WORLDSCULPT: GENERATING COMPOSITIONAL WORLDS FROM GROUNDED VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating large-scale cluttered scenes as individually editable object meshes is challenging. Existing compositional methods largely target simple scenes. Our key insight is that, with appropriate design, a single-object generative prior can scale to large-scale, densely cluttered scenes without scene-level training. We introduce **WorldSculpt**, which realizes this idea through canonical-space multi-view conditioning, training-time augmentation, and object-wise scene composition. Finetuned entirely on single-object data, it generates hundreds of mutually occluding objects as complete, spatially aligned meshes in a shared world frame. We further introduce a photorealistic benchmark of densely cluttered scenes with per-object annotations. Evaluations demonstrate superior geometric accuracy and observation alignment over prior approaches. We also demonstrate broader applicability by converting 3DGS worlds from Marble into compositional scenes.

1 INTRODUCTION

Generative models Xiang et al. (2025); Wan et al. (2025); HaCohen et al. (2026); Niu et al. (2024b; 2025); Shi et al. (2026a) such as Marble [1] can synthesize explorable 3D environments. Yet these worlds are typically represented as fused scene geometry or Gaussian splats, rather than collections of discrete, editable objects. Such representations do not directly support the object-level manipulation required by downstream applications including gaming, simulation, and robotics. A compositional representation should instead provide each object as an independent asset, with complete geometry and placement consistent with the observed scene.

Recent compositional methods Huang et al. (2025b); Meng et al. (2026); Shi et al. (2026b) generate scenes as collections of individual object assets. Much of this progress has focused on relatively small-scale settings, such as tabletop arrangements and indoor furniture layouts. In densely cluttered environments, however, objects heavily occlude one another, and the evidence needed to generate each object is distributed across multiple, individually incomplete observations. The challenge is therefore not merely to assemble more objects, but to recover complete object geometry from fragmented evidence while maintaining alignment with the scene. Single-object 3D generative priors offer a promising foundation for completing unobserved geometry Xiang et al. (2025); Zhao et al. (2025b), and Pixal3D Li et al. (2026b) has also discussed scene composition from single images. These developments motivate our central question: how can single-object priors be effectively adapted to densely cluttered large-scale scenes with multi-view observations?

Our key idea is simple: keep geometry learning object-centric while adapting its conditioning to scene observations. Given object grounding, scene generation can be decomposed into local generation problems whose outputs are composed in a shared world frame. Unlike clean object renders, observations within cluttered scenes contain view-dependent occlusion and limited image detail. Transferring a single-object prior therefore requires both *a geometrically consistent connection between scene observations and canonical-space generation*, and *adaptation to the incomplete and degraded evidence encountered in scenes*.

We introduce **WorldSculpt**, a framework that addresses these two requirements in simple yet effective ways. The framework first constructs an anchor-aligned canonical frame for each object and derives world-canonical-camera transformations that preserve its geometric relationship with the input views. Features from multiple observations are lifted into a common canonical voxel grid and fused

[1] https://marble.worldlabs.ai/

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

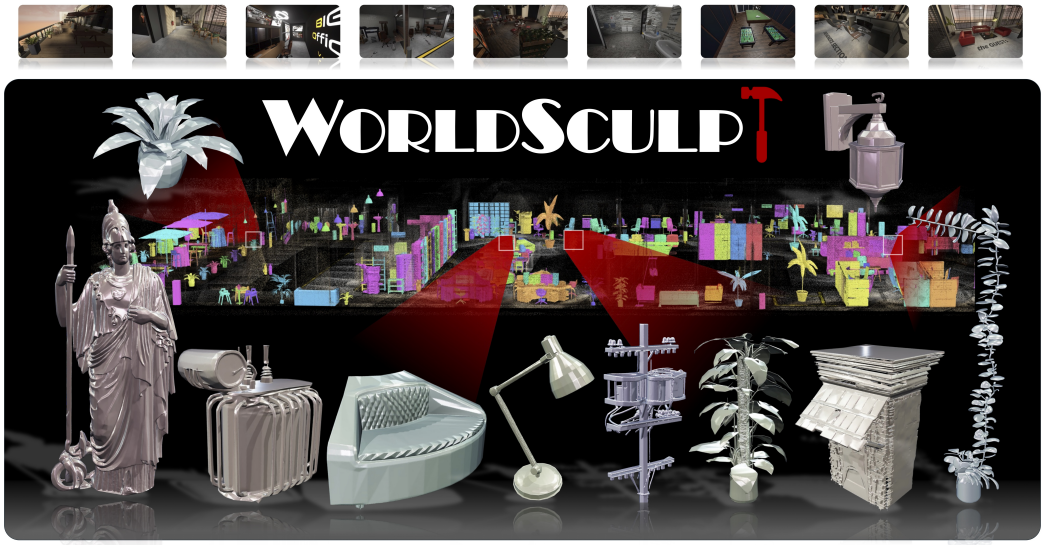

Figure 1: Given posed RGB images, 2D instance masks, and coarse 3D localization boxes, **WorldSculpt** generates densely cluttered scenes as collections of individually editable object meshes. Finetuned entirely on single objects, it generalizes to scenes containing hundreds of mutually occluding objects without scene-level training.

through a learned aggregator. The resulting representation conditions a pretrained Pixal3D generator, allowing complementary observations to guide the generation of complete geometry. Generated meshes are then mapped back to the shared world frame through their canonical-to-world transformations. To bridge the gap between object-centric training data and scene observations, we finetune the model on single objects with augmentations that simulate occlusion, camera inaccuracies, segmentation errors, and reduced resolution. This allows the model to learn from partial and degraded evidence without requiring scene-level geometry supervision, generating densely cluttered scenes containing hundreds of objects while remaining trained entirely at the level of individual objects.

To evaluate this approach at scale, we introduce UE-MeshyScene, a photorealistic benchmark of large indoor and outdoor environments containing hundreds of objects with substantial mutual occlusion. The benchmark provides precise per-object geometric annotations, enabling evaluation of both object geometry and spatial alignment. Experiments across four datasets demonstrate clear improvements over existing approaches. Ablations further highlight the importance of our designs. We also demonstrate broader applicability by converting Marble-generated 3DGS worlds into compositional collections of individual object meshes. In summary, our contributions are:

- We show that single-object generative priors could effectively support compositional generation of large-scale, densely cluttered scenes from multi-view observations, without scene-level training.
- We introduce WorldSculpt, a grounded compositional generation framework that connects complex scene observations to a single-object generative prior.
- We introduce *UE-MeshyScene*, a photorealistic benchmark of densely cluttered scenes with hundreds of objects and precise per-object geometric annotations.

2 Related Work

Images to 3D generation. Feed-forward models directly regress 3D representations from one or few images Hong et al. (2024); Wei et al. (2024); Xu et al. (2024a;b); Tang et al. (2024); Tochilkin et al. (2024); Boss et al. (2025). Another line first generates novel views with 2D diffusion priors Liu et al. (2023b; 2024b); Long et al. (2024); Wang et al. (2024); Liu et al. (2023a), before fusing them into 3D. Single-object 3D models Nichol et al. (2022); Jun & Nichol (2023); Zhao et al. (2023); Xiang et al. (2025); Zhao et al. (2025b); Li et al. (2026b) instead directly learn shape distributions using large-scale datasets Deitke et al. (2023b;a); Stojanov et al. (2021); Downs et al. (2022); Collins et al. (2022). Amodal reconstruction Agnew et al. (2021); Ozguroglu et al. (2024); Wu et al. (2025); Zhou & Tai (2025) seeks complete object geometry from occluded observations.

Compositional generation. Recent work jointly generates object geometry and spatial layouts from single or multiple images Liu et al. (2022); Zhao et al. (2025a); Lin et al. (2026b); Huang et al. (2025b); Zhang et al. (2026); Agarwal et al. (2025); Yin et al. (2026); Ardelean et al. (2025); Chen et al. (2026a;b); Meng et al. (2026); Shi et al. (2026b); Zadaianchuk et al. (2026). While these methods enable compositional 3D generation, they primarily focus on relatively small or simple scenes, such as tabletop arrangements and indoor furniture layouts.

Structures and motion from videos. Recovering object-level 3D representations from raw imagery relies on established components. Camera poses can be estimated from videos Schönberger & Frahm (2016); Schönberger et al. (2016); Pan et al. (2024); Teed & Deng (2021); Teed et al. (2023); Li et al. (2025); Huang et al. (2025a). Scene geometry can be inferred using depth and pointmap models Yang et al. (2024); Lin et al. (2025); Piccinelli et al. (2024); Wang et al. (2025b); Hu et al. (2025); Chen et al. (2025); Leroy et al. (2024); Murai et al. (2025); Zhang et al. (2025b); Yang et al. (2025); Wang et al. (2025a); Keetha et al. (2026); Wang et al. (2025c). Objects can be tracked through video segmentation and open-vocabulary detection Carion et al. (2025); Ren et al. (2024); Wu et al. (2024); Cheng et al. (2024b), and localized in 3D Brazil et al. (2023); Rukhovich et al. (2022); Zhang et al. (2025a); Lazarow et al. (2025); DeTone et al. (2026).

3 WorldSculpt

Problem setup. Given N posed images $\{I_n\}_{n=1}^{N}$ with known intrinsics $\{K_n\}_{n=1}^{N}$ and camera-to-world extrinsics $\{T_n^{\mathrm{cw}}\}_{n=1}^{N}$, we assume each object k has instance masks S_{kn} in views where it is visible Carion et al. (2025); Ravi et al. (2025); Cheng et al. (2023; 2024a); Cheng & Schwing (2022); Wu et al. (2024); Liu et al. (2024a); Ren et al. (2024); Cheng et al. (2024b) and a coarse world-space localization box B_k^{loc} Brazil et al. (2023); Rukhovich et al. (2022); Zhang et al. (2025a); Lazarow et al. (2025); DeTone et al. (2026); Lemeshko et al. (2026); Yang & Scherer (2019); Li et al. (2021); Wen et al. (2023). Estimating these inputs is well studied and outside our scope. Importantly, B_k^{loc} does not need to be oriented or precise, as we only use it for initialization. Robustness to annotation errors is evaluated in Sec. 4.4. For K objects, we seek the compositional representation

$$\begin{aligned} \mathcal{M} &= \{(\mathcal{M}_k^{\mathrm{c}}, T_k^{\mathrm{ow}})\}_{k=1}^{K}, \\ \mathcal{M}_k^{\mathrm{w}} &= T_k^{\mathrm{ow}}\,(\mathcal{M}_k^{\mathrm{c}}), \end{aligned} \tag{1}$$

where $\mathcal{M}_k^{\mathrm{c}}$ is a canonical-space object mesh, T_k^{ow} maps it to world space, and $\mathcal{M}_k^{\mathrm{w}}$ is the resulting world-space mesh. WorldSculpt maps each object's scene observations into an anchor-aligned virtual canonical frame, generates its mesh, and transforms it back to world space.

3.1 Anchor-aligned object canonicalization

Anchor view and canonical cube. Starting with the coarse localization box B_k^{loc}, we construct an anchor-aligned virtual canonical cube for Pixal3D. Let $\mathcal{I}_k$ contain the views observing object k. We choose an anchor $a_k \in \mathcal{I}_k$ randomly during training and as the eligible view with the largest foreground mask area at inference. For the canonical domain $\mathcal{V} = [-\frac{1}{2}, \frac{1}{2}]^3$, we use the box center c_k, initialize the scale s_k to its largest side length, and derive $R_k \in \mathrm{SO}(3)$ from the anchor camera orientation to match Pixal3D's canonical front-view convention. The canonical-to-world transformation is

$$T_k^{\mathrm{ow}} = \begin{bmatrix} s_k R_k & c_k \\ 0 & 1 \end{bmatrix}, \tag{2}$$

which maps $x \in \mathcal{V}$ to $x_k^{\mathrm{w}} = s_k R_k x + c_k$. This similarity transformation preserves object proportions through isotropic scaling. To accommodate imperfect localization boxes, we adjust s_k based on the projected cube bounds and observed masks.

Canonical observations and crop-aware projection. For each selected view $n \in \mathcal{I}_k$, we crop I_n to the projected cube extent, remove the background using S_{kn}, and resize to the model's input resolution, obtaining $\bar{I}_{kn}$. Let A_{kn} map original image coordinates to resized-crop coordinates, giving adjusted intrinsics $\bar{K}_{kn} = A_{kn} K_n$. A canonical voxel x, with homogeneous coordinate $\tilde{x} = [x^\top, 1]^\top$, projects into the crop as

$$\pi_{kn}(x) = \Pi\Big(\bar{K}_{kn}\,[I_3 \quad 0]\,(T_n^{\mathrm{cw}})^{-1}\,T_k^{\mathrm{ow}}\tilde{x}\Big), \tag{3}$$

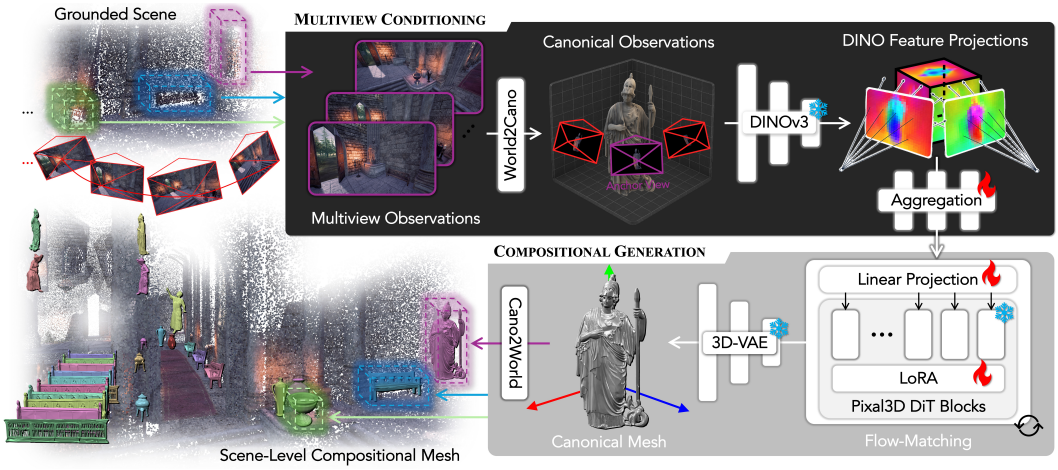

Figure 2: **Overview of WorldSculpt.** The anchor-aligned canonicalization stage constructs a virtual canonical cube for each object and derives crop-aware projections from its input views. Per-view DINOv3 features are lifted into canonical voxel grids and fused by multi-view aggregation. The fused features condition Pixal3D via projected conditioning and LoRA adaptation. Generated meshes are placed in the scene using canonical-to-world transformations.

where $\Pi([u, v, d]^\top) = (u/d, v/d)^\top$ is perspective division. Further details on anchor selection, cube adjustment, and crop construction are provided in Appendix B.

3.2 MULTI-VIEW CONDITIONED OBJECT GENERATION

Generative prior. We instantiate the per-object generator Φ with Pixal3D Li et al. (2026b), a native-3D model built on the TRELLIS.2 structured-latent backbone Xiang et al. (2026) with cascaded flow-matching stages Lipman et al. (2022). We use its first two geometry stages: the sparse-structure stage predicts occupancy on a 64^3 grid, and the shape stage generates a high-resolution latent over occupied voxels, subsequently decoded into a mesh.

Multi-view feature lifting. We independently encode each crop with DINOv3 Siméoni et al. (2025) and sample its dense features at projected voxel locations:

$$\begin{aligned} F_{kn} &= E_{\mathrm{DINO}}(\bar{I}_{kn}), \\ g_n(x) &= F_{kn}\left(\pi_{kn}(x)\right). \end{aligned} \tag{4}$$

This lifts per-view features into spatially aligned canonical 3D grids before cross-view aggregation.

Multi-view feature aggregation. We fuse lifted features from selected views $\mathcal{J}_k \subseteq \mathcal{I}_k$ using an IBRNet-style aggregator Wang et al. (2021); Schmid et al. (2026) that supports variable view counts. At each voxel x, each view's feature is concatenated with the cross-view mean $\mu(x)$ and element-wise variance. Two lightweight MLPs predict residual features $g'_n(x)$ and aggregation logits, respectively. Softmax-normalizing the logits across views gives weights $\alpha_n(x)$, yielding

$$g_{\mathrm{out}}(x) = \mu(x) + \sum_{n \in \mathcal{J}_k} \alpha_n(x) g'_n(x). \tag{5}$$

Applying this operation voxel-wise produces the 3D conditioning grid G_k. For the DINOv3 global feature, only the anchor crop $\bar{I}_{ka_k}$ is provided to Pixal3D.

Conditioning injection. At every transformer block in both geometry stages, we align G_k with the current 3D token locations, project it to the block feature dimension, and add it to the corresponding features. We freeze the original Pixal3D weights and adapt its DiT attention and per-block projection layers with LoRA Hu et al. (2022). Only the LoRA adapters and the multi-view aggregation MLPs are optimized. Details about feature extraction, aggregation architecture, and model configuration are detailed in Appendix C.

3.3 TRAINING

We train the sparse-structure and shape stages independently. For each object, we sample a variable number of conditioning views and randomly choose an anchor from them.

Table 2: **Quantitative comparison results on three public datasets.** $\bar{\text{CD}}_2$ is reported in 10^{-3} units; $\bar{\text{CD}}_1$ and EMD are reported in 10^{-2} units. F-scores are unscaled. The top three values in each column within each section are marked as 1st, 2nd, and 3rd. † Our single-view results use the same model without re-training.

Method	3D-Front						HouseCat6D						Toys4k-Scene					
	$\bar{\text{CD}}_2$	$\bar{\text{CD}}_1$	EMD	F@.05	F@.02	F@.01	$\bar{\text{CD}}_2$	$\bar{\text{CD}}_1$	EMD	F@.05	F@.02	F@.01	$\bar{\text{CD}}_2$	$\bar{\text{CD}}_1$	EMD	F@.05	F@.02	F@.01
Single-view input																		
SAM3D	7.91	8.71	10.62	0.680	0.364	0.184	9.01	7.93	9.07	0.764	0.447	0.250	9.14	9.09	9.99	0.670	0.356	0.194
RecGen	7.28	7.03	8.28	0.792	0.479	0.262	8.04	7.08	7.87	0.816	0.479	0.262	5.16	6.48	7.78	0.798	0.477	0.269
MIDI	25.93	16.06	15.06	0.441	0.231	0.118	30.07	17.83	16.24	0.378	0.178	0.090	20.14	14.15	13.51	0.468	0.231	0.122
SceneGen	27.76	16.80	15.69	0.420	0.207	0.105	29.25	17.14	15.32	0.408	0.199	0.102	19.19	13.64	12.92	0.502	0.254	0.134
SceneMaker	24.95	14.20	15.30	0.540	0.283	0.145	34.49	18.92	18.24	0.379	0.175	0.088	36.09	19.48	18.91	0.375	0.174	0.089
DepR	13.26	10.89	11.52	0.600	0.320	0.167	31.82	19.47	17.91	0.300	0.131	0.067	22.64	15.55	15.00	0.390	0.175	0.090
PixARMesh	6.42	6.63	9.57	0.786	0.569	0.367	6.51	8.16	11.49	0.681	0.370	0.195	11.15	11.10	12.82	0.543	0.258	0.135
SceneComplete	7.24	7.22	8.29	0.776	0.494	0.275	6.59	6.80	7.29	0.807	0.553	0.309	8.79	8.49	8.90	0.708	0.434	0.246
3D-Fixer	9.97	7.66	9.46	0.775	0.513	0.291	14.18	7.88	8.97	0.792	0.500	0.281	5.25	6.18	7.76	0.821	0.528	0.308
Cupid	48.46	14.53	15.50	0.621	0.395	0.223	35.38	11.39	12.37	0.729	0.465	0.259	8.19	8.07	8.92	0.712	0.404	0.224
Gen3DSR	11.58	10.44	11.23	0.614	0.284	0.134	10.04	9.59	9.61	0.646	0.323	0.163	13.12	10.68	10.98	0.609	0.300	0.151
OVOW	33.94	11.29	12.72	0.721	0.451	0.247	8.70	7.40	8.11	0.798	0.470	0.261	18.76	10.21	11.02	0.697	0.403	0.226
Ours†	2.14	3.30	6.08	0.931	0.805	0.598	2.22	3.90	4.60	0.910	0.718	0.489	2.12	3.84	5.32	0.905	0.714	0.509
Multi-view input																		
MV-SAM3D	6.23	7.40	9.73	0.751	0.450	0.241	13.32	11.32	12.43	0.591	0.295	0.151	7.04	7.46	9.08	0.755	0.439	0.246
Mix3R	14.03	10.34	11.39	0.643	0.413	0.232	16.51	11.40	11.38	0.620	0.368	0.199	9.44	8.66	8.97	0.687	0.435	0.254
Stream3D	1.20	3.28	5.37	0.951	0.750	0.468	0.82	3.01	4.31	0.970	0.753	0.477	2.70	4.72	5.74	0.881	0.611	0.375
ShapeR	8.07	4.13	6.83	0.924	0.823	0.626	1.26	1.88	4.77	0.973	0.943	0.823	8.38	7.81	9.41	0.746	0.460	0.275
Ours	0.71	1.87	4.70	0.970	0.914	0.799	0.18	1.23	2.66	0.998	0.957	0.862	0.40	1.30	3.02	0.989	0.952	0.865

Conditioning-view augmentation. To bridge clean object renders and partial, degraded scene observations, we apply an augmentation curriculum to conditioning input images. We independently sample four degradations: (i) occlusion using random 3D occluders between the camera and object; (ii) camera parameter perturbations; (iii) segmentation errors through perturbed and degraded mask boundaries; and (iv) downsampling input images. Full details are reported in Appendix D.

Implementation details. All training is performed using *16* NVIDIA H100 GPUs. We initialize both geometry stages from released Pixal3D checkpoints and finetune them on the TexVerse dataset Zhang et al. (2025c). Each stage is trained for 15k iterations. Batch sizes are 128 for sparse-structure and 32 for shape. Each iteration samples $1 \sim 20$ conditioning views. The total training time takes about 2 days. More details are provided in Appendix C.

Table 1: **Occlusion ratio of each dataset.** Details for calculation are in Appendix A.6.

Dataset	Visibility	Occlusion
Housecat6D	99.66%	0.34%
Toys4K-Scene	88.18%	11.82%
3D-FRONT	85.82%	14.18%
UE-MeshyScene	51.86%	48.14%

3.4 Generation and Scene Composition

The inference pipeline is illustrated in Figure 2. At inference, we process each object independently. Following Sections 3.1 and 3.2, WorldSculpt generates a canonical mesh $\mathcal{M}_k^{\text{c}}$, which is transformed into world coordinates using T_k^{ow} from Equation 2. The final compositional scene is

$$\mathcal{M}^{\text{w}} = \{T_k^{\text{ow}}(\mathcal{M}_k^{\text{c}})\}_{k=1}^{K}. \tag{6}$$

More details about the inference pipeline are provided in Appendix B.

4 Experiments

4.1 Dataset

We first assess object geometry and spatial alignment on public small-scale datasets, including 3D-FRONT Fu et al. (2021), Toys4k-Scene Stojanov et al. (2021), and HouseCat6D Jung et al. (2024). We then evaluate scalability to hundreds of objects on the proposed UE-MeshyScene dataset.

Small-scale scene. We use three complementary datasets. 3D-FRONT Fu et al. (2021) provides synthetic furniture scenes, from which we select 50 rooms for evaluation. Toys4k-Scene is constructed by arranging objects from Toys4k Stojanov et al. (2021) into cluttered layouts and rendering multi-view observations. It combines intricate object geometry with substantial inter-object occlusion while retaining complete per-object mesh ground truth. HouseCat6D Jung et al. (2024) provides real tabletop captures with scanned object meshes.

Large-scale scene. To evaluate generation results at larger scales, we introduce *UE-MeshyScene*, comprising six indoor and outdoor environments authored and rendered in Unreal Engine 5.8. The

Table 3: **Input-view visible and amodal silhouette IoU (%) on three public datasets.** The top three values within each input setting in each row are marked as 1st, 2nd, and 3rd. † Our single-view results use the same multi-view-trained model with a single input image.

Dataset	Single-view input													Multi-view input				
	SAM3D	RecGen	MIDI	SceGen	SceMak	DepR	PixARM	SceCom	3D-Fixer	Cupid	Gen3DSR	OVOW	Ours†	MV-SAM3D	Mix3R	Stream3D	ShapeR	Ours
Visible silhouette IoU																		
3D-Front	53.99	65.01	64.58	29.23	48.41	51.53	36.76	39.13	66.55	53.45	67.31	64.04	89.93	54.00	35.46	41.36	78.62	89.54
HouseCat6D	68.11	70.75	40.09	12.52	44.11	23.82	48.14	35.91	71.67	60.60	76.63	75.30	92.16	60.18	32.97	46.61	87.14	92.64
Toys4K-Scene	58.90	69.65	26.88	10.97	37.12	32.45	20.23	25.95	73.58	62.08	67.48	61.98	92.97	58.62	30.09	32.50	65.00	87.28
Amodal silhouette IoU																		
3D-Front	59.99	66.70	65.68	31.93	50.12	52.38	40.28	41.28	66.88	54.40	67.30	65.06	89.14	60.03	38.83	46.48	83.00	91.83
HouseCat6D	71.23	72.96	41.62	14.32	45.49	23.88	49.46	38.19	73.96	64.26	78.95	77.33	95.40	63.51	37.71	48.26	91.05	95.16
Toys4K-Scene	61.56	71.06	29.47	14.29	38.95	32.95	24.43	28.45	74.65	62.83	68.52	64.00	92.69	63.07	37.05	36.06	69.20	91.17

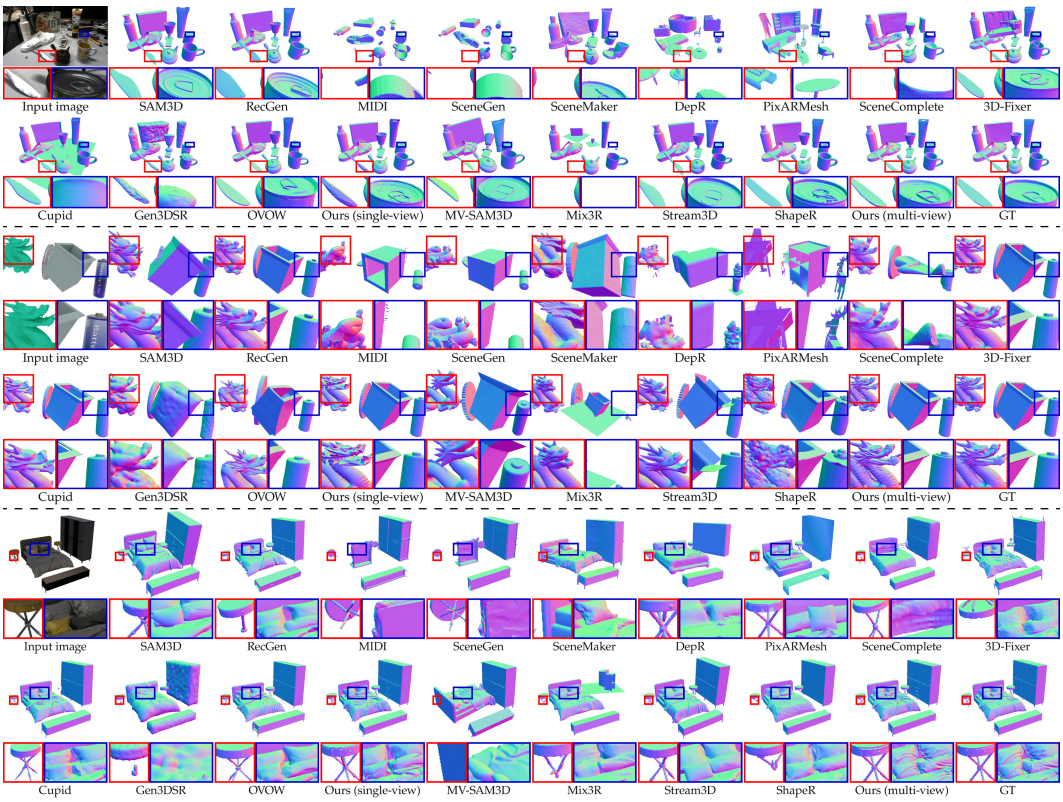

Figure 3: **Qualitative results on three public datasets.** From top to bottom: HouseCat6D, Toys4k-Scene, and 3D-FRONT. Our single-view results use the same model without re-training.

benchmark contains 2,299 objects, with $93 \sim 701$ objects per scene, arranged across shelves, desks, floors, and other cluttered regions with substantial mutual occlusion. Smooth camera trajectories yield $265 \sim 1{,}758$ frames per scene and 5,964 frames in total, rendered at 2560×1440 resolution. Annotations include camera parameters, instance masks, object-level 3D bounding boxes, and metric depth maps. Our method uses RGB images and grounding annotations but does not use depth. Table 10 summarizes scene statistics, and Figure 6 shows representative examples. As shown in Table 1, UE-MeshyScene has a much larger occlusion ratio than the other three datasets.

4.2 Small-scale scene evaluation

We compare WorldSculpt with state-of-the-art methods, evaluating compositional scene generation on 3 small-scale public datasets, including 3D-FRONT Fu et al. (2021), Toys4k-Scene Stojanov et al. (2021), and HouseCat6D Jung et al. (2024).

Compared methods. We compare WorldSculpt with 16 different baseline methods. The single-view baselines include SAM3D Chen et al. (2026b), RecGen Zadaianchuk et al. (2026), SceneGen Meng et al. (2026), MIDI Huang et al. (2025b), SceneMaker Shi et al. (2026b), DepR Zhao et al. (2025a), PixARMesh Zhang et al. (2026), SceneComplete Agarwal et al. (2025), 3D-Fixer Yin

Table 4: **Comparison on UE-MeshyScene dataset.** $\overline{\text{CD}}_2$ is reported in 10^{-3} units; $\overline{\text{CD}}_1$ and EMD are reported in 10^{-2} units. The top three values in each column are marked as 1st, 2nd, and 3rd.

Method	$\overline{\text{CD}}_2 \downarrow$	$\overline{\text{CD}}_1 \downarrow$	EMD↓	F@.05↑	F@.02↑	F@.01↑	IoU (visible)↑	IoU (amodal)↑
Mix3R	150.25	29.49	27.65	0.289	0.137	0.073	11.30	21.35
STREAM3D	8.63	8.33	9.19	0.703	0.421	0.242	33.98	52.64
MV-SAM3D	8.54	7.96	8.93	0.734	0.469	0.282	35.11	56.42
ShapeR	7.42	6.37	8.80	0.813	0.562	0.359	49.99	67.51
Ours	2.99	2.78	5.15	0.943	0.872	0.738	70.89	84.92

Table 5: **Quantitative ablation study on four datasets.** The top three values for each metric within each dataset are marked as 1st, 2nd, and 3rd.

	3D-FRONT						*HouseCat6D*					
Method	$\overline{\text{CD}}_2 \downarrow$	$\overline{\text{CD}}_1 \downarrow$	EMD ↓	F@.05 ↑	F@.02 ↑	F@.01 ↑	$\overline{\text{CD}}_2 \downarrow$	$\overline{\text{CD}}_1 \downarrow$	EMD ↓	F@.05 ↑	F@.02 ↑	F@.01 ↑
Pixal3D-MV + our assemb.	1.108	2.101	5.150	0.966	0.894	0.764	0.355	1.406	2.896	0.989	0.951	0.844
Ours w/ avg. aggregation	0.822	1.872	4.935	0.973	0.915	0.799	0.192	1.323	2.967	0.998	0.954	0.839
Ours w/o augmentation	1.614	2.317	5.370	0.955	0.893	0.773	0.330	1.363	2.998	0.991	0.947	0.855
Ours - full	0.707	1.866	4.700	0.970	0.914	0.799	0.176	1.232	2.662	0.998	0.957	0.862
	Toys4K-Scene						*UE-MeshyScene*					
Method	$\overline{\text{CD}}_2 \downarrow$	$\overline{\text{CD}}_1 \downarrow$	EMD ↓	F@.05 ↑	F@.02 ↑	F@.01 ↑	$\overline{\text{CD}}_2 \downarrow$	$\overline{\text{CD}}_1 \downarrow$	EMD ↓	F@.05 ↑	F@.02 ↑	F@.01 ↑
Pixal3D-MV + our assemb.	0.661	1.683	3.298	0.979	0.920	0.804	4.599	3.862	6.262	0.903	0.796	0.646
Ours w/ avg. aggregation	0.474	1.550	3.294	0.985	0.931	0.819	3.264	2.920	5.283	0.939	0.866	0.729
Ours w/o augmentation	0.635	1.566	3.419	0.977	0.928	0.838	6.495	4.360	6.934	0.891	0.796	0.653
Ours - full	0.401	1.295	3.017	0.989	0.952	0.865	2.988	2.784	5.145	0.943	0.872	0.738

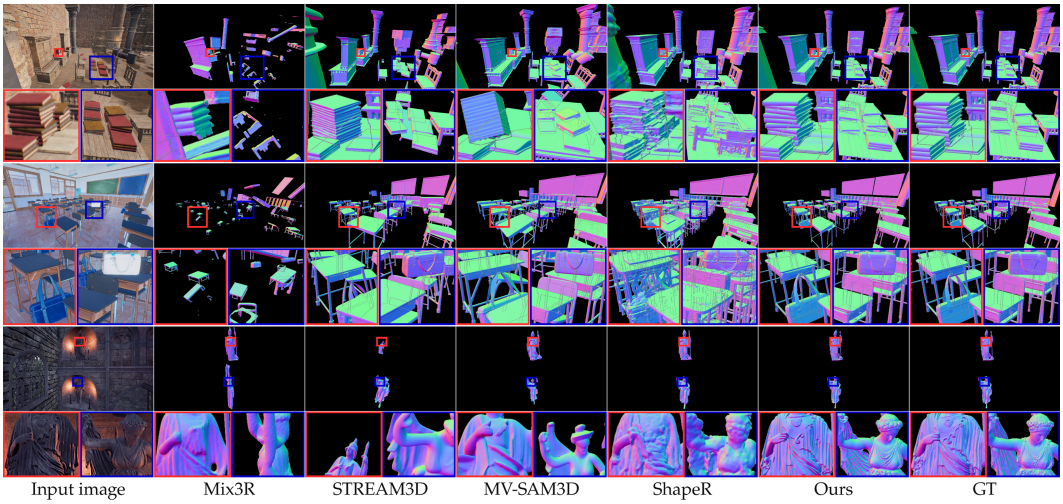

Figure 4: **Qualitative comparison results on the UE-MeshyScene dataset.**

et al. (2026), Cupid Huang et al. (2026), and Gen3DSR Ardelean et al. (2025). We also compare with OVOW Chen et al. (2026a), whose pipeline operates on videos. For static objects, however, OVOW generates mesh from a single selected anchor image, rather than jointly conditioning mesh generation on multiple observations. The multi-view baselines comprise MV-SAM3D Li et al. (2026a), Mix3R Lin et al. (2026a), Stream3D Zhou et al. (2026), and ShapeR Siddiqui et al. (2026), which exploit observations from multiple frames to generate object geometry. We provide all methods with GT annotations, including 2D instance masks, 3D boxes, and camera parameters. We also provide GT depth maps to methods that additionally require depth input.

Metrics. We evaluate each generated object against its corresponding ground-truth mesh in the shared world coordinate frame, jointly assessing geometric accuracy and spatial alignment. Geometric evaluation is normalized using each object's ground-truth bounding-box diagonal. We report the mean Chamfer distances, $\overline{\text{CD}}_p$ for $p \in \{1, 2\}$, Earth Mover's Distance (EMD), and F-scores at distance thresholds of 0.05, 0.02, and 0.01 in the normalized coordinate system. We additionally report input-view visible and amodal silhouette intersection over union (IoU). For visible IoU, all predicted instances are rendered jointly to account for inter-object occlusion, and the resulting visible masks are compared with ground-truth instance masks. For amodal IoU, each generated object and its corresponding GT mesh are rendered independently, without inter-object occlusion. Single-view methods are only evaluated on their single input frame, while multi-view methods are evaluated on all their input frames. Details are provided in Appendix E.

Table 6: **Robustness to grounding perturbations on four datasets.** For masks, mild/strong perturbations affect 5%/10% of the foreground; for boxes, they apply up to 5%/10% spatial perturbation; for cameras, they use rotation, translation, and FOV perturbations of $(3°, 0.03, 3\%)$ and $(5°, 0.05, 5\%)$, respectively, where translation is measured in canonical units. "w/ Aug." and "w/o Aug." denote models trained with and without the training-time augmentation.

Dataset		3D-Front				Housecat6D				Toys4K-Scene				UE-MeshyScene			
Metric		$\overline{\text{CD}}_2$	$\overline{\text{CD}}_1$	EMD	F@.05	$\overline{\text{CD}}_2$	$\overline{\text{CD}}_1$	EMD	F@.05	$\overline{\text{CD}}_2$	$\overline{\text{CD}}_1$	EMD	F@.05	$\overline{\text{CD}}_2$	$\overline{\text{CD}}_1$	EMD	F@.05
Clean		0.707	1.866	4.700	0.970	0.176	1.232	2.662	0.998	0.401	1.295	3.017	0.989	2.988	2.784	5.145	0.943
Mask (mild)	w/ Aug.	0.739	1.904	4.918	0.970	0.177	1.232	2.602	0.998	0.400	1.324	3.001	0.988	3.093	2.839	5.116	0.942
	w/o Aug.	1.630	2.414	5.738	0.951	0.612	1.652	3.304	0.980	1.148	1.862	3.792	0.969	6.634	4.146	7.790	0.882
Mask (strong)	w/ Aug.	0.754	1.885	4.936	0.970	0.178	1.252	2.710	0.998	0.452	1.296	3.153	0.990	3.129	2.852	5.181	0.942
	w/o Aug.	1.579	2.434	5.684	0.949	0.692	1.787	3.324	0.976	1.079	1.900	3.913	0.967	6.294	4.577	7.115	0.880
Box (mild)	w/ Aug.	0.696	1.926	4.826	0.968	0.179	1.232	2.847	0.998	0.438	1.258	2.920	0.991	3.066	2.850	5.290	0.941
	w/o Aug.	1.463	2.353	5.491	0.954	0.377	1.406	2.927	0.989	0.710	1.673	3.662	0.974	6.188	4.431	7.343	0.889
Box (strong)	w/ Aug.	0.855	2.043	5.216	0.967	0.182	1.249	2.718	0.998	0.452	1.298	3.015	0.989	3.086	2.883	5.316	0.940
	w/o Aug.	1.970	2.614	6.021	0.944	0.368	1.411	3.002	0.988	0.780	1.723	3.764	0.972	6.472	4.522	7.266	0.883
Cam (mild)	w/ Aug.	0.796	2.153	4.943	0.970	0.268	1.635	2.884	0.996	0.432	1.682	3.209	0.985	3.486	3.616	5.622	0.924
	w/o Aug.	4.018	3.565	6.526	0.924	0.676	2.094	3.433	0.981	1.445	2.383	4.141	0.961	11.352	7.207	8.420	0.792
Cam (strong)	w/ Aug.	0.878	2.504	5.039	0.965	0.484	2.184	3.366	0.986	0.554	2.048	3.476	0.981	4.982	4.613	6.307	0.892
	w/o Aug.	6.822	4.854	7.747	0.889	1.401	3.427	4.382	0.934	1.672	3.128	4.728	0.942	15.304	8.746	10.316	0.746

Table 7: **Single-view analysis on three datasets.** Better results within each dataset are bolded.

Method	3D-FRONT						HouseCat6D						Toys4K-Scene					
	$\overline{\text{CD}}_2$	$\overline{\text{CD}}_1$	EMD	F@.05	IoU_{vis}	IoU_{amo}	$\overline{\text{CD}}_2$	$\overline{\text{CD}}_1$	EMD	F@.05	IoU_{vis}	IoU_{amo}	$\overline{\text{CD}}_2$	$\overline{\text{CD}}_1$	EMD	F@.05	IoU_{vis}	IoU_{amo}
Pixal3D-SV + our assem.	4.16	4.83	7.61	0.88	0.86	0.83	3.10	5.04	5.70	0.87	0.86	0.89	3.54	4.80	6.01	0.88	0.77	0.81
Ours	**2.14**	**3.30**	**6.08**	**0.93**	**0.90**	**0.89**	**2.22**	**3.90**	**4.60**	**0.91**	**0.92**	**0.95**	**2.12**	**3.84**	**5.32**	**0.91**	**0.93**	**0.93**

Results. Quantitative results are reported in Tab. 2 and Tab. 3. WorldSculpt leads all geometric and silhouette metrics across the three datasets. Notably, even when given only a single input image, the same multi-view model outperforms all single-view baselines without retraining. Figure 3 shows the qualitative comparisons. Compared with existing methods, WorldSculpt better preserves fine-grained structures and surface details while achieving more accurate spatial alignment.

4.3 Large-scale scene evaluation

We compare WorldSculpt with state-of-the-art multi-view methods, evaluating compositional scene generation on the large-scale UE-MeshyScene dataset. At this scale, individual views capture only small parts of the scene, with complementary observations distributed across the camera trajectory. We therefore focus on methods that integrate multiple views.

Compared methods and metrics. Our baselines include Mix3R Lin et al. (2026a), Stream3D Zhou et al. (2026), MV-SAM3D Li et al. (2026a), and ShapeR Siddiqui et al. (2026). At inference time, we provide all methods with ground-truth camera intrinsics and extrinsics. We also provide ground-truth depth maps to methods that require depth input. Our method does not require depth as input. We follow the small-scale scene evaluation protocol, assessing object geometry and spatial alignment in the shared world frame. We report mean Chamfer distances ($p \in \{1, 2\}$), EMD, and F-scores at thresholds of 0.05, 0.02, and 0.01, using each object's ground-truth bounding-box diagonal for normalization. We also report input-view visible and amodal silhouette IoU results.

Results. Quantitative results are reported in Tab. 4. WorldSculpt consistently outperforms all competing methods across every geometric and silhouette metric on UE-MeshyScene, with particularly clear advantages under strict geometric thresholds like F-Score@.01. Figure 4 shows that WorldSculpt recovers more complete geometry and preserves finer structures under severe occlusion, while maintaining more accurate spatial alignment.

4.4 Model analysis

Ablation study. We ablate the main components of WorldSculpt in Tab. 5. Using the official multi-view checkpoint of Pixal3D with our scene composition pipeline performs consistently worse, showing that using our scene assembly paradigm alone is insufficient. Replacing our learned multi-view aggregation with simple averaging also degrades generation quality. Removing training-time augmentation significantly hurts the performance of our model, especially on cluttered scenes like UE-MeshyScene. Our full model overall achieves the best result. We also compare WorldSculpt with the single-view checkpoint of Pixal3D integrated into our scene assembly pipeline, using the same single-view inputs. WorldSculpt uses the same checkpoint as in multi-view evaluation, without

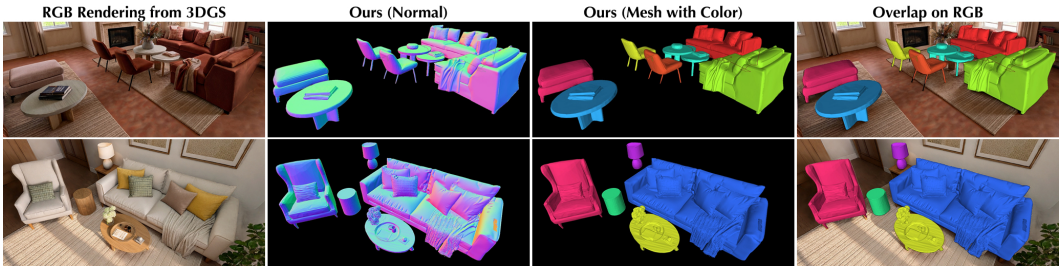

Figure 5: **Application to Marble-generated 3DGS worlds.** We apply WorldSculpt to convert the scene into a compositional representation of individual per-object meshes in a shared world frame.

additional single-view finetuning. As shown in Tab. 7, our method consistently outperforms this baseline across all geometric and silhouette metrics on the three datasets.

Robustness on grounding annotations. We further evaluate robustness to grounding annotations. Specifically, we add perturbations to instance masks, coarse localization boxes, and camera parameters. Details about perturbations are specified in Appendix D. As shown in Tab. 6, WorldSculpt remains largely stable under both mild and strong perturbations. Camera perturbations have a larger impact, consistent with the dependence of feature lifting on accurate projections, while training-time augmentation substantially mitigates this degradation. Without training-time augmentation, our model produces worse performance.

4.5 Application: Converting Generated 3D Worlds into Compositional Meshes

Generative world models such as Marble can synthesize photorealistic 3D environments represented as 3DGS. Although 3DGS supports high-quality free-viewpoint rendering, it does not explicitly provide editable object-level meshes. As a proof of concept, we apply WorldSculpt to Marble-generated worlds and convert them into compositional mesh scenes:

- **Render posed observations from the 3DGS.** We render RGB frames and depth maps from the 3DGS along a virtual trajectory with known camera intrinsics and extrinsics.
- **Obtain grounding annotations.** We apply SAM3 Carion et al. (2025) to obtain per-object masks. After that, masked depth pixels are back-projected to form world-space point clouds. We then merge the remaining observations and retain the dominant cluster using DBSCAN Ester et al. (1996). Its bounding box provides the coarse 3D localization.
- **Apply WorldSculpt.** The resulting posed RGB frames, masks, and coarse 3D boxes satisfy the input assumptions of Sec. 3. We therefore apply the WorldSculpt inference pipeline to generate individual object meshes in a shared world frame, as shown in Figure 5.

4.6 Limitations and Future Work

Although the model design improves robustness to moderate errors (as discussed in 4.4), large inaccuracies such as severely misplaced boxes or erroneous masks can still affect the generated geometry. Performance may also deteriorate under challenging capture conditions, such as low light Niu et al. (2023b; 2026a; 2023c;a) and motion-induced artifacts Ji et al. (2023; 2026); Kupyn et al. (2019); Niu et al. (2024a). Also, the current model generates geometry only. Our model could be extended to texture generation based on Pixal3D Li et al. (2026b) and TRELLIS.2 Xiang et al. (2026). Third, our formulation assumes static scenes. Dynamic scenes require modeling object motion through optical-flow Sun et al. (2018); Teed & Deng (2020); Ilg et al. (2017) or articulated representations Yang et al. (2023); Pavlakos et al. (2019); Niu et al. (2026b); Zhan et al. (2025).

5 Conclusion

We presented **WorldSculpt**, a framework that turns grounded multi-view observations into compositional 3D worlds by adapting a strong single-object generative prior. Despite being trained only on canonical individual objects, WorldSculpt generalizes at inference time to densely cluttered scenes with hundreds of objects, completing occluded geometry while remaining grounded in multi-view evidence. Our evaluation demonstrates that object-centric generative priors can scale beyond isolated assets toward editable, compositional representations of complex 3D worlds.

AI USE DISCLOSURE

We used generative AI tools for generating synthetic data, implementing part of the method, cleaning and reformatting dataset, and assisting with translation. We did not use generative AI tools for developing conceptual frameworks, designing or providing feedback on research methodology or experiments, and interpreting results. These were not applicable to this work: helping develop theoretical models, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, and supporting qualitative and thematic data analysis. All AI-assisted outputs were reviewed and verified by the authors, including manual check on whether there exist hallucinations, and whether the result match the claims. The authors take full responsibility for the final content of the submission.

ETHICS STATEMENT

This work focuses on object-centric 3D generation from multi-view visual observations. Our experiments do not involve human-subject studies or the collection of personally identifiable information. We acknowledge that techniques for generating detailed 3D environments could potentially raise privacy concerns when applied to imagery captured in private or sensitive spaces. Such applications should therefore respect applicable data-collection, consent, licensing, and privacy requirements. We do not target surveillance, identity inference, or other applications involving sensitive personal attributes.

REPRODUCIBILITY STATEMENT

We provide detailed information necessary to reproduce our method and experiments. The model architecture, multi-view conditioning mechanism, training procedure, and scene-level inference pipeline are described in Sec. 3. Details of the training and evaluation datasets are provided in Appendix A; canonicalization, cropping, and view selection are specified in Appendix B; architecture, optimization, sampling, and runtime configurations are reported in Appendix C; training-time augmentations and evaluation perturbations are described in Appendix D; and the complete evaluation protocol and metric definitions are provided in Appendix E. We will release the model and code to facilitate reproduction of the reported results.

REFERENCES

Aditya Agarwal, Gaurav Singh, Bipasha Sen, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. Scenecomplete: Open-world 3d scene completion in cluttered real world environments for robot manipulation. *IEEE Robotics and Automation Letters*, 11(1):482–489, 2025. 3, 6

William Agnew, Christopher Xie, Aaron Walsman, Octavian Murad, Yubo Wang, Pedro Domingos, and Siddhartha Srinivasa. Amodal 3d reconstruction for robotic manipulation via stability and connectivity. In *Conference on robot learning*, pp. 1498–1508. PMLR, 2021. 2

Andreea Ardelean, Mert Özer, and Bernhard Egger. Gen3dsr: Generalizable 3d scene reconstruction via divide and conquer from a single view. In *2025 International Conference on 3D Vision (3DV)*, pp. 616–626. IEEE, 2025. 3, 7

Mark Boss, Zixuan Huang, Aaryaman Vasishta, and Varun Jampani. Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16240–16250, 2025. 2

Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13154–13164, 2023. 3

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. 3, 9

Junhao Chen, Boran Zhang, Mingjin Chen, Henghaofan Zhang, Saining Zhang, Congcong Zhu, Hao Zhao, Ruqi Huang, Zhihao Li, and Yufei Wang. One video, one world: Turning monocular video into physical 4d scenes. *arXiv preprint arXiv:2606.31388*, 2026a. 3, 7

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22831–22840, 2025. 3

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7220–7232, 2026b. 3, 6

Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European conference on computer vision*, pp. 640–658. Springer, 2022. 3

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1316–1326, 2023. 3

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3151–3161, 2024a. 3

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16901–16911, 2024b. 3

Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21126–21136, 2022. 2

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023a. 2

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13142–13153, 2023b. 2

Daniel DeTone, Tianwei Shen, Fan Zhang, Lingni Ma, Julian Straub, Richard Newcombe, and Jakob Engel. Boxer: Robust lifting of open-world 2d bounding boxes to 3d. *arXiv preprint arXiv:2604.05212*, 2026. 3

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2553–2560. Ieee, 2022. 2

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pp. 226–231, 1996. 9

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10913–10922. IEEE, 2021. 5, 6

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026. 1

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*, volume 2024, pp. 50678–50702, 2024. 2

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022. 4

Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2005–2015, 2025. 3

Binbin Huang, Haobin Duan, Yiqun Zhao, Zibo Zhao, Yi Ma, and Shenghua Gao. Cupid: Generative 3d reconstruction via joint object and pose modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12741–12752, 2026. 7

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025a. 3

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23646–23657. IEEE, 2025b. 1, 3, 6

Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2462–2470, 2017. 9

Xiang Ji, Zhixiang Wang, Zhihang Zhong, and Yinqiang Zheng. Rethinking video frame interpolation from shutter mode induced degradation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12225–12234. IEEE, 2023. 9

Xiang Ji, Guixu Lin, Zhengwei Yin, Jiancheng Zhao, and Yinqiang Zheng. Moment-reenacting: Inverse motion degradation with cross-shutter guidance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. 9

Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 2

HyunJun Jung, Shun-Cheng Wu, Patrick Ruhkamp, Guangyao Zhai, Hannah Schieber, Giulia Rizzoli, Pengyuan Wang, Hongcheng Zhao, Lorenzo Garattoni, Sven Meier, et al. Housecat6d-a large-scale multi-modal category level 6d object perception dataset with household objects in realistic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22498–22508, 2024. 5, 6

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

Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8877–8886. IEEE, 2019. 9

Justin Lazarow, David Griffiths, Gefen Kohavi, Francisco Crespo, and Afshin Dehghan. Cubify anything: Scaling indoor 3d object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22225–22233, 2025. 3

Andrey Lemeshko, Bulat Gabdullin, Nikita Drozdov, Anton Konushin, Danila Rukhovich, and Maksim Kolodiazhnyi. Zoo3d: Zero-shot 3d object detection at scene level. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25820–25829, 2026. 3

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

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*, 2026a. 7, 8

Dong-Yang Li, Wang Zhao, Yuxin Chen, Wenbo Hu, Meng-Hao Guo, Fang-Lue Zhang, Ying Shan, and Shi-Min Hu. Pixal3d: Pixel-aligned 3d generation from images. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2026b. 1, 2, 4, 9

Kejie Li, Daniel DeTone, Yu Fan Steven Chen, Minh Vo, Ian Reid, Hamid Rezatofighi, Chris Sweeney, Julian Straub, and Richard Newcombe. Odam: Object detection, association, and mapping using posed rgb video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5998–6008, 2021. 3

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10486–10496, 2025. 3

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025. 3

Siyou Lin, Zhou Xue, Hongwen Zhang, Liang An, Dongping Li, Shaohui Jiao, and Yebin Liu. Mix3r: Mixing feed-forward reconstruction and generative 3d priors for joint multi-view aligned 3d reconstruction and pose estimation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2026a. 7, 8

Yuchen Lin, Chenguo Lin, Panwang Pan, Honglei Yan, Feng Yiqiang, Yadong Mu, and Katerina Fragkiadaki. Partcrafter: Structured 3d mesh generation via compositional latent diffusion transformers. *Advances in neural information processing systems*, 38:35387–35415, 2026b. 3

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The eleventh international conference on learning representations*, 2022. 4

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*, pp. 429–446. Springer, 2022. 3

Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36:22226–22246, 2023a. 2

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9298–9309, 2023b. 2

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024a. 3

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *International conference on learning representations*, volume 2024, pp. 27676–27697, 2024b. 2

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9970–9980, 2024. 2

Yanxu Meng, Haoning Wu, Ya Zhang, and Weidi Xie. Scenegen: Single-image 3d scene generation in one feedforward pass. In *2026 International Conference on 3D Vision (3DV)*, pp. 543–553. IEEE, 2026. 1, 3, 6

Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16695–16705, 2025. 3

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 2

Muyao Niu, Zhuoxiao Li, Yifan Zhan, Huy H Nguyen, Isao Echizen, and Yinqiang Zheng. Physics-based adversarial attack on near-infrared human detector for nighttime surveillance camera systems. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 8799–8807, 2023a. 9

Muyao Niu, Zhuoxiao Li, Zhihang Zhong, and Yinqiang Zheng. Visibility constrained wide-band illumination spectrum design for seeing-in-the-dark. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13976–13985. IEEE, 2023b. 9

Muyao Niu, Zhihang Zhong, and Yinqiang Zheng. Nir-assisted video enhancement via unpaired 24-hour data. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10744–10754. IEEE, 2023c. 9

Muyao Niu, Tong Chen, Yifan Zhan, Zhuoxiao Li, Xiang Ji, and Yinqiang Zheng. Rs-nerf: Neural radiance fields from rolling shutter images. In *European Conference on Computer Vision*, pp. 163–180. Springer, 2024a. 9

Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. In *European conference on computer vision*, pp. 111–128. Springer, 2024b. 1

Muyao Niu, Mingdeng Cao, Yifan Zhan, Qingtian Zhu, Mingze Ma, Jiancheng Zhao, Yanhong Zeng, Zhihang Zhong, Xiao Sun, and Yinqiang Zheng. Anicrafter: Customizing realistic human-centric animation via avatar-background conditioning in video diffusion models. *arXiv preprint arXiv:2505.20255*, 2025. 1

Muyao Niu, Mingze Ma, Yifan Zhan, Qingtian Zhu, Zhihang Zhong, Wei Guo, Chang Wen Chen, and Yinqiang Zheng. Toward robust and 3d-aware rgb-nir imaging in the dark. *arXiv preprint arXiv:2607.29684*, 2026a. 9

Muyao Niu, Yifan Zhan, Qingtian Zhu, Zhuoxiao Li, Wei Wang, Zhihang Zhong, Xiao Sun, and Yinqiang Zheng. Motion-aware animatable gaussian avatars deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40140–40151, 2026b. 9

Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3931–3940. IEEE Computer Society, 2024. 2

Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global Structure-from-Motion Revisited. In *European Conference on Computer Vision (ECCV)*, 2024. 3

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10975–10985, 2019. 9

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10106–10116, 2024. 3

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025. 3

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 3

Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2397–2406, 2022. 3

Katharina Schmid, Nicolas von Lützow, Jozef Hladkỳ, Angela Dai, and Matthias Nießner. Genrecon: Bridging generative priors for multi-view 3d scene reconstruction. *arXiv preprint arXiv:2605.23888*, 2026. 4

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 3

Shuwei Shi, Zhen Li, Muyao Niu, Chuanhao Li, Bo Zheng, Kaipeng Zhang, and Yinqiang Zheng. Surprise forcing: What to remember, when to skip in long video generation. *arXiv preprint arXiv:2607.18436*, 2026a. 1

Yukai Shi, Weiyu Li, Zihao Wang, Hongyang Li, Xingyu Chen, Ping Tan, and Lei Zhang. Scenemaker: Open-set 3D scene generation with decoupled de-occlusion and pose estimation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27146–27156, 2026b. 1, 3, 6

Yawar Siddiqui, Duncan Frost, Samir Aroudj, Armen Avetisyan, Henry Howard-Jenkins, Daniel DeTone, Pierre Moulon, Qirui Wu, Zhengqin Li, Julian Straub, et al. Shaper: Robust conditional 3d shape generation from casual captures. *arXiv preprint arXiv:2601.11514*, 2026. 7, 8

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, 2025. URL https://arxiv.org/abs/2508.10104. 4

Stefan Stojanov, Anh Thai, and James M Rehg. Using shape to categorize: Low-shot learning with an explicit shape bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1798–1808, 2021. 2, 5, 6

Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8934–8943, 2018. 9

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024. 2

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020. 9

Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021. 3

Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *Advances in Neural Information Processing Systems*, 36:39033–39051, 2023. 3

Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 2

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1

Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2021. 4

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10510–10522, 2025a. 3

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5261–5271, 2025b. 3

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025c. 3

Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. In *European conference on computer vision*, pp. 57–74. Springer, 2024. 2

Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Meshlrm: Large reconstruction model for high-quality meshes. *arXiv preprint arXiv:2404.12385*, 2024. 2

Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 606–617, 2023. 3

Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. General object foundation model for images and videos at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3783–3795, 2024. 3

Tianhao Wu, Chuanxia Zheng, Frank Guan, Andrea Vedaldi, and Tat-Jen Cham. Amodal3r: Amodal 3d reconstruction from occluded 2d images. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9181–9193. IEEE, 2025. 2

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21469–21480, 2025. 1, 2

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, et al. Native and compact structured latents for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14419–14429, 2026. 4, 9

Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024a. 2

Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. In *European Conference on Computer Vision*, pp. 1–20. Springer, 2024b. 2

Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21924–21935, 2025. 3

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 3

Shichao Yang and Sebastian Scherer. Cubeslam: Monocular 3-d object slam. *IEEE Transactions on Robotics*, 35(4):925–938, 2019. 3

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4212–4222. IEEE, 2023. 9

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. *arXiv preprint arXiv:2604.04406*, 2026. 3, 6

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. In *European Conference on Computer Vision*, pp. 498–519. Springer, 2026. 3, 6

Yifan Zhan, Qingtian Zhu, Muyao Niu, Mingze Ma, Jiancheng Zhao, Zhihang Zhong, Xiao Sun, Yu Qiao, and Yinqiang Zheng. Towards explicit exoskeleton for the reconstruction of complicated 3d human avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14259–14269, 2025. 9

Hanxue Zhang, Haoran Jiang, Qingsong Yao, Yanan Sun, Renrui Zhang, Hao Zhao, Hongyang Li, Hongzi Zhu, and Zetong Yang. Detect anything 3d in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5048–5059, 2025a. 3

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Forrester Cole, Deqing Sun, Ming-Hsuan Yang, et al. Monst3r: A simple approach for estimating geometry in the presence of motion. In *International Conference on Learning Representations*, volume 2025, pp. 82863–82886, 2025b. 3

Xiang Zhang, Sohyun Yoo, Hongrui Wu, Chuan Li, Jianwen Xie, and Zhuowen Tu. Pixarmesh: Autoregressive mesh-native single-view scene reconstruction. *arXiv preprint arXiv:2603.05888*, 2026. 3, 6

Yibo Zhang, Li Zhang, Rui Ma, and Nan Cao. Texverse: A universe of 3d objects with high-resolution textures, 2025c. URL https://arxiv.org/abs/2508.10868. 5

Qingcheng Zhao, Xiang Zhang, Haiyang Xu, Zeyuan Chen, Jianwen Xie, Yuan Gao, and Zhuowen Tu. Depr: Depth guided single-view scene reconstruction with instance-level diffusion. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5722–5733. IEEE, 2025a. 3, 6

Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. *Advances in neural information processing systems*, 36:73969–73982, 2023. 2

Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025b. 1, 2

Junwei Zhou and Yu-Wing Tai. Gena3d: Generative amodal 3d modeling by bridging 2d priors and 3d coherence. *arXiv preprint arXiv:2511.21945*, 2025. 2

Kaichen Zhou, Zeyang Bai, Xinhai Chang, Mengyu Wang, Paul Liang, and Fangneng Zhan. Stream3d: Sequential multi-view 3d generation via evidential memory. *arXiv preprint arXiv:2605.21472*, 2026. 7, 8

Table 8: Evaluation data inventory. Image dimensions are width × height.

Dataset	Scenes	Instances	Objects/scene	Frames/scene	Resolution
HouseCat6D	5	50	10	16	1096×852
Toys4K-Scene	50	305	$4 \sim 8$	16	4096×4096
3D-FRONT	50	215	$2 \sim 7$	11	768×768
UE-MeshyScene	6	$2,299$	$93 \sim 701$	$265 \sim 1758$	2560×1440

A Details for Training and Evaluation Data

A.1 TexVerse training data.

The training loader uses rendered single-object observations and precomputed, anchor-dependent latent targets from the refined TexVerse collection. It retains objects with at least two encoded anchor views and available conditioning renders. For the shape stage, the maximum token count over the available encoded views must not exceed 32,768. We use a total of 150,287 objects for model training. For each training sample, an available encoded anchor is selected at random. Its corresponding latent supplies the target coordinate frame. The number of conditioning images is sampled uniformly from $\{1, \ldots, \min(20, N)\}$, where N is the available rendered-view count. The anchor is always included; the other views are selected without replacement.

A.2 HouseCat6D.

The HouseCat6D benchmark contains 5 test scenes with ten instances each, spanning ten category labels. We use the official 5 test scenes. We select 16 frames evenly along each source sequence using rounded, equally spaced frame indices. RGB, instance masks, calibrated cameras, localization boxes, and depth maps are retained.

A.3 Toys4K-Scene.

The 50 scenes contain 305 instances drawn from 227 distinct assets from the Toys4K dataset. Each scene provides 16 views. For each scene, the camera is controlled to orbit with fixed lighting and collision avoidance enabled. Across the 50 scenes, camera elevation ranges from $11.42°$ to $69.39°$, orbit span from $25.51°$ to $179.52°$, and horizontal field of view from $40.52°$ to $59.56°$. As with HouseCat6D, RGB, instance masks, calibrated cameras, localization boxes, and depth maps are retained. The dataset will be released.

A.4 3D-FRONT.

We use the converted MIDI test-room subset, with 11 paired images and camera poses per retained room. The 215 instances comprise 98 cabinet/shelf/desk objects, 46 lights, 35 beds, 13 chairs, 10 tables, 9 objects labeled "Others," 3 sofas, and 1 pier/stool. Localization boxes are computed from complete reference meshes. As with HouseCat6D and Toys4K-Scene, RGB, instance masks, calibrated cameras, localization boxes, and depth maps are retained. The scene IDs selected for evaluation are listed in Table 9. For single view methods, we choose the same input image for each scene.

A.5 UE-MeshyScene.

The six scenes contain $5,964$ source frames in total; their individual sizes are listed in Table 10. Visual samples are shown in Figure 6. The rendering process uses Unreal Engine 5.8.0. The six source environments are `Hangar`, `Abandoned City`, `Cathedral`, `Office`, `Japanese School`, and `Desert Town`. Recorded camera trajectories are resampled uniformly in time, using linear position interpolation and normalized quaternion interpolation with hemisphere correction. We use a nominal 24-fps trajectory and a frame stride of five by default. The saved exposure biases are 0, -1, -1, -1, $+1$, and $+1$ EV, respectively. The current rendering script sets 32 engine warmup frames and 16 render warmup frames and disables motion blur. Camera positions, geometry, and

Table 9: The 50 rooms used for 3D-FRONT evaluation. Each scene ID includes the source scene UUID and room identifier.

No.	Scene ID
1	3dfront_fda72039-c926-4c5a-9c51-627a0ab5d3ca_SecondBedroom-1983
2	3dfront_fdb7f1f3-c28e-43f3-ac6b-fbb43396c015_MasterBedroom-14915
3	3dfront_fdb7f1f3-c28e-43f3-ac6b-fbb43396c015_SecondBedroom-12633
4	3dfront_fdc0b569-e71a-4cf9-9a16-a6f60bffc326_SecondBedroom-36167
5	3dfront_fdc30104-5d8a-4053-b9b8-606dca67b8c4_SecondBedroom-2439
6	3dfront_fdda8473-43ac-4012-9735-6b9ad6109b7d_MasterBedroom-30069
7	3dfront_fde0fdcc-7a6a-4e8f-a1bd-ff10cbb09b29_CloakRoom-61134
8	3dfront_fde5a857-8d3e-4f26-913d-901b182198be_DiningRoom-28111
9	3dfront_fdfe9b8c-f0ba-43e4-9fac-9d4f989180ba_MasterBedroom-12123
10	3dfront_fdfe9b8c-f0ba-43e4-9fac-9d4f989180ba_SecondBedroom-11694
11	3dfront_fe174e21-5e11-4004-8347-f4e5e2e7b30c_SecondBedroom-16460
12	3dfront_fe1c870e-fd97-4d8c-b0ef-ab646c06e3cc_MasterBedroom-13878
13	3dfront_fe31964d-c74a-40ef-9c49-4673dbbfafb5_MasterBedroom-10372
14	3dfront_fe3469c4-a466-4f84-b5d4-63f899ab0d74_Library-14218
15	3dfront_fe3721d3-1ae1-49c7-a049-ed1094649f20_MasterBedroom-5921
16	3dfront_fe3731be-633d-487d-a502-63a3c0ebfbc7_SecondBedroom-404
17	3dfront_fe4aa477-9203-4c50-a540-e1bbdf2d9fec_Aisle-1864
18	3dfront_fe4aa477-9203-4c50-a540-e1bbdf2d9fec_DiningRoom-2240
19	3dfront_fe4aa477-9203-4c50-a540-e1bbdf2d9fec_MasterBedroom-9360
20	3dfront_fe559d1c-7fdb-4193-ace1-63296877b3e4_CloakRoom-71921
21	3dfront_fe559d1c-7fdb-4193-ace1-63296877b3e4_Library-72012
22	3dfront_fe6717d3-31e0-41a2-83e6-98c29c3461df_SecondBedroom-1656
23	3dfront_fe94c17a-ba89-49ae-91e1-9173ded16545_MasterBedroom-136376
24	3dfront_fec83a3a-32df-449d-9ae4-7fa9f2433a51_Bedroom-113
25	3dfront_fecbd225-e2ad-4178-aa27-a5044d867fd0_MasterBedroom-37377
26	3dfront_fed3d417-97a3-44e0-ba86-b76c71672dc4_MasterBedroom-13989
27	3dfront_fed7d85b-9a53-4832-b6ed-6d76e5de5296_MasterBedroom-25568
28	3dfront_fef4c9c4-a340-4388-908b-53f9a866731f_Bedroom-8493
29	3dfront_ff08fb34-a3a4-495d-955b-a621cac7cba8_MasterBedroom-441
30	3dfront_ff39ebbf-6599-4edb-bec9-480ad1de6115_LivingRoom-211
31	3dfront_ff3e6faa-0224-40a4-83aa-498e31944844_MasterBedroom-94197
32	3dfront_ff3e6faa-0224-40a4-83aa-498e31944844_SecondBedroom-6442
33	3dfront_ff48ced7-689e-4292-a199-47435e73e3fa_Bedroom-6783
34	3dfront_ff48ced7-689e-4292-a199-47435e73e3fa_LivingDiningRoom-6761
35	3dfront_ff5c681b-7b7a-4f68-adfe-6e61c0451341_MasterBedroom-6887
36	3dfront_ff6079d0-41bb-449d-81e4-ec147d3b77a4_MasterBedroom-2297
37	3dfront_ff79b941-b656-41b9-a740-5647b52b2605_KidsRoom-39931
38	3dfront_ff7b42d9-0e58-4847-8d47-f793a11cd3bd_SecondBedroom-83506
39	3dfront_ff7ccf2d-0449-4367-aecc-3514af04afdd_Bedroom-11127
40	3dfront_ffa247bb-94fc-4f11-abb5-5681b09d1d81_Library-37712
41	3dfront_ffa406e2-6e48-4c38-8072-cffca7b2633b_Hallway-2369
42	3dfront_ffa6f9b3-4a73-435c-996a-f780833e77e1_MasterBedroom-39095
43	3dfront_ffb067ad-cf9a-4321-82ae-4e684c59ea3e_KidsRoom-5300
44	3dfront_ffb3d0af-142e-45b8-b6d7-176c4ec255a9_SecondBedroom-6425
45	3dfront_ffc7ddf1-827b-4154-9eee-5f8b497d0a12_Bedroom-860
46	3dfront_ffc7ddf1-827b-4154-9eee-5f8b497d0a12_Bedroom-871
47	3dfront_ffc8602d-3c06-4415-b5d0-2b0abc97f180_MasterBedroom-8283
48	3dfront_ffd7218c-06dc-4550-96ab-72705ec4f3aa_MasterBedroom-60990
49	3dfront_ffd98024-7200-429e-8b9a-1234a5937826_LivingRoom-360
50	3dfront_fff98d42-99a4-43fc-9639-5761cb4f87df_SecondBedroom-127961

Table 10: **Statistics for the UE-MeshyScene dataset.**

Scenes	*Hangar*	*Abandoned City*	*Cathedral*	*Office*	*Japanese School*	*Desert Town*	Total
Object Num.	171	93	145	678	511	701	2,299
View Num.	463	265	956	1268	1758	1254	5,964

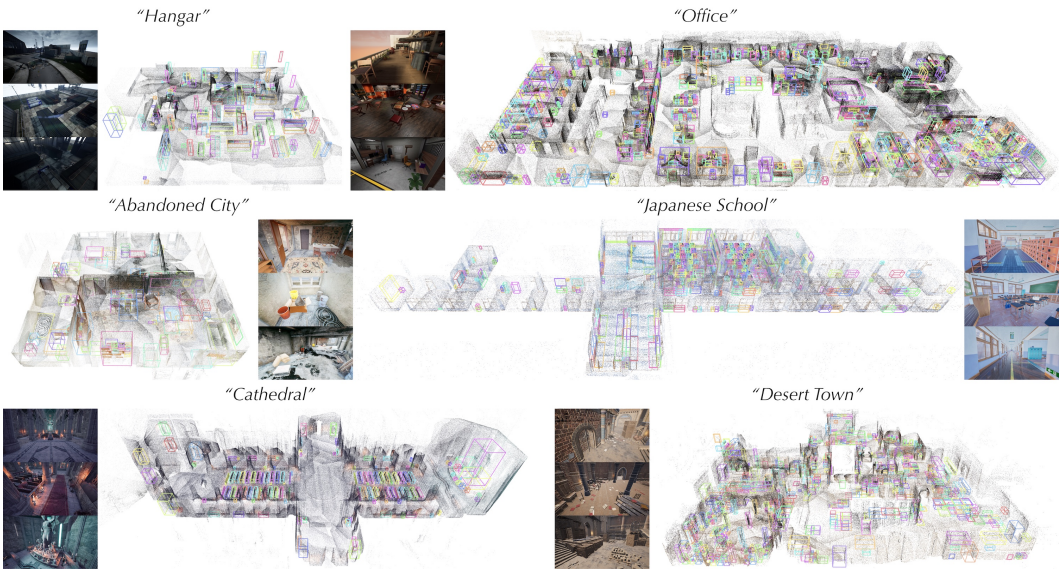

Figure 6: **Visualization of the UE-MeshyScene dataset.** Our benchmark contains six photorealistic environments spanning relatively orderly to extremely cluttered layouts, with up to 701 objects per scene. Each environment is captured along an orbiting camera trajectory and provides exact per-object ground truth, enabling world-frame evaluation under complex mutual occlusion.

depth share one scene normalization. Converted positive valid depths are stored as 16-bit integers with a scale of 1,000 in these normalized units.

A.6 Occlusion statistics

We quantify occlusion in image space using GT visible and amodal object masks. For object k in view n, let S_{kn}^{vis} denote its visible mask and S_{kn}^{amo} its full projected silhouette without inter-object occlusion. For a nonempty amodal mask, the visibility and occlusion ratios are

$$r_{kn}^{\text{vis}} = \frac{|S_{kn}^{\text{vis}}|}{|S_{kn}^{\text{amo}}|}, \qquad r_{kn}^{\text{occ}} = 1 - r_{kn}^{\text{vis}}, \tag{7}$$

where $|\cdot|$ denotes the foreground pixel count. Both masks correspond to the same camera view and image domain. These ratios measure projected visibility rather than the fraction of visible 3D surface area. We take the average of all objects from all views and report the number in Tab. 1.

B Anchor-Aligned Object Canonicalization

This section describes inference-time preprocessing, from calibrated scene observations and object annotations to independently generated meshes.

B.1 Anchor and initial cube.

We rank candidate frames by the number of foreground mask pixels, and select the largest-area candidate whose crop passes the geometric and mask-size checks. The canonical axes follow the selected camera's rotation, orthonormalized by SVD.

B.2 MASK-BASED CUBE EXPANSION.

For each eligible frame, we find an isotropic multiplier $a_v \geq 1$ for which the projected cube-corner bounding rectangle contains the visible-mask bounding rectangle. An upper bracket is found by doubling, followed by 24 bisection steps. The loop permits a final trial at 128. We use

$$s_k \leftarrow s_k \max\big(1, \text{percentile}_{95}\{a_v\}\big). \tag{8}$$

Using the 95th percentile tolerates outlier masks, so it does not guarantee containment in every frame.

B.3 CROPPING AND INTRINSICS.

The square crop is centered at the projection of the canonical origin. Its side length is the larger horizontal/vertical extent of projected cube corners, multiplied by 1.005 and rounded to integer pixels. Pixels outside the source image are zero-padded. For a crop beginning at (x_0, y_0) with side b, resized to $r = 1024$ ($r = 512$ for the sparse-structure stage), the image transformation and intrinsics are

$$A_{kv} = \begin{bmatrix} r/b & 0 & -rx_0/b \\ 0 & r/b & -ry_0/b \\ 0 & 0 & 1 \end{bmatrix}, \qquad K_{kv}^{\text{crop}} = A_{kv} K_v. \tag{9}$$

The principal point is retained even when it is off-center or outside the crop. Crops are resized with area interpolation, and RGBA is composited over black before feature extraction. We do not apply morphological erosion to alpha masks at inference.

B.4 VIEW SELECTION.

All eligible stored views are considered, with the selected anchor placed first. We cap conditioning at 20 views per object; beyond this cap, it keeps the anchor and selects other views by nonzero foreground area in their crops. The small-scale datasets (Housecat6D, Toys4k-Scene, and 3D-FRONT) remain below the cap. In the single-view experiment, the scene input image comes from the same image as the single-view baselines. We directly use the same multi-view-trained checkpoint; there is no additional single-view-specific finetuning.

C CONDITIONING ARCHITECTURE AND OPTIMIZATION

C.1 IMAGE FEATURES AND PROJECTION.

The frozen image encoder is DINOv3 ViT-L/16 (`dinov3-vitl16-pretrain-lvd1689m`), which produces 1,024-dimensional local descriptors. The sparse-structure stage uses 512×512 input images and a 16^3 conditioning grid. The shape stage uses 1024×1024 input images and a 64^3 conditioning grid. For each image, DINOv3 produces a 64×64 feature map with 1,024 channels. A frozen, RGB-guided NAF upsampler increases its spatial resolution to 512×512 while preserving the channel dimension. At each 3D grid location, we bilinearly sample both feature maps at locations corresponding to the same projected image coordinates and concatenate the resulting descriptors into a 2,048-dimensional feature. These per-view features are subsequently fused by the multi-view aggregator.

C.2 LEARNED RESIDUAL AGGREGATION.

For projected features $z_{kv} \in \mathbb{R}^C$ at a fixed voxel and view-validity weights $m_v \in \{0, 1\}$, define

$$\mu = \frac{\sum_v m_v z_{kv}}{\max(1, \sum_v m_v)}, \qquad \sigma^2 = \max\left(0, \frac{\sum_v m_v z_{kv}^2}{\max(1, \sum_v m_v)} - \mu^2\right). \tag{10}$$

The two MLPs take $[z_{kv}, \mu, \sigma^2]$ and predict a residual h_v and a scalar logit a_v. Fusion is

$$z_k^{\text{fused}} = \mu + \sum_v \text{softmax}_v(a_v) h_v, \tag{11}$$

Table 11: Training configuration.

Parameter	Sparse-structure	Shape
Input image side	512	1,024
Condition-grid side	16	64
Projected feature channels	1,024	2,048
Denoiser blocks / width / heads	30 / 1,536 / 12	30 / 1,536 / 12
GPUs	16	16
Batch per GPU / gradient accumulation	8 / 2	2 / 2
Global batch	128	32
Optimizer	AdamW	AdamW
Learning rate / weight decay	10^{-4} / 0	10^{-4} / 0
Adam (β_1, β_2) / ϵ	$(0.9, 0.95)$ / 10^{-8}	$(0.9, 0.95)$ / 10^{-8}
Checkpoint Iteration	15,000	15,000
Conditioning Dropout	0.1	0.1
Time sampling	logit-normal $(1, 1)$	uniform $[0, 1]$
Precision	BF16 AMP	BF16 AMP

where invalid-view logits are set to -10^9. The residual MLP has widths $3C \to C \to C \to C$, with ReLU after its first two layers. The weight MLP has widths $3C \to C \to 1$ and one ReLU. Zero-initializing the last residual layer makes initial fusion equal to the masked mean.

C.3 Global conditioning tokens.

At inference time, our model uses global conditioning tokens from the anchor view only, while aggregating local projected features across views.

C.4 Finetuned components.

We initialize the sparse-structure stage and shape stage from the corresponding Pixal3D single-view checkpoints. LoRA has rank and scaling parameter 32. It is applied to attention query/key/value and output projections and to the per-block projected-conditioning linear layers (`to_qkv`, `to_q`, `to_kv`, `to_out`, `proj_linear`). Base denoiser weights, image encoders, and decoders remain frozen; the aggregation MLPs are trained in full. The LoRA parameter counts are $21,135,360$ for the sparse-structure stage and $22,118,400$ for the shape stage. The aggregators add $8,393,729$ and $33,564,673$ parameters, respectively, giving $29,529,089$ and $55,683,073$ trainable parameters in these components.

C.5 Flow-matching objective.

For a target latent x_0 and $\epsilon \sim \mathcal{N}(0, I)$, training uses

$$x_t = (1-t)x_0 + [\sigma_{\min} + (1-\sigma_{\min})t]\epsilon, \quad v^* = (1-\sigma_{\min})\epsilon - x_0, \quad \mathcal{L} = \mathbb{E}\|v_\theta(x_t, t, c) - v^*\|_2^2, \tag{12}$$

with $\sigma_{\min} = 10^{-5}$. Sparse-structure times are sampled as $t = \mathrm{sigmoid}(u)$ for $u \sim \mathcal{N}(1, 1)$; shape times are uniform. The optimizer has no configured learning-rate scheduler. Adaptive gradient clipping uses a maximum norm of 1 and the 95th-percentile clipping configuration. Python, NumPy, and PyTorch training RNGs are seeded by distributed rank; the holdout split independently uses seed 42.

C.6 Inference sampling.

Both geometry stages use the checkpoint pipeline's Euler guidance-interval sampler with 12 steps and guidance strength 7.5. The sparse-structure stage uses guidance rescale 0.7 and time rescale 5; the shape stage uses 0.5 and 3. The guidance interval is $[0.6, 1.0]$ in both stages. The random seed is 42.

Table 12: Training augmentation settings after warmup.

Augmentation	Settings
Occlusion	3D occlusion with probability 0.8, otherwise 2D occlusion; up to 75% foreground removal for 2D masks and 12–75% anchor coverage for 3D occluders.
Camera	Non-anchor rotation up to 5°, translation up to 0.05 canonical units, and relative horizontal-FOV perturbation within $\pm 5\%$.
Mask boundary	Remove up to 25% of an inner boundary band of width approximately $0.012W$, where W is the conditioning-image width.
Resolution	A shared resize factor sampled uniformly from $[0.25, 1]$ for all views of an object.

C.7 Runtime Efficiency.

On NVIDIA H100 GPUs, the denoising time averages 6.53 seconds per object across all four datasets. Across the 105 small-scale evaluation scenes, the average cumulative denoising time is 37.3 seconds per scene. Extrapolating from the UE-MeshyScene average of 6.43 seconds per object gives approximately 75 minutes for a scene with 701 objects.

D Training-Time Augmentations and Grounding Perturbations

D.1 Training augmentations.

We apply four types of conditioning augmentation: occlusion, camera perturbation, mask-boundary noise, and resolution degradation. Each is independently enabled with probability 0.5 per training sample. Scheduled perturbation strengths increase linearly over the first $3{,}000$ iterations:

$$a(t) = a_{\max} \min(1, t/3000). \tag{13}$$

All augmentations modify only the conditioning observations; the supervision target remains unchanged. Table 12 summarizes the specific settings.

Occlusion. We sample up to six 3D occluders from a pool of $130{,}400$ TexVerse objects and project their geometry consistently across views to mask the target observations. Alternatively, 2D occlusion removes foreground regions using irregular noise patterns, polygons, ellipses, rectangular holes, or edge truncation. Occluders affect the conditioning images and masks but are not added to the target geometry.

Camera perturbation. We perturb camera rotations, translations, and horizontal fields of view in canonical space, while keeping the anchor camera fixed. Rotation axes and translation directions are sampled independently. These perturbations affect only the cameras used for conditioning, leaving the images and target coordinate frame unchanged.

Mask-boundary noise. We introduce irregular boundary errors by removing pixels from a narrow band inside the foreground mask. A fractal-noise field determines which pixels are removed, with the removal fraction sampled up to the scheduled bound. This produces spatially varying boundary damage rather than uniform morphological erosion.

Resolution degradation. We downsample conditioning images and restore their original dimensions to simulate reduced image quality. The resize factor is shared across an object's views. Downsampling uses randomly selected bilinear, box, or Lanczos interpolation, followed by bilinear or bicubic upsampling. RGB and alpha are processed separately, and the degraded alpha is then applied to RGB.

D.2 Evaluation perturbations.

Mask, box, and camera experiments alter separate aspects of the input. The same deterministic perturbed input is supplied to the augmented and no-augmentation checkpoints.

- **Mask $q\%$:** randomly alter $q\%$ of the original foreground pixels, using irregular noise patterns, polygons, ellipses, rectangular holes, or edge truncation.
- **Box $q\%$:** randomly change the position and scale of the localization box within $[1 - q/100, 1 + q/100]$ of the box's size.
- **Camera:** each view independently receives the camera-noise operation above. The 3 and 5 settings correspond to maxima $(3^\circ, 0.03, 3\%)$ and $(5^\circ, 0.05, 5\%)$, respectively.

E Comparison Inputs and Evaluation Protocol

E.1 Shared single-image selection.

We choose a scene frame by first maximizing the number of instances with at least 64 mask pixels, then using relative visible area and total area to break ties, and finally preferring the earlier frame. Each single-image baseline uses the same frame as input.

E.2 Reference geometry and normalization.

We score individual object assets rather than concatenating the entire scene. For object k, let P_k and G_k be area-weighted surface samples of prediction and reference. Let $d_k = \|b_k^{\max} - b_k^{\min}\|_2$ be the diagonal of its annotated GT world AABB. Both point sets are divided by the same d_k.

E.3 Distance metrics.

Write $d(p, G) = \min_{g \in G} \|p - g\|_2$ for nearest-neighbor Euclidean distance after the shared diagonal normalization. We report

$$\mathrm{CD}_p(P, G) = \frac{1}{|P|} \sum_{x \in P} d(x, G)^p + \frac{1}{|G|} \sum_{y \in G} d(y, P)^p, \qquad p \in \{1, 2\}. \tag{14}$$

The subscript 1 means unsquared Euclidean distance. Precision and recall at threshold τ are the fractions of predicted and reference samples, respectively, whose nearest-neighbor distance is strictly less than τ. We use $F_\tau = 2PR/(P + R)$, or zero when $P + R = 0$, with $\tau \in \{0.05, 0.02, 0.01\}$.

The evaluator samples $32,768$ points from each mesh with seed 0. EMD uses independent subsets of $2,048$ points from these clouds and the exact Hungarian assignment under Euclidean cost:

$$\mathrm{EMD}(P', G') = \min_{\pi} \frac{1}{2048} \sum_{i=1}^{2048} \|p'_i - g'_{\pi(i)}\|_2. \tag{15}$$

CD_2 is displayed in 10^{-3} units, CD_1 and EMD in 10^{-2} units, and F-scores without scaling.

E.4 Visible and amodal silhouette IoU.

For an object and camera, IoU is the intersection of predicted and reference masks divided by their union. A zero union is excluded; a one-sided empty mask scores zero. Scoring uses the original camera calibration and image resolution, with no mask dilation or resize. Visible predictions are jointly depth-rendered with other included objects, and are compared with the provided visible instance masks. Amodal evaluation independently renders each predicted object and its reference mesh, removing inter-object occlusion while retaining image boundaries and clipping planes.