SIM2OCC: SCALING INDOOR OCCUPANCY PRETRAINING WITH UNIFIED SYNTHETIC SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Indoor occupancy models benefit from large-scale supervised learning, but existing real-world datasets provide limited volumetric supervision and are costly to expand while maintaining consistent geometry, semantic labels, and instance identities. We introduce **Sim2Occ**, an automated framework that turns furnished virtual interiors into a unified source of synthetic occupancy supervision for pretraining and benchmarking. At its core is an entity-consistent scene field that resolves mesh ownership and completes enclosed object bodies, keeping geometry, semantic labels, and instance identities consistent within room-bounded supervision domains. Scene enrichment and room-adaptive view selection produce diverse image–occupancy pairs, with alignment checked against rendered depth. We provide nested semantic training sets of 20k, 50k, and 100k observations, alongside panoptic and embodied occupancy benchmarks. Across five baselines, increasing the synthetic training scale improves both geometric and semantic occupancy prediction, with gains of 5.32–9.37 mIoU from 20k to 100k. When fine-tuned on real-world datasets, models pretrained on our benchmark achieve semantic completion gains of up to 4.30 mIoU on NYUv2 and 3.47 on OccScanNet. Open-vocabulary occupancy also benefits, gaining 4.21 mIoU on NYUv2. Larger single-frame training sources also improve embodied occupancy prediction, while panoptic baselines demonstrate instance-aware completion from the shared annotations. Sim2Occ establishes a scalable resource for studying transferable indoor occupancy representations.

1 INTRODUCTION

Indoor occupancy learning aims to recover occupied and free space from visual observations, including regions hidden from view. This involves inferring the extent of an occluded bed, distinguishing neighboring chairs as separate instances, and integrating successive observations into a coherent room representation. Recent work has advanced monocular semantic completion (Cao & De Charette, 2022; Yu et al., 2024a; Qian et al., 2026; Zhou et al., 2026a), open-vocabulary occupancy (Vobecky et al., 2023; Zhou et al., 2026b), and embodied 3D occupancy prediction (Wu et al., 2025). These advances motivate a shared volumetric supervision source that aligns image observations with scene geometry, semantic categories, and physical instance identities. Scaling this source provides a foundation for occupancy pretraining across architectures and tasks.

Real-world datasets such as NYUv2 and OccScanNet provide established benchmarks for indoor semantic occupancy (Silberman et al., 2012; Song et al., 2017; Yu et al., 2024a). Extending real-world supervision to more scenes requires additional capture, reconstruction, and annotation (Dai et al., 2017; Yeshwanth et al., 2023; Chang et al., 2017). These demands increase when targets must cover occluded volumes, distinguish physical instances, and remain consistent across observations. Simulation offers a complementary route: a furnished scene exposes geometry, object placement, appearance, and camera calibration, allowing many supervised observations to be generated from a common source. This provides a scalable source for occupancy pretraining before fine-tuning on real-world data.

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

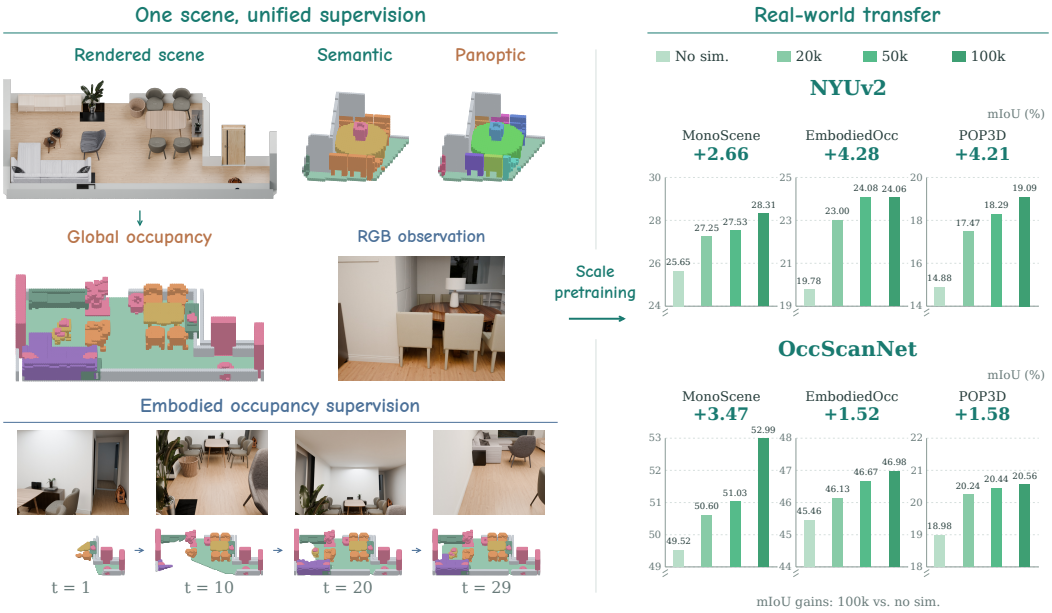

Figure 1: **From unified synthetic supervision to real-world transfer.** Left: a shared scene field provides semantic, panoptic, and embodied occupancy supervision. Right: synthetic pretraining improves mIoU after fine-tuning on NYUv2 and OccScanNet (Table 4).

Converting renderable interiors into useful occupancy supervision requires more than voxelizing their meshes. Indoor assets are organized for rendering: separate meshes for cabinet panels and drawers must be assigned to one physical entity, while independently supported accessories retain distinct identities. Closed furniture bodies require volumetric completion, whereas the spaces between chair legs and within open shelving must remain empty. Room boundaries define the completion domain: occluded objects within the observed room remain targets, while hidden geometry in neighboring rooms is excluded. Selected observations must provide informative scene coverage and maintain geometric alignment between images and occupancy targets. These coupled requirements motivate a scene-level construction process that keeps rendered appearance, occupied geometry, semantic labels, and physical identities consistent across observations.

We introduce **Sim2Occ**, an automated framework for generating indoor occupancy supervision at scale from furnished virtual interiors. Its entity-consistent scene field jointly encodes completed geometry, semantic categories, and physical instance identities. Category-compatible materials and support-constrained asset placement enrich the scenes while preserving coherent household layouts. Room-adaptive camera sampling balances spatial coverage and scene content, and calibrated depth comparisons select aligned image–occupancy pairs. From this shared field, we derive semantic and panoptic targets for individual observations and cumulative room targets for embodied occupancy, supporting pretraining and benchmarking across tasks.

We study how synthetic data scale affects occupancy learning and **real-world transfer**. The semantic benchmark provides nested 20k, 50k, and 100k training sets and a fixed 20k test set. Increasing the source from 20k to 100k improves mIoU for all five evaluated baselines, including both voxel and Gaussian representations. Synthetic pretraining improves mIoU for MonoScene, EmbodiedOcc, and the open-vocabulary model POP3D after fine-tuning on NYUv2 and OccScanNet. With pretraining and fine-tuning iterations matched across source scales, larger sources generally provide stronger initialization. Larger single-frame sources also improve embodied occupancy prediction. Panoptic evaluation demonstrates instance-level completion using the shared annotations. Figure 1 summarizes the connection from virtual scenes to scalable supervision and real-world transfer.

Our contributions are threefold:
A scalable indoor occupancy supervision framework. Sim2Occ automatically generates aligned image–occupancy pairs from furnished virtual interiors by combining structured scene composition, entity-consistent volumetric annotation, and room-adaptive observation selection.

Evidence for scalable synthetic pretraining. Larger synthetic sources improve occupancy learning across architectures and generally provide stronger initialization for real-world fine-tuning.

A unified benchmark family. Nested semantic training sets, a panoptic benchmark, and room episodes support semantic, open-vocabulary, panoptic, and embodied occupancy research through a shared annotation source.

2 Related work

Occupancy learning. Semantic scene completion predicts occupied geometry and semantic labels. SSCNet uses depth (Song et al., 2017), MonoScene and ISO use monocular RGB (Cao & De Charette, 2022; Yu et al., 2024a), and OccDepth supports stereo or RGB-D inputs (Miao et al., 2023). SplatSSC and GPOcc introduce Gaussian representations for indoor completion (Qian et al., 2026; Zhou et al., 2026a), while EmbodiedOcc integrates sequential RGB observations for embodied occupancy prediction (Wu et al., 2025). Outdoor methods explore voxel, tri-plane, and Gaussian representations (Li et al., 2023; Huang et al., 2023; 2024), supported by dense occupancy benchmarks such as Occ3D and OpenOccupancy (Tian et al., 2023; Wang et al., 2023). We study how a shared synthetic indoor source supports pretraining across occupancy architectures.

Synthetic interiors and volumetric supervision. SUNCG, Hypersim, Structured3D, InteriorNet, and 3D-FRONT provide synthetic indoor scenes and annotations (Song et al., 2017; Roberts et al., 2021; Zheng et al., 2020; Li et al., 2018; Fu et al., 2021a); Replica supplies photorealistic reconstructions of captured spaces (Straub et al., 2019). Habitat and Habitat 2.0 support embodied simulation (Savva et al., 2019; Szot et al., 2021), ProcTHOR and Infinigen Indoors generate environments procedurally (Deitke et al., 2022; Raistrick et al., 2024), and BlenderProc provides modular rendering pipelines (Denninger et al., 2019). Our construction turns furnished scene assets into occupancy supervision by resolving physical ownership, completing enclosed volumes, defining room-bounded support, and selecting aligned observations.

Open-vocabulary and panoptic scene understanding. CLIP aligns image and language representations (Radford et al., 2021); OpenScene, ConceptFusion, and OpenMask3D extend language queries to 3D points, maps, and instances (Peng et al., 2023; Jatavallabhula et al., 2023; Takmaz et al., 2023). POP3D, OVO, and LegoOcc predict language-queryable occupancy from images (Vobecky et al., 2023; Tan et al., 2023; Zhou et al., 2026b). Panoptic reconstruction jointly recovers geometry, semantic labels, and object identities in indoor scenes (Dahnert et al., 2021; Chu et al., 2023), while PaSCo and IPFormer study panoptic scene completion (Cao et al., 2024; Gross et al., 2026). CarlaOcc supplies instance-centric supervision for driving scenes (Feng et al., 2026). Our shared indoor field provides semantic and panoptic targets for individual observations and cumulative room targets for embodied occupancy.

Transferable 3D pretraining. Domain randomization supports real-world transfer through varied synthetic renderings (Tobin et al., 2017). PointContrast and DepthContrast learn transferable 3D features through self-supervision (Xie et al., 2020; Zhang et al., 2021); ULIP-2 and Uni3D scale multimodal 3D representation learning (Xue et al., 2024; Zhou et al., 2024). YouTube-Occ uses online videos for indoor occupancy pretraining (Chen et al., 2026). We study how scaling synthetic occupancy supervision affects image-to-occupancy learning and real-world transfer.

3 Sim2Occ: A scalable source of indoor occupancy supervision

Sim2Occ turns furnished virtual interiors into occupancy supervision through three automated stages: scene composition, entity-consistent volumetric annotation, and room-adaptive observation selection (Figure 2). Given an authored layout H_0 and an asset and material library $\mathcal{L}$, the scene composer $\mathcal{S}$ produces an enriched house H. The annotation compiler $\mathcal{C}$ builds its scene field $\mathcal{A}_H$, and the sampler $\mathcal{V}$ uses H and $\mathcal{A}_H$ to propose a set of camera poses $\mathcal{T}_R$ for each room R:

$$H = \mathcal{S}(H_0, \mathcal{L}), \qquad \mathcal{A}_H = \mathcal{C}(H), \qquad \mathcal{T}_R = \mathcal{V}(H, \mathcal{A}_H, R). \tag{1}$$

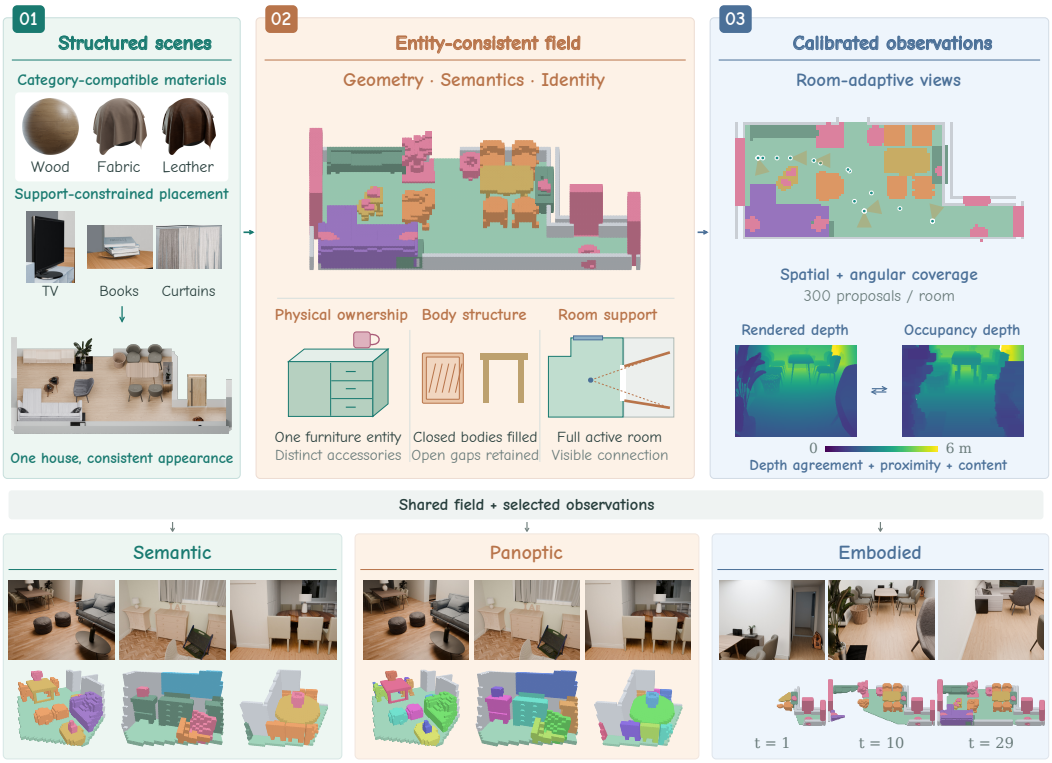

Figure 2: **The Sim2Occ construction framework.** Structured scene composition, entity-consistent annotation, and room-adaptive observation selection produce aligned RGB–occupancy pairs. A shared scene field supplies semantic and panoptic targets, together with cumulative room targets for embodied occupancy.

For each candidate pose $T \in \mathcal{T}_R$, we render RGB and depth and extract occupancy targets from $\mathcal{A}_H$ under the same camera calibration. Local targets span a $4.8 \times 4.8 \times 2.88\,\text{m}$ volume at 8 cm voxel resolution. Depth-guided quality selection retains aligned image–occupancy pairs for export. The shared field supplies consistent geometry, semantic categories, and physical identities for task-specific target extraction.

3.1 Structured Scene Composition

Category-compatible appearance. We retain the architectural layouts of 3D-FRONT and the furniture geometry of 3D-FUTURE (Fu et al., 2021a;b), and enrich appearance through a curated material library. Material assignments follow object categories and are coordinated within each house: matching furniture sets share materials, floor finishes remain coherent, and walls and ceilings use neutral finishes. Furniture colors are selected to remain distinct from nearby floors and walls, while independent accessories retain their source appearance. These house-level assignments introduce appearance variation across scenes and keep each entity visually consistent across observations.

Support-constrained content enrichment. We add televisions, doors, window treatments, and household objects to enrich categories that are sparse in the original layouts. Placement follows the geometry of supporting surfaces and openings: tabletop objects require suitable support and clearance, wall-mounted assets align with the wall and face the room interior, and doors and window treatments fit their openings. Size, orientation, and collision checks determine admissible placements, while house-level budgets limit the number of additions. Each accepted insertion is incorporated into the shared scene used for rendering and volumetric annotation, and remains fixed across observations of the house. Appendix B details the retained resources and placement rules.

3.2 ENTITY-CONSISTENT VOLUMETRIC ANNOTATION

Physical ownership. We use physical entities as the units of annotation. Source hierarchy, component categories, and geometric relations define an ownership map $\pi : \mathcal{M} \to \mathcal{E}$ from mesh components to physical entities. Structural panels of a cabinet share one identity, independently supported accessories remain separate, and repeated placements of the same CAD model receive distinct identities. Components associated with one window opening form a single window entity, while architectural trim is assigned to the wall. The geometry of entity e and the scene annotation field are represented as

$$\mathcal{G}_e = \bigcup_{a:\, \pi(a)=e} \mathcal{G}_a, \qquad \mathcal{A}_H(v) = \big(o(v), c(e(v)), i(e(v))\big), \tag{2}$$

Here $\mathcal{G}_a$ is the geometry of component a, $e(v)$ denotes the entity owning occupied voxel v, and o, c, and i denote occupancy, semantic category, and stable house-level identity, respectively. Category and identity readouts apply to occupied voxels, with free space retained within the valid supervision domain. Original category metadata is preserved for task-specific taxonomy mappings.

Topology-conditioned completion. Volumetric completion follows each entity's structure. Window assemblies are consolidated into continuous, wall-aligned volumes bounded by their openings. Enclosed furniture bodies, such as mattresses, sofas, and closed storage, receive conservative interior completion, while gaps beneath tables and chairs and within open shelving are preserved. Each added voxel inherits its owner's category and identity, keeping completed geometry consistent with semantic and instance supervision.

Room-bounded support. The primary room's non-convex footprint and architectural boundaries define the completion domain. Occluded objects within this domain remain targets, while hidden geometry in neighboring rooms is excluded. Ordinary framed doorways preserve room separation. Across broad, unframed connections, the primary room retains its full completion target and the connected region contributes depth-supported visible surfaces. This preserves visible continuity between connected spaces, such as living and dining areas.

3.3 ROOM-ADAPTIVE OBSERVATION AND QUALITY SELECTION

Room-adaptive view proposals. Camera sampling follows each room's navigable space and scene content. Candidate positions cover central and peripheral regions, with viewing directions guided by furniture, architectural, and room-context anchors. We combine central exploration, moderate-tilt context, and downward context pools in a fixed ratio, with camera positions and viewing directions adapted to each room's geometry.

Coverage-balanced selection. Within each pool, greedy selection favors new position–orientation combinations and underrepresented targets. Given the selected pose set $\mathcal{P}$, a candidate pose T receives the score

$$s(T \mid \mathcal{P}) = D_\phi(T, \mathcal{P}) + \frac{\lambda_c}{1 + n_c(T, \mathcal{P})} + \frac{\lambda_p}{1 + n_p(T, \mathcal{P})} + q_R(T, \mathcal{P}), \tag{3}$$

Here D_ϕ is the capped distance to the nearest selected pose in position–orientation feature space. The counts n_c and n_p record how often the candidate's anchor category and position have been selected, reducing repeated coverage, while q_R balances regional coverage and scene content. We retain 300 candidate poses per room for rendering and quality evaluation. The same poses are used for RGB and depth rendering and occupancy target extraction.

Depth-guided quality selection. At each pixel u, we compare rendered depth $d_r(u)$ with occupancy first-hit depth $d_o(u)$ under the same camera calibration:

$$r(u) = d_o(u) - d_r(u). \tag{4}$$

Negative residuals indicate occupancy ahead of the rendered surface, while positive residuals indicate occupancy behind it. Missing intersections indicate absent occupancy support along rays

Table 1: **Indoor occupancy benchmarks.** Counts denote individual observations or room/scene episodes × views. Global target extents vary by scene.

Dataset	Domain	Observations	Local	Embodied	Instance
Local volume: $4.8 \times 4.8 \times 2.88$ m Voxel resolution: 8 cm					
NYUv2 (Silberman et al., 2012; Song et al., 2017)	Real	1,449	✓	×	×
OccScanNet (Yu et al., 2024a)	Real	65,519	✓	×	×
EmbodiedOcc-ScanNet (Wu et al., 2025)	Real	674×30	✓	✓	×
Sim2Occ Semantic	Synthetic	120,000	✓	×	×
Sim2Occ Panoptic	Synthetic	4,905	✓	×	✓
Sim2Occ Embodied	Synthetic	500×30	✓	✓	×

with valid rendered depth. View selection combines these checks with the extent of connected mismatch regions, depth validity, camera proximity, and scene content. Room-level ranking then favors interiors with consistently usable views and varied spatial content before task-specific curation. Appendix C specifies the sampling and quality criteria.

3.4 Unified benchmarks and training scales

All benchmarks derive their targets from the shared scene field $\mathcal{A}_H$. Within a shared valid domain, fine-category, coarse-category, and panoptic targets have the same occupied support. For a taxonomy map m,

$$Y_c(v) = m(Y_f(v)), \qquad \mathrm{supp}(Y_f) = \mathrm{supp}(Y_c) = \mathrm{supp}(Y_p). \tag{5}$$

Here Y_f, Y_c, and Y_p denote fine-category, coarse-category, and panoptic targets, and support denotes occupied voxels within the valid domain. Changing semantic granularity preserves occupied geometry, while physical instance identities remain fixed in the shared field.

Semantic and open-vocabulary occupancy. The semantic benchmark provides nested training sets of 20k, 50k, and 100k observations with a shared test set. All scales use 11 occupied categories plus free space. Retained entity categories support finer taxonomies, while the same geometry can be paired with image–language features for open-vocabulary occupancy.

Panoptic occupancy. Thing voxels inherit physical instance identities, while walls, floors, and ceilings are stuff. Curation prioritizes coherent entities, repeated instances, and diverse object configurations, selecting views in which individual instances are clearly distinguishable.

Embodied occupancy. Following EmbodiedOcc (Wu et al., 2025), embodied occupancy prediction updates scene-level occupancy from sequential RGB observations and camera poses. Each episode pairs 30 ordered views with a global room target. Let $\Omega(I_j, T_j)$ denote the supported region of observation j. At step t, we form the cumulative target over the union of observation supports:

$$\Omega_t = \bigcup_{j=1}^{t} \Omega(I_j, T_j), \qquad Y_t = \mathcal{A}_H\big|_{\Omega_t}. \tag{6}$$

Evaluation uses the accumulated support at the end of each episode.

House-level train/test separation is maintained across all benchmarks and pretraining-to-evaluation pipelines. Table 1 summarizes benchmark sizes and task coverage.

4 Experiments

We evaluate how scaling synthetic occupancy supervision affects occupancy learning and real-world transfer across architectures. We also examine how the shared source supports open-vocabulary, embodied, and panoptic occupancy.

4.1 EVALUATION SETUP

We evaluate MonoScene (Cao & De Charette, 2022), EmbodiedOcc (Wu et al., 2025), SplatSSC (Qian et al., 2026), and the DPT variant of GPOcc (Zhou et al., 2026a) for semantic occupancy, and adapt POP3D (Vobecky et al., 2023) for open-vocabulary occupancy. Synthetic experiments use nested 20k, 50k, and 100k training sets with a common 20k test set. For real-world transfer, we fine-tune MonoScene, EmbodiedOcc, and POP3D on NYUv2 and OccScanNet from either synthetic or standard initialization. Across source scales, pretraining iterations are matched per method and fine-tuning iterations per method–target-dataset pair. SC IoU measures occupied geometry, and SSC mIoU averages class-wise IoU over the 11 occupied categories.

Label quality. Table 2 compares local semantic continuity and depth alignment at 8 cm resolution. S_{sc} measures same-class support in six-connected neighborhoods, A_{12} first-surface alignment within 12 cm, and F_{24} the fraction of evaluated rays with occupancy more than 24 cm ahead of the reference surface. Sim2Occ achieves the lowest F_{24} at 9.08%, less than half that of either real dataset. Its A_{12} lies between NYUv2 and OccScanNet, exceeding the latter by 11.16. All three datasets exhibit high local semantic continuity.

Table 2: **Occupancy annotation quality (%).** S_{sc}: local semantic continuity; A_{12}: first-surface alignment; F_{24}: premature occupancy.

Dataset	$S_{\text{sc}}\uparrow$	$A_{12}\uparrow$	$F_{24}\downarrow$
NYUv2	**99.9957**	**52.53**	23.70
OccScanNet	99.9315	38.11	21.22
Sim2Occ	99.9744	49.27	**9.08**

Table 3: **Scaling synthetic occupancy learning.** Nested 20k, 50k, and 100k training splits share a fixed 20k test set. POP3D uses open-vocabulary decoding without 3D semantic category labels. Bold marks the best scale within each method; tint identifies 100k. Scores are percentages.

Method	Split	SC IoU	Ceiling	Floor	Wall	Window	Chair	Bed	Sofa	Table	TV	Furniture	Objects	mIoU
MonoScene	20k	48.72	34.3	84.7	26.5	29.0	20.6	57.0	48.8	20.4	19.9	35.5	22.0	36.24
	50k	53.78	38.2	86.8	32.6	33.8	25.3	61.7	53.1	28.2	26.8	42.5	27.0	41.44
	100k	**58.15**	**42.9**	**88.2**	**36.4**	**35.5**	**29.2**	**67.4**	**58.3**	**31.9**	**31.4**	**49.9**	**30.5**	**45.61**
EmbodiedOcc	20k	36.22	20.5	70.0	12.4	12.6	16.2	49.6	39.0	17.1	17.0	31.1	17.4	27.53
	50k	40.35	21.8	76.7	14.3	15.0	19.6	57.0	46.4	21.3	18.4	36.4	20.5	31.58
	100k	**44.54**	**24.7**	**80.5**	**17.3**	**17.3**	**23.2**	**61.6**	**53.5**	**25.4**	**21.2**	**41.4**	**24.7**	**35.54**
SplatSSC	20k	37.42	21.5	67.3	9.7	10.7	13.1	50.8	35.4	16.6	15.2	30.6	14.3	25.92
	50k	40.59	**25.4**	71.2	12.4	15.9	15.9	54.4	41.5	20.5	19.5	34.3	17.1	29.83
	100k	**43.55**	24.8	**75.0**	**14.1**	**18.3**	**19.4**	**59.8**	**47.7**	**23.4**	**22.0**	**39.6**	**20.3**	**33.12**
GPOcc	20k	35.50	24.7	35.5	13.0	15.1	15.1	46.7	40.1	19.9	20.7	34.0	18.2	25.72
	50k	39.49	27.2	43.7	14.1	16.4	19.5	55.1	46.2	23.9	24.2	40.5	21.6	30.21
	100k	**42.48**	**30.3**	**47.0**	**17.1**	**20.6**	**22.7**	**58.3**	**50.4**	**26.2**	**26.8**	**42.6**	**24.5**	**33.32**
POP3D	20k	47.13	22.6	58.3	16.8	18.2	7.4	35.9	27.6	12.1	10.1	18.8	0.1	20.73
	50k	52.00	27.3	**61.0**	21.2	22.2	8.0	41.7	34.2	**14.4**	17.1	21.2	0.4	24.43
	100k	**55.63**	**28.5**	58.7	**23.5**	**23.8**	**9.6**	**45.6**	**37.4**	14.2	**22.4**	**22.1**	**0.6**	**26.05**

4.2 SCALING SYNTHETIC OCCUPANCY LEARNING

Both SC IoU and mIoU increase consistently across the three training scales for all five baselines (Table 3). Scaling the source from 20k to 100k yields gains of 5.32–9.37 mIoU. This trend spans voxel and Gaussian representations, showing that larger synthetic sources benefit multiple occupancy architectures.

For MonoScene, gains extend across all 11 occupied categories. Scaling from 20k to 100k yields gains of 11.5 IoU for both table and television and 14.4 IoU for furniture. SC IoU also rises from 48.72 to 58.15, indicating improvements in occupied geometry alongside category-level semantics. Together, the results establish a clear empirical scale trend for learning from Sim2Occ.

Table 4: **Synthetic pretraining for real-world transfer.** Models are fine-tuned after standard initialization (No sim.) or synthetic pretraining. Δ denotes absolute gains over No sim. Bold marks the best score per method and dataset; shaded rows mark 100k.

		NYUv2				OccScanNet			
Method	Source split	SC IoU	Δ SC IoU	mIoU	Δ mIoU	SC IoU	Δ SC IoU	mIoU	Δ mIoU
MonoScene	No sim.	41.58	—	25.65	—	59.75	—	49.52	—
	20k	42.39	+0.81	27.25	+1.60	60.48	+0.73	50.60	+1.08
	50k	43.98	+2.40	27.53	+1.88	60.88	+1.13	51.03	+1.51
	100k	**44.54**	+2.96	**28.31**	+2.66	**62.44**	+2.69	**52.99**	+3.47
EmbodiedOcc	No sim.	30.68	—	19.78	—	53.76	—	45.46	—
	20k	36.81	+6.13	23.00	+3.22	54.55	+0.79	46.13	+0.67
	50k	38.25	+7.57	**24.08**	+4.30	55.11	+1.35	46.67	+1.21
	100k	**39.24**	+8.56	24.06	+4.28	**55.39**	+1.63	**46.98**	+1.52
POP3D	No sim.	40.64	—	14.88	—	60.57	—	18.98	—
	20k	42.18	+1.54	17.47	+2.59	63.65	+3.08	20.24	+1.26
	50k	44.46	+3.82	18.29	+3.41	63.80	+3.23	20.44	+1.46
	100k	**45.04**	+4.40	**19.09**	+4.21	**63.97**	+3.40	**20.56**	+1.58

4.3 Scaling synthetic pretraining for real-world transfer

Synthetic pretraining improves both SC IoU and mIoU for all three methods on NYUv2 and OccScanNet after fine-tuning (Table 4). The gains hold at every evaluated source scale. Peak gains over standard initialization reach 4.30 mIoU for EmbodiedOcc on NYUv2 and 3.47 mIoU for MonoScene on OccScanNet. These gains demonstrate that pretraining on Sim2Occ provides effective initialization for real-world occupancy learning across methods and target datasets.

With pretraining and fine-tuning iteration budgets each matched across source scales, larger synthetic sources generally yield stronger real-world transfer. For MonoScene on OccScanNet, the gain over standard initialization grows from 1.08 mIoU with 20k pretraining observations to 3.47 mIoU with 100k. EmbodiedOcc also improves with scale on OccScanNet; on NYUv2, its mIoU remains similar at 50k and 100k, while SC IoU continues to increase. The same positive scale trend holds for POP3D on both datasets. Together, these results demonstrate the value of scaling synthetic supervision for real-world occupancy learning.

4.4 Open-vocabulary occupancy

The pretraining benefits extend beyond fixed-category semantic prediction. POP3D learns occupied geometry and language-aligned voxel features without 3D semantic category labels and uses text queries for semantic prediction. Scaling the synthetic source from 20k to 100k yields gains of 5.32 mIoU and 8.50 SC IoU (Table 3). After real-world fine-tuning, 100k synthetic pretraining yields gains of 4.21 mIoU on NYUv2 and 1.58 mIoU on OccScanNet over standard initialization (Table 4). Quantitative evaluation uses the shared 11-class query set, while qualitative examples in the appendix illustrate additional concepts such as pillows and curtains. The same synthetic occupancy source therefore supports both semantic and open-vocabulary pretraining.

4.5 From single-frame pretraining to embodied occupancy

We evaluate embodied occupancy on Sim2Occ room episodes following the EmbodiedOcc-ScanNet format. EmbodiedOcc updates learned global memory, while GPOcc uses incremental fusion; both start from their corresponding single-frame pretrained models. For both methods, SC IoU and mIoU increase consistently across the three source scales (Table 5). From 20k to 100k, mIoU rises from 34.45 to 37.19 for EmbodiedOcc and from 36.85 to 40.67 for GPOcc. Larger synthetic sources therefore improve accumulated room-level completion as well as single-frame occupancy.

Table 5: **Embodied occupancy.** SC IoU/mIoU (%). Source denotes the single-frame pretraining set size.

Method	Source	SC IoU	mIoU
EmbodiedOcc	20k	50.19	34.45
	50k	51.33	35.77
	100k	52.21	37.19
GPOcc	20k	49.75	36.85
	50k	50.46	38.03
	100k	52.61	40.67

Table 6: **Training strategies on NYUv2 with EmbodiedOcc.** Bold marks the best result for each metric.

Method	Training strategy	SC IoU↑	mIoU↑
Baseline	Standard	30.68	19.78
HASSC	Aux. loss	30.42	19.54
GaussRender	Aux. loss	30.96	19.22
Pri3D	Pre-training	29.01	17.94
YouTube-Occ	Pre-training	31.85	20.18
Sim2Occ-20k	Pre-training	36.81	23.00
Sim2Occ-50k		38.25	**24.08**
Sim2Occ-100k		**39.24**	24.06

4.6 Effectiveness of synthetic pretraining

Table 6 compares auxiliary-loss and pretraining strategies (Wang et al., 2024; Chambon et al., 2025; Hou et al., 2021) for EmbodiedOcc on NYUv2. Sim2Occ pretraining yields gains of 6.13–8.56 SC IoU and 3.22–4.30 mIoU over standard initialization. Even 20k observations outperform the compared strategies on both metrics, demonstrating effective initialization for real-world occupancy learning.

Table 7: **Indoor panoptic completion (%).** Panoptic metrics use PaSCo-compatible aggregation. Bold marks the best result in each column.

	All				Things			Stuff			Completion	
Method	PQ†	PQ	SQ	RQ	PQ	SQ	RQ	PQ	SQ	RQ	SC IoU	mIoU
Panoptic-FlashOCC	13.67	11.69	48.96	17.23	6.93	50.17	11.59	24.40	45.75	32.26	43.54	26.60
SparseOcc	19.49	16.10	54.99	23.05	12.06	58.39	19.43	26.89	45.92	32.72	49.88	33.63
MonoScene (Query)	24.43	21.59	59.57	30.35	16.07	56.13	24.67	36.30	68.74	45.49	52.69	34.14
MonoScene (C/O)	**25.34**	**22.94**	**67.50**	**31.69**	**16.49**	**65.26**	**24.69**	**40.15**	**73.48**	**50.36**	**59.75**	**41.22**

4.7 Indoor panoptic completion

We evaluate four models: MonoScene with center–offset (C/O) and PaSCo-style query heads (Cao et al., 2024), SparseOcc (Liu et al., 2024), and Panoptic-FlashOCC (Yu et al., 2024b). Checkpoints are selected using the same house-disjoint validation protocol. Table 7 reports Panoptic Quality (PQ), Segmentation Quality (SQ), and Recognition Quality (RQ) overall and separately for things and stuff, together with PQ†, SC IoU, and mIoU.

The center–offset head achieves the highest scores across all reported metrics, including 22.94 PQ and 41.22 mIoU. For thing classes, it achieves 65.26 SQ compared with 56.13 for the query head, while RQ is similar (24.69 vs. 24.67), indicating higher geometric overlap for matched instances. It also achieves higher stuff PQ (40.15 vs. 36.30). The shared annotation field therefore supports comparison of instance grouping and scene completion across panoptic occupancy architectures under a common indoor protocol.

5 Conclusion

We introduced Sim2Occ, an automated framework and benchmark family for indoor occupancy learning. Its entity-consistent scene field unifies geometric, semantic, and instance supervision from furnished virtual interiors. Larger synthetic sources improve occupancy learning across architectures and generally provide stronger initialization for fine-tuning on NYUv2 and OccScanNet. These scale benefits extend to open-vocabulary and embodied occupancy, while panoptic evaluation demonstrates instance-level completion from the shared annotations. Sim2Occ thus provides a scalable resource for learning transferable indoor occupancy representations.

REPRODUCIBILITY STATEMENT

The appendix details annotation conventions, dataset splits, baseline adaptations, and evaluation protocols, and reports per-class results. Dataset manifests record camera calibration, source entities, and split membership. We will release code, dataset documentation, and benchmark configurations to support reproducibility.

ETHICS STATEMENT

Sim2Occ is generated from authored virtual interiors without recording identifiable occupants. It primarily covers furnished residential environments. Use in other indoor settings or safety-critical systems requires evaluation of scene coverage and prediction reliability.

STATEMENT OF AI ASSISTANCE

AI-assisted tools were used to improve the grammar, clarity, and readability of the manuscript. They also supported code development. The authors take full responsibility for the scientific content and final manuscript.

REFERENCES

Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3981–3991. IEEE, 2022.

Anh-Quan Cao, Angela Dai, and Raoul De Charette. Pasco: Urban 3d panoptic scene completion with uncertainty awareness. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14554–14564. IEEE, 2024.

Loick Chambon, Eloi Zablocki, Alexandre Boulch, Mickael Chen, and Matthieu Cord. Gaussrender: Learning 3d occupancy with gaussian rendering. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27010–27020. IEEE, 2025.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

Haoming Chen, Lichen Yuan, Tianfang Sun, Jingyu Gong, Zhizhong Zhang, Xin Tan, Yanyun Qu, and Yuan Xie. Youtube-occ: Learning indoor 3d semantic occupancy prediction from youtube videos. In *European Conference on Computer Vision*, pp. 470–489. Springer, 2026.

Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12472–12482. IEEE, 2020.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1280–1289. IEEE, 2022.

Tao Chu, Pan Zhang, Qiong Liu, and Jiaqi Wang. Buol: A bottom-up framework with occupancy-aware lifting for panoptic 3d scene reconstruction from a single image. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4937–4946. IEEE, 2023.

Manuel Dahnert, Ji Hou, Matthias Nießner, and Angela Dai. Panoptic 3d scene reconstruction from a single rgb image. *Advances in Neural Information Processing Systems*, 34:8282–8293, 2021.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2432–2443. IEEE, 2017.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. procthor: Large-scale embodied ai using procedural generation. *Advances in neural information processing systems*, 35:5982–5994, 2022.

Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Youssef Zidan, Dmitry Olefir, Mohamad Elbadrawy, Ahsan Lodhi, and Harinandan Katam. Blenderproc. *arXiv preprint arXiv:1911.01911*, 2019.

Yi Feng, Zizhan Guo, Yu Ma, Hanli Wang, Rui Fan, et al. An instance-centric panoptic occupancy prediction benchmark for autonomous driving. *arXiv preprint arXiv:2603.27238*, 2026.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10913–10922. IEEE, 2021a.

Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129 (12):3313–3337, 2021b.

Markus Gross, Aya Fahmy, Danit Niwattananan, Dominik Muhle, Rui Song, Daniel Cremers, and Henri Meeß. Ipformer: Visual 3d panoptic scene completion with context-adaptive instance proposals. *Advances in Neural Information Processing Systems*, 38:4989–5014, 2026.

Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning? In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5673–5682. IEEE, 2021.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9223–9232. IEEE, 2023.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. In *European Conference on Computer Vision*, pp. 376–393. Springer, 2024.

Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, et al. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*, 2023.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9396–9405. Ieee, 2019.

Wenbin Li, Sajad Saeedi, John McCormac, Ronald Clark, Dimos Tzoumanikas, Qing Ye, Yuzhong Huang, Rui Tang, and Stefan Leutenegger. Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. *arXiv preprint arXiv:1809.00716*, 2018.

Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9087–9098. IEEE, 2023.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Haisong Liu, Yang Chen, Haiguang Wang, Zetong Yang, Tianyu Li, Jia Zeng, Li Chen, Hongyang Li, and Limin Wang. Fully sparse 3d occupancy prediction. In *European Conference on Computer Vision*, pp. 54–71. Springer, 2024.

Ruihang Miao, Weizhou Liu, Mingrui Chen, Zheng Gong, Weixin Xu, Chen Hu, and Shuchang Zhou. Occdepth: A depth-aware method for 3d semantic scene completion. *arXiv preprint arXiv:2302.13540*, 2023.

Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–824. IEEE, 2023.

Rui Qian, Haozhi Cao, Tianchen Deng, Shenghai Yuan, and Lihua Xie. Splatssc: Decoupled depth-guided gaussian splatting for semantic scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8520–8528, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, et al. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21783–21794. IEEE, 2024.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10892–10902. IEEE, 2021.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9338–9346. IEEE, 2019.

Yuheng Shi, Minjing Dong, and Chang Xu. Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23487–23497. IEEE, 2025.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 190–198. IEEE, 2017.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in neural information processing systems*, 34:251–266, 2021.

Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023.

Zhiyu Tan, Zichao Dong, Cheng Zhang, Weikun Zhang, Hang Ji, and Hao Li. Ovo: Open-vocabulary occupancy. *arXiv preprint arXiv:2305.16133*, 2023.

Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in neural information processing systems*, 36:64318–64330, 2023.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

Antonin Vobecky, Oriane Siméoni, David Hurych, Spyridon Gidaris, Andrei Bursuc, Patrick Pérez, and Josef Sivic. Pop-3d: Open-vocabulary 3d occupancy prediction from images. *Advances in Neural Information Processing Systems*, 36:50545–50557, 2023.

Song Wang, Jiawei Yu, Wentong Li, Wenyu Liu, Xiaolu Liu, Junbo Chen, and Jianke Zhu. Not all voxels are equal: Hardness-aware semantic scene completion with self-distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14792–14801. IEEE, 2024.

Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17804–17813. IEEE, 2023.

Yuqi Wu, Wenzhao Zheng, Sicheng Zuo, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Embodiedocc: Embodied 3d occupancy prediction for vision-based online scene understanding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26360–26370. IEEE, 2025.

Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *European conference on computer vision*, pp. 574–591. Springer, 2020.

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27081–27091. IEEE, 2024.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12–22. IEEE, 2023.

Hongxiao Yu, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Monocular occupancy prediction for scalable indoor scenes. In *European Conference on Computer Vision*, pp. 38–54. Springer, 2024a.

Zichen Yu, Changyong Shu, Qianpu Sun, Yifan Bian, Xiaobao Wei, Jiangyong Yu, Zongdai Liu, Dawei Yang, Hui Li, and Yan Chen. Panoptic-flashocc: An efficient baseline to marry semantic occupancy with panoptic via instance center. *arXiv preprint arXiv:2406.10527*, 2024b.

Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10232–10243. IEEE, 2021.

Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *European Conference on Computer Vision*, pp. 519–535. Springer, 2020.

Changqing Zhou, Yueru Luo, and Changhao Chen. Generalizing visual geometry priors to sparse gaussian occupancy prediction. *arXiv preprint arXiv:2602.21552*, 2026a.

Changqing Zhou, Yueru Luo, Han Zhang, Zeyu Jiang, and Changhao Chen. Monocular open vocabulary occupancy prediction for indoor scenes. *arXiv preprint arXiv:2602.22667*, 2026b.

Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *International Conference on Learning Representations*, volume 2024, pp. 46766–46782, 2024.

Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

A ANNOTATION UNITS AND SEMANTIC OWNERSHIP

Appearance assignments and physical instance identities remain fixed across observations of each house. Rooms define completion domains, and calibrated observations provide local targets. Semantic and panoptic targets share occupied geometry within the same valid domain, while embodied occupancy targets accumulate observation support across views. All local targets span a $4.8 \times 4.8 \times 2.88$ m volume at 8 cm voxel resolution.

Source mesh components are grouped into physical entities, whose categories are then mapped to the evaluation taxonomy (Table S1). Repeated placements of the same chair model receive distinct instance identities, whereas structural components of one cabinet share an identity. Changing semantic granularity preserves these entity boundaries and instance identities.

Table S1: **Semantic classes and annotation units.** The 11 occupied evaluation classes comprise three stuff and eight thing categories.

Evaluation class	Panoptic type	Annotation unit
Ceiling, floor, wall	Stuff	Architectural surfaces; wall trim is assigned to wall.
Window	Thing	One window assembly per opening, including frame, glazing, and window treatments.
Chair, bed, sofa	Thing	One placed entity, including its structural components.
Table	Thing	Body and legs; independent tabletop items remain separate.
Television	Thing	One television entity; independent computers belong to objects.
Furniture	Thing	One storage or furniture entity.
Objects	Thing	Independent items, including computers, mirrors, and doors.

B SCENE ENRICHMENT AND ENTITY-CONSISTENT OCCUPANCY

3D-FRONT layouts and 3D-FUTURE geometry (Fu et al., 2021a;b) are compiled into a shared metric scene. Placement identifiers link rendered meshes to semantic and instance records. Material assignments and inserted assets remain fixed across observations of each house. Each accepted insertion is incorporated into the shared scene used for rendering and volumetric annotation.

B.1 CATEGORY-COMPATIBLE MATERIALS

The retained library contains 100 category–material pairings drawn from 70 distinct source materials (Table S2). Candidates are reviewed both as source previews and on scene geometry under indoor lighting. Figures S1 and S2 show representative resources and their rendered appearance.

Table S2: **Retained material library.** Counts describe available assignments, not occurrence in the benchmark.

Application	Assignments	Appearance families
Floor	11	Residential wood, laminate, and parquet.
Chair / stool	11	Neutral textiles and leather.
Bed	20	White, beige, gray, and muted blue upholstery.
Sofa	18	Neutral woven fabrics and leather.
Table	20	Wood veneers and matte finishes.
Storage furniture	20	Wood, white, and gray finishes.
General pools	100	70 distinct source materials.
Functional-room floor	3	Tile and terrazzo, managed separately.
Wall / ceiling	1 policy	Smooth near-white appearance.

General floors share a house-level finish, while functional rooms use a designated hard-surface pool. Matching furniture sets share materials, and furniture colors are selected to remain visually distinct from nearby walls and floors. Independent accessories retain their source appearance, preserving distinctions such as books on a tabletop. Components with ambiguous material roles retain their original appearance. Texture scale remains consistent in world units.



Figure S1: **Representative materials by application.** Source IDs identify texture resources; PH and ACG denote Poly Haven and ambientCG. Table S2 gives the complete pool sizes.

B.2 Supported assets and fitted openings

Supplementary assets enrich categories that are sparse in the original layouts. The retained pool contains 11 televisions, seven doors, and 63 household objects across nine kinds; window treatments use a separate opening-level policy (Table S3). Figures S3 and S4 pair representative assets with in-room renders.

Table S3: **Supplementary assets and placement constraints.** Placement depends on support, clearance, and orientation.

Asset kind	Entries	Placement and category
Television	11	Furniture or wall; television category.
Residential door	7	Fitted doorway; objects category.
Bag / basket	6 / 6	Furniture or floor support; distinct objects.
Books / bottle / box	12 / 5 / 5	Unoccupied tabletop or shelf.
Computer	4	Desk or low furniture; objects category.
Lamp	9	Compatible support; source appearance retained.
Picture	13	Wall-attached, image-facing side toward the room.
Indoor bin	3	Floor support at residential scale.
Curtain	11	Reviewed BlenderKit assets fitted to openings.
Glass-window candidate	2	Reviewed assets fitted to openings.
Blind / glass construction	Existing	Retained fitted constructions.

Inserted objects require exposed support at a suitable height, a contained footprint, collision clearance, and an inward-facing orientation. Existing accessories take precedence. Televisions remain smaller than their supporting furniture and parallel to it or to the mounting wall; bedroom placement favors views toward the bed. Television insertion is attempted once per eligible living/dining room and with probability 0.6 in bedrooms. Infeasible insertions are omitted, and house-level budgets prevent overcrowding.

Window treatments are assigned to eligible physical windows with target proportions of 40% curtains, 40% blinds, and 20% glass. Reviewed curtains and fitted blind or glass constructions align with the opening and supporting wall; doors likewise fit their doorway. Window targets consolidate

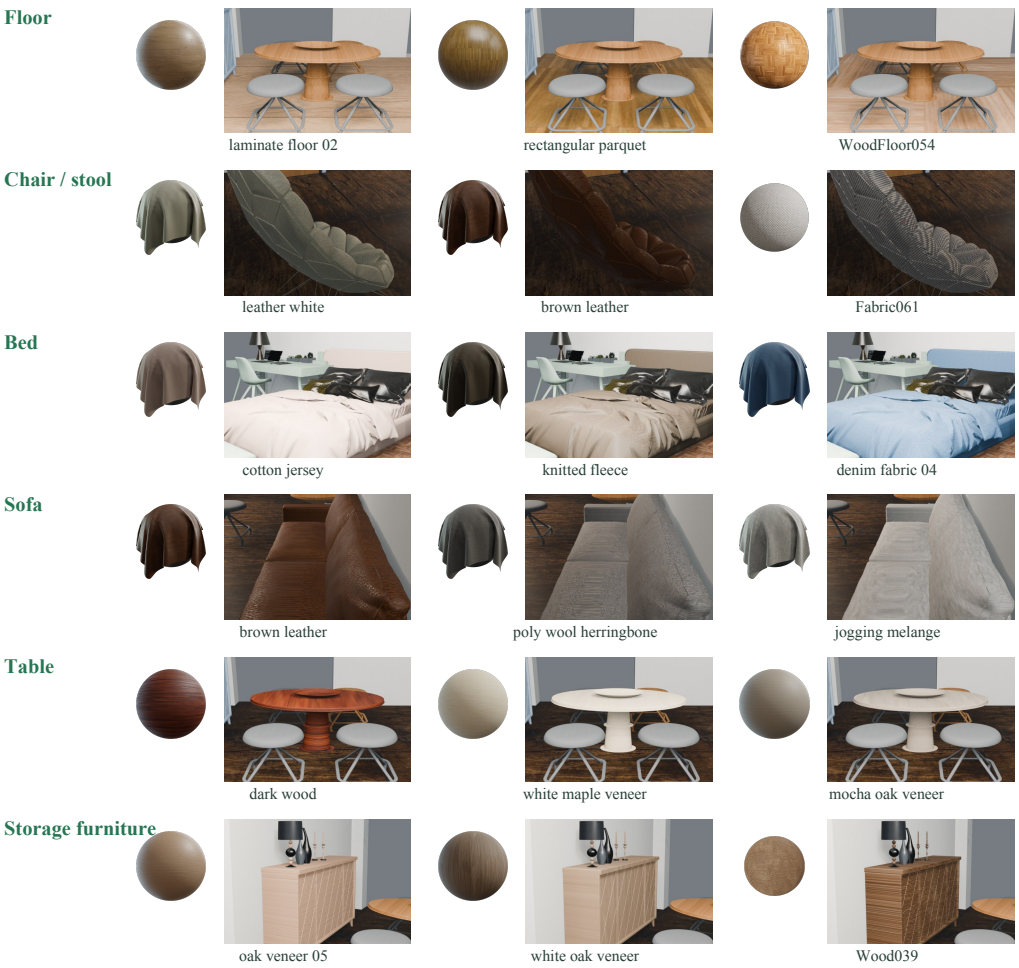

Figure S2: **Material variation on scene geometry.** Material previews are paired with in-room renders. Within each row, scene geometry and camera viewpoint remain fixed while the material varies.

deep or layered geometry into continuous, wall-aligned volumes bounded by the openings, filling gaps left by frame-only geometry. Frame, glazing, and treatment share one window category and identity.

B.3 Entity integrity and room completion support

Structural furniture components share their owner's semantic category and instance identity, while independently supported or attached items retain separate identities. Wall trim is assigned to the wall category, with its geometry preserved. Enclosed furniture bodies receive conservative interior completion, while gaps within open shelving, beneath furniture, and between legs are preserved. For surface support V_e and admissible body interior B_e,

$$V_e^\star = V_e \cup \big(\mathrm{Fill}(V_e) \cap B_e\big), \qquad (c(v), i(v)) = (c(e), i(e)) \quad \forall v \in V_e^\star. \tag{S1}$$

Here, $\mathrm{Fill}(V_e)$ generates candidate interior voxels, and B_e restricts completion according to the entity's body structure.

The primary room retains its non-convex footprint, structural boundaries, and occluded interior content. Opaque boundaries and ordinary framed doorways exclude hidden neighboring interiors.

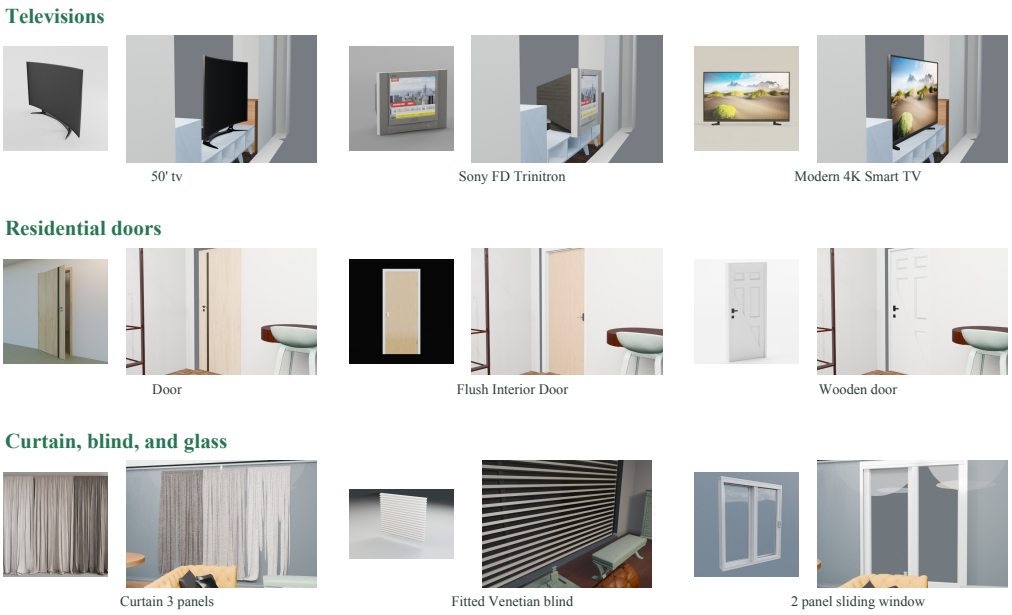

Figure S3: **Television, door, and window-treatment diversity.** Asset previews are paired with in-room renders under the support, orientation, and opening-fit constraints.

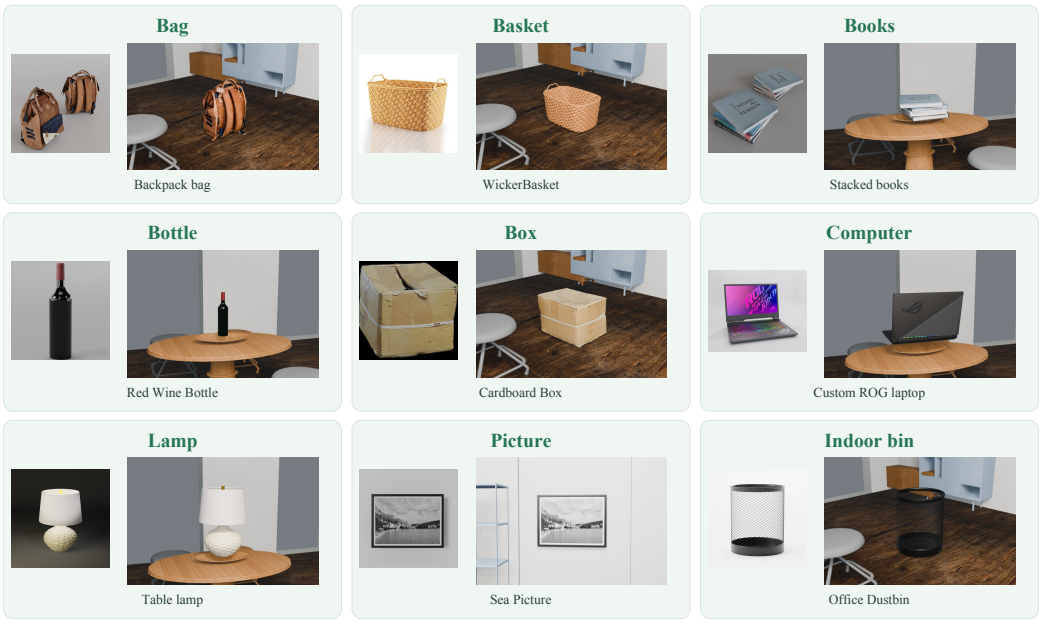

Figure S4: **Household object diversity.** Asset previews and scene renders cover bags, baskets, books, bottles, boxes, computers, lamps, pictures, and bins. Independent placements retain their appearance and receive distinct identities.

Broad unframed connections admit visible first surfaces from connected regions:

$$U(T) = U_R \cup \bigcup_{R' \in \mathcal{N}_{\text{open}}(R)} V_{R',T}, \tag{S2}$$

where U_R denotes the full completion support of the primary room, $\mathcal{N}_{\text{open}}(R)$ contains rooms connected to R through broad, unframed openings, and $V_{R',T}$ denotes depth-supported first-surface occupancy in room R' at pose T. This preserves visible continuity across connected living/dining spaces. The same support applies to coarse semantics, entity categories, and panoptic labels.

Free and invalid label encoding. The ownership expression in Equation 2 applies to occupied voxels. Valid free voxels have occupancy $o(v) = 0$, category label 0, and instance ID 0. Invalid voxels are excluded from supervision: their category sentinel is 255 in the coarse semantic readout and 65535 in the 16-bit source-category and entity-category readouts. Their exported instance ID is 0. The panoptic readout also uses instance ID 0 for stuff (wall, floor, and ceiling), and positive IDs for occupied things. Thus, the semantic validity and occupancy masks distinguish invalid voxels, free space, and stuff; instance ID 0 alone does not distinguish them. Invalid voxels have no supervised binary occupancy value.

B.4 Resource provenance and research-use permissions

Table S4 summarizes the provenance, roles, and licenses of the resources used to construct Sim2Occ. Asset records retain source identifiers, URLs, and per-asset licenses; BlenderKit records also identify the contributing creators. Generated observations and annotations remain subject to the applicable upstream terms.

Table S4: **Resource provenance and licensing for scientific research.** Source resources and author-generated benchmark outputs are listed separately.

Resource	Source and role	License and permitted scope
3D-FRONT	Tianchi / Alibaba; furnished layouts, architectural geometry, and source appearance.	3D-FRONT Use License Agreement: scientific research and distribution of research derivatives; source attribution and agreement retention required. Original dataset redistribution and commercial use are prohibited.
3D-FUTURE	Tianchi / Alibaba; furniture meshes, textures, and category metadata.	3D-FUTURE Use License Agreement: scientific research and distribution of research derivatives under the same attribution, agreement-retention, and non-commercial restrictions.
Poly Haven	Additional physically based materials and texture maps.	CC0: use, modification, and redistribution permitted, including research use.
ambientCG	Additional physically based materials and texture maps.	CC0 1.0: use, modification, and redistribution permitted, including research use.
BlenderKit	Creator-authored televisions, doors, window treatments, and household objects.	Per-asset CC0 or Royalty Free. Rendering and incorporation into projects are permitted; Royalty Free assets cannot be redistributed as standalone or extractable source assets.
Rendered RGB and depth	Generated by our pipeline from the assembled scenes and sampled cameras.	Research derivatives subject to the upstream scene and asset terms; source credits and applicable notices accompany distribution.
Occupancy labels	Generated semantic, entity-category, and panoptic targets, including embodied occupancy supervision.	Author-generated annotations subject to applicable upstream terms. Their distribution does not grant rights to redistribute the underlying source assets.

The benchmark release is scoped to derived observations, occupancy annotations, and supporting metadata for scientific research. It excludes upstream scene files, meshes, texture libraries, and Blender asset files. The 3D-FRONT and 3D-FUTURE agreements accompany research derivatives, together with source attribution; BlenderKit asset-specific terms are retained.

C OBSERVATION DESIGN AND GEOMETRIC QUALITY

C.1 COVERAGE AND USABLE VIEWS

Farthest-point sampling distributes camera positions over central and peripheral navigable regions. Furniture and structural anchors guide viewing directions. Central exploration, moderate-tilt context, and downward context form a 40/40/20 mixture, with varying height and pitch. The same selected poses are used for RGB and depth rendering and occupancy target extraction (Figure S5).

Pose features and selection score. For horizontal camera position (x, y) in meters, yaw ψ, and pitch θ, Equation 3 uses

$$\phi(T) = [x/0.45,\ y/0.45,\ 1.8\cos\psi,\ 1.8\sin\psi,\ \theta/(\pi/18)]\,, \tag{S3}$$

where angles are in radians. Height and roll vary across proposals but are omitted from this distance feature. The novelty term is

$$D_\phi(T, \mathcal{P}) = \min\left\{8, \min_{T' \in \mathcal{P}} \|\phi(T) - \phi(T')\|_2\right\}, \tag{S4}$$

with $D_\phi(T, \varnothing) = 8$. The cap is in feature-space units. We set $\lambda_c = 2.0$ and $\lambda_p = 1.5$. The counts n_c and n_p refer to previously selected poses with the same anchor category and candidate-position index, respectively, and reset for each pool. Central-exploration proposals share one anchor-category code and have distinct position indices.

The region- and content-dependent term is

$$q_R(T, \mathcal{P}) = b_R(T, \mathcal{P}) + 0.15\min\{K(T), 6\} + 0.2\min\{N_{\text{occ}}(T)/2500, 1.5\}\,, \tag{S5}$$

where $K(T)$ counts occupied categories present in the valid local target and $N_{\text{occ}}(T)$ counts its occupied voxels. For a pool selecting N poses, the central-region target is $\text{round}(0.7N)$ and the peripheral target is the remainder. The bonus b_R is 1 while the candidate's region is below its target and -4 otherwise. Hence $-4 \le q_R \le 2.2$; the region targets are soft preferences. Greedy selection maximizes the score over unselected candidates, with candidate order breaking ties. Planning-valid candidates form the pool when sufficient; otherwise, the remaining proposals are included before selection and subjected to subsequent rendered-view quality checks.

Counts through the sampling stages. The two context pools share up to 180 navigable positions, reduced when fewer cells are available. Positions use a 70/30 central/peripheral split (126/54 at 180 positions). Each context pool generates seven poses per position, yielding at most 1,260 proposals per pool. Selection retains 120 poses from the central-exploration pool, 120 moderate-tilt poses, and 60 downward-context poses. These 300 poses per room are rendered for quality evaluation.

Rendered and occupancy first-hit depths are compared in camera-axis meters. Signed residuals distinguish premature occupancy from deeper surfaces, while missing intersections identify unsupported visible regions. These checks combine with proximity, exposure, and scene content (Table S5); Figure S6 illustrates distinct rejection modes. Wall-dominated views remain eligible when other categories and useful spatial context remain visible.

Table S5: **Frame and room quality criteria.** Geometry and scene content are considered jointly.

Criterion	Evidence
Depth support	Valid depth and comparable target geometry.
Camera proximity	Near-surface prevalence, central obstruction, and shallow depth quantiles.
Surface correspondence	First-hit agreement and missing visible non-window support.
Premature occupancy	Occupancy ahead of observed surfaces and connected error extent.
Scene information	Semantic diversity, dominant planes, and depth variation.
Image exposure	Severe underexposure or overexposure.
Room reliability	Usable-view fraction, structural failures, and at least 250 eligible observations.
View refinement	Geometric quality and scene content within each room.

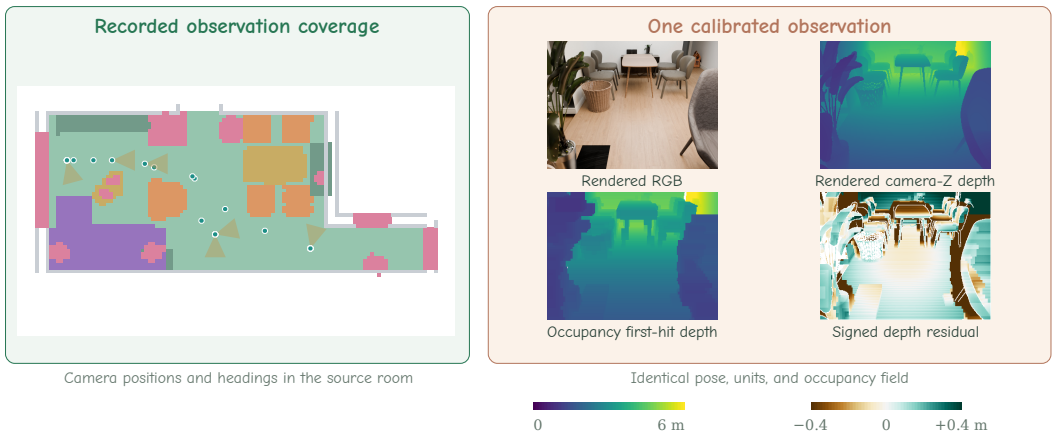

Figure S5: **Observation coverage and alignment.** Camera positions and headings are paired with RGB, rendered depth, occupancy first-hit depth, and signed residual. The two depth panels share a 0–6 m scale.

Window transparency and opening-level targets receive specific treatment during curation. The cross-dataset measurements below use one definition for all categories. Room ranking combines view reliability with room area and furniture content; the semantic benchmark retains 480 rooms with 250 eligible observations each. Final selection jointly targets source-pool counts of 100/100/50 and central/intermediate/peripheral spatial counts of 125/87/38. Minimum-cost flow selects 250 quality-approved views, balancing quota deviations, view quality, and near-duplicate penalties. The quotas are soft targets; rejected views remain excluded.

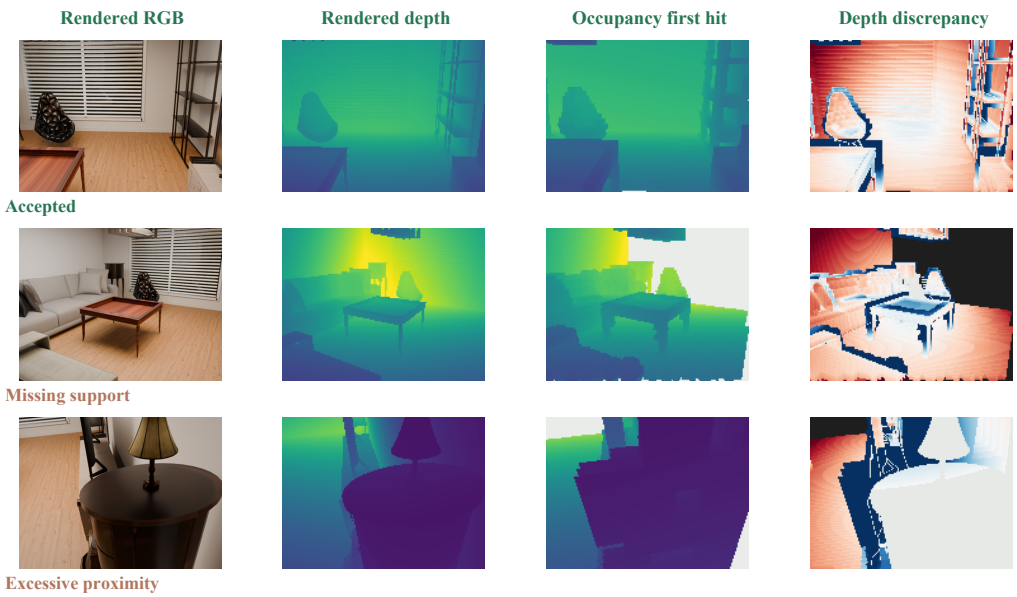

Figure S6: **Quality-selection examples.** A retained view is compared with views rejected for missing support or excessive proximity. RGB, rendered depth, first-hit depth, and residual share calibration within each row. Depth spans 0–6 m; residuals span ± 0.6 m (blue: nearer, red: deeper), with black indicating no hit and gray excluded pixels.

C.2 CROSS-DATASET QUALITY METRICS

Table 2 evaluates original NYUv2 train/test data, all published OccScanNet observations, and the 120k semantic release. For occupied voxel v and its six-connected neighborhood $\mathcal{N}_6(v)$ within valid camera support, semantic continuity is

$$S_{\mathrm{sc}} = 1 - \frac{\sum_{v \in V_{\mathrm{occ}}} \mathbf{1}[\nexists w \in \mathcal{N}_6(v) : c(w) = c(v)]}{|V_{\mathrm{occ}}|}. \tag{S6}$$

It measures the fraction of occupied voxels with at least one same-class neighbor within the valid camera field of view. Depth scores use eligible rays $\mathcal{R}$ with valid, locally stable reference depth inside the target volume. Stability requires a 3×3 depth range of at most $\max(0.12, 0.04 d_r)$ m; endpoints within 0.12 m of the volume boundary are excluded. With first-hit indicator $h(u)$ and residual $r(u)$ from Equation 4,

$$A_{12} = \frac{1}{|\mathcal{R}|} \sum_{u \in \mathcal{R}} \mathbf{1}[h(u) \wedge |r(u)| \le 0.12], \tag{S7}$$

$$F_{24} = \frac{1}{|\mathcal{R}|} \sum_{u \in \mathcal{R}} \mathbf{1}[h(u) \wedge r(u) < -0.24]. \tag{S8}$$

No-hit rays remain in the denominator. Counts are pooled across rays in common physical units. Eligible-ray coverage is 59.78% for NYUv2, 72.14% for OccScanNet, and 81.87% for Sim2Occ; these differences reflect the available reference-depth support.

D BENCHMARK CURATION AND DATASET CHARACTERISTICS

All benchmarks preserve house-level train/test separation. The semantic benchmark emphasizes reliable views, the panoptic benchmark coherent instances, and the embodied occupancy benchmark cumulative room coverage.

D.1 SEMANTIC COMPLETION AT CONTROLLED SCALES

The semantic benchmark contains 120,000 observations from 480 rooms in 454 houses. Nested training subsets $D_{20k} \subset D_{50k} \subset D_{100k}$ share one 20k test set. Tables S6 and S7 report partitions and room types. Bedrooms and living/dining rooms dominate; all retained room types meet the same quality criteria.

Table S6: **Frozen benchmark partitions.** Panoptic fitting and validation subdivide the public train/development release. The embodied benchmark groups observations into 30-view room episodes.

Benchmark	Partition	Observations	Rooms	Houses
Semantic	Train 20k	20,000	80	78
	Train 50k	50,000	200	192
	Train 100k	100,000	400	374
	Test	20,000	80	80
Panoptic	Public train/development	4,075	730	598
	Fitting	3,682	654	538
	Internal validation	393	76	60
	Test	830	134	115
Embodied	Train	12,000	400	374
	Test	3,000	100	93

The median observation contains six occupied classes. Figure S7 distinguishes category occurrence from voxel share: objects appear in 95.12% of the 100k training views but occupy 8.34% of labeled volume; televisions account for 13.45% of views and 0.56% of volume. Frequent appearance therefore need not imply abundant voxel supervision.

Table S7: **Room-type composition.** Semantic and embodied occupancy entries count rooms; panoptic entries count observations, reflecting its selection of a small number of strong views from many rooms.

	Semantic rooms		Panoptic	Embodied subset	
Room type	Train 100k	Test	All observations	Train	Test
Bedroom	182	36	2,413	228	51
Living / dining	182	36	2,275	150	44
Study	20	4	121	11	3
Circulation	8	2	33	6	2
Kitchen	8	2	0	5	0
Utility	0	0	63	0	0
Total	400	80	4,905	400	100

D.2 Panoptic selection for coherent and distinguishable entities

Panoptic observations are selected from the broader production pool, with at most 24 views per room. Selection favors intact, distinguishable entities under level or moderately downward views, reliable depth, and sufficient image support. Fragmented labels, heavy truncation, and foreground obstructions are excluded. Both repeated furniture and mixed household-object arrangements are retained; the median observation contains four readable entities, including three main-furniture instances.

D.3 Episode selection for embodied occupancy

Episodes for embodied occupancy follow EmbodiedOcc's 30-view interface (Wu et al., 2025). For occupied room reference V_R and the union of observation supports Ω_{30}, target coverage is

$$C_R = \frac{|V_R \cap \Omega_{30}|}{|V_R|}. \tag{S9}$$

This measures target support, not prediction IoU. Episodes require $C_R \geq 0.80$, floor coverage of at least 0.70, useful incremental support from at least half the views, and usable observations throughout. The 500 retained rooms have median target coverage of 89.23%/89.91% and floor coverage of 88.70%/89.44% in train/test.

E Baseline architectures and evaluation

E.1 Single-frame and embodied occupancy evaluation

Synthetic experiments use nested 20k, 50k, and 100k training sets with a common 20k test set. Real-world transfer compares standard and synthetic initialization followed by training on NYUv2 or OccScanNet. Across the three source scales, pretraining uses a fixed number of optimizer updates for each method, and fine-tuning uses a fixed number for each method–target-dataset pair. The two stages have separately matched iteration budgets, allowing source-data scale to be compared at fixed optimization lengths. POP3D learns occupied geometry and language-aligned voxel features using a frozen Trident teacher (Shi et al., 2025), without 3D semantic category supervision. Text queries decode semantics within predicted occupancy; quantitative evaluation uses the 11 occupied classes.

Embodied occupancy evaluation follows EmbodiedOcc-ScanNet's 30-view room episodes (Wu et al., 2025). EmbodiedOcc updates global memory, whereas GPOcc uses incremental fusion (Zhou et al., 2026a). Both start from their corresponding single-frame models and are evaluated over the union of the 30 observation supports.

E.2 Two panoptic extensions of MonoScene

Both heads use MonoScene's image encoder, FLoSP projection, and contextual 3D U-Net (Cao & De Charette, 2022), taking one RGB image and calibrated voxel projections. The models are trained

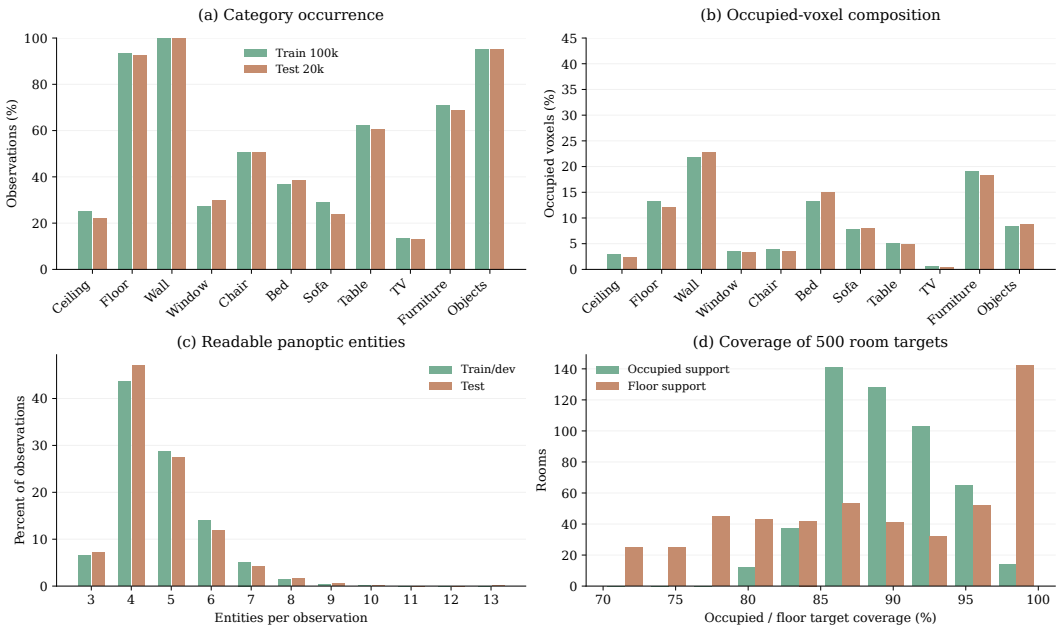

Figure S7: **Benchmark characteristics.** (a) Semantic category occurrence and (b) occupied-voxel composition in the semantic train/test sets. (c) Readable-entity counts in panoptic observations. (d) Cumulative occupied and floor coverage across the 500 rooms in the embodied benchmark. Each panel summarizes the corresponding frozen benchmark membership.

independently on the same house-disjoint partition. Figure S8 contrasts center–offset grouping with PaSCo-style query-mask decoding over the shared feature interface.

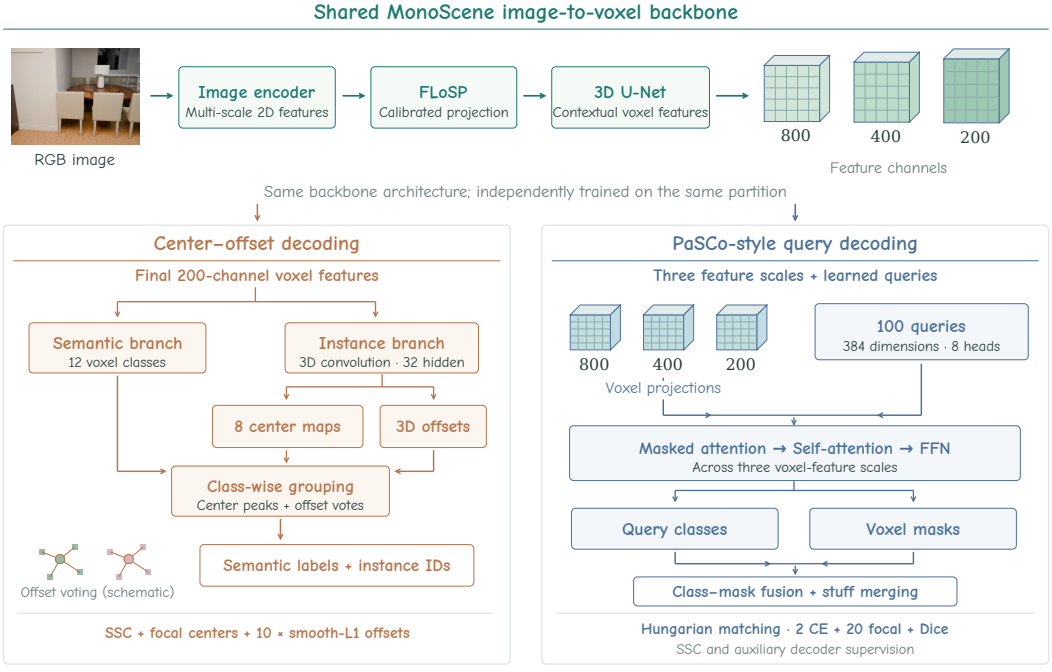

Figure S8: **Two panoptic extensions of MonoScene.** Center–offset decoding groups voxels through category-specific centers and offsets. PaSCo-style decoding uses multi-scale features and learned queries to predict classes and masks. Both heads retain semantic completion supervision.

Center–offset grouping. Following center–offset grouping (Cheng et al., 2020; Yu et al., 2024b), we retain the semantic branch and add 3D thing-category center heatmaps and voxel-to-center offsets. Each target center is the occupied voxel nearest its instance centroid. Localized Gaussian heatmaps and per-instance-normalized offsets supervise the branches:

$$\mathcal{L}_{\text{ctr}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{geo}} + \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{heat}} + 10\mathcal{L}_{\text{off}}. \tag{S10}$$

The first three terms are MonoScene's semantic and scaling objectives (Cao & De Charette, 2022); heatmaps use the Gaussian-weighted focal variant of CenterNet (Zhou et al., 2019), and offsets use smooth-L_1. At inference, thing voxels vote toward predicted centers and join the nearest center of the same category. Stuff is not subdivided into instances.

Query-mask grouping. The second head adapts PaSCo's transformer decoder (Cao et al., 2024) to dense indoor voxel features. One hundred learned queries interact with three feature scales through masked cross-attention (Cheng et al., 2022) and self-attention. Query q_j predicts a class distribution and a mask

$$M_j(v) = \sigma\big(\psi(q_j)^\top f(v)\big)\,, \tag{S11}$$

where $\psi(q_j)$ is a mask embedding and $f(v)$ a projected voxel feature. Bipartite matching uses class agreement, focal-mask cost (Lin et al., 2017), and Dice overlap. Matched queries optimize

$$\mathcal{L}_{\text{set}} = 2\mathcal{L}_{\text{query-CE}} + 20\mathcal{L}_{\text{focal}} + \mathcal{L}_{\text{Dice}}, \tag{S12}$$

together with semantic completion and intermediate prediction losses. At inference, class confidence and mask probability determine voxel ownership. Weak queries are removed, thing segments retain separate identities, and same-class stuff regions merge.

E.3 Completion and panoptic metrics

SC IoU pools occupied categories; SSC mIoU averages the IoU of the 11 occupied classes. Counts are accumulated over the evaluated split. Excluded targets are omitted, while valid free voxels penalize false occupancy.

Panoptic Quality (Kirillov et al., 2019) uses category-specific matching with voxel-set IoU strictly above 0.5. For matched segments TP_c, unmatched predictions FP_c, and unmatched targets FN_c,

$$\mathrm{PQ}_c = \frac{\sum_{(p,g)\in \mathrm{TP}_c} \mathrm{IoU}(p,g)}{|\mathrm{TP}_c| + \frac{1}{2}|\mathrm{FP}_c| + \frac{1}{2}|\mathrm{FN}_c|}, \tag{S13}$$

$$\mathrm{SQ}_c = \frac{\sum_{(p,g)\in \mathrm{TP}_c} \mathrm{IoU}(p,g)}{|\mathrm{TP}_c|}, \qquad \mathrm{RQ}_c = \frac{|\mathrm{TP}_c|}{|\mathrm{TP}_c| + \frac{1}{2}|\mathrm{FP}_c| + \frac{1}{2}|\mathrm{FN}_c|}. \tag{S14}$$

PQ combines recognition and overlap; SQ measures matched-segment overlap and RQ segment recognition. A class without a match receives zero SQ. Counts are pooled before averaging over 11 All, eight Thing, or three Stuff classes. The group mean of PQ need not equal the product of mean SQ and RQ.

We use PaSCo-compatible aggregation (Cao et al., 2024): a positive-overlap stuff prediction is exempt from false-positive counting, although a true-positive match still requires IoU above 0.5. $\mathrm{PQ}^{\dagger}$ combines ordinary thing PQ with each stuff class's mean positive-overlap segment IoU. This segment statistic differs from dataset-level semantic IoU. Evaluation and visualization share valid target support without restricting predictions to ground-truth occupied voxels.

F Additional quantitative analysis

Tables 3 and S10 report per-category results for single-frame and embodied occupancy prediction. Category gains remain uneven: POP3D's objects IoU stays below 1% across the three synthetic scales.

F.1 TRANSFER TO REAL INDOOR OBSERVATIONS

Tables S8 and S9 give complete per-class transfer results. With 100k initialization, POP3D's NYUv2 window IoU increases from 5.3 to 10.6; its OccScanNet ceiling and wall IoU rise from 8.2 to 16.6 and 28.5 to 31.1. MonoScene improves across all 11 OccScanNet categories. These profiles complement the aggregate gains in Table 4.

Table S8: **Complete NYUv2 transfer results.** Source split denotes simulation pretraining followed by real-domain fine-tuning. No sim. uses the standard initialization. Bold marks the best reported value for each method.

Method	Split	SC IoU	Ceiling	Floor	Wall	Window	Chair	Bed	Sofa	Table	TV	Furniture	Objects	mIoU
	No sim.	41.58	8.5	93.4	11.5	10.2	13.1	45.9	34.2	14.3	12.5	26.3	12.3	25.65
	20k	42.39	11.1	93.2	12.7	11.3	14.1	47.3	37.4	14.9	**16.9**	28.1	12.8	27.25
MonoScene	50k	43.98	10.6	**93.7**	13.6	12.9	15.7	48.6	37.9	15.3	12.6	28.3	**13.7**	27.53
	100k	**44.54**	**11.7**	93.6	**13.9**	**13.0**	**16.6**	**48.8**	**38.4**	**16.6**	15.3	**30.0**	13.7	**28.31**
	No sim.	30.68	5.9	71.5	5.9	10.3	9.7	38.0	24.9	9.7	15.6	16.1	10.0	19.78
	20k	36.81	6.7	79.5	6.6	10.6	11.2	42.7	32.0	11.8	**18.3**	23.4	10.0	23.00
EmbodiedOcc	50k	38.25	**7.2**	82.9	**10.1**	**11.0**	12.0	42.9	33.9	**11.9**	17.7	24.5	**10.9**	**24.08**
	100k	**39.24**	7.1	**87.0**	8.7	10.3	**12.9**	**43.3**	**35.0**	11.5	13.3	**24.8**	10.8	24.06
	No sim.	40.64	0.2	68.9	6.0	5.3	0.0	33.1	27.4	7.1	0.0	13.9	1.9	14.88
	20k	42.18	**6.3**	**80.1**	8.7	8.7	0.0	31.3	23.4	**9.7**	0.0	**19.8**	4.1	17.47
POP3D	50k	44.46	0.7	72.0	9.9	8.6	2.6	35.3	30.7	9.1	8.8	19.2	**4.5**	18.29
	100k	**45.04**	4.9	70.4	**11.2**	**10.6**	**7.3**	**38.2**	**31.5**	7.4	**9.8**	15.2	3.6	**19.09**

Table S9: **Complete OccScanNet transfer results.** Source split denotes simulation pretraining followed by real-domain fine-tuning. No sim. uses the standard initialization. Bold marks the best reported value for each method.

Method	Split	SC IoU	Ceiling	Floor	Wall	Window	Chair	Bed	Sofa	Table	TV	Furniture	Objects	mIoU
	No sim.	59.75	38.8	60.2	41.9	38.4	46.9	61.8	66.4	50.0	37.4	56.8	46.2	49.52
	20k	60.48	42.1	61.3	42.7	39.8	48.0	63.3	67.1	50.2	37.8	57.3	47.1	50.60
MonoScene	50k	60.88	41.7	61.0	44.4	39.7	46.2	64.0	64.9	51.8	41.7	57.5	48.4	51.03
	100k	**62.44**	**43.9**	**62.8**	**45.4**	**42.6**	**49.0**	**64.8**	**68.1**	**52.9**	**44.0**	**59.7**	**49.8**	**52.99**
	No sim.	53.76	39.4	50.1	42.1	33.3	41.1	55.3	61.2	43.8	37.3	54.1	42.4	45.46
	20k	54.55	39.6	50.9	42.2	33.3	42.2	56.7	62.3	45.1	37.6	54.7	42.9	46.13
EmbodiedOcc	50k	55.11	**39.8**	51.8	42.8	34.1	42.5	56.6	62.7	45.6	38.6	**55.3**	43.7	46.67
	100k	**55.39**	**39.8**	**52.2**	**43.2**	**34.4**	**42.9**	**56.9**	**62.8**	**45.8**	**39.4**	**55.3**	**44.0**	**46.98**
	No sim.	60.57	8.2	36.4	28.5	19.9	12.4	**36.1**	33.3	18.0	7.0	**4.4**	**4.5**	18.98
	20k	63.65	13.9	38.1	30.8	21.3	13.8	36.0	34.9	18.4	**8.2**	3.4	3.7	20.24
POP3D	50k	63.80	15.5	**38.7**	30.9	21.2	**14.2**	35.3	34.8	18.9	7.9	3.6	4.0	20.44
	100k	**63.97**	**16.6**	38.5	**31.1**	**21.7**	13.4	36.0	**35.7**	**19.0**	6.8	3.6	3.9	**20.56**

F.2 EMBODIED OCCUPANCY PREDICTION BY CATEGORY

The main text reports aggregate embodied occupancy scores in Table 5 to keep the scale comparison compact. Table S10 gives the complete SC IoU, per-category IoU, and mIoU breakdown for the

same 100 held-out rooms and 30-view episodes. The two methods use matched single-frame source scales and the same accumulated room support.

Table S10: **Detailed embodied occupancy results.** Results on 100 held-out rooms use the final accumulated support of 30-view episodes. Source sizes refer to the corresponding single-frame pretraining sets. Values are percentages.

Method	Source	SC IoU	Ceiling	Floor	Wall	Window	Chair	Bed	Sofa	Table	TV	Furniture	Objects	mIoU
EmbodiedOcc	20k	50.19	30.1	46.4	37.3	32.8	19.1	59.1	44.3	21.0	21.4	45.4	22.1	34.45
	50k	51.33	30.9	48.0	38.2	34.6	21.0	60.7	46.7	23.0	20.0	47.5	22.9	35.77
	100k	52.21	30.2	49.3	39.0	34.8	25.1	62.9	49.1	26.1	20.5	48.2	23.7	37.19
GPOcc	20k	49.75	33.8	49.5	37.5	29.6	21.7	60.8	50.1	24.4	22.1	50.3	25.4	36.85
	50k	50.46	33.8	54.5	36.8	28.3	22.8	64.3	52.5	26.8	21.9	51.2	25.5	38.03
	100k	52.61	35.0	55.9	39.4	29.9	26.1	66.7	56.5	28.5	27.4	54.0	27.8	40.67

F.3 Source composition and construction ablations

We examine source composition and data construction using EmbodiedOcc pretraining followed by NYUv2 fine-tuning under a common protocol.

Table S11: **Room diversity and observation density.** The first three rows contain 20k source observations. Scores are percentages.

Source rooms	Views / room	Source observations	SC IoU↑	mIoU↑
80	250	20k	36.81	23.00
200	100	20k	37.28	22.77
400	50	20k	37.10	22.92
200	250	50k	38.25	24.08
400	250	100k	39.24	24.06

Table S12: **Data construction ablations.** All variants use the same 80 rooms and 20k observations. Scores are mIoU (%).

Source variant	mIoU↑
Generic geometry (G2)	21.34
Random proposals (C1)	21.53
No frame curation (Q0)	21.68
Full Sim2Occ	23.00

At a fixed budget of 20k observations, redistributing views across 80, 200, and 400 rooms yields a maximum difference of 0.23 mIoU (Table S11). With the room sets fixed, increasing observations per room from 100 to 250 for 200 rooms and from 50 to 250 for 400 rooms yields gains of 1.31 and 1.14 mIoU, respectively. These results show that additional observations within existing rooms provide further gains in real-world transfer.

Construction ablations. Each ablation changes one component (Table S12). G2 replaces furniture completion and window normalization with generic per-entity solid voxelization, retaining the original images and valid masks. C1 samples legal camera positions and yaw uniformly while retaining annotation and quality selection. Q0 randomly selects original candidate views without frame-level curation, preserving room selection. The full pipeline outperforms the three ablated variants by 1.32–1.66 mIoU.

G Dataset examples and qualitative results

G.1 Appearance, semantics, and physical identity

Figure S9 compares coarse semantics, entity categories, and physical instances on shared occupied geometry. Repeated chairs, tables, and storage units illustrate how category labels and placement identities differ. Panels throughout this section use stored labels or saved predictions with matched geometric views.

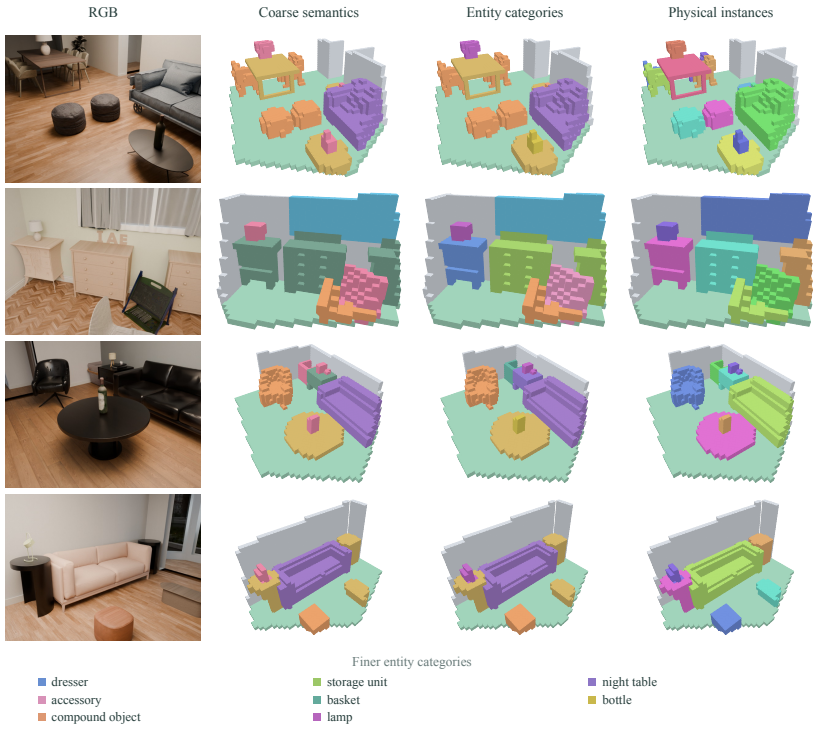

Figure S9: **Shared geometry across annotation readouts.** RGB observations are paired with coarse semantics, entity categories, and physical instances. Semantic colors encode categories; instance colors distinguish placements.

G.2 Accumulating room targets across observations

Figure S10 shows cumulative ground-truth targets from above in two held-out rooms. Occupied-target coverage grows from 9.61% to 89.97% in the living/dining room and from 5.96% to 89.90% in the bedroom.

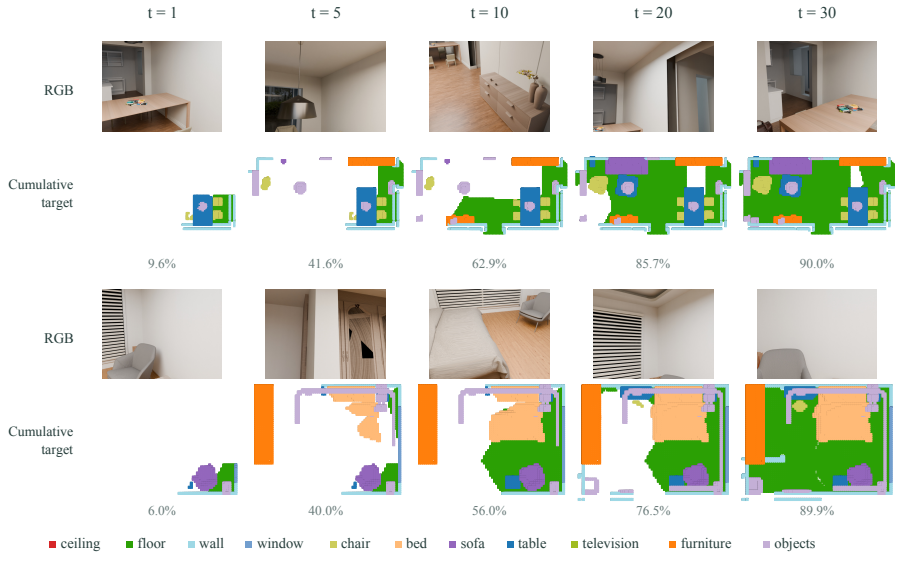

Figure S10: **Cumulative room targets.** Two episodes pair RGB observations at steps 1, 5, 10, 20, and 30 with accumulated ground truth in a fixed overhead view. Ceilings and upper wall sections are hidden to expose room interiors. Coverage is measured against the full occupied-room reference.

G.3 SEMANTIC COMPLETION AT INCREASING TRAINING SCALES

Figure S11 illustrates more continuous wall recovery with larger training splits, alongside persistent foreground semantic confusion.

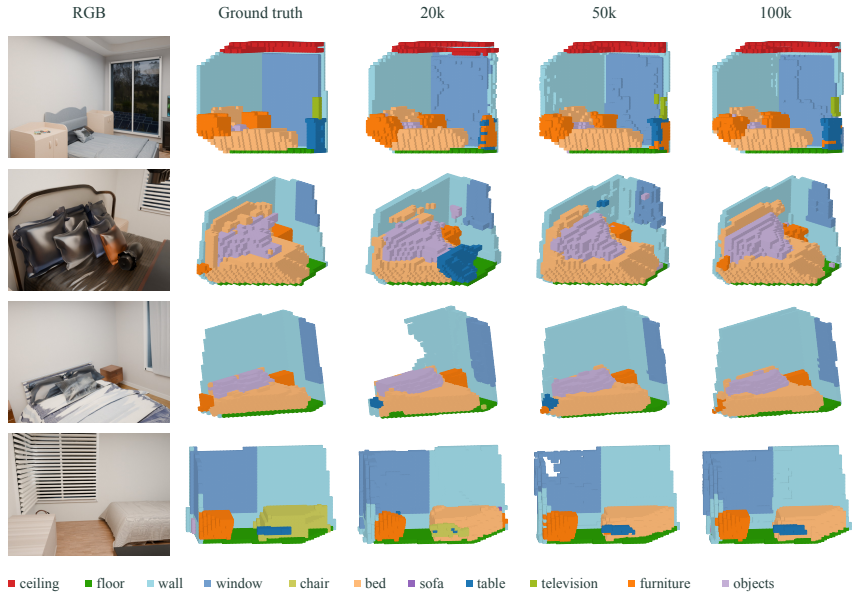

Figure S11: **MonoScene on the synthetic benchmark.** Four examples compare RGB, ground truth, and predictions from nested 20k, 50k, and 100k training splits on a common test set.

G.4 SYNTHETIC INITIALIZATION FOR REAL-WORLD OBSERVATIONS

Synthetic initialization improves the room envelope in the first NYUv2 example and the table–seating distinction in the first OccScanNet example (Figure S12), complementing Table 4.

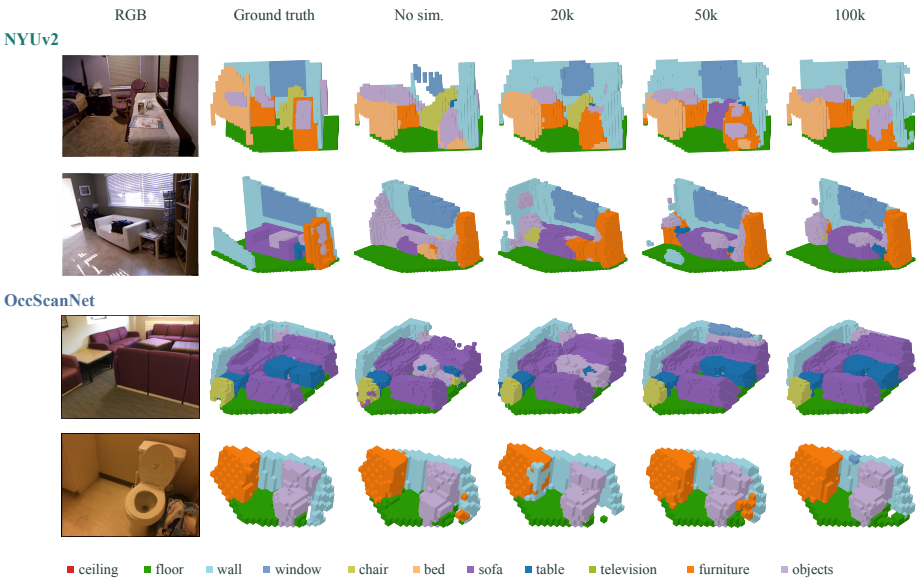

Figure S12: **MonoScene transfer.** NYUv2 and OccScanNet examples compare target-domain training without simulation ("No sim.") against 20k, 50k, and 100k synthetic initialization followed by target-domain fine-tuning.

G.5 Open-vocabulary interpretation of occupied space

Figure S13 uses queries such as armchair, curtain, pillow, and lamp to distinguish concepts within broader semantic groups.

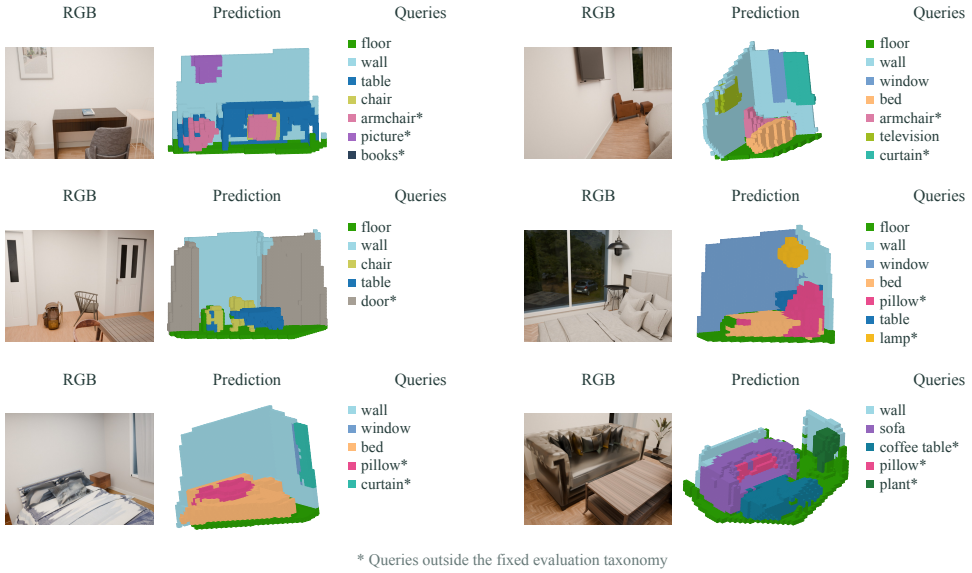

Figure S13: **POP3D open-vocabulary predictions.** Six examples pair RGB, occupancy predictions, and query legends. Starred queries lie outside the fixed evaluation taxonomy.

G.6 Instance-level occupancy predictions

Figure S14 compares MonoScene with the center–offset head against ground truth in eight held-out scenes, including repeated chairs, tables, and storage units.

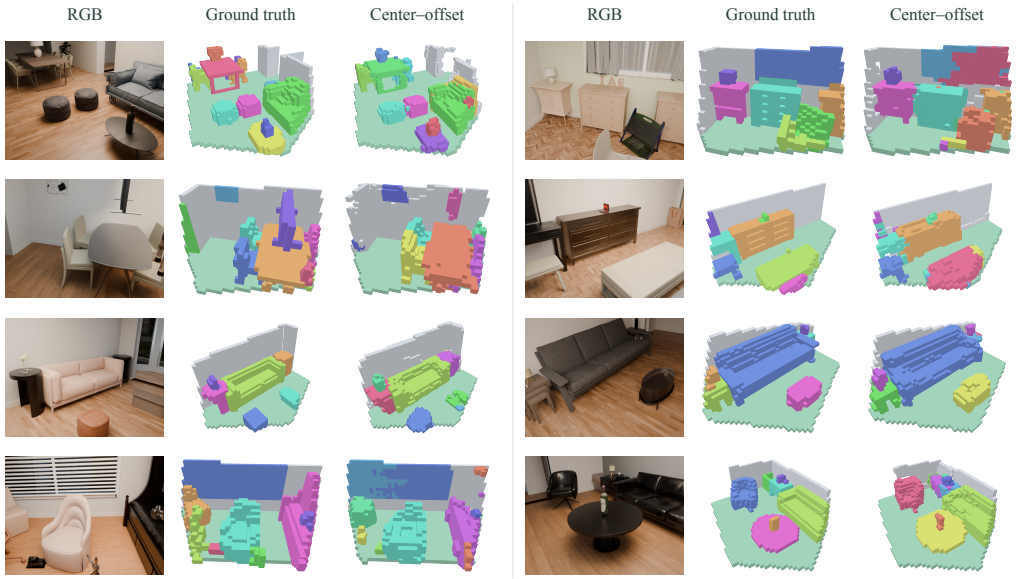

Figure S14: **MonoScene center–offset panoptic predictions.** Each row contains two examples with RGB, ground truth, and center–offset predictions. Ground truth and predictions use the same view and scale. Same-category predicted and target instances with IoU above 0.5 share display colors; unmatched predictions use separate colors.

G.7 EMBODIED OCCUPANCY PREDICTIONS ACROSS SUCCESSIVE OBSERVATIONS

Figure S15 follows two models receiving the same ordered observations. New views expand the reconstructed room and reveal major furniture. Differences between learned memory and incremental fusion persist around furniture boundaries and incompletely recovered floor regions.

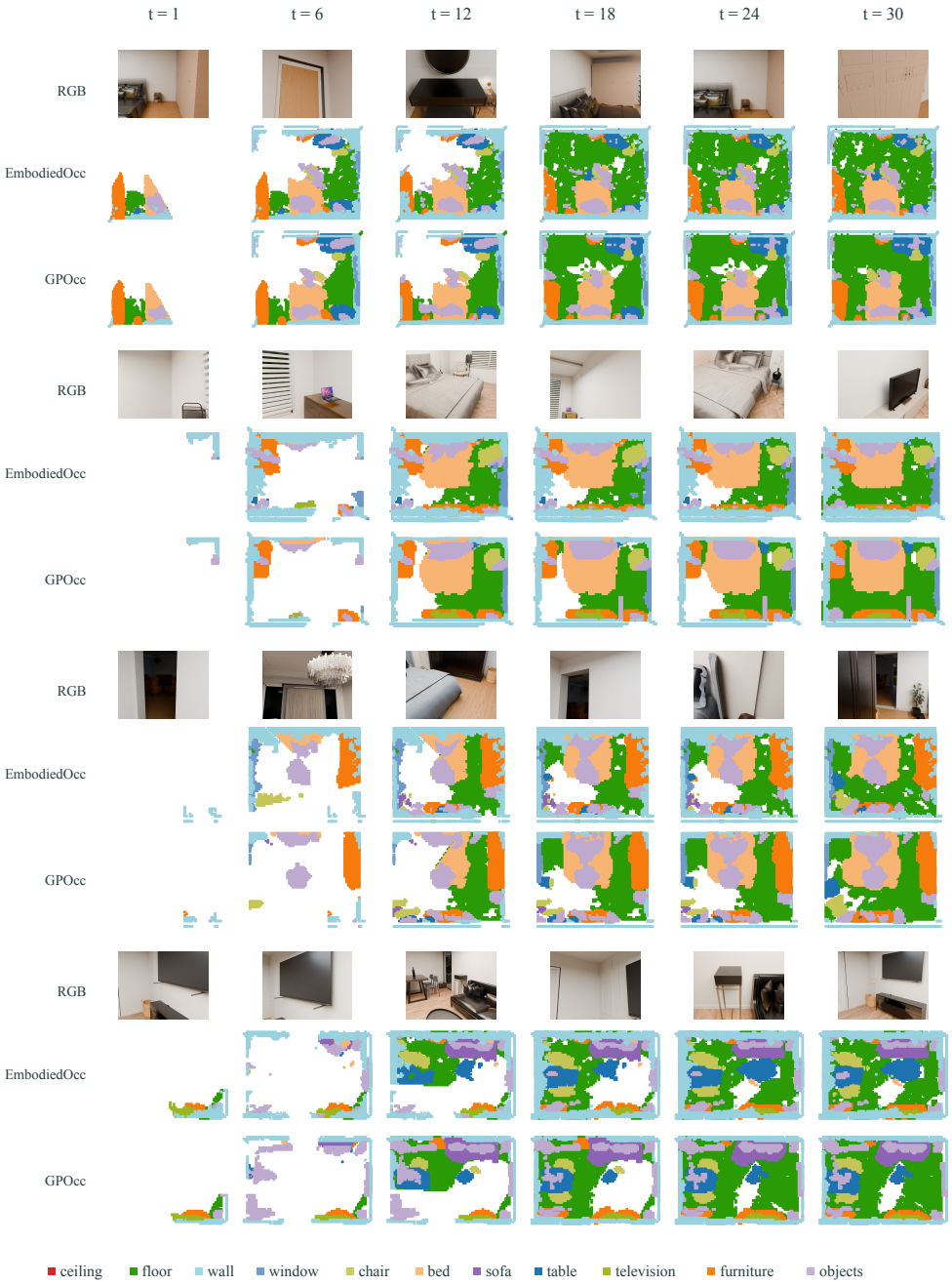

Figure S15: **Embodied occupancy predictions in four held-out rooms.** Each block follows observations 1, 6, 12, 18, 24, and 30, with RGB above EmbodiedOcc and GPOcc cumulative predictions. Both models use 50k single-frame initialization. The room view remains fixed, with ceilings hidden to expose the interior.