CoVoGen: Coarse Voxels for Compositional 3D Scene Generation

Anonymous authors
Paper under double-blind review



Figure 1: **Compositional scene generation from posed RGB-D views.** Left: RGB-D inputs and instance masks, colored as on the right. Center: generated meshes in two connected rooms, one per object. Right: individual meshes. Walls and floor are reference geometry (Appendix C.4).

Abstract

Compositional 3D scene generation from sparse observations requires complete geometry and coherent spatial arrangement. We introduce CoVoGen, a scene-first framework that extracts coarse object voxels from a completed shared scene to guide component refinement. Given posed RGB-D views with instance masks, a scene predictor completes coarse occupancy in world coordinates. A learned membership classifier extracts each target's coarse voxels from this shared prediction. Object structure refinement combines these voxels with neighboring occupancy and aligned observations to predict finer occupancy. A frozen pretrained geometry generator converts the refined occupancies into meshes in the shared scene. On four benchmarks comprising 1,129 scenes and 6,880 target objects, CoVoGen achieves the highest F@5 and Mean F for objects and scenes in our main comparison using the same input views, improving scene F@5 by up to 4.57 percentage points over the best baseline. These results support compositional scene generation through scene-derived coarse structure and object refinement.

1 Introduction

From posed RGB-D views and instance masks, we generate separate object meshes in shared world coordinates. Figure 1 illustrates the goal: completing individual geometry while preserving spatial relationships across rooms, from furniture arrangements to tableware.

Compositional reconstruction combines pretrained shape priors (Xiang et al., 2025b;a) with observed points, estimated fragments, or boxes derived from them (Siddiqui et al., 2026; Niu et al.,

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.

2026; Yin et al., 2026). These conditions describe each object's visible surface, which a few views cover only partly; beyond it, geometry comes from the image and the shape prior, or from a coarse completion of the object's fragment with scene context (Yin et al., 2026), so evidence is thinnest exactly where completion matters most. Scene-level generation (Schmid et al., 2026) offers a different source: shared structure completed across all views and objects, but not separated into objects.

CoVoGen uses **co**arse **vo**xels for compositional 3D scene **gen**eration (Figure 2). It first completes shared scene occupancy, including walls and floors, then uses learned membership to extract each target's coarse voxels. These voxels extend beyond visible surfaces and, together with neighboring geometry and aligned observations, condition object structure refinement. Frozen TRELLIS.2 converts the refined occupancies into meshes assembled in the shared scene.

Across 3D-FRONT, ARSG, SAGE, and ScanNet with MetaScenes references (1,129 scenes; 6,880 targets), CoVoGen leads four baselines in two-view object and scene F@5 and Mean F (Tables 1 and 2), and its scene F@5 exceeds ShapeR at 1, 2, 4, and 8 views (Table 3). Ablations locate the gain where evidence is scarce: scene-derived coarse voxels raise object Mean F by 7.01 points on ARSG, and by 17.75 points where under a quarter of the surface is visible (Figure 6); object structure refinement raises object IoU on every benchmark (Table 7).

Our contributions are: (i) CoVoGen, a framework extracting coarse object occupancies from a completed shared scene through learned membership; and (ii) object structure refinement using these coarse occupancies and aligned observations to guide a frozen pretrained geometry generator.

2 Related Work

Compositional scene reconstruction. Early methods predict layout and shapes (Nie et al., 2020; Liu et al., 2022) or retrieve and align CAD models (Avetisyan et al., 2019; Gao et al., 2024). Generative methods align objects to estimated geometry (Ardelean et al., 2025; Zhou et al., 2024; Yao et al., 2025), or optimize decomposed object fields per scene with 2D diffusion guidance (Ni et al., 2025). Generation can also couple features across objects (Huang et al., 2025; Meng et al., 2026), jointly predict shape and placement (SAM 3D Team et al., 2025; Zadaianchuk et al., 2026), or combine per-view generative predictions with physical constraints (Li et al., 2026). SimRecon (Xia et al., 2026a) optimizes projected views for object completion and uses scene graphs for physical assembly; ReplicateAnyScene (Dong et al., 2026) combines object discovery, generation, and visual–spatial alignment. For single-image reconstruction, 3D-RE-GEN (Sautter et al., 2025) uses background constraints, while SimuScene (Lee et al., 2026) refines shape and layout with physics feedback. Geometric conditions guide object generation within the scene: ShapeR (Siddiqui et al., 2026) uses posed images and sparse SLAM points within object boxes; WorldSculpt (Niu et al., 2026) conditions generation in coarsely localized object frames; and 3D-Fixer (Yin et al., 2026) coarsely completes estimated fragments with scene context to infer object bounds before fine geometry generation. Gaussian Object Carver (Liu et al., 2024) completes partial object geometry extracted from a Gaussian scene reconstruction. GARDEN (Sun et al., 2026) uses generated meshes to condition foreground–background point classification. CoVoGen extracts coarse object voxels from completed shared scene occupancy through learned membership, providing structural conditions for component refinement and mesh synthesis.

Scene-level reconstruction and completion. Fusion (Newcombe et al., 2011) and learned scene reconstruction or completion (Murez et al., 2020; Sun et al., 2021; Dai et al., 2020; 2021; Wang et al., 2024; 2025) recover scene surfaces without producing separate, complete object assets. CC3D (Bahmani et al., 2023) synthesizes compositional scenes conditioned on semantic layouts. GenRecon (Schmid et al., 2026) extends TRELLIS.2 to posed multi-view scenes through projected image conditioning and overlapping chunks. Our scene predictor builds on its sparse-structure stage. HIVE-3D (Zang et al., 2026) generates a coarse scene from a single image, retrieves components through image segmentation and attention, refines them hierarchically, and estimates their scale and registration. Using posed RGB-D evidence, our membership classifier extracts object occupancies that condition structure refinement alongside neighboring geometry and observations.

Pretrained 3D shape priors. Image-conditioned reconstruction (Hong et al., 2024) and generative models (Zhang et al., 2024; Wu et al., 2024; Tencent Hunyuan3D Team, 2025) provide object

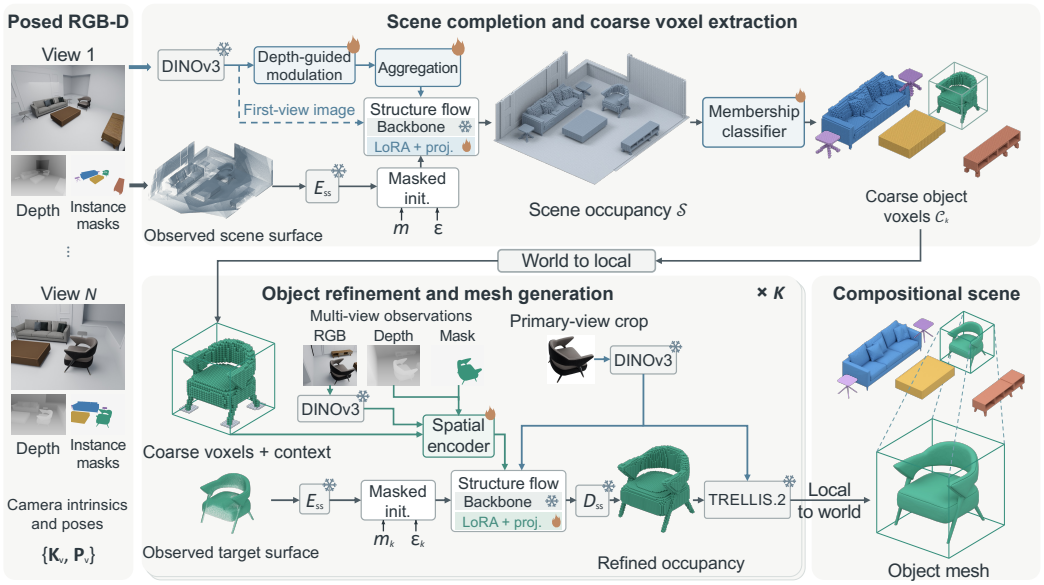

Figure 2: Overview of CoVoGen. The scene predictor completes occupancy including walls and floor. Learned membership extracts coarse object voxels, which guide object structure refinement together with neighboring geometry and observations. Frozen TRELLIS.2 converts the refined occupancies into scene components. Snowflakes and flames indicate frozen and trainable components.

shape priors. Structured-latent models (Xiang et al., 2025b;a) separate sparse structure from detailed geometry, while Amodal3R (Wu et al., 2025) extends such priors to partially occluded objects. Block3D (Cui et al., 2026) uses block-wise diffusion for text-conditioned shape generation, while AutoPartGen (Chen et al., 2025) autoregressively generates parts and discovers decompositions. CoVoGen uses pretrained object generation as a frozen geometry stage for scene generation.

3 CoVoGen: Compositional Scene Generation

3.1 Problem Formulation and Overview

Problem formulation. Given N scene views and K target objects, the observations $\mathcal{O} = \left\{I_v, D_v, \mathbf{K}_v, \mathbf{P}_v, \{M_{vk}\}_{k=1}^{K}\right\}_{v=1}^{N}$ comprise an RGB image I_v, its observed camera-z depth D_v, the intrinsics $\mathbf{K}_v$, the world-to-camera transform $\mathbf{P}_v$, and the visible mask M_{vk} of object k in view v; object identities are associated across views. We generate a compositional 3D scene of object meshes $\{\widehat{\mathcal{M}}_k\}_{k=1}^{K}$ in shared world coordinates, recovering geometry, placement, and scale.

Method overview. Figure 2 summarizes CoVoGen. The scene predictor completes coarse occupancy $\mathcal{S}$ in world coordinates, including background structures such as walls and floors. The membership classifier extracts each object's coarse occupancy $\mathcal{C}_k \subseteq \mathcal{S}$ by selecting its occupied scene cells; only target objects enter the final mesh collection.

Object structure refinement uses $\mathcal{C}_k$, neighboring scene geometry, and observations to predict finer occupancy $\widehat{\mathcal{S}}_k$ in a local object chunk. We pass this occupancy and the primary-view image condition c_k to the pretrained TRELLIS.2 geometry generator G_{geo} (Xiang et al., 2025a):

$$\widehat{\mathcal{M}}_k = \mathcal{T}_k\Big(G_{\text{geo}}(\widehat{\mathcal{S}}_k, c_k)\Big)\,. \tag{1}$$

Here $\mathcal{T}_k$ maps local chunk coordinates to world coordinates; G_{geo} retains its pretrained weights.

3.2 Scene Structure Prediction

We predict scene occupancy in overlapping chunks placed by the observed geometry, each a 64^3 occupancy grid with a 16^3 latent grid, building on the sparse-structure stage of GenRecon (Schmid et al., 2026): its pretrained backbone, image-feature projection, and conditioning injection are retained, with observed depth added to multi-view aggregation and flow initialization.

Depth-aware multi-view aggregation. Image projection alone can assign foreground appearance to an occluded 3D location. At each latent-grid query q, we compare projected occupancy-cell depths with observed depth to distinguish free space, observed surface, space behind the first surface, and missing depth. Attention summarizes these ray states, depth residuals, viewing directions, and surface appearance into a feature correction $\Delta h_v(q)$ and a view-logit correction $\Delta \ell_v(q)$ for view v (Appendix B.2). With $f_v(q)$, $h_v(q)$, and $\ell_v(q)$ the inherited projected feature, residual, and aggregation logit, depth-aware aggregation over projectively valid views $\mathcal{V}(q)$ is

$$\begin{aligned} \alpha_v(q) &= \frac{\exp(\ell_v(q) + \Delta\ell_v(q))}{\sum_{u\in\mathcal{V}(q)} \exp(\ell_u(q) + \Delta\ell_u(q))}, \\ a(q) &= \sum_{v\in\mathcal{V}(q)} \alpha_v(q)\left[f_v(q) + h_v(q) + \Delta h_v(q)\right], \end{aligned} \tag{2}$$

so depth reweights views and corrects the aggregated feature. The condition a enters the flow backbone through residual projections, alongside the retained first-view image cross-attention.

Flow initialization from observed surfaces. To initialize completion, we back-project valid depth into each chunk, voxelize the visible surface into a binary grid O_{vis}, and encode it with the frozen sparse-structure encoder, $z_{\text{vis}} = E_{\text{ss}}(O_{\text{vis}})$. With $m(q)$ indicating whether the occupancy block of q contains an observed surface cell, the flow source is $z_{\text{src}} = m \odot z_{\text{vis}} + (1-m) \odot \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$, and the flow updates all latent cells under the image and spatial conditions. Training uses the same source with a linear path to the complete structure latent (Appendix B), and inference integrates from z_{src}. The frozen sparse-structure decoder D_{ss} converts the generated latent into occupancy, and the union of the chunks on the shared scene lattice forms $\mathcal{S}$.

3.3 Coarse Object Occupancy Extraction

The membership classifier extracts coarse object occupancies from the completed scene $\mathcal{S} = \{\mathbf{x}_i\}$ by assigning occupied scene-cell centers to target objects. The selected voxels provide the object geometry used to condition structure refinement.

Object membership estimation. Appearance and ray-state features at occupied cells are aggregated across views and processed by a sparse encoder–decoder into a shared feature pyramid. Object queries combine mask-pooled image features with scene features on visible support: cells hit by masked depth. Queries attend to the pyramid from coarse to fine, guided by visible-support evidence, relative positions, and preceding membership predictions, and exchange information through gated self-attention (Appendix B.3). Final queries and scene features yield cell–object logits ℓ_{ik}:

$$p_{ik} = \text{sigmoid}(\ell_{ik}), \qquad \mathcal{C}_k = \{\mathbf{x}_i \in \mathcal{S} \mid p_{ik} \geq \tau\}, \tag{3}$$

with τ selected on validation data. $\mathcal{C}_k$ denotes coarse object occupancy, also called target support. Cells without a mask hit remain eligible through scene context, allowing the classifier to select predicted unobserved geometry. A coarse cell may support multiple objects.

Local chunk construction. Empty supports yield no mesh. We enclose each nonempty $\mathcal{C}_k$ in a scene-aligned cube with a margin on every side (Appendix B.4). The cube defines the refinement chunk and the local-to-world map $\mathcal{T}_k$; its axes stay aligned with the scene, so occupancy represents object orientation, and the margin allows completion beyond the coarse voxels.

3.4 Spatially Conditioned Object Refinement

We condition the prediction of finer object structure, a 64^3 grid in the local chunk, on the extracted coarse voxels by adapting the pretrained TRELLIS.2 sparse-structure flow (Xiang et al., 2025a) on

its 16^3 latent grid. The primary-view object crop provides the image condition c_k, and a spatial encoder combines the coarse object occupancy and neighboring scene geometry with multi-view observations.

Local spatial conditioning. Nearby occupied scene cells outside $\mathcal{C}_k$ form the context support $\mathcal{N}_k$; both supports are transformed by $\mathcal{T}_k^{-1}$. At each latent-cell center $\mathbf{u}$, target/context overlap indicators and local position define a query $e_k(\mathbf{u})$. Its footprint is projected into each selected full-scene image to sample frozen image features, depth evidence, and the target mask. Depth locates the first observed surface along each ray; the mask identifies whether that surface belongs to the target. Query-conditioned attention aggregates samples within each view and then across views into a spatial condition $g_k(\mathbf{u})$ (Appendix B).

Conditional structure generation. The coarse object and neighboring voxels provide spatial conditions, while masked observed depth defines the flow source. As in Section 3.2, we back-project, voxelize, and encode the target's visible surface from the selected views in the object frame; latent cells marked as observed receive the surface encoding, the others noise, and all cells evolve during sampling.

The flow keeps its image cross-attention to c_k. We add low-rank adaptation (LoRA) (Hu et al., 2022) updates to its self-attention and cross-attention projections and inject spatial conditions after each Transformer block. At flow time t, block $\mathcal{B}_n$ updates hidden tokens H_k^n as

$$H_k^{n+1}(\mathbf{u}) = \mathcal{B}_n(H_k^n, t, c_k)\,(\mathbf{u}) + A_n(g_k(\mathbf{u}))\,, \tag{4}$$

where each A_n is a zero-initialized affine projection. The frozen decoder D_{ss} yields $\widehat{\mathcal{S}}_k$, which supplies the finer object structure for mesh generation through Eq. 1.

3.5 Training and Inference

We train the scene predictor, membership classifier, and object structure refinement module separately on the same types of RGB-D observations as at inference: the two structure flows with a flow-matching loss from observed-surface sources to encoded complete occupancies, and the classifier with focal (Lin et al., 2017) and Dice (Milletari et al., 2016) losses (Appendix B). Refinement is trained with ground-truth target/context supports and frames. Image encoders, structure autoencoders, and the downstream generators stay frozen. At inference the primary view is the one with the largest observed target-surface coverage; trained parameters, data, architecture, and settings appear in Appendix C.

4 Experiments

4.1 Experimental Setup

Benchmarks and inputs. We evaluate on fixed cohorts from 3D-FRONT (Fu et al., 2021) (925 scenes and 3,584 target objects), ARSG (Yin et al., 2026) (100/802), SAGE (Xia et al., 2026b) (80/2,227), and ScanNet (Dai et al., 2017) with MetaScenes references (Yu et al., 2025) (24/267), 1,129 scenes and 6,880 objects in total. Following 3D-Fixer (Yin et al., 2026), we use MetaScenes assets as the reference geometry for ScanNet evaluation. ScanNet is held out entirely; the other datasets supply upstream training data, with evaluation scenes disjoint from training and validation (Appendix C).

Compared methods. We compare against RecGen (Zadaianchuk et al., 2026), MV-SAM3D (Li et al., 2026), WorldSculpt (Niu et al., 2026), and ShapeR (Siddiqui et al., 2026), report the one-view MIDI (Huang et al., 2025) and 3D-Fixer (Yin et al., 2026), and compare ShapeR at 1, 2, 4, and 8 views. Methods share observations and targets but keep their own conditioning (Appendix A); outside 3D-FRONT, MIDI's output is placed by one scene-wide similarity transform fitted to first-view depth.

Metrics. We report symmetric Chamfer distance (CD, lower is better) and surface F-score (higher is better) at 1/2/5/10 cm (0.01/0.02/0.05/0.10 released ARSG units), with Mean F their average.

Object metrics compare identity-matched meshes in world coordinates; scene metrics compare target unions without object identities. Failures receive zero F-score; CD uses outputs valid for every compared configuration (Table 8).

4.2 Quantitative Comparisons

Tables 1–3 summarize quantitative comparisons across benchmarks and input-view budgets. Complete results and paired statistical analyses are provided in Appendix A.

Table 1: **Comparisons on 3D-FRONT, ARSG, and SAGE.** Object and scene metrics: CD (cm; 10^{-2} released units for ARSG), F@5 (5 cm; 0.05 ARSG units), and Mean F (average of F@1/2/5/10), F-scores in %. Best results are bold.

	3D-FRONT						ARSG						SAGE					
	Objects			Scene			Objects			Scene			Objects			Scene		
Method	CD↓	F@5	Mean F	CD↓	F@5	Mean F	CD↓	F@5	Mean F	CD↓	F@5	Mean F	CD↓	F@5	Mean F	CD↓	F@5	Mean F
Input views: 1																		
MIDI	10.19	56.60	44.77	8.05	61.65	49.81	60.06	5.82	5.08	41.41	9.50	8.23	72.96	6.41	5.73	39.99	13.58	11.89
3D-Fixer	8.19	71.00	60.21	11.74	69.40	57.68	21.86	42.20	34.12	15.07	46.99	37.96	5.54	73.54	64.09	8.87	66.76	57.50
Input views: 2																		
RecGen	9.42	51.52	41.18	12.80	42.07	34.64	5.09	70.43	55.62	6.10	65.29	51.65	4.57	74.97	61.31	7.85	64.24	51.64
MV-SAM3D	10.21	57.52	47.21	14.98	46.56	38.98	6.53	65.38	54.08	7.94	62.02	51.28	8.23	64.29	54.67	12.07	53.77	45.63
WorldSculpt	3.17	86.44	73.71	3.88	82.32	71.30	2.88	85.91	73.82	3.55	83.65	71.12	2.69	87.26	77.69	5.01	83.26	72.97
ShapeR	3.06	87.73	76.23	**3.46**	84.22	73.32	**2.02**	91.21	84.06	2.59	88.64	80.21	2.02	90.89	85.37	3.99	85.51	78.54
Ours	**2.86**	**89.36**	**77.41**	3.76	**84.81**	**73.91**	2.12	**93.01**	**86.66**	**2.03**	**91.45**	**84.10**	**1.93**	**92.70**	**85.76**	**2.95**	**89.18**	**80.32**

Results on synthetic scenes. CoVoGen achieves the highest F@5 and Mean F for objects and scenes across 3D-FRONT, ARSG, and SAGE in the two-view comparison (Table 1), with scene F@5 gains over ShapeR of 2.81 points on ARSG and 3.67 on SAGE. The ARSG advantage persists when ShapeR receives image-generated object descriptions (Appendix A.3). These results support shared scene structure as a geometric condition for object generation.

Generalization to captured scenes. This advantage extends to ScanNet + MetaScenes, where CoVoGen achieves the best results for objects and scenes (Table 2). Its scene F@5 exceeds ShapeR's by 4.57 points, the largest gain across the four benchmarks, while scene CD decreases from 15.63 to 13.81 cm. Additional validation against captured depth is provided in Appendix A.7.

Table 2: **Comparisons on ScanNet + MetaScenes.** Object and scene metrics: CD in cm and F-scores in %. Best results are bold.

	Objects (world coordinates)			Scene (target-object union)		
Method	CD ↓	F@5	Mean F	CD ↓	F@5	Mean F
Input views: 1						
MIDI	35.84	7.39	6.81	36.67	14.71	12.81
3D-Fixer	14.31	35.92	29.77	24.64	39.26	31.93
Input views: 2						
RecGen	12.84	44.19	36.12	22.91	38.36	31.19
MV-SAM3D	12.84	44.68	37.31	18.57	40.92	34.20
WorldSculpt	12.18	42.86	35.19	19.58	36.99	30.42
ShapeR	10.06	56.15	46.06	15.63	48.75	40.77
Ours	**9.65**	**56.93**	**46.56**	**13.81**	**53.32**	**43.63**

Effect of the number of input views. Both methods improve as the number of input views increases from one to eight (Table 3). CoVoGen maintains higher scene F@5 than ShapeR across all four benchmarks and all evaluated view counts. On ARSG, the single-view gain reaches 3.6 points. This improvement is consistent with shared scene structure providing geometric information beyond the visible surfaces of individual objects.

Table 3: **Effect of input-view count.** Object/scene CD (cm; 10^{-2} ARSG units) and F@5 (%); paired best values are bold. Common CD subset: Appendix A.2.

		3D-FRONT				ARSG				SAGE				ScanNet + MetaScenes			
		Objects		Scene		Objects		Scene		Objects		Scene		Objects		Scene	
Method	Input views	CD↓	F@5	CD↓	F@5	CD↓	F@5	CD↓	F@5	CD↓	F@5	CD↓	F@5	CD↓	F@5	CD↓	F@5
ShapeR	1	3.74	84.15	**4.61**	80.02	4.36	77.61	5.47	74.61	2.66	87.28	5.01	80.45	**12.15**	**51.50**	18.60	45.55
Ours	1	**3.61**	**86.18**	4.83	**81.01**	**3.68**	**81.91**	**4.93**	**78.17**	**2.47**	**89.55**	**3.56**	**85.32**	12.45	48.84	**16.77**	**47.88**
ShapeR	2	3.01	87.73	**3.46**	84.22	**2.08**	91.21	2.59	88.64	2.05	90.89	3.99	85.51	10.56	56.15	15.63	48.75
Ours	2	**2.80**	**89.36**	3.76	**84.81**	2.10	**93.01**	**2.03**	**91.45**	**1.94**	**92.70**	**2.95**	**89.18**	**10.35**	**56.93**	**13.81**	**53.32**
ShapeR	4	2.12	92.34	2.28	90.05	0.97	97.27	1.25	95.25	1.39	94.35	2.47	90.66	**8.60**	**61.78**	11.63	54.79
Ours	4	**1.80**	**93.45**	**2.11**	**91.11**	**0.83**	**97.47**	**1.00**	**96.30**	**1.28**	**95.66**	**2.16**	**92.24**	8.80	61.75	**11.53**	**56.21**
ShapeR	8	1.96	93.54	2.16	91.01	0.81	98.14	1.06	96.24	**0.89**	97.00	**1.28**	94.81	**7.30**	65.05	**8.74**	57.89
Ours	8	**1.67**	**94.12**	**1.93**	**91.95**	**0.69**	**98.22**	**0.83**	**97.07**	**0.89**	**97.77**	1.32	**95.51**	7.52	**66.39**	8.98	**61.18**

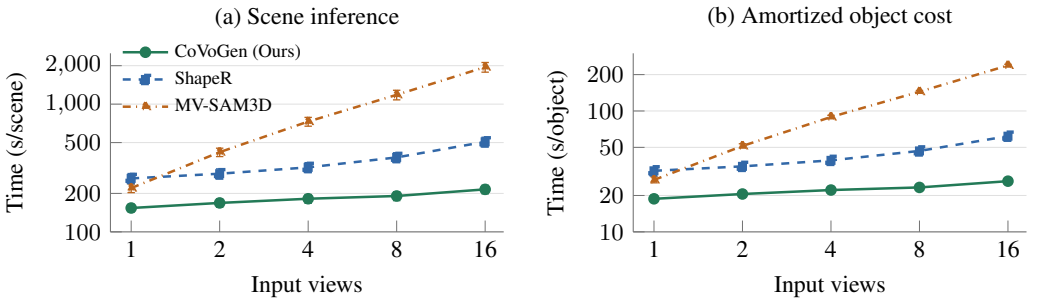

Figure 3: **Inference time versus input views.** Scene (a) and amortized object (b) time; log axes and 95% intervals (Appendix C.4).

Inference efficiency. On the common ARSG cohort (Figure 3), CoVoGen is 1.69–2.36× faster than ShapeR and 1.43–9.07× faster than MV-SAM3D. Its cost grows only 40% from one to sixteen views, because added views enter only through the conditioning while the flows and mesh generation run once per scene and object; MV-SAM3D generates per view and grows nearly ninefold (Appendix C.4).



(a) Study + bedroom (b) Living room + bedroom

Figure 4: **Additional multi-room scene generation.** (a) Study and bedroom. (b) Living room and bedroom. Colors identify individual generated objects. Walls and floor are reference geometry (Appendix C.4).

4.3 Qualitative Comparisons

Figure 4 extends the compositional output to connected, furnished rooms. Figure 5 shows CoVoGen matching the reference where ShapeR leaves gaps, smooths relief, or fuses thin parts: the pedestal, tufted headboard, petals, and carved owl eye stay complete and distinct. MV-SAM3D keeps coarse shapes but smooths the eye and enlarges or displaces objects.



Figure 5: **Qualitative comparison on 3D-FRONT (a, b) and ARSG (c, d).** One scene per row; columns: input view, MV-SAM3D, ShapeR, CoVoGen (Ours), and the reference (GT). Insets enlarge the circled regions; clocks give serial per-scene time on one A800. CoVoGen preserves geometry lost by ShapeR and scale and placement distorted by MV-SAM3D.

4.4 Ablation Studies

We ablate the scene-derived coarse structure on ARSG with two input views (30 scenes and 243 targets; voxel extraction on all 100 scenes and 802 targets) and object structure refinement on all four benchmarks (details in Appendix A.4).

Table 4: **Effect of the target condition.** Refinement conditioned on visible-depth versus scene-derived target voxels; final meshes on 30 ARSG scenes (243 targets). CD in 10^{-2} released units; Mean F in %.

	Objects		Scene	
Target condition	CD↓	Mean F	CD↓	Mean F
Visible-depth voxels	2.62	80.46	3.25	77.08
Scene-derived coarse voxels	**2.34**	**87.47**	**1.80**	**85.03**

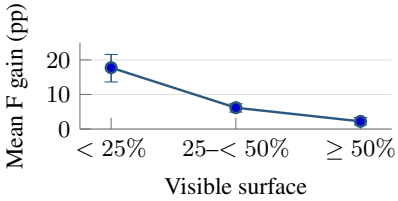

Figure 6: **Gain by visibility.** Object Mean F gain of scene-derived voxels (Table 4) by visible reference surface fraction; paired 95% intervals.

Scene-derived coarse structure. Visible-depth voxels stop at observed surfaces, providing no structural evidence for unobserved target geometry; scene-derived voxels extend into it and raise object and scene Mean F by 7.01 and 7.95 points (Table 4). Gains grow with occlusion (Figure 6), from 2.23 points for mostly visible targets to 17.75 below one-quarter visibility.

Table 5: **Effect of learned membership.** All 100 ARSG scenes, 802 targets; scores in %.

Extraction rule	Prec.	Rec.	F1	IoU
Bounding box	65.94	72.74	67.47	52.72
Learned membership	**68.51**	**73.93**	**69.89**	**55.66**

Table 6: **Effect of depth-aware aggregation.** 30 ARSG scenes, 243 targets; scores in %.

View aggregation	Prec.	Rec.	F1	IoU
Image only	59.05	63.23	58.77	43.84
Depth-aware	**67.03**	**74.47**	**69.39**	**54.90**

Table 7: **Effect of object structure refinement.** Full model versus passing the coarse object voxels directly to the frozen generator (w/o refinement); Obs.: observed depth alone, an unranked reference. Shift is mean centroid correction in cm (10^{-2} ARSG units); Rot. is median rotation in degrees.

	Object IoU ↑			**Scene F@5 (%) ↑**			**Shift ↓**		**Rot. ↓**	
Dataset	Obs.	w/o refinement	Full	Obs.	w/o refinement	Full	w/o refinement	Full	w/o refinement	Full
3D-FRONT	0.407	0.449	**0.628**	60.64	83.79	**84.81**	3.0	**2.3**	0.7	**0.3**
ARSG	0.600	0.542	**0.738**	78.08	90.11	**91.45**	2.1	**1.8**	0.9	**0.4**
SAGE	0.553	0.396	**0.700**	71.99	87.36	**89.18**	4.0	**2.5**	1.0	**0.5**
ScanNet	0.210	0.191	**0.260**	40.34	**57.10**	53.32	22.9	**20.0**	27.4	**12.5**

Learned membership extraction. Learned membership uses appearance, ray-state, and scene features to separate target support from neighboring geometry, improving voxel F1 over box selection by 2.43 points (Table 5). This also improves final-mesh object and scene Mean F by 1.84 and 1.10 points (Appendix A.4).

Depth-aware scene aggregation. Table 6 disables the scene predictor's depth-derived feature and logit corrections at inference. Image projection alone assigns foreground appearance to cells behind the first observed surface, and the completed scene inherits these errors: coarse voxel precision, recall, and F1 fall by 7.98, 11.24, and 10.62 points.

Object structure refinement. Refinement predicts finer occupancy in the object's local grid (Table 7), raising object IoU on every benchmark and shrinking the rigid correction that aligns objects to their references. Scene F@5 improves on the synthetic benchmarks; on ScanNet, refinement trades the recall of the inflated unrefined voxels for 5.6 points of precision (Table 11).

4.5 Limitations

CoVoGen relies on supplied camera poses, depth, and instance masks with consistent identities across views, rather than jointly estimating these inputs. Integrating input estimation with scene generation would extend its applicability to unconstrained image collections. Our current formulation focuses on geometry and does not jointly model object appearance. Extending scene conditioning to appearance synthesis is a direction for future work.

5 Conclusion

We presented CoVoGen, a framework for compositional 3D scene generation from posed RGB-D observations. Its central idea is to derive geometric conditions for individual objects from a completed shared scene. Learned membership extracts coarse object occupancies, which guide structure refinement and subsequent mesh generation. Experiments on synthetic and captured scenes demonstrate the effectiveness of this design, while controlled ablations support the value of scene-derived structure when objects are only partially observed. By connecting shared scene completion with object refinement, CoVoGen produces separate object meshes in a common coordinate system while keeping the pretrained geometry generator unchanged.

AI USE STATEMENT

In this work, we used generative AI tools to design and provide feedback on the evaluation methodology and experiments; to implement methods, including model, training, and evaluation code and the adapters that run the compared methods; to clean and reformat datasets, including benchmark cohorts, manifests, and data conversion; to interpret results by drafting analyses that the authors checked against the computed numbers; and to assist with translation of the authors' working notes into the manuscript. We have not used generative AI tools to develop the conceptual framework, to propose or refine the hypotheses, or to generate synthetic data sets, and formulating or proving mathematical claims and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to create and edit code, to create and modify figures, to draft and edit parts of the paper for readability, to search for and summarize related literature, and to format references. We have reviewed all AI-assisted work. All reported numbers are computed by the evaluation pipeline from saved model outputs, with cohort denominators and aggregates recomputed from the per-scene results; AI-generated code was reviewed and tested by the authors; citations were checked against the cited papers; and all AI-drafted text was revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses existing synthetic and captured-scene datasets and does not involve new data collection with human participants. Use of captured indoor scenes should respect the source datasets' access conditions and privacy restrictions. Generated completions infer unobserved geometry and should not be treated as evidence of the actual contents of an unobserved space or as verified geometry for safety-critical use.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix B describe the architecture and objectives. Appendix C specifies the training partitions, optimization settings, evaluated checkpoints, and inference configuration. Appendix A provides the complete quantitative results, metric definitions, and aggregation protocol. Code, exact data partitions, and evaluation manifests will be released.

REFERENCES

Andreea Ardelean, Mert Özer, and Bernhard Egger. Gen3DSR: Generalizable 3D scene reconstruction via divide and conquer from a single view. In *International Conference on 3D Vision (3DV)*, 2025. URL https://arxiv.org/abs/2404.03421.

Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X. Chang, and Matthias Nießner. Scan2CAD: Learning CAD model alignment in RGB-D scans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. URL https://arxiv.org/abs/1811.11187.

Sherwin Bahmani, Jeong Joon Park, Despoina Paschalidou, Xingguang Yan, Gordon Wetzstein, Leonidas Guibas, and Andrea Tagliasacchi. CC3D: Layout-conditioned generation of compositional 3D scenes. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://arxiv.org/abs/2303.12074.

Minghao Chen, Jianyuan Wang, Roman Shapovalov, Tom Monnier, Hyunyoung Jung, Dilin Wang, Rakesh Ranjan, Iro Laina, and Andrea Vedaldi. AutoPartGen: Autoregressive 3D part generation and discovery. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/e19b6f65791e350347bcff8a3955cb5b-Abstract-Conference.html.

Bowen Cui, Weijie Wang, Zeyu Zhang, Yefei He, Mingda Lin, Haoyu Zhao, Yuanyu He, Donny Y. Chen, Feng Chen, and Bohan Zhuang. Block3D: Efficient text-to-3D generation via block-wise diffusion. *arXiv preprint arXiv:2608.19567*, 2026. URL https://arxiv.org/abs/2608.19567.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. URL https://arxiv.org/abs/1702.04405.

Angela Dai, Christian Diller, and Matthias Nießner. SG-NN: Sparse generative neural networks for self-supervised scene completion of RGB-D scans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL https://arxiv.org/abs/1912.00036.

Angela Dai, Yawar Siddiqui, Justus Thies, Julien Valentin, and Matthias Nießner. SPSG: Self-supervised photometric scene generation from RGB-D scans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. URL https://arxiv.org/abs/2006.14660.

Mingyu Dong, Chong Xia, Mingyuan Jia, Weichen Lyu, Long Xu, Zheng Zhu, and Yueqi Duan. ReplicateAnyScene: Zero-shot video-to-3D composition via textual-visual-spatial alignment. *arXiv preprint arXiv:2604.10789*, 2026. URL https://arxiv.org/abs/2604.10789.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3D-FRONT: 3D furnished rooms with layouts and semantics. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. URL https://arxiv.org/abs/2011.09127.

Daoyi Gao, Dávid Rozenberszki, Stefan Leutenegger, and Angela Dai. DiffCAD: Weakly-supervised probabilistic CAD model retrieval and alignment from an RGB image. *ACM Transactions on Graphics*, 43(4), 2024. URL https://arxiv.org/abs/2311.18610.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3D. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2311.04400.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2106.09685.

Zehuan Huang, Yuan-Chen Guo, Xingqiao An, Yunhan Yang, Yangguang Li, Zi-Xin Zou, Ding Liang, Xihui Liu, Yan-Pei Cao, and Lu Sheng. MIDI: Multi-instance diffusion for single image to 3D scene generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2412.03558.

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

Baicheng Li, Dong Wu, Jun Li, Shunkai Zhou, Zecui Zeng, Lusong Li, and Hongbin Zha. MV-SAM3D: Adaptive multi-view fusion for layout-aware 3D generation. *arXiv preprint arXiv:2603.11633*, 2026. URL https://arxiv.org/abs/2603.11633.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017. URL https://arxiv.org/abs/1708.02002.

Haolin Liu, Yujian Zheng, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Towards high-fidelity single-view holistic reconstruction of indoor scenes. In *European Conference on Computer Vision (ECCV)*, 2022. URL https://arxiv.org/abs/2207.08656.

Liu Liu, Xinjie Wang, Jiaxiong Qiu, Tianwei Lin, Xiaolin Zhou, and Zhizhong Su. Gaussian object carver: Object-compositional Gaussian splatting with surfaces completion. *arXiv preprint arXiv:2412.02075*, 2024. URL https://arxiv.org/abs/2412.02075.

Yanxu Meng, Haoning Wu, Ya Zhang, and Weidi Xie. SceneGen: Single-image 3D scene generation in one feedforward pass. In *International Conference on 3D Vision (3DV)*, 2026. URL https://arxiv.org/abs/2508.15769.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision (3DV)*, 2016. URL https://arxiv.org/abs/1606.04797.

Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3D scene reconstruction from posed images. In *European Conference on Computer Vision (ECCV)*, 2020. URL https://arxiv.org/abs/2003.10432.

Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011.

Junfeng Ni, Yu Liu, Ruijie Lu, Zirui Zhou, Song-Chun Zhu, Yixin Chen, and Siyuan Huang. Decompositional neural scene reconstruction with generative diffusion prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2503.14830.

Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3DUnderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL https://arxiv.org/abs/2002.12212.

Muyao Niu, Jixuan He, Ruihan Yu, Lian Fu, Yonghao Yu, Zheng-Hui Huang, Yifan Zhan, Fengbo Lan, Yongtao Ge, Yinqiang Zheng, Kaipeng Zhang, and Zhixiang Wang. WorldSculpt: Generating compositional worlds from grounded videos. *arXiv preprint arXiv:2609.05416*, 2026. URL https://arxiv.org/abs/2609.05416.

Qwen Team. Qwen3.8-27B model card. https://huggingface.co/Qwen/Qwen3.8-27B, 2026.

SAM 3D Team, Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jiawei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. SAM 3D: 3Dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025. URL https://arxiv.org/abs/2511.16624.

Tobias Sautter, Jan-Niklas Dihlmann, and Hendrik P. A. Lensch. 3D-RE-GEN: 3D reconstruction of indoor scenes with a generative framework. *arXiv preprint arXiv:2512.17459*, 2025. URL https://arxiv.org/abs/2512.17459.

Katharina Schmid, Nicolas von Lützow, Jozef Hladký, Angela Dai, and Matthias Nießner. GenRecon: Bridging generative priors for multi-view 3D scene reconstruction. *arXiv preprint arXiv:2605.23888*, 2026. URL https://arxiv.org/abs/2605.23888.

Yawar Siddiqui, Duncan Frost, Samir Aroudj, Armen Avetisyan, Henry Howard-Jenkins, Daniel DeTone, Pierre Moulon, Qirui Wu, Zhengqin Li, Julian Straub, Richard Newcombe, and Jakob Engel. ShapeR: Robust conditional 3D shape generation from casual captures. *arXiv preprint arXiv:2601.11514*, 2026. URL https://arxiv.org/abs/2601.11514.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Jiahao Sun, Dingkun Wei, Zehong Shen, Hongyu Zhou, Yujun Shen, and Liang Li. GARDEN: Gravity-aligned reconstruction of disentangled ENvironments from RGB images. *arXiv preprint arXiv:2606.03921*, 2026. URL https://arxiv.org/abs/2606.03921.

Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. NeuralRecon: Real-time coherent 3D reconstruction from monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. URL https://arxiv.org/abs/2104.00681.

Tencent Hunyuan3D Team. Hunyuan3D 2.0: Scaling diffusion models for high resolution textured 3D assets generation. *arXiv preprint arXiv:2501.12202*, 2025. URL https://arxiv.org/abs/2501.12202.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2503.11651.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D vision made easy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2312.14132.

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3D: Scalable image-to-3D generation via 3D latent diffusion transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2405.14832.

Tianhao Wu, Chuanxia Zheng, Frank Guan, Andrea Vedaldi, and Tat-Jen Cham. Amodal3R: Amodal 3D reconstruction from occluded 2D images. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. URL https://arxiv.org/abs/2503.13439.

Chong Xia, Kai Zhu, Zizhuo Wang, Fangfu Liu, Zhizheng Zhang, and Yueqi Duan. SimRecon: SimReady compositional scene reconstruction from real videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026a. URL https://arxiv.org/abs/2603.02133.

Hongchi Xia, Xuan Li, Zhaoshuo Li, Qianli Ma, Jiashu Xu, Ming-Yu Liu, Yin Cui, Tsung-Yi Lin, Wei-Chiu Ma, Shenlong Wang, Shuran Song, and Fangyin Wei. SAGE: Scalable agentic 3D scene generation for embodied AI. *arXiv preprint arXiv:2602.10116*, 2026b. URL https://arxiv.org/abs/2602.10116.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, and Jiaolong Yang. Native and compact structured latents for 3D generation. *arXiv preprint arXiv:2512.14692*, 2025a. URL https://arxiv.org/abs/2512.14692.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b. URL https://arxiv.org/abs/2412.01506.

Kaixin Yao, Longwen Zhang, Xinhao Yan, Yan Zeng, Qixuan Zhang, Wei Yang, Lan Xu, Jiayuan Gu, and Jingyi Yu. CAST: Component-aligned 3D scene reconstruction from an RGB image. *ACM Transactions on Graphics*, 44(4), 2025. URL https://arxiv.org/abs/2502.12894.

Ze-Xin Yin, Liu Liu, Xinjie Wang, Wei Sui, Zhizhong Su, Jian Yang, and Jin Xie. 3D-Fixer: Coarse-to-fine in-place completion for 3D scenes from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://arxiv.org/abs/2604.04406.

Huangyue Yu, Baoxiong Jia, Yixin Chen, Yandan Yang, Puhao Li, Rongpeng Su, Jiaxin Li, Qing Li, Wei Liang, Song-Chun Zhu, Tengyu Liu, and Siyuan Huang. MetaScenes: Towards automated replica creation for real-world 3D scans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2505.02388.

Andrii Zadaianchuk, Leonardo Barcellona, Lennard Schuenemann, Christian Gumbsch, Zehao Wang, Muhammad Zubair Irshad, Fabien Despinoy, Rahaf Aljundi, Stratis Gavves, and Sergey Zakharov. Reconstruction by generation: 3D multi-object scene reconstruction from sparse observations. *arXiv preprint arXiv:2604.27106*, 2026. URL https://arxiv.org/abs/2604.27106.

Bin Zang, Wenting Zheng, Xiaoliang Luo, Zhiyuan Fang, Shi Li, Lvchun Wang, Wei Yu, Yi Zhao, Tian Xie, Yuchi Huo, and Rengan Xie. HIVE-3D: Hierarchical voxel enhancement for high-quality 3D scene generation. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2607.13468.

Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. CLAY: A controllable large-scale generative model for creating high-quality 3D assets. *ACM Transactions on Graphics*, 43(4), 2024. URL https://arxiv.org/abs/2406.13897.

Junsheng Zhou, Yu-Shen Liu, and Zhizhong Han. Zero-shot scene reconstruction from single images with deep prior assembly. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2410.15971.

A Evaluation Protocol and Complete Results

This appendix specifies the comparison sets, metric estimators, and ablation protocols, followed by the complete per-threshold results.

A.1 Benchmarks and Comparison Settings

Evaluation covers 1,129 scenes and 6,880 target objects, including 24 ScanNet-v2 scenes with 267 MetaScenes targets. The ScanNet subset comprises 24 scenes whose MetaScenes reference assets agree closely with the scans, which keeps the reference geometry reliable. The 1, 2, 4, and 8 view comparisons use a shared input-view order and the same target set. Training partitions are specified in Appendix C. Methods use the preprocessing described in Section 4.1 and retain their own conditioning: ShapeR and WorldSculpt derive object boxes from observed points. MIDI, trained on 3D-FRONT, generates directly in its released test frame, which is the source of our 3D-FRONT cohort; on the other benchmarks, one scene-wide similarity transform fitted to first-view observed depth places its output in world coordinates. Bold marks the best value within each matched view-budget group.

ShapeR configuration. ShapeR uses the released 019-0 denoiser and 088-0 autoencoder with bfloat16 inference and quality-preset sampling: 50 steps and 8,192 latent tokens, restricted to the input-view budget of each comparison. Its point condition is back-projected from valid target-mask depth in the selected views. We center the observed point bounding box in the model's gravity-aligned coordinate system and scale its longest half-extent to 0.9. Images are converted to grayscale and processed with the released camera-consistent rotation and crop/pad functions. Quantitative evaluation retains crop background and supplies instance masks separately. The main comparisons use the text condition `a 3D object`; Appendix A.3 additionally evaluates per-object descriptions. Mesh extraction uses UDF level 0.375, followed by floating-component removal and simplification only when the mesh exceeds 125,000 faces. The inverse observed-point normalization maps each mesh to world coordinates.

Qualitative comparison settings. Figure 5 uses all saved observations of each separately curated scene: 11 views for the two 3D-FRONT rooms and 30 for the two ARSG scenes. ShapeR receives foreground-masked RGB for this comparison.

A.2 Metrics and Aggregation

Surface distances. Object scores draw 100,000 area-uniform samples from the generated mesh and 100,000 from its reference (for 3D-FRONT, the official reference surface point set) and use symmetric nearest-neighbor distances between the two sample sets. Scene scores draw 100,000 samples over each target union, allocating samples to objects in proportion to surface area, and measure each sample to the closest triangle of the other surface with a bounding-volume hierarchy. The object and scene tracks therefore use distinct distance estimators and weighting: object scores preserve instance identity, while scene scores measure the area-weighted target union. CD is the mean of the two directed average distances.

Thresholded scores. At threshold d, precision is the fraction of predicted samples within d of the reference, recall is the reverse fraction, and F-score is their harmonic mean (zero when both fractions are zero). Thresholds are 1, 2, 5, and 10 cm, or 0.01, 0.02, 0.05, and 0.10 released units on ARSG. Mean F is the arithmetic mean of these four F-scores. Each score is computed per object or scene and then averaged over the corresponding cohort.

Aggregation sets. F-scores include every scene or target in the specified evaluation cohort. An empty object receives zero object F-score and adds no predicted surface to the scene union; its reference surface remains in the scoring target. CD uses the outputs valid for every configuration in the corresponding comparison (Table 8). The baseline tables include the single-view references in this intersection. On 3D-FRONT, Table 1 uses the subset of Table 16, which spans all 13 method/view configurations (3,451 objects and 925 scenes), so the rows the two tables share are identical. Table 3 instead uses the eight Ours/ShapeR view configurations, yielding a different CD

subset: 3,580/925, 797/100, 2,217/80, and 262/24 objects/scenes on 3D-FRONT, ARSG, SAGE, and ScanNet + MetaScenes.

Table 8: **Evaluation denominators for the benchmark comparisons.** F-score includes every scene or target in the specified evaluation cohort; CD uses the common valid-output subset, on 3D-FRONT over all configurations of Table 16.

Benchmark	Object F-score	Object CD	Scene F-score	Scene CD
3D-FRONT	3584	3451	925	925
ARSG	802	795	100	100
SAGE	2227	2182	80	80
ScanNet + MetaScenes	267	227	24	24

A.3 Paired Comparisons with ShapeR

Table 9 reports CoVoGen minus ShapeR for the two-view comparisons in Tables 1 and 2. We use 10,000 paired bootstrap resamples of scenes within each benchmark. Object scores pool the object sums and counts of the sampled scenes; scene scores average their scene values. F-scores retain all targets, including failures, and CD retains the same common valid subsets as the main tables. Differences are computed before rounding. The 2.5th and 97.5th percentiles give pointwise 95% intervals, without adjustment for simultaneous coverage across comparisons. These intervals describe variation across scenes for the evaluated outputs.

CoVoGen achieves higher estimated F@5 and Mean F for both objects and scenes across all four benchmarks. For scene reconstruction on ARSG, SAGE, and ScanNet, the pointwise 95% intervals support higher F@5 and Mean F and lower CD for CoVoGen than for ShapeR.

Table 9: **Paired differences from ShapeR.** CoVoGen minus ShapeR at two input views, with pointwise 95% scene-bootstrap intervals in brackets. CD is in cm (10^{-2} released units for ARSG); F-score differences are percentage points.

Dataset	Scope	ΔCD $\downarrow$	ΔF@5 $\uparrow$	ΔMean F $\uparrow$
3D-FRONT	Object	$-0.20\,[-0.33, -0.06]$	$+1.63\,[1.18, 2.10]$	$+1.17\,[0.77, 1.59]$
	Scene	$+0.30\,[0.09, 0.54]$	$+0.59\,[-0.05, 1.22]$	$+0.59\,[0.02, 1.16]$
ARSG	Object	$+0.10\,[-0.34, 0.71]$	$+1.80\,[1.08, 2.53]$	$+2.60\,[1.79, 3.41]$
	Scene	$-0.57\,[-0.81, -0.33]$	$+2.81\,[1.79, 3.91]$	$+3.89\,[2.75, 5.05]$
SAGE	Object	$-0.09\,[-0.32, 0.18]$	$+1.81\,[1.04, 2.59]$	$+0.40\,[-0.41, 1.24]$
	Scene	$-1.04\,[-1.69, -0.32]$	$+3.66\,[2.56, 4.74]$	$+1.78\,[0.66, 2.87]$
ScanNet	Object	$-0.40\,[-1.14, 0.25]$	$+0.78\,[-2.32, 3.87]$	$+0.50\,[-2.15, 3.07]$
	Scene	$-1.82\,[-3.30, -0.38]$	$+4.57\,[1.39, 7.95]$	$+2.86\,[0.09, 5.77]$

Image-derived text conditions for ShapeR. CoVoGen outperforms ShapeR with both generic prompts and image-derived object descriptions (Table 10). We evaluate the latter setting on all 100 ARSG scenes and 802 targets, using Qwen3.8-27B to describe the visible category and shape from foreground-masked crops of the two input views (Qwen Team, 2026). The descriptions use no reference geometry or category labels and receive no manual correction; the remaining inference settings and evaluation protocol are unchanged. CoVoGen achieves higher F@5 and Mean F for both objects and scenes. Against ShapeR with generated descriptions, its F@5 gains are 1.51 points for objects (95% CI: 0.84–2.21) and 2.52 points for scenes (1.54–3.60), showing that the advantage persists with more informative text conditioning.

Table 10: **ARSG comparison with image-derived text for ShapeR.** Qwen descriptions replace the generic prompt, while the remaining ShapeR inference is unchanged. Two input views; F-scores (%) retain all 802 targets and 100 scenes.

Method / text condition	Objects (802)		Scenes (100)	
	F@5 ↑	Mean F ↑	F@5 ↑	Mean F ↑
ShapeR / generic text	91.21	84.06	88.64	80.21
ShapeR / Qwen descriptions	91.50	84.35	88.93	80.48
CoVoGen	**93.01**	**86.66**	**91.45**	**84.10**

A.4 Ablation Protocols

Table 11: **Effect of object structure refinement.** Scene/Refine/Gen. denote scene completion and extraction, object refinement, and frozen generation. Distances are in cm (10^{-2} released units for ARSG); IoU uses 5 cm surface voxels (0.05 ARSG units). Best generated results are bold; observed depth is an unranked reference. Placement uses shared valid objects (Appendix A.4).

	Stages			Objects			Scene			Placement	
Variant	Scene	Refine	Gen.	F@5	Mean F	IoU	F@5	P@5	R@5	Shift	Rot. (°)
3D-FRONT											
Observed depth only	–	–	–	70.32	63.55	0.407	60.64	100.00	45.03	–	–
w/o refinement	✓	–	✓	86.00	67.61	0.449	83.79	81.68	**87.40**	3.0	0.7
Full model	✓	✓	✓	**89.36**	**77.41**	**0.628**	**84.81**	**88.57**	82.75	**2.3**	**0.3**
ARSG											
Observed depth only	–	–	–	80.73	75.50	0.600	78.08	100.00	64.82	–	–
w/o refinement	✓	–	✓	90.94	75.05	0.542	90.11	88.55	**92.30**	2.1	0.9
Full model	✓	✓	✓	**93.01**	**86.66**	**0.738**	**91.45**	**94.25**	89.17	**1.8**	**0.4**
SAGE											
Observed depth only	–	–	–	79.29	73.30	0.553	71.99	100.00	57.00	–	–
w/o refinement	✓	–	✓	83.39	66.80	0.396	87.36	82.97	**92.71**	4.0	1.0
Full model	✓	✓	✓	**92.70**	**85.76**	**0.700**	**89.18**	**90.57**	88.34	**2.5**	**0.5**
ScanNet + MetaScenes											
Observed depth only	–	–	–	48.34	39.12	0.210	40.34	77.84	28.31	–	–
w/o refinement	✓	–	✓	49.12	40.31	0.191	**57.10**	63.31	**54.35**	22.9	27.4
Full model	✓	✓	✓	**56.93**	**46.56**	**0.260**	53.32	**68.92**	45.57	**20.0**	**12.5**

Observed-surface baseline and IoU. The observed-surface row of Table 7 back-projects, for each target, the valid depth pixels inside its instance mask in the two input views (pixel centers, camera-axis depth, the released intrinsics and poses), connects neighboring pixels into two triangles per pixel quad, discards triangles with an edge longer than 5 cm (0.05 released units on ARSG), and concatenates the two view meshes. The resulting meshes use the same coordinate frame, estimators, and aggregation as the generated meshes. The object IoU column voxelizes 100,000 area-uniform samples from each predicted and reference object at 5 cm (0.05 released units on ARSG) and reports the intersection over union of their occupied voxel sets. Empty outputs receive zero IoU.

Rigid placement diagnostic. The last two columns of Table 11 quantify the rigid correction needed to align each generated object to its reference. This is a diagnostic separate from the world-coordinate CD and F-scores; the fitted transforms are never applied to the meshes used in the main comparisons. We draw 6,000 area-uniform fitting points from each surface and run point-to-point ICP, with rotation and translation but no scale, from two initializations: identity and a translation aligning the sample centroids. Each run uses at most 60 nearest-neighbor/Kabsch iterations and stops when the change in mean squared correspondence distance is below 10^{-12} in the input coordinate units squared. We select the transform with the smallest symmetric fitting CD, also retaining the untransformed object as a candidate. With R, t the selected transform and c the centroid of an independent 100,000-point prediction sample, the shift is $\|Rc + t - c\|_2$ and the rotation is $\arccos((\mathrm{tr}(R) - 1)/2)$, with the argument clipped to $[-1, 1]$. Shift is averaged and rotation is reported as a median over objects with valid outputs in both generated variants. This surface-based

diagnostic uses 3,583 objects on 3D-FRONT, 802 on ARSG, 2,222 on SAGE, and 265 on ScanNet + MetaScenes. These surface-based corrections depend on object shape and symmetry as well as placement.

Scene-derived conditioning on ARSG. We compare observed-surface voxels with scene-derived coarse object voxels as conditions for object structure refinement (Tables 4 and 12). The observed-surface condition is obtained from masked input depth, while the scene-derived condition is extracted from predicted scene occupancy. Both conditions use the same trained refinement module and frozen geometry generator, with all other inputs and inference settings held fixed. Evaluation covers 30 ARSG scenes containing 243 objects, with two input views per scene.

Scene-derived conditioning improves object Mean F by 7.01 percentage points (95% CI: 5.85–8.26) and scene Mean F by 7.95 points (6.23–9.74). The scene score evaluates the union of the target meshes, as in the main benchmark. This comparison measures the contribution of the coarse structural condition to final reconstruction.

Table 12: **Effect of scene-derived conditioning on final reconstruction.** Two input views, 30 ARSG scenes and 243 targets. All targets and scenes are valid for both conditions and contribute to every metric. CD is in 10^{-2} released units; F-scores are percentages.

Target voxel condition	CD ↓	F@1	F@2	F@5	F@10	Mean F
Objects (243)						
Visible depth	2.62	68.22	77.63	85.30	90.69	80.46
Scene-derived coarse voxels	**2.34**	**74.17**	**85.93**	**93.40**	**96.39**	**87.47**
Scene target unions (30)						
Visible depth	3.25	63.55	73.63	82.60	88.54	77.08
Scene-derived coarse voxels	**1.80**	**69.62**	**82.33**	**92.06**	**96.09**	**85.03**

Coarse object support accuracy. Table 13 compares predicted supports with ground-truth (GT) membership on the predicted scene and GT coarse object voxels. All variants share the scene lattice, processing chunks, observations, and models, and omit explicit neighborhood conditioning. Reconstruction quality rises with the accuracy of the coarse support.

Table 13: **Effect of coarse object support accuracy.** 30 ARSG scenes, 243 targets, and two input views. Mean F (%) retains all targets, including empty predictions.

Target support	Object Mean F ↑	Scene Mean F ↑
Predicted scene + predicted membership	87.28	85.02
Predicted scene + GT membership	89.59	87.70
GT coarse object voxels	**92.31**	**90.11**

Coarse object voxel extraction. Table 5 evaluates extraction on all 100 ARSG scenes and 802 targets. Visible depth voxelizes valid masked observations from the two input views without dilation. Box selection retains predicted scene cells within the observed object bounds expanded by one scene voxel on each side. Learned extraction uses the membership classifier; both extraction rules operate on the same completed scene. The complete reference surface is voxelized by triangle–cell intersection on the same scene lattice, including reference geometry outside predicted chunks. Precision, recall, F1, and IoU compare these occupied cell sets and are averaged per object, retaining empty predictions. These strict coarse surface-occupancy scores differ from the mesh F-scores and sampled IoU in Table 7. Learned extraction improves voxel F1 over box selection by 2.43 points (95% CI: 1.95–2.90).

Final-mesh effect of membership extraction. On the same 30-scene, 243-target subset as the conditioning comparison, we pass box-selected or learned coarse supports through the same trained refinement and geometry modules, changing only the extraction rule at inference. Table 14 shows that the improvement extends to the final meshes: learned membership raises object Mean F by 1.84 points (95% CI: 1.10–2.59) and scene Mean F by 1.10 points (0.37–1.83).

Table 14: **Final-mesh effect of learned membership extraction.** The two rules extract target support from the same completed scene. Two input views, 30 ARSG scenes and 243 targets. F-scores (%) retain all targets, including one empty box support. CD uses 242 common valid objects and all 30 scenes, in 10^{-2} released units.

Extraction rule	CD ↓	F@1	F@2	F@5	F@10	Mean F
Objects (243)						
Bounding box	1.91	72.45	84.13	91.39	94.56	85.63
Learned membership	**1.66**	**74.17**	**85.93**	**93.40**	**96.39**	**87.47**
Scene target unions (30)						
Bounding box	2.16	68.71	81.40	90.87	94.74	83.93
Learned membership	**1.80**	**69.62**	**82.33**	**92.06**	**96.09**	**85.03**

Depth aggregation. On the same 30-scene, 243-target subset as the conditioning comparison, we disable the depth-derived feature and logit residuals during upstream inference. Observed depth still supplies the refinement initialization. Table 6 scores the resulting coarse object voxels: depth aggregation improves F1 by 10.62 points (95% CI: 9.20–12.07) and IoU by 11.06 points (9.71–12.41).

Visibility strata. Figure 6 groups the same 243 final-mesh comparisons by reference surface visibility in the two input views. Visibility uses 30,000 area-weighted reference samples per target, projected into both images and checked against instance masks and camera-z depth, with tolerance of half a scene voxel plus two depth-quantization steps. The intervals $[0, 0.25)$, $[0.25, 0.5)$, and $[0.5, 1]$ contain 49, 102, and 92 objects, respectively. Their object Mean F gains are 17.75, 6.17, and 2.23 points, with paired 95% intervals of [13.63, 21.58], [4.90, 7.33], and [1.19, 3.33]. All targets are retained in this descriptive stratification.

Paired uncertainty. The intervals above use 10,000 paired bootstrap resamples of scenes within each comparison set. Object metrics pool object sums and counts in each resample, while scene metrics average scene scores. Differences are computed before rounding, and intervals are pointwise 95% intervals.

A.5 Membership and Sampling Sensitivity

Figure 7 evaluates threshold sensitivity and repeated sampling on 10 ARSG scenes (71 targets). Final-mesh scores vary little over the tested thresholds, and coarse-voxel conditioning improves reconstruction in all three paired draws. The validation-selected $\tau = 0.70$ remains the default. Table 15 reports assignment quality and multiple memberships on all 100 ARSG scenes (802 targets): predicted memberships reach 90.17 F1 against the ownership labels, only 1.42% of selected cells are shared with a wrong owner, and no target has an empty support at any tested threshold.

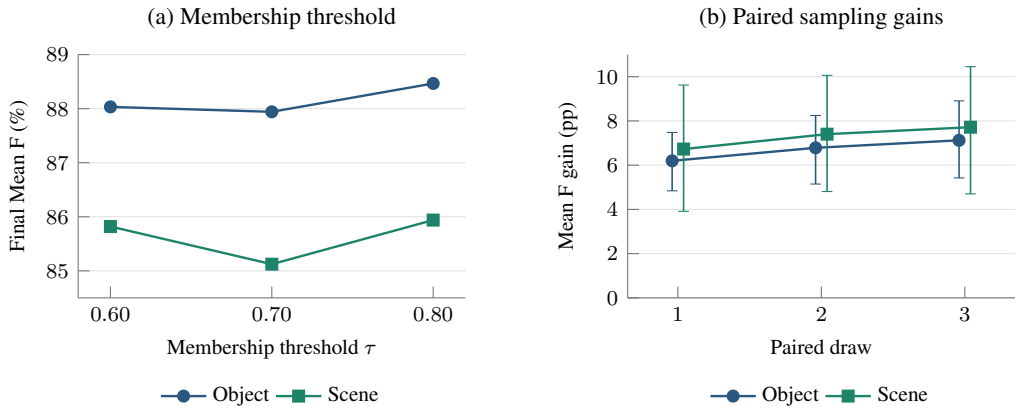

Figure 7: **Threshold sensitivity and sampling consistency.** Two input views on the same predetermined 10 ARSG scenes (71 targets). (a) Final-mesh Mean F across membership thresholds. (b) Scene-derived coarse minus visible-depth conditioning over three paired refinement and mesh-generation draws, with upstream predictions shared. Error bars are pointwise 95% scene-bootstrap intervals.

Table 15: **Membership quality.** Two input views, 100 ARSG scenes and 802 targets; $\tau = 0.70$ unless specified. Scores and rates are percentages.

(a) Conditional membership: object averages

Precision	F1	IoU
85.34	90.17	83.72

(b) Multiple assignments and empty supports

Statistic	Count / total	Rate
Erroneous multiple assignments	56,725 / 3,987,454	1.42
Empty supports (each tested τ)	0 / 802	0.00

Assignment scores use predicted scene cells and the training ownership labels (center-to-surface distance ≤ 1.25 scene-voxel widths); they differ from the geometric voxel scores in Table 5. Erroneous sharing counts selected cells with multiple owners and at least one incorrect label. Empty supports are checked at $\tau = 0.60/0.70/0.80$.

A.6 Full Per-Threshold Results

Tables 16, 17, 18, and 19 report CD and F@1/2/5/10 for both scopes on all four benchmarks; Table 1 shows the F@5 and Mean F columns of the training-domain benchmarks and Table 2 the corresponding ScanNet results.

Table 16: **Complete quantitative results on 3D-FRONT.** All configurations use 925 scenes and 3,584 targets. CD is in cm; F@1/2/5/10 use 1/2/5/10 cm thresholds (%, ↑). Mean F averages the four F-scores. Best results within each input-view group are bold; CD uses a shared subset of 3,451 objects and 925 scenes.

	Objects (world coordinates)						Scene (target-object union)					
Method	CD ↓	F@1	F@2	F@5	F@10	Mean F	CD ↓	F@1	F@2	F@5	F@10	Mean F
Input views: 1												
MIDI	10.19	16.65	32.06	56.60	73.78	44.77	8.05	22.29	37.36	61.65	77.94	49.81
3D-Fixer	8.19	36.31	54.69	71.00	78.83	60.21	11.74	32.98	49.96	69.40	78.40	57.68
ShapeR	3.79	47.10	67.68	84.15	91.16	72.52	**4.61**	45.51	62.91	80.02	**88.32**	69.19
Ours	**3.67**	**48.11**	**69.93**	**86.18**	**92.12**	**74.08**	4.83	**46.00**	**64.17**	**81.01**	88.12	**69.82**
Input views: 2												
RecGen	9.42	14.00	26.50	51.52	72.72	41.18	12.80	11.91	21.22	42.07	63.35	34.64
MV-SAM3D	10.21	22.16	36.41	57.52	72.77	47.21	14.98	18.86	29.18	46.56	61.31	38.98
WorldSculpt	3.17	46.82	68.24	86.44	93.35	73.71	3.88	47.80	64.29	82.32	90.79	71.30
ShapeR	3.06	51.35	72.65	87.73	93.20	76.23	**3.46**	49.80	67.96	84.22	**91.31**	73.32
Ours	**2.86**	**51.81**	**74.14**	**89.36**	**94.32**	**77.41**	3.76	**50.63**	**69.26**	**84.81**	90.93	**73.91**
Input views: 4												
ShapeR	2.16	**57.99**	79.44	92.34	95.97	81.44	2.28	56.33	75.71	90.05	94.99	79.27
Ours	**1.83**	57.29	**80.00**	**93.45**	**97.02**	**81.94**	**2.11**	**58.20**	**77.11**	**91.11**	**95.45**	**80.47**
Input views: 8												
ShapeR	1.99	**61.19**	**82.16**	93.54	96.52	**83.35**	2.16	58.55	77.87	91.01	95.34	80.69
Ours	**1.70**	59.58	81.76	**94.12**	**97.29**	83.19	**1.93**	**60.98**	**79.35**	**91.95**	**95.85**	**82.03**

Table 17: **Complete quantitative results on ARSG.** CD is in 10^{-2} released units; F@1/2/5/10 use thresholds of 0.01, 0.02, 0.05, and 0.10 released units. F-scores and Mean F are percentages (↑). Best results are bold. Denominators are specified in Appendix A.

	Objects (world coordinates)						Scene (target-object union)					
Method	CD ↓	F@1	F@2	F@5	F@10	Mean F	CD ↓	F@1	F@2	F@5	F@10	Mean F
Input views: 1												
MIDI	60.06	1.27	2.53	5.82	10.70	5.08	41.41	2.33	4.34	9.50	16.75	8.23
3D-Fixer	21.86	13.33	23.88	42.20	57.08	34.12	15.07	15.56	27.00	46.99	62.29	37.96
Input views: 2												
RecGen	5.09	24.11	42.12	70.43	85.82	55.62	6.10	21.01	36.69	65.29	83.60	51.65
MV-SAM3D	6.53	27.94	43.16	65.38	79.81	54.08	7.94	25.99	40.08	62.02	77.04	51.28
WorldSculpt	2.88	48.36	67.96	85.91	93.07	73.82	3.55	44.86	63.73	83.65	92.22	71.12
ShapeR	**2.02**	68.13	81.83	91.21	95.06	84.06	2.59	61.70	76.51	88.64	94.00	80.21
Ours	2.12	**72.60**	**84.76**	**93.01**	**96.27**	**86.66**	**2.03**	**68.02**	**81.28**	**91.45**	**95.65**	**84.10**

Table 18: **Complete quantitative results on SAGE.** CD is in cm; F@1/2/5/10 use 1/2/5/10 cm thresholds. F-scores and Mean F are percentages (↑). Best results are bold. Denominators are listed in Table 8.

	Objects (world coordinates)						Scene (target-object union)					
Method	CD ↓	F@1	F@2	F@5	F@10	Mean F	CD ↓	F@1	F@2	F@5	F@10	Mean F
Input views: 1												
MIDI	72.96	1.49	2.81	6.41	12.19	5.73	39.99	3.55	6.36	13.58	24.07	11.89
3D-Fixer	5.54	42.38	57.56	73.54	82.87	64.09	8.87	35.78	50.46	66.76	76.99	57.50
Input views: 2												
RecGen	4.57	31.95	50.49	74.97	87.85	61.31	7.85	22.84	38.12	64.24	81.37	51.64
MV-SAM3D	8.23	31.77	45.35	64.29	77.26	54.67	12.07	24.66	35.91	53.77	68.19	45.63
WorldSculpt	2.69	57.19	73.36	87.26	92.94	77.69	5.01	51.19	67.60	83.26	89.84	72.97
ShapeR	2.02	**72.68**	83.53	90.89	94.37	85.37	3.99	**62.82**	75.17	85.51	90.66	78.54
Ours	**1.93**	70.14	**84.27**	**92.70**	**95.94**	**85.76**	**2.95**	60.83	**77.24**	**89.18**	**94.04**	**80.32**

Table 19: **Complete quantitative results on ScanNet + MetaScenes.** The cohort is the 24-scene subset of Section 4.1 (267 targets). CD is in cm; F@1/2/5/10 use 1/2/5/10 cm thresholds. F-scores and Mean F are percentages (↑). Best results are bold. Denominators are listed in Table 8.

	Objects (world coordinates)						Scene (target-object union)					
Method	CD ↓	F@1	F@2	F@5	F@10	Mean F	CD ↓	F@1	F@2	F@5	F@10	Mean F
Input views: 1												
MIDI	35.84	1.86	3.32	7.39	14.70	6.81	36.67	4.27	7.34	14.71	24.91	12.81
3D-Fixer	14.31	12.45	20.86	35.92	49.86	29.77	24.64	13.70	22.56	39.26	52.22	31.93
Input views: 2												
RecGen	12.84	15.09	25.68	44.19	59.54	36.12	22.91	12.31	21.25	38.36	52.85	31.19
MV-SAM3D	12.84	16.98	27.16	44.68	60.44	37.31	18.57	15.58	24.46	40.92	55.86	34.20
WorldSculpt	12.18	14.31	24.27	42.86	59.34	35.19	19.58	11.79	19.88	36.99	53.01	30.42
ShapeR	10.06	22.02	35.28	56.15	70.77	46.06	15.63	20.04	30.89	48.75	63.41	40.77
Ours	**9.65**	**22.10**	**35.89**	**56.93**	**71.31**	**46.56**	**13.81**	**20.62**	**33.58**	**53.32**	**66.99**	**43.63**

A.7 Consistency with Captured Depth

We additionally evaluate the two-view reconstructions against six held-out ScanNet depth observations per scene on the same 24-scene, 267-target benchmark cohort. The references are captured depths, with instance masks projected from the annotated scan, rather than rendered MetaScenes surfaces. Predictions are evaluated in their original world coordinates without fitting to the validation observations.

Surface coverage measures the fraction of valid target-mask pixels whose predicted first-hit depth agrees with the captured depth within a given tolerance; missing predictions count as misses. *Free-space violation* measures the fraction of predicted first-hit pixels with valid captured depth that lie in front of the observed surface by more than that tolerance, including predictions outside target masks. Surfaces behind the captured depth may be occluded and do not count as free-space violations. We combine pixel counts over the six views within each scene, then average the scene rates.

Table 20 shows that refinement raises coverage and lowers free-space violation at every tolerance, with the largest gains at the tighter tolerances. These measurements evaluate consistency with visible captured surfaces.

Table 20: **Consistency with held-out captured depth.** Two input views on 24 ScanNet scenes (267 targets); six additional views per scene provide measured depth. Rates (%) are averaged over scenes at 1, 2, and 5 cm tolerances; Δ is full minus without refinement in percentage points, with paired 95% scene-bootstrap intervals. Better values are bold.

Configuration	1 cm	2 cm	5 cm
Surface coverage ↑			
w/o refinement	11.86	25.57	59.13
Full model	**24.91**	**42.05**	**62.77**
Δ (95% CI)	+13.05 [9.33, 17.07]	+16.48 [11.92, 20.99]	+3.63 [0.68, 6.51]
Free-space violation ↓			
w/o refinement	66.08	56.17	29.39
Full model	**36.09**	**25.40**	**11.69**
Δ (95% CI)	−29.99 [−35.08, −24.84]	−30.77 [−35.72, −25.65]	−17.71 [−21.19, −14.13]

B Method Details

B.1 Scene Chunking and Occupancy Merging

At inference, chunk placement uses a subsample of the foreground depth points back-projected from the input views. Percentile bounds of these points define the horizontal extent and the vertical range $[h_-, h_+]$ in the scene's up-aligned coordinates. All chunks in a scene have side length $L =$

$\kappa(h_+ - h_-)$, with an extent factor κ slightly above one, and vertical center $h_- + L/2$. Along each horizontal axis, we cover the padded extent with the smallest number of regularly spaced chunks whose center spacing is an integer multiple of $L/16$ and satisfies the minimum overlap. The centers are symmetric about the extent's midpoint; chunks with insufficient observed points are omitted.

Each chunk is predicted independently and decoded at 64^3 resolution. Its occupied cells are translated onto a common lattice of spacing $L/64$; the center-spacing constraint makes all offsets integral on this lattice. The scene support is the union of the occupied lattice indices, with duplicates removed. Thus overlapping chunks contribute to the same scene cells without averaging their latents or logits.

B.2 Depth-Aware Scene Aggregation

Within each latent-grid query cell, ray evidence is evaluated at its 4^3 constituent occupancy-cell centers. For each center and view, its camera-space depth z_v is compared with the observed depth D_v at the projected pixel. Let a be the occupancy-cell side length in camera units and let $f_{x,v}, f_{y,v}$ be focal lengths in pixels at the depth-map resolution. The depth tolerance is

$$\epsilon_v = \lambda_{\text{vox}} a + \lambda_{\text{pix}} r_v, \qquad r_v = \frac{D_v}{2}\sqrt{f_{x,v}^{-2} + f_{y,v}^{-2}}, \tag{5}$$

where r_v is the half-diagonal of a pixel's footprint at the observed depth. A valid depth sample indicates free space if $z_v - D_v < -\epsilon_v$, an observed surface if $|z_v - D_v| \le \epsilon_v$, and space behind the first surface otherwise. An invalid projection or missing depth has a separate state. Space behind the first surface remains unobserved. The fine-cell tokens combine these states, depth residuals, viewing directions, confidence, relative positions, and appearance features retained at observed surfaces. Within each view, the coarse image feature attends to these tokens, and the resulting summary, augmented with observation and state statistics, predicts the feature correction $\Delta h_v(q)$ and the view-logit correction $\Delta \ell_v(q)$ of Eq. 2. When fine-cell projections are unavailable, the inherited aggregation is retained. Both the appearance mean and the feature residuals use the corrected view weights.

B.3 Object Membership Estimation

Shared scene features. At each occupied scene cell, we project frozen image features into the input views and apply the ray-state rule of Eq. 5. Appearance is retained at observed surfaces; an encoded geometric descriptor also carries the depth relation, residual, and validity. A learned score on appearance and geometry weights the views, producing one feature per cell. A four-level sparse encoder–decoder builds the shared feature pyramid. Each downsampling step pools the existing children of a cell using projected mean and maximum features and a learned count embedding. Upsampling gathers parent features back to the original child coordinates and fuses skip features, preserving the predicted scene support.

Object queries. For each object and view, image patch features are pooled with mask-coverage weights and augmented with the viewing direction from the observed object anchor and the mask coverage. A Transformer pools these view tokens with a learned aggregation token. Its output is added to a projection of the scene features pooled over the object's visible support and an embedding of support statistics, then normalized to initialize the object query. The anchor is the confidence-weighted centroid of occupied cells hit by the object's masked depth. If no scene cell is hit, the pooled 3D feature is zero and anchor-dependent terms are disabled; the image query remains available.

Query updates and membership. The decoder visits the scene pyramid from coarse to fine. After layer normalization of query q_k and scene feature f_i, each head computes

$$A_{ki} = \frac{\langle W_Q q_k, W_K f_i \rangle}{\sqrt{d_h}} + b_{ik}^{\text{obs}} + b_{ik}^{\text{rel}} + b_{ik}^{\text{prev}}, \tag{6}$$

where d_h is the head width. The observed-support term is a bounded, nonnegative sigmoid-MLP output, zero for cells without a hit. Its inputs are support mass, view count, mean and maximum confidence, and a hit indicator. The relative-position term is a bounded tanh-MLP of the cell's

displacement from the object anchor, encoded by sinusoidal features, signed-log coordinates, direction, and log distance. From the second layer onward, b^{prev} is a learned sigmoid gate times the clipped log probability of the preceding prediction; preceding fine-cell logits are max-pooled to the current scale and detached. These finite biases guide attention without excluding unobserved cells. Cross-attention is followed by gated inter-object self-attention and a feed-forward update.

After every decoder layer, a shared head computes the scaled cosine similarity of separately projected and normalized fine-cell features and object queries, plus learned evidence and relative-position terms and a scalar bias. Independent sigmoids produce the membership probabilities; there is no softmax competition between objects. All layers receive the supervision in Eq. 11, and the final layer defines the supports used in Section 3.3.

B.4 Local Chunk Construction

Let $\mathbf{a}_k, \mathbf{b}_k$ denote the cell-volume bounds of a nonempty coarse occupancy $\mathcal{C}_k$, including half-cell extents. Adding margin δ on each side (Appendix C) and padding the shorter axes gives the local-to-scene mapping for $\mathbf{u} \in [-\frac{1}{2}, \frac{1}{2}]^3$:

$$\mathcal{T}_k(\mathbf{u}) = s_k\mathbf{u} + \mathbf{t}_k, \qquad \mathbf{t}_k = \tfrac{1}{2}(\mathbf{a}_k + \mathbf{b}_k), \qquad s_k = \max_{d\in\{1,2,3\}} (b_{k,d} - a_{k,d}) + 2\delta. \tag{7}$$

The cube defines the refinement chunk; its axes remain aligned with the scene, so occupancy represents object orientation. The margin allows completion beyond the coarse voxels, and $\mathcal{T}_k$ returns the generated mesh to world coordinates.

B.5 Flow Formulation

Both flows use the source construction of Section 3.2 and the interpolation

$$z_t = (1-t)z_{\text{gt}} + [\sigma_{\min} + (1-\sigma_{\min})t]\, z_{\text{src}}, \qquad t \in [0,1], \tag{8}$$

where z_{gt} is the complete training structure latent and $\sigma_{\min} = 10^{-5}$ is a small endpoint-noise constant. At inference, integration starts from z_{src} at $t = 1$ and proceeds toward $t = 0$. For object refinement, the visible structure and the complete training target are expressed in the object frame.

B.6 Spatial Conditioning for Object Refinement

For each selected view, a 3×3 stencil is sampled over the projected footprint of the latent cell. At each sample, frozen image features are combined with the depth evidence and the target-mask channel. Let $f_{vkj}(\mathbf{u})$ be the fused feature at stencil sample j in view v. Query-conditioned attention first aggregates samples within each view and then combines the view summaries:

$$\begin{aligned} b_{vk}(\mathbf{u}) &= \sum_{j=1}^{9} \beta_{vkj}(\mathbf{u})\psi(f_{vkj}(\mathbf{u}))\,, \\ g_k(\mathbf{u}) &= \eta\left(e_k(\mathbf{u}) + \sum_{v\in\mathcal{V}_k} \gamma_{vk}(\mathbf{u}) b_{vk}(\mathbf{u})\right), \end{aligned} \tag{9}$$

where ψ and η are learned feature mappings and β and γ are learned attention weights normalized over valid image projections; the within-view weights use the query and the sampled features, and the cross-view weights use the query and the view summaries. Specifically, with query e_k, a projected key $\phi(f_{vkj})$, and spatial width d_g, the sample logits are $\langle e_k, \phi(f_{vkj})\rangle/\sqrt{d_g}$; the view logits are $w^{\mathsf{T}} \tanh(b_{vk} + e_k)$. Both use masked softmax, with zero image contribution when all projections are invalid. The query combines local position, target/context overlap indicators, and latent-cell side length. This branch classifies depth with tolerance $\epsilon_0 + s_k/(2R)$, where R is the latent-grid resolution and s_k the object-frame side length. Its evidence also includes the target mask and the back-projected point's offset from the query. Unlike the scene aggregation, it retains appearance from valid image projections even when their depth is missing, with depth validity encoded explicitly. The image crop condition and the full-scene spatial features are encoded separately; patch tokens from the full scene are not used as a substitute for the primary crop. The spatial condition depends only on the supports, the local frame, and the observations, so it is computed once per object and reused at every flow step.

View selection. Within each predicted object frame, the primary view maximizes the number of distinct observed target-surface cells on a 256^3 grid computed from target-mask depth; the remaining admitted views are ordered by additional surface coverage. For object refinement, view selection precedes the construction of both conditioning branches and of the visible-surface source.

B.7 Training Objectives

For scene prediction and object refinement, with χ the module's image and spatial conditions, d the number of latent elements, and z_t from Eq. 8, each flow minimizes

$$\mathcal{L}_{\text{flow}} = \mathbb{E}\left[\frac{1}{d}\left\|v_\theta(z_t, t; \chi) - \big((1-\sigma_{\min})z_{\text{src}} - z_{\text{gt}}\big)\right\|_2^2\right], \tag{10}$$

where the expectation covers training examples, source noise, time, and condition sampling. Scene prediction samples $\text{logit}(t) \sim \mathcal{N}(1,1)$, while object refinement samples t uniformly on $[0,1]$. For membership estimation, let Y contain the binary object assignments and $p^{(r)}$ the sigmoid predictions after decoder layer r; all three decoder layers are supervised with

$$\mathcal{L}_{\text{mem}} = \sum_{r=1}^{3} \lambda_r \left[\mathcal{L}_{\text{focal}}(p^{(r)}, Y) + \mathcal{L}_{\text{Dice}}(p^{(r)}, Y)\right], \qquad (\lambda_1, \lambda_2, \lambda_3) = (0.25, 0.25, 1). \tag{11}$$

The focal term uses exponent 2 and averages positive and negative cells separately for each object, giving the two groups equal weight; the soft Dice term measures overlap over all occupied scene cells. Both terms are averaged over objects within each scene and then over scenes.

C Implementation Details

C.1 Architecture Configuration

Table 21 specifies the network dimensions. All image branches use the frozen DINOv3 (Siméoni et al., 2025) ViT-L/16 pretrained on LVD-1689M, with 512×512 inputs and 32×32 spatial tokens of width 1,024. For component refinement and mesh synthesis, the primary crop condition retains the full 1,029-token sequence; spatial sampling uses the 1,024 patch tokens. Both structure flows use an 8×16^3 latent and a 64^3 occupancy decoder. The scene predictor starts from GenRecon's sparse-structure checkpoint; object refinement starts from the 1.3B sparse-structure flow of TRELLIS.2-4B. Their frozen structure encoder and decoder are the released TRELLIS-image-large `ss_enc_conv3d_16l8` and `ss_dec_conv3d_16l8` models.

Table 21: **Network architecture.** Widths are feature dimensions. Backbone dimensions are listed jointly for the two structure flows; each stage has its own trained conditioning modules. Optimization and sampling are specified in Tables 23 and 24.

Module	Component	Configuration
Both structure flows	Transformer backbone	30 blocks; width 1,536; 12 heads; image condition width 1,024
Scene predictor	Fine-cell depth attention	width 128; 4 heads; appearance width 64; ray-state embedding width 16
Membership	Per-view features	appearance width 128; geometry width 32
	Sparse encoder–decoder	widths 64/96/128/192; encoder blocks 2/2/3/3; decoder blocks 2/2/2; output width 256
	Query view encoder	2 layers; width 256; 8 heads; feed-forward expansion 2
	Query decoder	3 layers; width 256; 8 heads; feed-forward expansion 4; downsampling factors 8/4/2
	Membership head	width 256; cosine scale initialized to 10 and bounded to $[1, 100]$
Object refinement	Spatial encoder	width 128; query-to-sample dot-product attention and learned view scoring
	Spatial injection	independent zero-initialized affine $128 \to 1{,}536$ projection after each of the 30 blocks

Spatial settings. Chunk placement uses at most 10,000 uniformly subsampled foreground depth points. Horizontal bounds use percentiles 0.5 and 99.5; vertical bounds use percentiles 0.2 and 99.97. Horizontal padding is $3.5L/16$ per side, adjacent chunks overlap by at least $L/4$, and each retained chunk contains at least 30 sampled points. The scene predictor resizes depth maps to 504×504; the surface-band coefficients in Eq. 5 are $\lambda_{\text{vox}} = \sqrt{3}/2$ and $\lambda_{\text{pix}} = 1/2$. The membership classifier applies the same rule at scene-cell centers. Its attention bias bound is 4, and relative-position sinusoidal wavelengths are $(0.125, 0.25, 0.5, 1, 2, 4)$ in the scene coordinate units. Local-frame bounds are expanded by two scene-voxel widths per side; context contains non-target scene cells within two grid steps of a target cell. Object refinement uses a 3×3 footprint stencil and $\epsilon_0 = 0.01$ in the depth coordinate units. The frozen geometry stage uses the TRELLIS.2 1024 configuration, receiving all refined occupied cells and the same primary-view crop condition.

C.2 Training Data and Partitions

Table 22 gives the data used for the evaluated checkpoints. The scene predictor is trained on 31,640 chunks from 11,387 3D-FRONT and SAGE scenes, with 1,736 validation chunks from 619 scenes. Its 3D-FRONT partition separates houses; SAGE uses a 90/5/5 scene split while retaining the existing held-out scenes. The membership and refinement modules additionally use ARSG. Object refinement uses 20,166 training objects from 7,224 scenes and 2,120 validation objects from 803 scenes. Training and validation partitions are defined for each module. All 1,129 evaluation scenes are disjoint from these partitions, and no ScanNet data is used for training.

Table 22: **Training and development-validation data.** Counts are scenes for the scene predictor and membership classifier, and target objects for object refinement. All entries refer to the evaluated checkpoints.

Module (unit)	Partition	3D-FRONT	ARSG	SAGE
Scene predictor (scenes)	Train	4,475	–	6,912
	Validation	235	–	384
Membership classifier (scenes)	Train	4,111	2,621	1,828
	Validation	457	292	203
Object refinement (objects)	Train	9,543	8,244	2,379
	Validation	930	919	271

C.3 Optimization and Sampling

Table 23 lists the training configuration of the three learned modules. Scene chunks are shuffled without replacement per epoch. The classifier uses asset-balanced anchor sampling, selecting three to five visible object queries when available while retaining the full scene support. View counts are sampled uniformly from the available counts between one and eight. The classifier and refinement clip the gradient norm at 1; the scene flow uses a 95th-percentile adaptive threshold over the last 1,000 norms, capped at 1. During the first 2,000 steps the scene aggregator linearly transitions its appearance-mean term from the inherited mean to the depth-weighted mean of Eq. 2. In object refinement the image and spatial conditions are dropped jointly.

Table 23: **Training configuration of the three learned modules.** The pretrained flow weights, image encoders, sparse-structure autoencoders, and geometry and texture generators stay frozen. LoRA is applied to the attention projections of both flows with rank 8 and $\alpha = 16$. Condition modes retain the observed-surface flow source in all cases; scene probabilities describe the retained conditions.

Setting	Scene predictor	Membership classifier	Object refinement
Trained parameters	aggregator, depth branch, conditioning projections, LoRA	all	spatial adapter, LoRA
Objective	flow matching (Eq. 10)	focal + Dice per decoder layer (Eq. 11)	flow matching (Eq. 10)
Time sampling	$\mathrm{logit}(t) \sim \mathcal{N}(1, 1)$	–	$t \sim \mathcal{U}[0, 1]$
Supervision frame	ground-truth scene occupancy	ground-truth scene occupancy with object assignments	ground-truth object and context supports
Learning rate	10^{-4} (LoRA 3×10^{-5})	5×10^{-5}	10^{-4} (LoRA 3×10^{-5})
Schedule	2,000 warmup, constant	500 warmup, cosine to 10^{-6}	2,000 warmup, constant
Input views	1–8	1–8	1–8
Condition modes	image+spatial 0.7 / spatial 0.2 / none 0.1	no dropout	joint dropout 0.1
Evaluated checkpoint	36,000 steps (EMA)	7,000 steps	12,000 steps
Effective batch	64 chunks	64 scenes	64 objects
AdamW betas	$(0.9, 0.95)$	$(0.9, 0.95)$	$(0.9, 0.95)$
Weight decay	matrix weights 0.01; bias/norm/LoRA 0	0.05	spatial 0.01; LoRA 0
EMA decay	0.9995	disabled	disabled

C.4 Inference Configuration and Cost

Table 24 summarizes the inference settings.

Table 24: **Inference configuration.** Settings for the main two-view budget. Scene prediction and membership are shared across the target objects; structure refinement and geometry generation operate per object.

Stage	Model	Steps	Guidance	Conditions and output
Scene structure	scene predictor	12 Euler	3	views and observed-surface source; scene occupancy
Object membership	classifier	–	–	threshold $\tau = 0.70$ on validation; coarse object voxels
Object structure	refinement module	25 Euler	1	image and spatial conditions; 64^3 occupancy per object
Object geometry	frozen TRELLIS.2 (1024 configuration)	12	7.5	refined occupied cells and the 512×512 primary-view crop; mesh mapped from the local chunk to the scene

Stage-wise inference time. Table 25 reports the cost of CoVoGen at the main two-view budget. Each scene is processed sequentially on one NVIDIA A800 80 GB GPU. We measure synchronized wall time after warmup, with each stage's models resident on the GPU. Timing includes observation preprocessing, conditioning, generation, and native mesh hole filling, and excludes model loading and unloading, file I/O, mesh export, and evaluation. Object conditioning includes local-chunk preparation and image-feature extraction. Scene-level times sum the work for all target objects; the amortized object cost divides the total time by the total number of targets, including their share of scene prediction and membership. The complete pipeline takes 168.48 s per scene, or 20.57 s per target on this cohort.

Table 25: **Stage-wise inference cost of CoVoGen.** Two-view ARSG inputs on one A800 GPU. Statistics use the 47 scenes (385 targets) with complete outputs in the 50-scene timing cohort. Each stage's scene time includes all targets; object costs amortize that time over the targets.

Stage	Mean s/scene	Amortized s/object	Share (%)
Observation preprocessing	0.37	0.05	0.2
Scene structure prediction	56.43	6.89	33.5
Object membership	1.00	0.12	0.6
Object conditioning	31.12	3.80	18.5
Object structure refinement	39.73	4.85	23.6
Object geometry generation	39.83	4.86	23.6
Total	168.48	20.57	100.0

For the multi-view timing comparison in Figure 3, scene prediction and object conditioning admit all supplied views, including at the 16-view budget; geometry generation retains its primary-view crop. All curves use one common set of scenes with complete outputs across the compared methods and view budgets. Confidence intervals resample scenes jointly across methods and budgets; object costs divide the resampled total time by the resampled target count.

Multi-room display. Figure 1 uses a constructed two-room scene with 46 object instances and a pool of 215 posed RGB-D views. Object refinement uses 25–117 admitted views per target with 512×512 image inputs. Walls and floor are reference geometry; all furniture meshes are generated. The four input panels are frames from the same view pool, shown with their instance masks tinted in the object colors of the render and with observed depth on a light-to-dark scale; the overview uses a perspective camera with the near walls and the room partition cut away, and the four isolated meshes at the right are the same generated meshes rendered from their own viewpoints above a shadow-catching ground plane.

Figure 4 uses two further constructed layouts: a bedroom and study with 45 objects, and a living room and bedroom with 44. Their view pools contain 230 and 214 posed RGB-D frames, with 22–99 and 24–102 admitted views per target, respectively. Both use 512×512 model inputs. All furniture and tabletop objects are newly generated; reference walls and doorframes show the room boundaries. Near walls are cut away, with their wall-mounted artwork omitted together with the wall. All generated outputs are retained; these display choices do not affect quantitative results.

Qualitative scene timing. The times in Figure 5 are measured separately for its four scenes using the same input observations and model settings as the displayed results: all 11 observations of each 3D-FRONT room and all 30 of each ARSG scene. We measure one repeat after warmup and sum synchronized single-A800 inference costs across all original target objects. These totals represent serial inference cost, rather than the elapsed time of parallel object processing. They include observation preprocessing, conditioning, generation, placement, and required geometry postprocessing; CoVoGen also includes the official TRELLIS.2 hole filling and remeshing used in the figure. Model loading and unloading, file I/O, mesh export, rendering, and evaluation are excluded.

C.5 Training Cost, Scene Size, and Memory

Table 26 summarizes the cost of training the three learned modules to the checkpoints used in our experiments. Figure 8 shows how the measured inference costs vary with target count and scene-chunk count on the same two-view cohort.

Table 26: **Training cost to the evaluated checkpoints.** Approximate A800 GPU-hours from recorded production-training spans, including scheduled validation and checkpointing. Pretrained-backbone training, data preparation, and development trials are excluded.

Module	Training steps	A800 GPUs	GPU-hours
Scene predictor	36,000	8	423
Membership classifier	7,000	8	127
Object refinement	12,000	$8 \rightarrow 16$	555
Total	–	–	**1,105**

Object refinement uses eight GPUs for the first 4,000 updates and sixteen for the remaining 8,000; costs are weighted by the active GPU count. All GPUs are NVIDIA A800 80 GB.

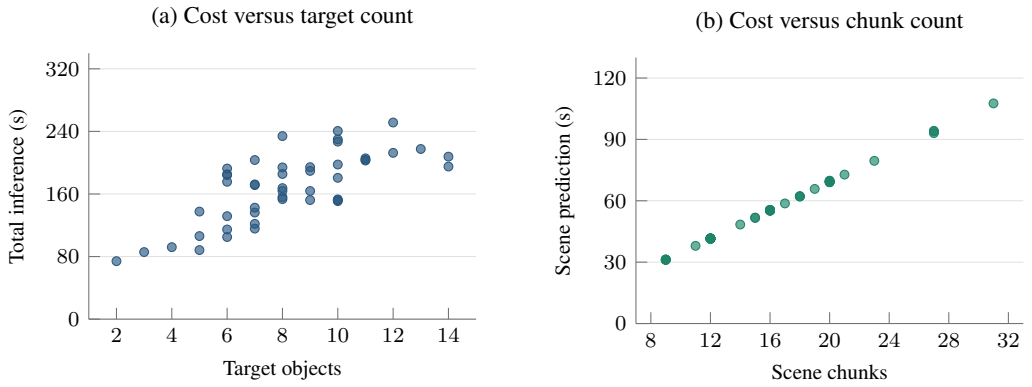

Figure 8: **Inference cost across scene sizes.** Each point is one of the same 47 two-view ARSG scenes (385 targets) used in Table 25, processed sequentially on one A800. (a) Complete-pipeline time versus target count. (b) Scene-prediction time versus the number of processed scene chunks. Scene content varies across points; these are observed costs rather than a controlled expansion of one scene. Timing uses the same exclusions as Table 25.

Peak inference memory. Table 27 reports device memory for five two-view ARSG scenes selected at the minimum, median, and maximum target counts and the minimum and maximum scene-chunk counts of the 50-scene timing cohort. Modules run sequentially, releasing their GPU models between stages. We sample device memory using NVML at nominal 10 ms intervals and retain the maximum across all inference stages for each scene. The highest observed peak is 10.90 GiB, reached during mesh generation; including model loading and warmup leaves this maximum unchanged.

Table 27: **Peak GPU memory across scene sizes.** Each row is one two-view ARSG scene processed on a single A800. Peaks cover the complete inference pipeline.

Target objects	Scene chunks	Device peak (GiB)
2	15	5.95
4	8	7.21
6	31	7.92
8	20	8.52
14	12	10.90

Target counts include every target of each scene.

D COMPLETE MULTI-ROOM GENERATIONS



Figure 9: **Living room and dining room: complete generated scene and all 46 object instances.** The overview shows the spatial arrangement; below, every generated instance appears once from a selected view, grouped by room and colored consistently with the overview. Individual meshes are scaled for legibility. Walls and floor are reference geometry. Display settings follow Appendix C.4.

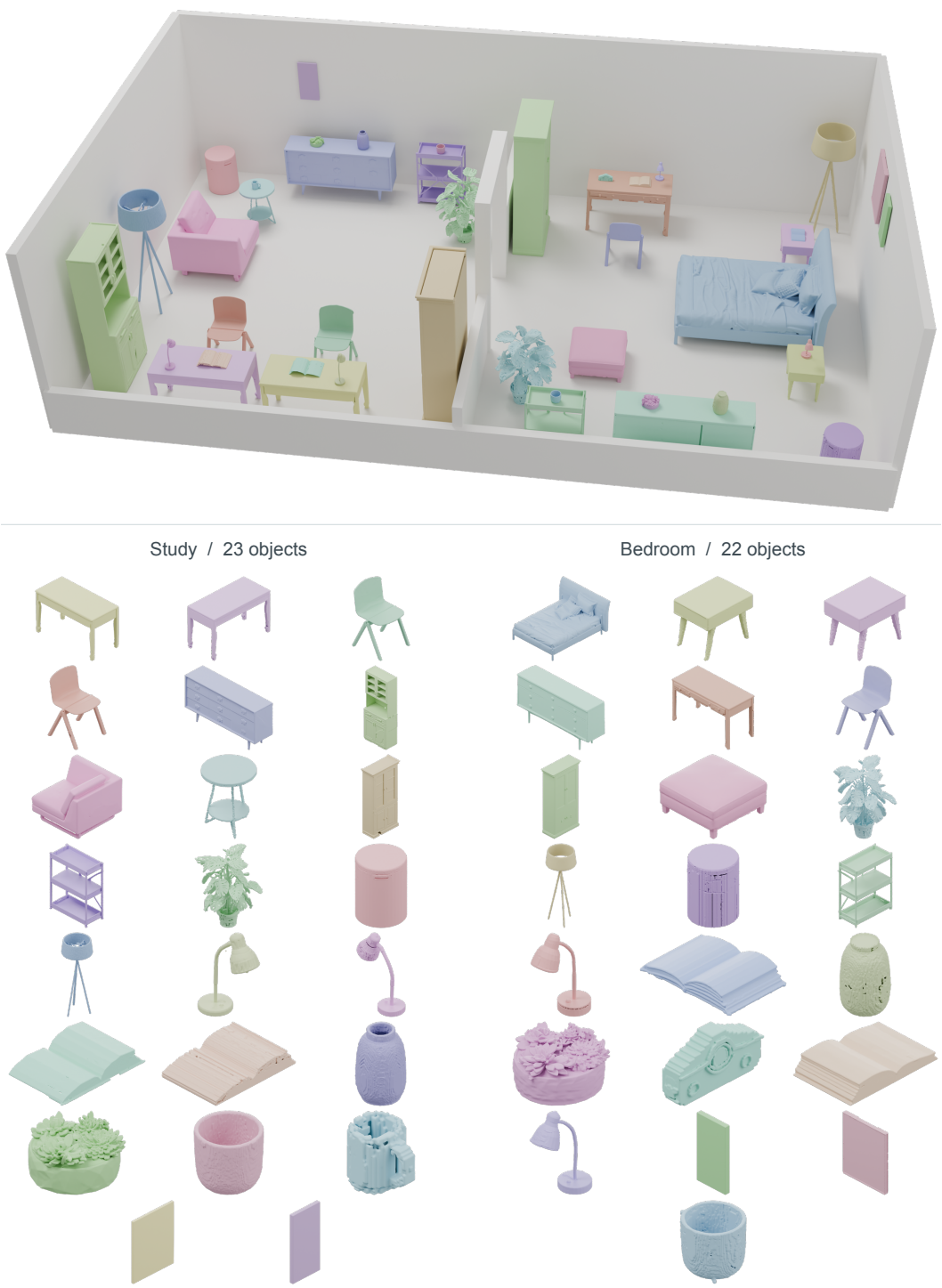

Figure 10: **Study and bedroom: complete generated scene and all 45 object instances.** The overview shows the spatial arrangement; below, every generated instance appears once from a selected view, grouped by room and colored consistently with the overview. Individual meshes are scaled for legibility. Wall-mounted artwork hidden by the overview's architectural cutaway is included below. Reference architecture and display processing follow Appendix C.4.

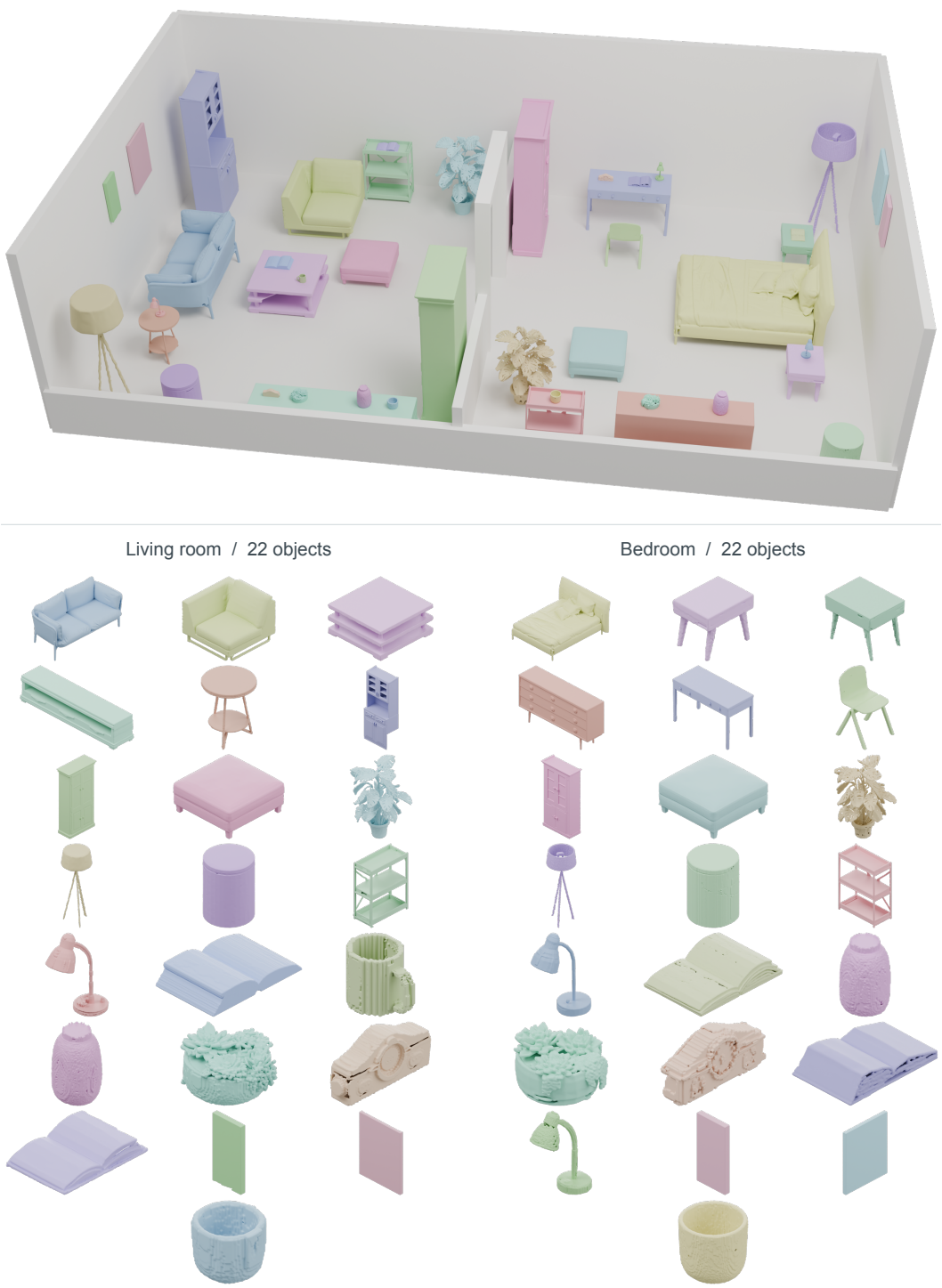

Figure 11: **Living room and bedroom: complete generated scene and all 44 object instances.** The overview shows the spatial arrangement; below, every generated instance appears once from a selected view, grouped by room and colored consistently with the overview. Individual meshes are scaled for legibility. Reference architecture and display processing follow Appendix C.4.