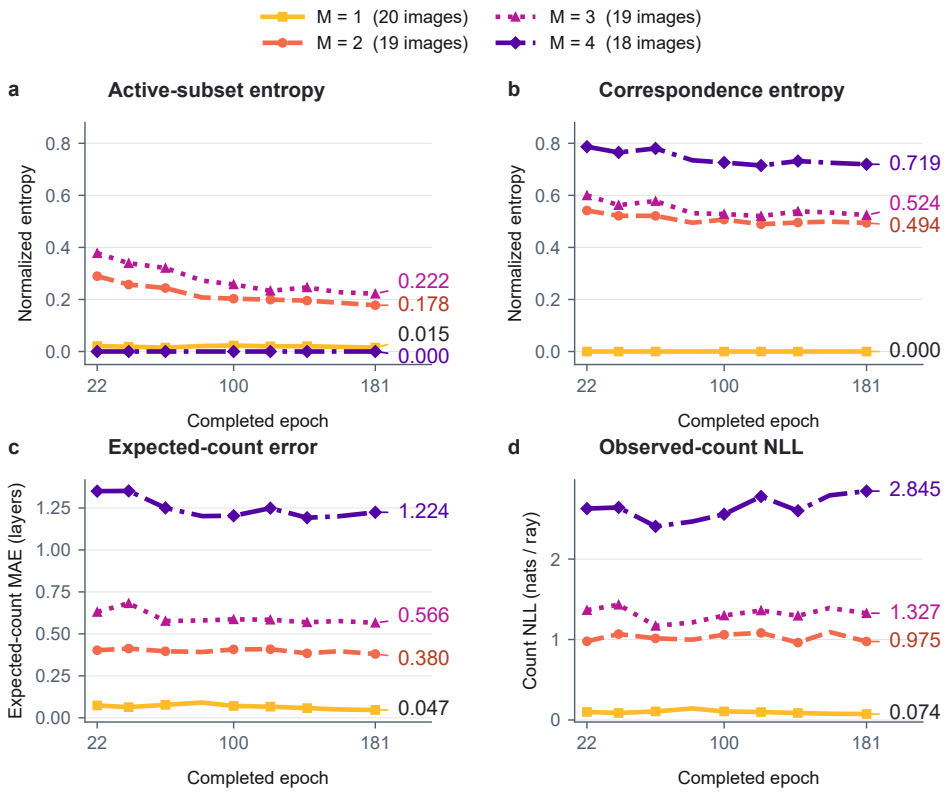


Figure S23: **Lower assignment uncertainty need not improve the count log score.** An auxiliary-regularized ExactMB trajectory (seed 6, weight .05), separate from the primary study, evaluated on the fixed rays with distinct supplied targets in Section D.2. (a) Active-subset entropy. (b) Conditional-correspondence entropy. Both are normalized per ray by their respective combinatorial maxima. Subset entropy at $M = 4$ and correspondence entropy at $M = 1$ are zero by construction. (c) Absolute expected-count error. (d) Pre-decoder observed-count NLL. Lines connect nine measured stages without smoothing. Means weight images equally within cardinality. All stages follow the same sampled rays and supplied targets within this single training run.

Assignment uncertainty and count scores. Figure S23 tracks these normalized entropies alongside expected-count error and count NLL. From first to last stage, subset entropy falls .289 $\rightarrow$ .178 at $M = 2$ and .378 $\rightarrow$ .222 at $M = 3$, while correspondence entropy changes .542 $\rightarrow$ .494 and .600 $\rightarrow$ .524. At $M = 4$, subset entropy is identically zero, but correspondence remains .787 $\rightarrow$ .719. Lower entropy describes more concentrated assignments, but count likelihood need not improve in parallel. For $M = 4$, expected-count MAE falls $1.350 \rightarrow 1.224$ layers while count NLL rises $2.627 \rightarrow 2.845$ nats/ray. Paired changes are $-.126 \pm .415$ layers and $+.217 \pm 1.575$ nats. At $M = K$, these scores are $\sum_j (1 - q_j)$ and $-\sum_j \log q_j$: their different penalties permit opposite trends. The image-level dispersion describes variation within this single training run. Geometric evaluation is also needed because correspondence depends on the learned localization scales.

These entropy trends are insensitive to the numerical treatment of saturated probabilities. Without saved logits, sigmoid endpoints are moved to the nearest interior float32 value. Clipping instead to $[10^{-6}, 1 - 10^{-6}]$ changes image-equal mean absolute subset and correspondence entropies by at most 7.52×10^{-7} and 1.42×10^{-9} nats, respectively. No sampled observed-count probability is zero.

Aggregate summaries obscure multilayer uncertainty. To assess the effect of population averaging, we extend the analysis beyond the sampled distinct-target rays above to all 6,635,520 pixels in the same 20 archived rasters. Only 8.67% have at least two supplied targets. From epochs 22 to

181, the fraction of pixel–slot probabilities with $q_j \le .05$ or $q_j \ge .95$ rises $89.90 \to 93.30\%$ overall, versus $39.41 \to 60.41\%$ on multilayer rays. Pooling obscures uncertainty on multilayer rays.

Localization scales show a related aggregation effect. For posterior-co-occupied slot pairs, let the multiplicative scale gap be $G_\beta = \exp(\mathbb{E}[|\log \beta_j - \log \beta_k|])$, with expectations weighted by joint posterior occupancy across pixels and pairs. This gap narrows $1.513 \to 1.311$, while the fraction of pair weight with both scales at most .101 m rises $22.34 \to 53.79\%$. The model's lower bound is .1 m. Conditioning instead on at least one scale above .101 m gives gaps $1.703 \to 1.795$, without comparable narrowing. Pair statistics draw on 19 images with multilayer targets. Thus, aggregate narrowing accompanies concentration near the scale floor, rather than comparable narrowing among pairs with larger scales. Empirical calibration and convergence call for separate diagnostics.

D.2.1 Candidate Geometry Versus Selection and Filtering

We next connect these uncertainty summaries to geometric recovery by separating the depths the network proposes from those the decoder retains. Let D_{all} contain all finite native centers and D_{emit} the thresholded, depth-filtered, last-retained-gap output, preserving native component indices. At tolerance ϵ, $m_\epsilon(\mathcal{T}, D)$ is the maximum number of one-to-one matches with residual at most ϵ. One candidate cannot recover multiple targets. On occupied rays,

$$\begin{aligned} R_{\text{proposal}} &= m_\epsilon(\mathcal{T}, D_{\text{all}})/M, \\ R_{\text{emit}} &= m_\epsilon(\mathcal{T}, D_{\text{emit}})/M, \\ N_{\text{extra}} &= |D_{\text{emit}}| - m_\epsilon(\mathcal{T}, D_{\text{emit}}). \end{aligned} \tag{S28}$$

Complete geometric recovery requires both the correct output count and a distinct match for every target:

$$\text{SetOK}_\epsilon = \mathbf{1}[|D_{\text{emit}}| = M \ \wedge \ m_\epsilon(\mathcal{T}, D_{\text{emit}}) = M]. \tag{S29}$$

Proposal recall measures how many targets the native-center pool can explain before selection, using ground-truth one-to-one matching. Unmatched emissions include both redundancy and mislocalization, so their interpretation differs from annotation-defined OverPred. These same-ray diagnostics use absolute tolerance, whereas the primary rank-wise scores use relative error on a different cohort.

We fix $\epsilon = .05$ m before inspection and test $.02/.10$ m sensitivity. Decoding uses strict $q > .01$, depth $> .02$ m, and a strict .02 m last-retained gap. Each decoding stage deletes candidates without moving centers. Differences in recall between successive stages measure losses at the existence gate, depth floor, and gap filter. Together with emitted recall and the fraction of targets missed by all candidates, these losses sum to one. Figure S24 accounts for this budget on the same fixed rays.

Missing candidate geometry dominates the late four-target deficit. At the last stage of this auxiliary-regularized run, four-target proposal recall at .05 m is 29.83% and emitted recall 27.63%. All-candidate matching misses 70.17% of target returns, versus .224 percentage points lost at the gate, 1.984 at the gap rule, and none at the depth floor. The dominant bottleneck is missing accurate candidates, rather than loss during .01 gating or subsequent gap filtering.

Proposal recall improves by 2.95 ± 12.59 points and emitted recall by 1.54 ± 10.90 points across the 18 paired four-target images. Proposal/emit recall increase in $13/12$ images, decline in two images each, and tie in the others. Meanwhile gap-stage loss rises $.80 \to 1.98$ points and unmatched emissions fall $2.487 \to 2.313$ per ray. Better candidate recovery and additional pruning therefore coexist, even though the number of emitted depths decreases. At $M = 1$, proposal recall instead falls $38.67 \to 34.38\%$ while emitted recall changes $25.23 \to 25.94\%$: the aggregate gate-loss budget shrinks even though all-candidate matching recovers a smaller share of targets.

Count agreement and geometric recovery can move in opposite directions. On four-target rays, exact output count falls $64.37 \to 52.38\%$, but $\text{SetOK}_{.05}$ rises $.121 \to .686\%$. Although geometric recovery remains rare, its paired mean change is $+.565 \pm 1.824$ points, with three positive images, one negative, and 14 ties. These opposing trends motivate evaluating count agreement together with geometry, and highlight reliable four-surface recovery as an open challenge. At .02 m, emitted recall changes $13.44 \to 13.88\%$ and complete geometric recovery remains zero. At .10 m they change $41.14 \to 42.57\%$ and $4.42 \to 5.35\%$. Absolute recovery levels depend on geometric tolerance, but exact count and simultaneous geometric recovery remain different evaluation events.

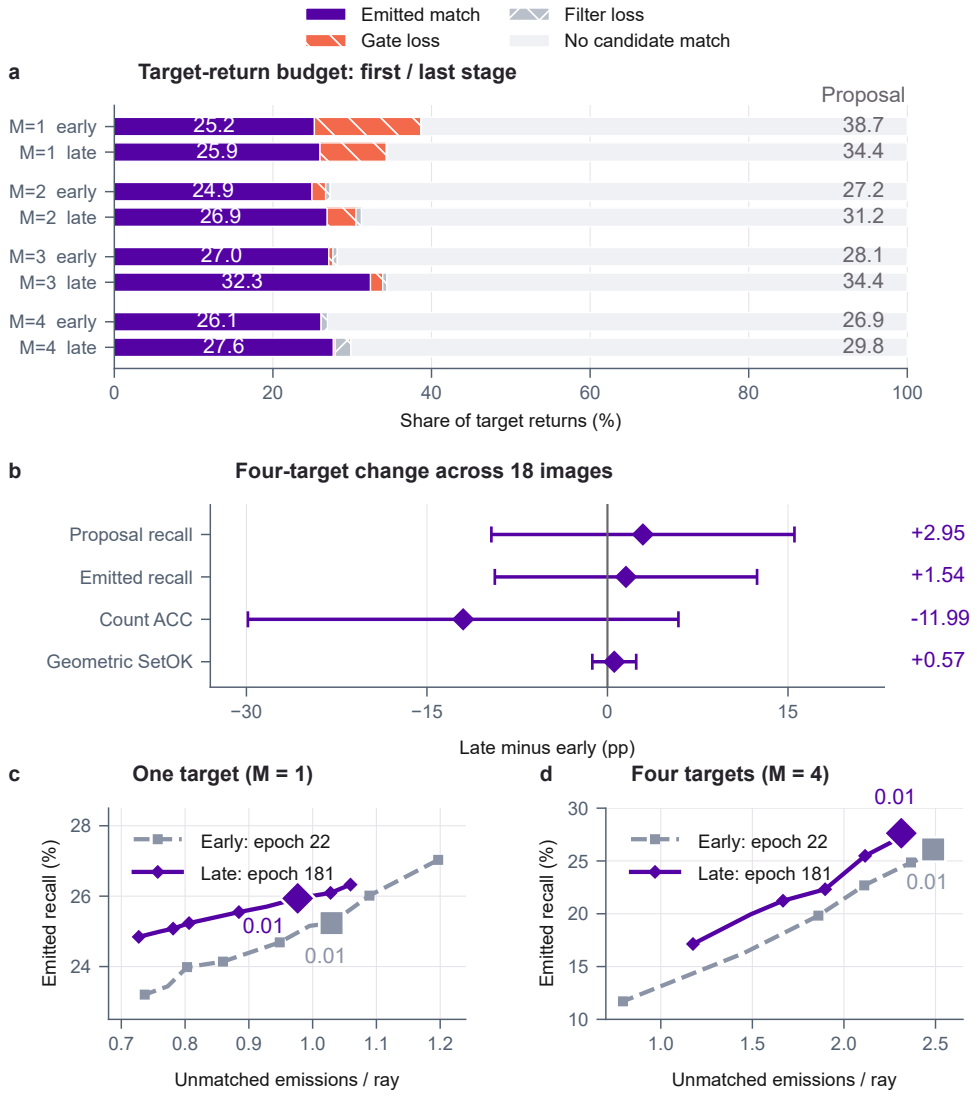


Figure S24: **Candidate geometry, emitted support, and count accuracy evolve differently on the same rays.** ExactMB (seed 6) with auxiliary regularization, matching tolerance $\epsilon = .05$ m, and the fixed rays with distinct supplied targets defined in Section D.2. (a) First/last archived stages (epochs 22/181). Stacked bars account for the full target-return budget: emitted matches, gate losses, depth-floor/gap losses, and the deficit of all-candidate one-to-one matching. Numbers inside solid emitted-match segments give emitted recall. The right column gives proposal recall. All quantities are averaged equally across images using identical rays. (b) Within-image last-minus-first changes for four-target rays. Symbols and whiskers show mean $\pm$ sample SD across 18 images rather than training seeds. (c) First/last threshold curves for $M = 1$. (d) First/last threshold curves for $M = 4$. Only the existence threshold varies across nine fixed values from .001 to .9. Large symbols mark $\tau = .01$. Curves connect recall and unmatched-emission measurements across thresholds.

For targets separated by more than $\max(2\epsilon, .02\,\mathrm{m})$, the .05 m four-target sensitivity retains only 137 rays in ten images. The .10 m sensitivity has just ten rays in three images. These small strata motivate broader sampling to compare closely spaced and well-separated target surfaces.

To distinguish geometric learning from score changes, note that with fixed centers, a common strictly increasing score transformation preserves all subsets obtainable by thresholding. Changes in proposal

recall in Figure S24 therefore reflect evolving candidate geometry beyond score remapping. These curves characterize score and geometry changes, motivating targeted causal and calibration analyses. The principal threshold remains fixed throughout the analysis.

D.3 Cardinality Weighting and Retained Support

We finally examine population weighting and fallback in aggregate evaluation. This separate study compares 16 configuration families: 14 with seeds 7/61 and two with seed 61 only. Latest-800k summaries average seeds within families, then weight families equally. Thresholds $.01/.1/.3/.5/.7/.9$ use separate inference passes at the same checkpoints, not re-thresholding one prediction tensor or independent replications. Selected operating points characterize this split without held-out calibration.

First, cardinality weighting changes the apparent benefit of an operating point. One-layer rays contribute 87.488% of occupied pixels. Moving from the .5 to the .3 evaluation pass, micro exact-count accuracy decreases by .343 points, while cardinality-macro accuracy increases by 1.730 points and accuracy at $c = 4$ increases by 7.602 points. All 16 families share these directions. The preferred global operating point therefore depends on the cardinality weighting.

Second, shallower-depth fallback changes which predictions contribute to error. From $\tau = .01$ to .9, L7 own-slot support falls $69.31 \rightarrow 27.29\%$, while last-visible support remains $100.00 \rightarrow 99.52\%$ because an absent deep slot inherits the last surviving shallower prediction. Fallback AbsRel consequently appears to improve $17.15 \rightarrow 14.98\%$, whereas own-slot AbsRel ends slightly worse ($21.86 \rightarrow 22.05\%$). Fallback therefore entangles depth quality with selective omission, making the retained-support denominator essential to interpretation. These family-level trends describe the separate geometry study without changing the primary decoder.

E Output and Membership Conventions

Multilayer methods differ in their predicted geometry and valid-output selection. Table S9 distinguishes variable-count decoding from explicit component-wise presence and absence modeling.

Table S9: **Visible layers, depth hypotheses, and amodal intersections.** The cited variants differ in their native outputs, supervised geometry, and treatment of unused capacity.

Method	Native output	Target geometry	Count and membership
MDA (Bian et al., 2026)	K weighted depth components	Boundary hypotheses; transparent-layer extension	Categorical alternatives; sigmoid weights permit coexisting transparent depths
DepthFocus (Min et al., 2026)	One stereo depth map per scalar query	Focus-selected surface	One layer per query; no simultaneous ray-set null/count law
LayeredDepth baselines (Wen et al., 2025)	K pixel-aligned scalar depth maps	Coexisting visible layers	Fixed rank/query identity; no learned per-map null/count law
SeeGroup (Wen & Deng, 2026)	Four recurrent Laplace components	Coexisting visible layers	Variable count after validity and gap filtering; no component-wise Bernoulli empty event
World Tracing (Zhang et al., 2026)	Six front-to-back camera-space XYZ maps	Visible and generated occluded intersections	Fixed stack; final real point is forward-filled
LaRI (Li et al., 2026)	Ordered XYZ maps and stopping index	Visible, unseen, and back-facing intersections	Learned stopping gives a $0, \ldots, L$ valid prefix
Depth Any Seen: ExactMB	Metric depth centers, localization scales, and presence	Annotated coexisting visible returns	$0, \ldots, K$ optional components; Bernoulli absence and an injective likelihood

"Per ray" refers to one input pixel before multiview fusion. Fixed-capacity outputs can yield variable counts through validity masks, stopping, or repeated-point removal. ExactMB's normalized law uses untruncated densities before positive-depth filtering and deterministic decoding; ordered assignment additionally fixes correspondence. Across these comparisons, annotations define visibility, and recovery from a single image may be ambiguous.