ShapeR: Robust Conditional 3D Shape Generation from Casual Captures

Yawar Siddiqui Duncan Frost Samir Aroudj Armen Avetisyan
Henry Howard-Jenkins Daniel DeTone Pierre Moulon Qirui Wu† Zhengqin Li
Julian Straub Richard Newcombe Jakob Engel

Meta Reality Labs Research Simon Fraser University†

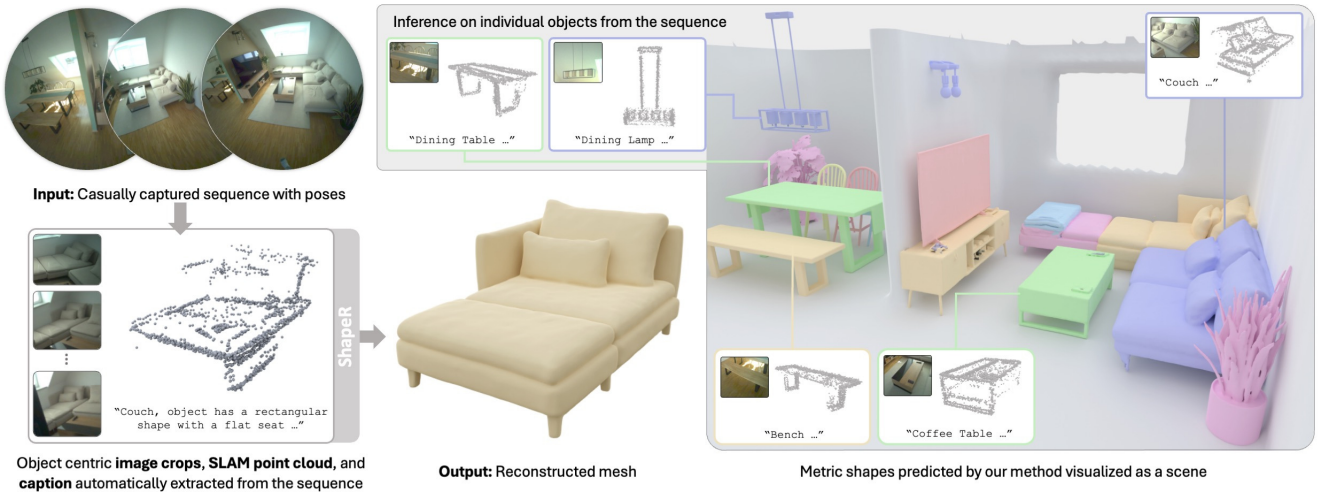

Figure 1. ShapeR introduces a novel approach to metric shape generation. Given an input image sequence, preprocessing extracts per-object metric sparse SLAM points, images, poses, and captions using off-the-shelf methods. A rectified flow transformer operating on VecSet latents conditions on these multimodal inputs to generate a shape code, which is decoded into the object's mesh. (Right) By applying the model object-centrically to each detected object, we obtain a metric reconstruction of the entire scene.

Abstract

Recent advances in 3D shape generation have achieved impressive results, but most existing methods rely on clean, unoccluded, and well-segmented inputs. Such conditions are rarely met in real-world scenarios. We present ShapeR[1]*, a novel approach for conditional 3D object shape generation from casually captured sequences. Given a image sequence, we leverage off-the-shelf visual-inertial SLAM, 3D detection algorithms and VLMs to extract for each object, a set of sparse SLAM points, posed multi-view images, and machine-generated captions. A rectified flow transformer trained to effectively condition on these modalities then generates high-fidelity metric 3D shapes. To ensure robustness to the challenges of casually captured data, we employ a range of techniques including on-the-fly compositional augmentations, a curriculum training scheme spanning object- and scene-level datasets, and strategies to handle background clutter. Additionally, we introduce a new evaluation benchmark comprising 178 in the wild objects across 7 real-world scenes with geometry annotations. Experiments show that ShapeR significantly outperforms existing approaches in this challenging setting, achieving an improvement of 2.7× in Chamfer distance compared to SoTA.*

1. Introduction

3D reconstruction is a longstanding challenge in computer vision, essential for understanding and interacting with the physical world. Scene-centric methods typically reconstruct entire scenes as single entities [20, 53, 62, 72, 88, 93], but produce monolithic representations, often with limited resolution and missing surfaces in unobserved areas. Object-centric reconstruction [2, 34, 50, 57, 81, 84] instead focuses on recovering individual objects within a scene, enabling more detailed and complete results.

Recent advances in object-level generative models [42, 80–82, 92], enabled by improved architectures [24, 29, 61], large-scale 3D datasets [21, 22], and better shape representations [82, 90], have rapidly advanced object-centric shape generation. These models produce high-fidelity shapes from clean, well-segmented, and unoccluded inputs. However, their performance drops significantly in casual capture settings, *i.e.*, real-world scenarios with natural, non-scanning trajectories where users move freely and captures often include occlusions, background clutter, sensor noise, low resolution, and suboptimal views (Fig. 2)

[1] facebookresearch.github.io/ShapeR

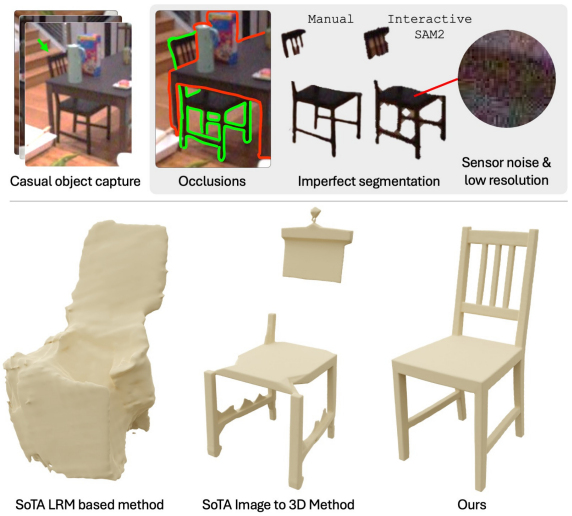

Figure 2. (Top) Objects captured in casual settings pose challenges like clutter, poor viewpoints, low resolution, noise, motion blur, and occlusions that are difficult to segment, even interactively. (Bottom) State-of-the-art 3D models often fail in these scenarios, while ShapeR remains robust and effective.

To address these challenges, we propose ShapeR, a large-scale rectified flow model for robust 3D shape generation from casually captured sequences. ShapeR is designed to leverage complementary information from multiple modalities, including sparse metric point clouds, multi-view posed images, and machine-generated captions. Given an input sequence, we first use off-the-shelf SLAM [27] to obtain sparse point clouds and camera poses. Next, we apply 3D instance detection [72] to extract object-centric crops from both images and point clouds, and generate text captions using vision-language models [52]. These multimodal cues condition a flow-matching [47] transformer, which is trained to denoise latent VecSets [90] that can be decoded into complete 3D shapes.

To improve robustness, we apply extensive on-the-fly augmentations across all input modalities during training. Unlike prior work that relies on explicit 2D segmentation [42, 44, 80, 82], ShapeR learns to implicitly segment objects within images by utilizing the 3D instance points. Training is conducted in a two stage curriculum learning setup: the first stage uses large and diverse object-centric datasets with objects in isolation, where we address the limitations of these contrived settings through extensive point and image augmentations. The second stage employs synthetic scene data [4], which covers fewer categories but offers more realistic scenarios. This captures diverse object combinations that single-object datasets cannot model due to combinatorial complexity.

For evaluation, we introduce a new dataset of in-the-wild sequences with paired posed multi-view images, SLAM point clouds, and individually complete 3D shape annotations for 178 objects across 7 diverse scenes. In contrast to existing real-world 3D reconstruction datasets which are either captured in controlled setups [23, 40] or have merged object and background geometries or incomplete shapes [7, 87], this dataset is designed to capture real-world challenges like occlusions, clutter, and variable resolution and viewpoints to enable realistic, in-the-wild evaluation.

We believe ShapeR represents a key step toward unifying generative 3D shape modeling [42, 80–82, 92] and metric 3D scene reconstruction [62, 68, 72, 88, 93]: ShapeR produces complete, high-fidelity object shapes at appropriate level of detail for each object, while preserving real-world metric consistency. We will release all code, model weights and the ShapeR evaluation dataset. In summary, our contributions are:

- A rectified flow model for robust 3D metric shape generation from casually captured sequences, trained with a robust pipeline that combines sparse point clouds, posed images, on-the-fly cross-modal augmentations, and a two-stage curriculum for effective generalization.
- An evaluation dataset of causally captured sequences with paired images, SLAM points, and 3D shape annotations for 178 objects across seven scenes, enabling systematic evaluation under realistic conditions.

2. Related Works

Non Object-centric Reconstruction. Surface reconstruction has been widely studied using both learned and optimization-based methods [19, 20, 35, 37, 58, 69]. Recent approaches such as NeRF [53], 3DGS [39], and their extensions [5, 6, 11, 54, 89] achieve high-fidelity view synthesis but prioritize appearance over geometric accuracy. SDF-based implicit methods [43, 76, 77, 85, 86] improve geometric faithfulness while maintaining view quality. Feed-forward methods [55, 72, 73, 75, 93] directly predict global scene geometry from posed images, reducing optimization overhead. However, these methods reconstruct scenes as a single surface, leaving individual objects incomplete under occlusion. In contrast, ShapeR performs explicit object-level reconstruction from sequences, producing complete geometry for each object.

Conditional Object Reconstruction. Early work explored class-specific reconstruction models [15, 51, 60, 62] conditioned on images or point clouds. Later methods, such as Dreamfusion [63] and its extensions [12, 36, 46, 49, 78], used 2D diffusion models for text-conditioned shape generation, moving beyond fixed classes. Large Reconstruction Models [32] and follow-ups [30, 44, 70, 74] scaled image-to-3D reconstruction and integrated mesh generation, texturing, and relightable assets, relying on 2D diffusion priors. With large-scale datasets [21], native 3D diffusion approaches [42, 45, 80, 82, 91, 92] have further improved fidelity. However, most methods require clean,

well-segmented inputs and lack metric grounding from single images, and even amodal approaches [81] struggle in real-world scenarios. ShapeR differs by leveraging multimodal conditioning with sparse metric point clouds, posed images, and captions, enabling robust, metrically accurate reconstruction under occlusion, clutter, and viewpoint variation.

Object-centric Scene Reconstruction. Early approaches addressed object-centric scene reconstruction through joint detection and completion [16, 33, 67] or CAD model retrieval [2, 3, 41], but often resulted in incomplete or mismatched geometry. Later methods [17, 48, 57] reconstructed individual objects and scene layouts from single views, but were typically limited to specific classes. Recent work [1, 34, 50, 56, 83, 84] leverages diffusion priors, open-vocabulary detection, and generative models to improve per-object geometry and scene assembly, but often depends on high-quality 2D instance segmentation. While ShapeR focuses on object-centric rather than joint scene reconstruction, it generates 3D metric shapes conditioned on point cloud crops from off-the-shelf detectors, which can be composed for scene-level reconstruction. Unlike prior methods that degrade with machine-generated segments in real-world scenarios, ShapeR remains robust to imperfect segmentation and challenging, casual capture conditions.

3. Method

ShapeR performs generative, object-centric 3D reconstruction from image sequences by leveraging multimodal inputs and robust training strategies. First, a sparse 3D point cloud and camera poses are extracted using an off-the-shelf visual-inertial SLAM method [27]. Object instances are then identified via a 3D instance detection method [72], leveraging both SLAM points and posed images. For each detected object, its sparse points, the images in which it appears, 2D projections of its 3D points in those images, and a machine-generated caption from a vision-language model [52] are extracted. These multimodal inputs condition a 3D rectified flow matching model, which denoises a latent VecSet [90] and decodes it to produce the object's 3D shape (Fig. 3). The use of multimodal conditioning, along with heavy on-the-fly compositional augmentations and curriculum training, ensures the robustness of ShapeR in real-world scenarios.

3.1. Multimodally Conditioned Flow Matching

Following recent advances in 3D generative modeling [42, 82, 91, 92], ShapeR formulates object-centric shape generation as a rectified flow process that denoises latent representations learned by a 3D VAE.

3D Variational Autoencoder. We adopt the Dora [13] variant of VecSets [90] as our latent autoencoder. Given a mesh

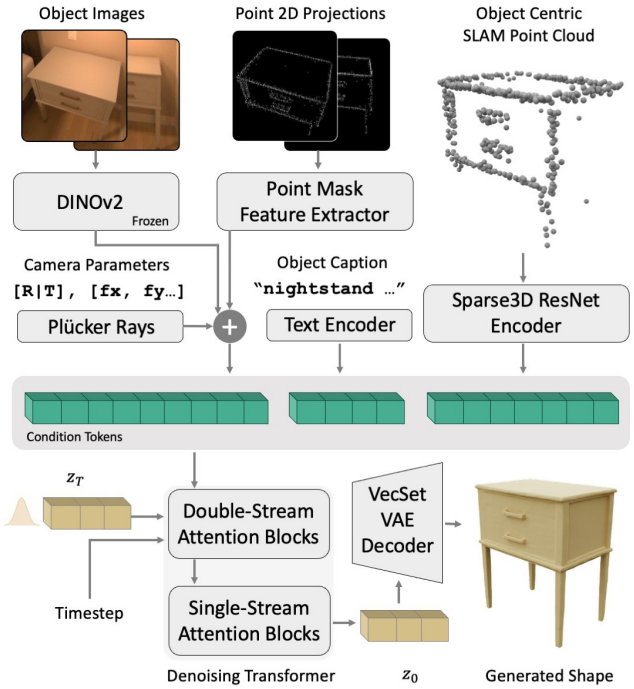

Figure 3. The ShapeR denoising transformer, built on the FLUX DiT architecture, denoises latent VecSets by conditioning on multiple modalities: posed images, SLAM points, captions, and the 2D projections of SLAM points observed in those input images. SLAM points are encoded with a sparse 3D ResNet, images using a frozen DINOv2 backbone, poses using Plücker encodings, and projection masks via a 2D convolutional network. The denoised latent is decoded into a SDF, from which the final object shape is extracted using marching cubes.

S, two point clouds are sampled: (i) uniformly distributed surface points capturing overall geometry and (ii) edge-salient points capturing fine detail. These are separately cross-attended, downsampled, concatenated, and further processed through self-attention to produce a latent code $z \in \mathbb{R}^{L \times d}$, where L is variable in $\{256, 512, \ldots, 4096\}$ and feature width $d = 64$. The decoder D predicts signed distance values $s = D(z, x)$ for a grid of query points $x \in \mathbb{R}^3$ through cross-attention with the processed latent sequence. The VAE is trained using

$$\mathcal{L}_{\text{VAE}} = ||s - s_{GT}||_2^2 + \beta \mathcal{L}_{\text{KL}}\Big(q(z|S) \;||\; \mathcal{N}(0, I)\Big). \quad (1)$$

Rectified Flow Model. The latent distribution $z \sim q(z|S)$ serves as the target distribution for flow matching. A denoising transformer f_θ is trained to transport Gaussian noise $z_1 \sim \mathcal{N}(0, I)$ to the latent manifold z_0, conditioned on multimodal cues (C):

$$\dot{z}_t = f_\theta(z_t, t, C), \quad t \in [0, 1]. \quad (2)$$

The training objective minimizes the expected squared error

Figure 4. Incorporating SLAM points significantly enhances robustness. These points provide a complementary geometric signal to posed images, encoding aggregated shape information across the entire sequence.

between the model-predicted and true transport velocity:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,z_t,C}\big[||f_\theta(z_t,t,C) - (z_0 - z_1)||_2^2\big]. \quad (3)$$

We employ a FLUX.1-like dual-single-stream transformer [8], where the first four dual layers cross-attend to text tokens and subsequent dual and single layers to image and point tokens. Similar to [42, 92], positional embeddings are omitted. Dual-stream outputs are concatenated and subsequently processed by several self-attention layers. Both dual and single stream blocks are modulated with timestep and CLIP [64] text embeddings.

Condition Encoding. Condition inputs $C = \{C_{\text{pts}}, C_{\text{img}}, C_{\text{txt}}\}$ comprise of the 3D SLAM points, images, and captions respectively. For C_{pts}, a ResNet [31] style 3D sparse-convolutional encoder downscales the point features into a token stream. For C_{img}, a frozen DINOv2 [59] backbone extracts image tokens, concatenated with Plücker ray encodings of the corresponding camera poses. The object's 3D points observed in their respective frames are projected to 2D to form binary point masks, which are processed by a 2D convolutional extractor and concatenated with DINO and Plücker tokens. For C_{txt}, captions are tokenized with a frozen T5 encoder [65] and a CLIP [64] text encoder. Notably, no segmentation masks are used; the object of interest is learned implicitly from the 3D point tokens and 2D projected point mask information.

3.2. Two-Stage Curriculum Learning Setup

As a class-agnostic generative model, ShapeR must learn priors across diverse categories. In the first stage, we train on a large-scale object-centric dataset containing over 600K meshes of diverse semantic categories created by 3D artists. To simulate noisy, real-world inputs, we apply extensive augmentations to all modalities (Fig. 5, left), including background compositing, occlusion overlays, visibility fog, resolution degradation, and photometric perturbations on images. For SLAM points, we simulate partial trajectories, a diverse range of point dropout strategies, Gaussian noise, and point occlusion. These augmentations are applied on-the-fly in the data loader in a compositional manner, yielding a virtually infinite stream of unique training samples. While this stage teaches the model general shape priors, it does not fully reflect the complexity of real captures. Hence, we fine-tune the model on a second dataset consisting of object crops extracted from Aria Synthetic Environments [4]. Although this dataset is less diverse, it exhibits realistic occlusions, inter-object interactions, and SLAM noise patterns (Fig. 5, right).

3.3. Inference

Given a posed image sequence $I = I^1, \ldots, I^K$ and corresponding camera intrinsics & extrinsics $\Pi = \Pi^1, \ldots, \Pi^K$, we first compute sparse metric point clouds P by tracking and triangulating high-gradient image regions similarly to [26]. This provides both 3D point positions and their visibility association across frames, represented as $P_{I^k} \subseteq P$, denoting the subset of points observed in frame I^k. An instance detection model [72] is applied on the posed images and point cloud to predict 3D bounding boxes for object instances. For each object i, the corresponding point set $P_i \subset P$ is refined within its bounding box using SAM2 [66] to remove spurious samples from neighboring instances. Using the point–frame association P_{I^k}, we identify all frames where object i is visible and select a fixed number N of representative frames I_i. For each selected frame I_i^j, the points $P^i \cap P_{I_i^j}$ are projected onto the image plane to generate binary masks M_i, approximating the object's silhouette in that view. A vision–language model [52] is then prompted on each object's representative image to generate a descriptive caption T_i. The complete conditioning set for object i is thus $C_i = \{P_i, I_i, \Pi_i, M_i, T_i\}$. Before generation, each object's point cloud P_i is normalized to the normalized device coordinate cube $[-1, 1]^3$. The flow-matching model predicts the object's shape within this normalized space, and the reconstructed mesh is rescaled back to the original metric coordinate system of P_i, ensuring physically accurate dimensions. Sampling proceeds by integrating the learned flow:

$$z_1 \sim \mathcal{N}(0, I), \quad z_{t-\Delta t} = z_t + \Delta t\, f_\theta(z_t, t, C_i), \quad (4)$$

with midpoint sampling. The final mesh is reconstructed as

$$\hat{S}_i = \text{Rescale}\big(\text{MarchingCubes}(D(z_0)), P_i\big), \quad (5)$$

producing metrically consistent, fully reconstructed meshes for each detected object i, aligned with the real-world scale and placement of the input sequence.

Implementation Details. The point cloud is derived from images using SLAM or SfM; specifically, we use semi-dense point clouds from Project Aria's Machine Perception

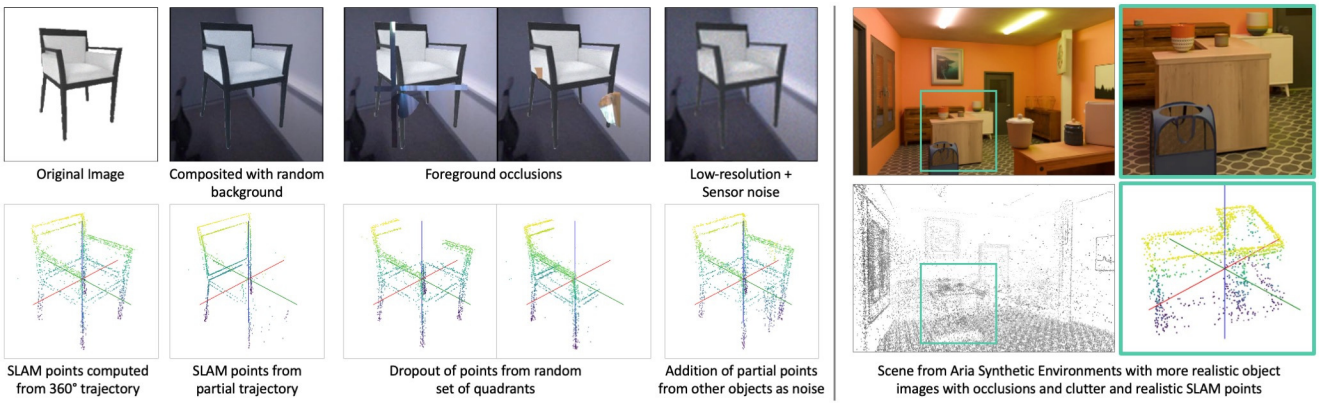

Figure 5. (Left) We pretrain on 600K object meshes with extensive, compositional augmentations across all modalities, simulating realistic backgrounds via image compositing, and introducing diverse occlusions and noise in both images and SLAM points. (Right) We then fine-tune on object-centric crops from Aria Synthetic Environment scenes, which feature realistic image occlusions, SLAM point cloud noise, and inter-object interactions.

Services [28], obtained via a visual-inertial SLAM system with Aria's monochrome cameras and IMUs. During training, conditioning is performed using two randomly sampled views per object, while inference uses up to sixteen selected views at a resolution of 280×280 pixels from Aria Mono scene SLAM cameras. Additional details are provided in the Appendix Sec. C.

4. Experiments

We evaluate and ablate key components of ShapeR on a novel high quality dataset against nine leading 3D reconstruction and generation methods, grouped by the type of input they exploit and the nature of their reconstruction task.

ShapeR Evaluation Dataset. While several datasets exist for benchmarking 3D reconstruction [7, 10, 18, 21, 23, 25, 40, 87], most are limited in realism or completeness. Synthetic datasets such as ShapeNet [10] and Objaverse [21, 22] offer large-scale coverage but lack real-world complexity. Controlled datasets like DTC [23], GSO [25], and StanfordORB [40] focus on isolated tabletop objects in studio settings. In-the-wild datasets such as ScanNet [18], ScanNet++ [87], and ARKitScenes [7] provide realistic imagery but lack complete object-level 3D geometry for evaluation (Fig. 10). To address these gaps, we introduce the ShapeR Evaluation Dataset, designed to benchmark reconstruction under challenging, casual capture conditions. The dataset contains seven casually-captured recordings from distinct cluttered scenes annotated with 178 diverse high quality object shapes. It covers a wide range of categories, from large objects like furniture to smaller items such as remotes, toasters, and tools as can be seen in Figs. 6, 7, 11 and 12. For each sequence, we provide multi-view images, calibrated camera parameters, SLAM point clouds, and machine-generated object captions. Each annotated object also includes a complete reference mesh generated using internal image-to-3D modeling methods under ideal conditions, which we manually refined and realigned for geometric and pose consistency. More details are provided in the supplementary. All quantitative and qualitative evaluations in the following sections are conducted on this dataset. Evaluations on further datasets are in the Appendix Sec. B.

Method	CD↓ $_{\times 10^2}$	NC↑	F1↑
EFM3D [72]	13.82	0.614	0.276
FStereo [79]	6.483	0.677	0.435
LIRM [44]	8.047	0.683	0.384
DP-Recon [56]	8.364	0.661	0.436
w/o SLAM Points	4.514	0.765	0.486
w/o Point Augmentation	3.276	0.805	0.667
w/o Image Augmentation	3.397	0.778	0.649
w/o Two Stage Training	3.053	0.801	0.689
w/o Point Mask Prompting	2.568	**0.813**	0.701
ShapeR	**2.375**	0.810	**0.722**

Table 1. Comparison on ShapeR evaluation dataset against posed multiview to 3D approaches, and an ablation of components.

Metrics. We evaluate the reconstructed geometry with 3 complementary metrics following prior works [62, 69, 70]: Chamfer ℓ_2 Distance (CD), Normal Consistency (NC) and F-score (F1) at 1% threshold. All metrics are computed in the normalized coordinate space.

4.1. Results

Posed Multi-view to 3D. We compare against EFM3D [72], TSDF fusion with FoundationStereo depths [79], DP-Recon [56], and LIRM [44]. These methods take posed images and predict metric 3D geometry. For monolithic mesh predictors such as EFM3D and FoundationStereo-based fusion, we extract object instances by cropping the predicted mesh using ground-truth geom-

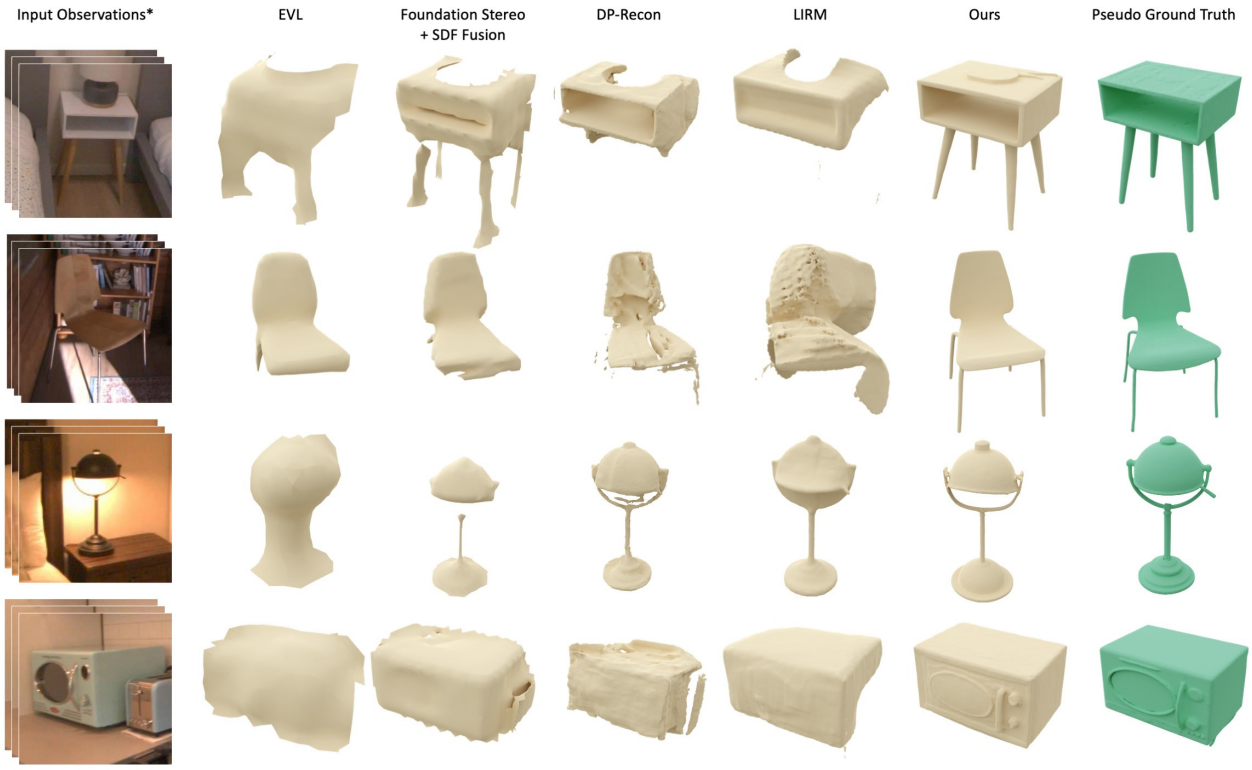

Figure 6. Qualitative comparison on the ShapeR evaluation dataset against posed multiview-to-3D methods. For scene-centric fusion approaches (EVL, Foundation Stereo), ground-truth meshes are used to segment individual object shapes. For methods relying on image segmentation masks (DP-Recon, LIRM), we employ SAM2, prompted with bounding boxes, to generate input image masks.

etry as guidance. For DP-Recon and LIRM, which rely on 2D object segmentations, we provide SAM2-generated masks. As shown in Tab. 1 and Fig. 6, monolithic scene reconstruction methods produce incomplete objects due to occlusions, while segmentation-based methods degrade under imperfect masks. ShapeR, by contrast, reconstructs complete, metric shapes without requiring segmentation inputs, remaining robust across casual captures.

Foundation Image to 3D. We also evaluate against recent large-scale image-to-3D generative models including TripoSG [42], Direct3DS2 [80], Hunyuan3D-2.0 [92], and Amodal3R [81]. Hunyuan3D-2.0, TripoSG and Direct3DS2 are trained to predict shapes from one or multiple unposed views and perform well under idealized, clean conditions with minimal occlusion. Amodal3R, which extends TRELLIS [82], improves robustness by reasoning about occluded regions and generating amodal completions. We found that for non-standard viewpoints common in casual captures, their single-view versions are significantly more competitive than the multi-view ones, so we report results using the single-view setting. To ensure their optimal performance, we manually select views with clear object visibility and use interactive SAM2-based segmentations, while ShapeR operates fully automatically using multiple posed views. Our method achieves metrically consistent, complete, and robust reconstructions without any manual intervention as shown in Tab. 2 and Fig. 7.

Method	ShapeR Win Rate↑
TripoSG	86.67%
Amodal3R	86.11%
Direct3DS2	88.33%
Hunyuan3D-2.0	81.11%

Table 2. Percentage of users who prefer our method over the image-to-3d baselines over 660 responses. Our generated meshes are preferred significantly more often.

Image to Scene Layout. We also compare against scene-level reconstruction methods, MIDI3D [34] and SceneGen [50], which predict multiple object geometries and spatial layout. MIDI3D uses a single image, while SceneGen takes multiple views; both require interactive instance segmentation. Although effective in simplified settings, these methods struggle with realistic, cluttered scenes, often yielding inconsistent object scales and layouts (Fig. 8). In contrast, ShapeR reconstructs objects automatically with consistent scale and layout. Comparison to the recent SAM3D Objects is provided in Appendix Sec. B.

4.2. Ablation Study of ShapeR Components

Effect of SLAM Points. We evaluate the impact of adding SLAM points as an input modality. As shown in Tab. 1

Figure 7. Qualitative comparison against foundation image-to-3D models. For these baselines, we manually select a view with clear object visibility and use interactive SAM2-based segmentations to provide optimal input. In contrast, ShapeR operates fully automatically on multiple posed views and preprocessed inputs, requiring no manual intervention.

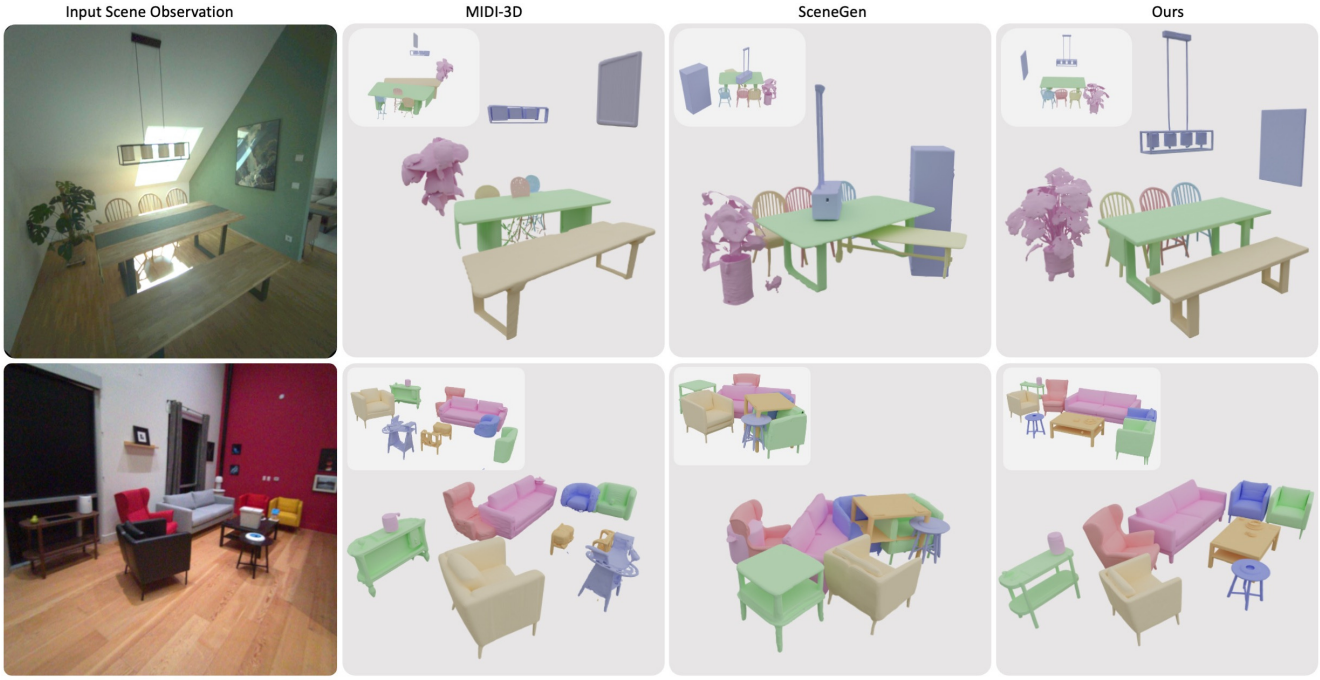

Figure 8. Comparison with image-to-scene methods. MIDI uses a single image and SceneGen uses four views, both with manual object segmentations. These approaches struggle with object scale and arrangement, while ShapeR reconstructs each object metrically and independently, maintaining consistent scale and layout across the scene and without interactive segmentation.

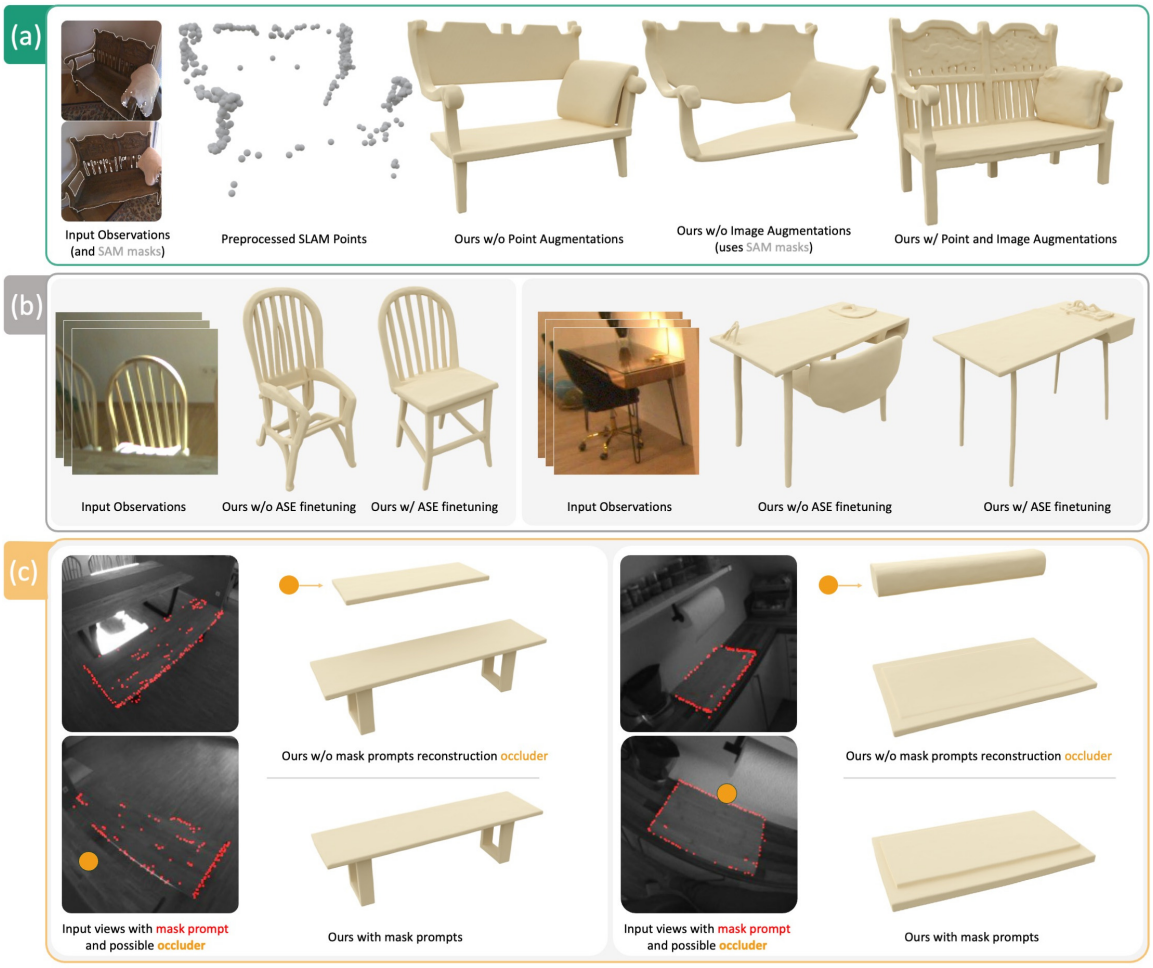

Figure 9. Ablations of components. (a) Without point augmentations, the model overfits to point inputs, missing geometry in regions without points. Image augmentations address occlusions and incomplete objects crops. Omitting background composition requires pre-segmentation, which can introduce noisy masks and prediction errors. (b) Fine-tuning on scene-centric crops improves robustness in challenging scenarios over object-centric training alone. (c) Prompting DINO features with 2D point projections clarifies which object to reconstruct in cluttered scenes, reducing confusion from nearby objects and improving reconstruction accuracy.

and Fig. 4, while image-only inputs yield reasonable reconstructions, incorporating SLAM points significantly improves robustness by providing complementary geometric information that encodes aggregated shape across the entire sequence, especially benefiting cases with weak visual cues.

Effect of Augmentations. Tab. 1 and Fig. 9(a) show that both point cloud and image augmentations are critical for robust real-world performance. Removing either leads to degraded reconstructions under noise and partial observations. The variant without image augmentation relies on explicit foreground segmentation, similar to foundation image-to-3D models, and therefore struggles with noisy masks, underscoring the importance of synthetic occlusion and background augmentation over mask dependence.

Effect of Two-stage Curriculum Training. Fine-tuning on a more realistic scene dataset substantially improves robustness as shown in Tab. 1 and Fig. 9(b). This confirms that combining large-scale object-centric pretraining with realistic scene fine-tuning provides strong generalization to casual captures.

Effect of 2D Point Mask Prompting. Without the 2D point mask cues, our method sometimes reconstructs adjacent objects. Using 2D point masks to guide DINO features mitigates this issue and leads to cleaner reconstructions, as illustrated in Fig. 9(c) and Tab. 1.

5. Conclusion

We introduce ShapeR, a multimodally conditioned rectified flow model for robust 3D shape generation from casually captured sequences. By leveraging posed images, sparse SLAM points, and textual cues, ShapeR reconstructs objects accurately and completely without explicit segmentation. Large-scale training, extensive augmentations, and a

two-stage curriculum enable strong generalization to real-world scenarios. We also present the ShapeR Evaluation Dataset as a benchmark for object-centric reconstruction under casual capture. ShapeR advances scalable and automatic 3D reconstruction in natural environments.

References

[1] 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)*, pages 616–626. IEEE, 2025. 3

[2] Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X Chang, and Matthias Nießner. Scan2cad: Learning cad model alignment in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 2614–2623, 2019. 1, 3

[3] Armen Avetisyan, Angela Dai, and Matthias Nießner. End-to-end cad model retrieval and 9dof alignment in 3d scans. In *Proceedings of the IEEE/CVF International Conference on computer vision*, pages 2551–2560, 2019. 3

[4] Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, et al. Scenescript: Reconstructing scenes with an autoregressive structured language model. In *European Conference on Computer Vision*, pages 247–263. Springer, 2024. 2, 4

[5] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5855–5864, 2021. 2

[6] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19697–19705, 2023. 2

[7] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 2, 5

[8] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025. 4

[9] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 13

[10] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 5

[11] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European conference on computer vision*, pages 333–350. Springer, 2022. 2

[12] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22246–22256, 2023. 2

[13] Rui Chen, Jianfeng Zhang, Yixun Liang, Guan Luo, Weiyu Li, Jiarui Liu, Xiu Li, Xiaoxiao Long, Jiashi Feng, and Ping Tan. Dora: Sampling and benchmarking for 3d shape variational auto-encoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16251–16261, 2025. 3

[14] Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jia-Wei 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, 2025. 13, 14, 15, 16

[15] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6970–6981, 2020. 2

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

[17] Manuel Dahnert, Angela Dai, Norman Müller, and Matthias Nießner. Coherent 3d scene diffusion from a single rgb image. *Advances in Neural Information Processing Systems*, 37:23435–23463, 2024. 3

[18] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 5, 13

[19] Angela Dai, Christian Diller, and Matthias Nießner. Sg-nn: Sparse generative neural networks for self-supervised scene completion of rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 849–858, 2020. 2

[20] Angela Dai, Yawar Siddiqui, Justus Thies, Julien Valentin, and Matthias Nießner. Spsg: Self-supervised photometric scene generation from rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1747–1756, 2021. 1, 2

[21] 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, 2023. 1, 2, 5

[22] 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*, pages 13142–13153, 2023. 1, 5

[23] Zhao Dong, Ka Chen, Zhaoyang Lv, Hong-Xing Yu, Yunzhi Zhang, Cheng Zhang, Yufeng Zhu, Stephen Tian, Zhengqin Li, Geordie Moffatt, et al. Digital twin catalog: A large-scale photorealistic 3d object digital twin dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 753–763, 2025. 2, 5, 13, 14, 16, 18

[24] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[25] 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)*, pages 2553–2560. IEEE, 2022. 5, 13

[26] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *European conference on computer vision*, pages 834–849. Springer, 2014. 4

[27] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017. 2, 3

[28] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 5, 13

[29] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 1

[30] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin Brualla, Pratul Srinivasan, Jonathan Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *Advances in Neural Information Processing Systems*, 37:75468–75494, 2024. 2

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[32] 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. *arXiv preprint arXiv:2311.04400*, 2023. 2

[33] Ji Hou, Angela Dai, and Matthias Nießner. Revealnet: Seeing behind objects in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2098–2107, 2020. 3

[34] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23646–23657, 2025. 1, 3, 6, 14

[35] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568, 2011. 2

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

[37] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, 2006. 2

[38] 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. *arXiv preprint arXiv:2509.13414*, 2025. 15

[39] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[40] Zhengfei Kuang, Yunzhi Zhang, Hong-Xing Yu, Samir Agarwala, Elliott Wu, Jiajun Wu, et al. Stanford-orb: a real-world 3d object inverse rendering benchmark. *Advances in Neural Information Processing Systems*, 36:46938–46957, 2023. 2, 5, 13

[41] Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. Mask2cad: 3d shape prediction by learning to segment and retrieve. In *European Conference on Computer Vision*, pages 260–277. Springer, 2020. 3

[42] Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *arXiv preprint arXiv:2502.06608*, 2025. 1, 2, 3, 4, 6

[43] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2023. 2

[44] Zhengqin Li, Dilin Wang, Ka Chen, Zhaoyang Lv, Thu Nguyen-Phuoc, Milim Lee, Jia-Bin Huang, Lei Xiao, Yufeng Zhu, Carl S Marshall, et al. Lirm: Large inverse rendering model for progressive reconstruction of shape, materials and view-dependent radiance fields. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 505–517, 2025. 2, 5, 14, 16, 18

[45] Zhihao Li, Yufei Wang, Heliang Zheng, Yihao Luo, and Bihan Wen. Sparc3d: Sparse representation and construction for high-resolution 3d shapes modeling. *arXiv preprint arXiv:2505.14521*, 2025. 2

[46] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler,

Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[47] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 2

[48] 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*, pages 429–446. Springer, 2022. 3

[49] 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, 2023. 2

[50] Yanxu Meng, Haoning Wu, Ya Zhang, and Weidi Xie. Scenegen: Single-image 3d scene generation in one feedforward pass. *arXiv preprint arXiv:2508.15769*, 2025. 1, 3, 6, 14

[51] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 2

[52] AI Meta. Llama 4: Multimodal intelligence, 2025. 2, 3, 4

[53] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2

[54] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 2

[55] 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*, pages 414–431. Springer, 2020. 2

[56] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6022–6033, 2025. 3, 5, 14, 15, 17

[57] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 55–64, 2020. 1, 3, 13

[58] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG)*, 32 (6):1–11, 2013. 2

[59] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4

[60] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[61] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 1

[62] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *European Conference on Computer Vision*, pages 523–540. Springer, 2020. 1, 2, 5

[63] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[64] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 4

[65] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 4

[66] 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. *arXiv preprint arXiv:2408.00714*, 2024. 4

[67] Martin Runz, Kejie Li, Meng Tang, Lingni Ma, Chen Kong, Tanner Schmidt, Ian Reid, Lourdes Agapito, Julian Straub, Steven Lovegrove, et al. Frodo: From detections to 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14720–14729, 2020. 3

[68] Mohamed Sayed, John Gibson, Jamie Watson, Victor Prisacariu, Michael Firman, and Clément Godard. Simplerecon: 3d reconstruction without 3d convolutions. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. 2

[69] Yawar Siddiqui, Justus Thies, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. Retrievalfuse: Neural 3d scene reconstruction with a database. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12568–12577, 2021. 2, 5

[70] Yawar Siddiqui, Tom Monnier, Filippos Kokkinos, Mahendra Kariya, Yanir Kleiman, Emilien Garreau, Oran Gafni, Natalia Neverova, Andrea Vedaldi, Roman Shapovalov, et al. Meta 3d assetgen: Text-to-mesh generation with high-quality geometry, texture, and pbr materials. *Advances in Neural Information Processing Systems*, 37:9532–9564, 2024. 2, 5

[71] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl

Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 13, 14, 15, 17

[72] Julian Straub, Daniel DeTone, Tianwei Shen, Nan Yang, Chris Sweeney, and Richard Newcombe. Efm3d: A benchmark for measuring progress towards 3d egocentric foundation models. *arXiv preprint arXiv:2406.10224*, 2024. 1, 2, 3, 4, 5

[73] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15598–15607, 2021. 2

[74] 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*, pages 1–18. Springer, 2024. 2

[75] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2

[76] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2

[77] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3295–3306, 2023. 2

[78] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36: 8406–8441, 2023. 2

[79] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5249–5260, 2025. 5

[80] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Yajie Bao, Jiachen Qian, Siyu Zhu, Xun Cao, Philip Torr, et al. Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention. *arXiv preprint arXiv:2505.17412*, 2025. 1, 2, 6, 14

[81] Tianhao Wu, Chuanxia Zheng, Frank Guan, Andrea Vedaldi, and Tat-Jen Cham. Amodal3r: Amodal 3d reconstruction from occluded 2d images. *arXiv preprint arXiv:2503.13439*, 2025. 1, 3, 6, 14

[82] 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 Computer Vision and Pattern Recognition Conference*, pages 21469–21480, 2025. 1, 2, 3, 6

[83] Zesong Yang, Bangbang Yang, Wenqi Dong, Chenxuan Cao, Liyuan Cui, Yuewen Ma, Zhaopeng Cui, and Hujun Bao. Instascene: Towards complete 3d instance decomposition and reconstruction from cluttered scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7771–7781, 2025. 3

[84] Kaixin Yao, Longwen Zhang, Xinhao Yan, Yan Zeng, Qixuan Zhang, Lan Xu, Wei Yang, Jiayuan Gu, and Jingyi Yu. Cast: Component-aligned 3d scene reconstruction from an rgb image. *ACM Transactions on Graphics (TOG)*, 44(4): 1–19, 2025. 1, 3

[85] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in neural information processing systems*, 34:4805–4815, 2021. 2

[86] Lior Yariv, Peter Hedman, Christian Reiser, Dor Verbin, Pratul P Srinivasan, Richard Szeliski, Jonathan T Barron, and Ben Mildenhall. Bakedsdf: Meshing neural sdfs for real-time view synthesis. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–9, 2023. 2

[87] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 2, 5, 13, 14, 15, 17

[88] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35:25018–25032, 2022. 1, 2

[89] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19447–19456, 2024. 2

[90] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions On Graphics (TOG)*, 42(4):1–16, 2023. 1, 2, 3

[91] 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 (TOG)*, 43(4):1–20, 2024. 2, 3

[92] 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*, 2025. 1, 2, 3, 4, 6, 14

[93] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12786–12796, 2022. 1, 2

ShapeR: Robust Conditional 3D Shape Generation from Casual Captures

Supplementary Material

In this appendix, we provide additional details on the ShapeR evaluation dataset, further experimental results, including results on additional datasets, expanded implementation details of our method, and a discussion of its limitations.

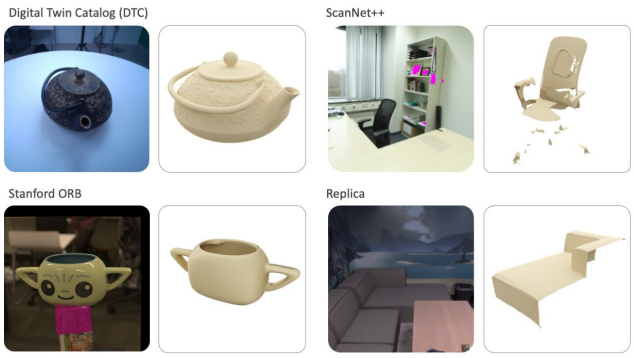

Figure 10. Comparison of 3D reconstruction datasets. DTC [23] and StanfordORB [40] offer controlled studio captures of isolated objects, while ScanNet++ [87] and Replica [71] provide realistic scenes but lack complete ground-truth shapes. The ShapeR evaluation dataset features casually captured sequences with complete meshes for geometric evaluation (see Figs. 11 and 12).

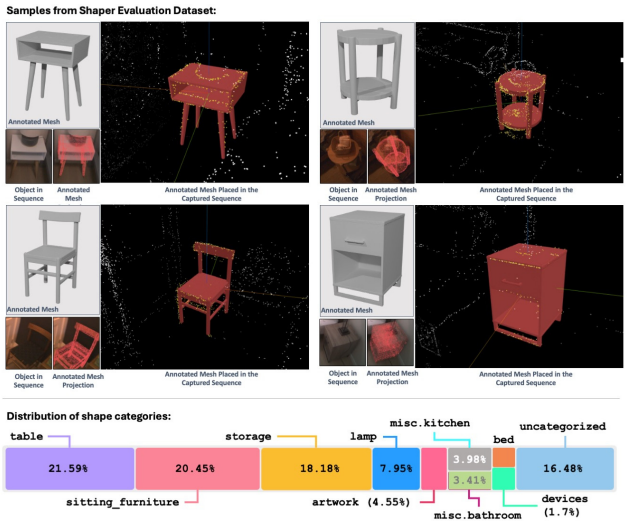

Figure 11. (Top) Examples from the ShapeR evaluation dataset. Each sub-image shows the annotated ground-truth mesh, a representative frame containing the object, the mesh placed within the sequence, and the projection of the mesh onto the image. (Bottom) Distribution of object shapes categories in the ShapeR evaluation set, covering 178 objects across 7 sequences

A. ShapeR Evaluation Dataset

Existing real-world 3D datasets for object reconstruction can be classified into two broad categories. Some, like Digital Twin Catalog [23], StanfordORB [57] and Google Scanned Objects [25] provide complete 3D shape geometry, but only in highly controlled setups. Here, objects are the central focus, placed on uncluttered, disoccluded tabletops, and captured in studio-like conditions (see Fig. 10 left). These datasets typically feature relatively small objects. Others, like ScanNet [18], ScanNet++ [87], Matterport3D [9] offer realistic scene arrangements, with clutter and occlusions captured casually. However, these are not suitable for object-centric evaluation, as the target geometry, usually obtained by 3D scanning, is incomplete in occluded for unobservable regions (see Fig. 10 right). The ShapeR Evaluation Dataset addresses these limitations by providing complete mesh geometry annotations for a selected set of objects, while maintaining casual capture conditions.

As shown in Fig. 12, sequences are recorded using Project Aria [28] Gen 1 or Gen 2 glasses, with the annotator casually walking through the scene and collecting images from the device's RGB and CV cameras. Aria Machine Perception Services [28] are then used to extract SLAM points and camera parameters from the sequence. For a selected set of objects, we obtain 3D shape annotations by moving each object to an area free of clutter and occlusions, capturing a high-resolution image, and manually segmenting it. A state-of-the-art image-to-3D model is then used to generate the 3D geometry. This geometry is manually verified for plausibility and aligned to the object's position in the original casual sequence using a web interface. This interface allows annotators to reposition and rigidly deform the shape in 3D space, guided by SLAM points from the sequence. Annotators further verify placement and dimensions by projecting the mesh into the original sequence images.

In total, we annotate 178 objects across 7 real indoor sequences, spanning a range of categories. Fig. 11 shows sample objects and the distribution of categories in the dataset.

B. Additional Experiments

In this section, we provide additional evaluations of ShapeR across a variety of datasets and tasks. We include comparisons against SegmentAnything 3D Objects [14], assessments on ScanNet++ [87] and Replica [71], results on the Digital Twin Catalog [23] (DTC), analysis of robustness trends, and demonstrations of monocular image-to-3D reconstruction.

Figure 12. To obtain pseudo-ground truth geometry for an object in the sequence (left), we first place the object in isolation to avoid clutter and occlusion, and capture a high-quality, uncluttered image. We then apply segmentation and image-to-3D modeling to generate the object's geometry (mid). This geometry is manually aligned and inserted back into the original casual sequence using a web annotation interface, verified by matching 2D projections to image silhouettes and by checking alignment with the sequence's point cloud (right).

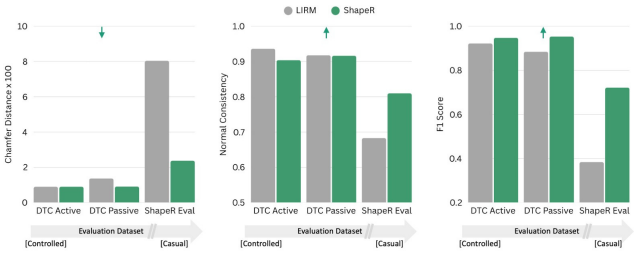

Figure 13. DTC Active, DTC Passive, and ShapeR Evaluation datasets represent a progression from highly controlled capture setups (DTC Active), to slightly less controlled environments (DTC Passive), and finally to casual, real-world scenes (ShapeR Evaluation). As the datasets become more challenging, baseline method metrics deteriorate, while ShapeR remains comparatively stable. Notably, the increase in scene casualness is not linear; ShapeR Evaluation is significantly more challenging than DTC Passive.

Comparison against SegmentAnything 3D Object [14]. SAM 3D Objects was very recently released and addresses the single image-to-3D reconstruction task using *interactive* segmentation. This approach marks a significant improvement in shape quality compared to previous image-to-scene methods like MIDI3D [34] and SceneGen [50], as well as single image-to-3D models such as Hunyuan3D [92], Amodal3R [81], and Direct3DS2 [80].

However, SAM 3D Objects is fundamentally limited by its reliance on single images. As a result, the reconstructed shapes are not metrically accurate. When scenes become more cluttered and contain multiple objects, the method struggles: layout, shape quality, aspect ratios, and relative scales all deteriorate, as shown in Fig. 16. In contrast, ShapeR leverages multiple posed views and additional modalities (such as SLAM points) to *automatically* reconstruct objects with metric accuracy and robust layout, even in casual, cluttered environments, while having only ever been trained on synthetic data. This multimodal approach enables ShapeR to maintain high-quality, metrically consistent reconstructions and object arrangements without interaction, outperforming single image-based methods in challenging real-world scenarios.

Evaluation on Scannet++ [87] and Replica [71]. Fig. 18 and Tab. 3 present a comparison of ShapeR on third-party casually captured datasets. For these experiments, we follow the protocol of DP-Recon [56], using their six ScanNet++ scenes and seven Replica scenes for evaluation. Since these datasets do not provide complete 3D geometry for evaluation (Figs. 10 and 18), we report only recall-based metrics. Notably, ShapeR produces complete reconstructions, often surpassing the ground-truth scans in terms of completeness, as the ground-truth meshes lack geometry in occluded regions.

Evaluation on Digital Twin Catalog (DTC) [23]. Fig. 17 and Tab. 4 show a comparison of ShapeR against LIRM [44] on the controlled capture datasets DTC Active and DTC Passive. Both datasets contain approximately 100 sequences each, with objects placed on a tabletop, free from occlusions and clutter. The passive variant allows for more free user movement, making it more casual compared to the active variant, where the user circles the object. As highlighted in Tab. 4, ShapeR matches state-of-the-art LIRM quality on the highly controlled active set and surpasses it on the more casual passive variant. Additionally, ShapeR produces sharper details on both datasets, as illustrated in Fig. 17.

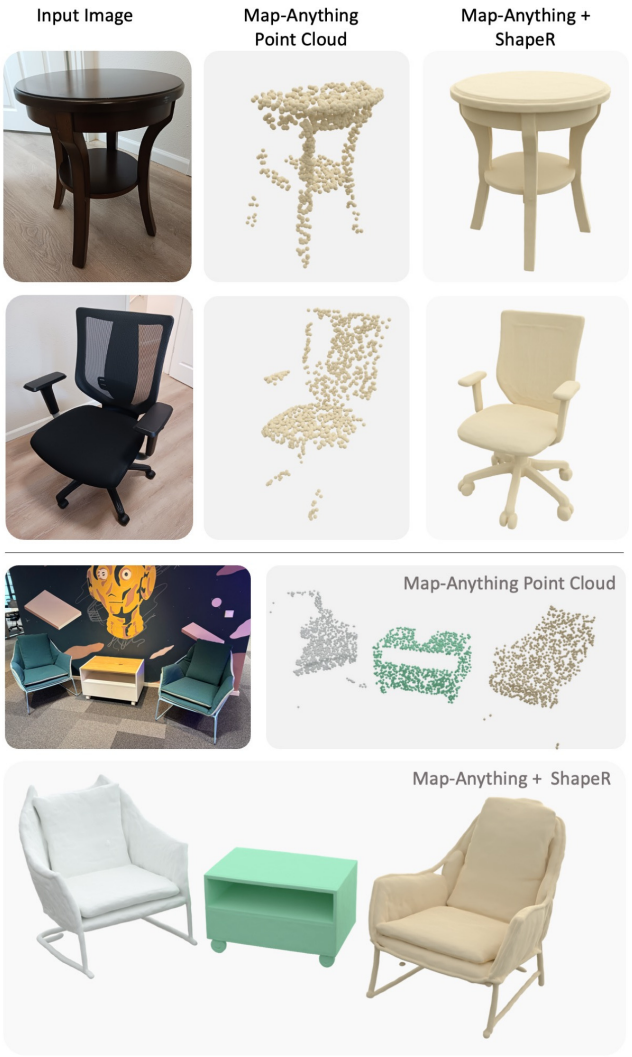

Figure 14. Single image to metric 3D with ShapeR. While ShapeR is trained to leverage posed multi-view signals, it can be configured for single-image 3D reconstruction without retraining by using a metric point cloud and camera estimator such as MapAnything [38]. This enables ShapeR to generate metrically accurate 3D shapes from a monocular image.

Figure 15. ShapeR limitations. (a) Low image fidelity or limited views lead to incomplete or low-detail reconstructions. (b) Closely stacked or attached objects can cause meshes to include parts of adjacent structures, even when the point associated with these structures are not in the input (c) ShapeR relies on upstream 3D detection; missed or inaccurate detections result in unrecoverable objects.

Table 3. Reconstruction performance comparison on ScanNet++ [87] and Replica [71] datasets against DPRecon [56]. We use six scenes from ScanNet++ and seven scenes from Replica as processed by DPRecon. Note that chamfer distance, normal consistency and recall (R) are calculated in one direction, *i.e.* only point present on ground truth meshes are used for evaluation, due to tha lack of incomplete meshes present in these datasets.

Methods	ScanNet++			Replica		
	CD$\times 10^2 \downarrow$	NC$\uparrow$	R$\uparrow$	CD$\times 10^2 \downarrow$	NC$\uparrow$	R$\uparrow$
DPRecon [56]	7.69	0.73	0.45	4.65	0.75	0.57
ShapeR	**1.09**	**0.84**	**0.91**	**1.77**	**0.84**	**0.82**

Robustness Trends. DTC Active, DTC Passive, and the ShapeR evaluation dataset represent a non-linear progression from highly controlled to markedly more complex and casual capture setups. As shown in Fig. 13, ShapeR demonstrates significantly greater robustness to increased scene casualness compared to baseline methods such as LIRM, maintaining high reconstruction quality even as the capture conditions become more challenging.

Monocular Image-to-3D. While ShapeR is trained using multiple posed views and SLAM points extracted from them, it can also be applied to monocular images to produce metric 3D shapes without retraining by leveraging approaches like MapAnything [38]. As illustrated in Fig. 14, ShapeR can condition on a single image and its associated point cloud (obtained from MapAnything) to reconstruct both individual objects and entire scenes. Further improvements are possible by fine-tuning the model on real data collected in this monocular setup, as demonstrated in recent works [14].

C. Implementation Details

The 3D VAE encoder consists of 8 transformer layers and the decoder of 16 layers, each with a hidden width of 768, 12 attention heads. The VAE is trained for 200K steps with an effective batch size of 640 across 64 NVIDIA H100

Figure 16. Comparison with SAM 3D Objects [14]. SAM 3D Objects takes a single image and interactive object segments to produce non-metric 3D shapes, which are generally accurate but may exhibit minor hallucinations (e.g., predicting five lamps instead of four) and poor object placement. In contrast, ShapeR leverages posed images from a sequence to generate metrically accurate geometry and consistently well-placed objects.

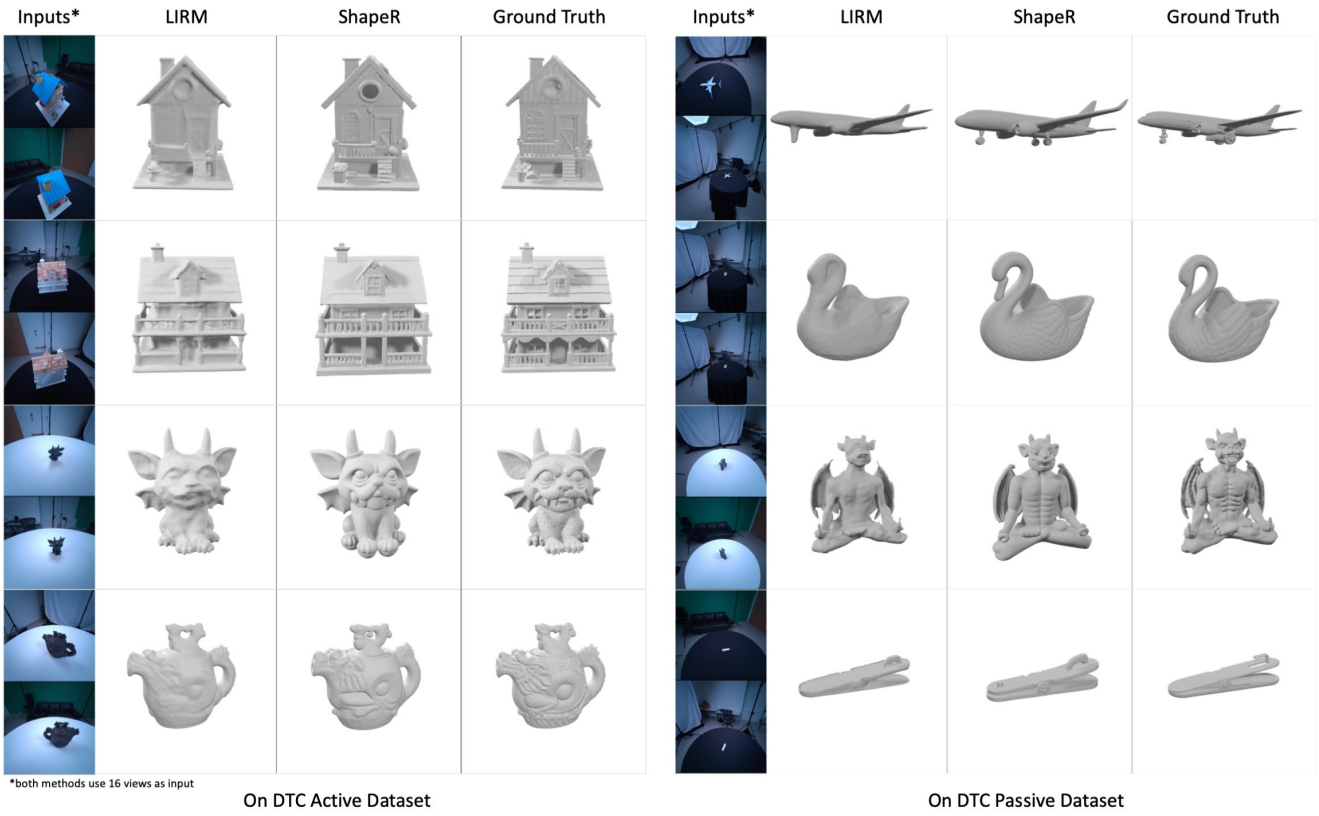

Figure 17. Comparison against LIRM [44] on DTC [23] Active and Passive sequences. Both setups feature tabletop objects without clutter or occlusions; however, Passive sequences allow more free user movement, while Active sequences involve the user circling the object. ShapeR performs competitively on Active sequences and surpasses LIRM on the slightly more casual Passive sequences.

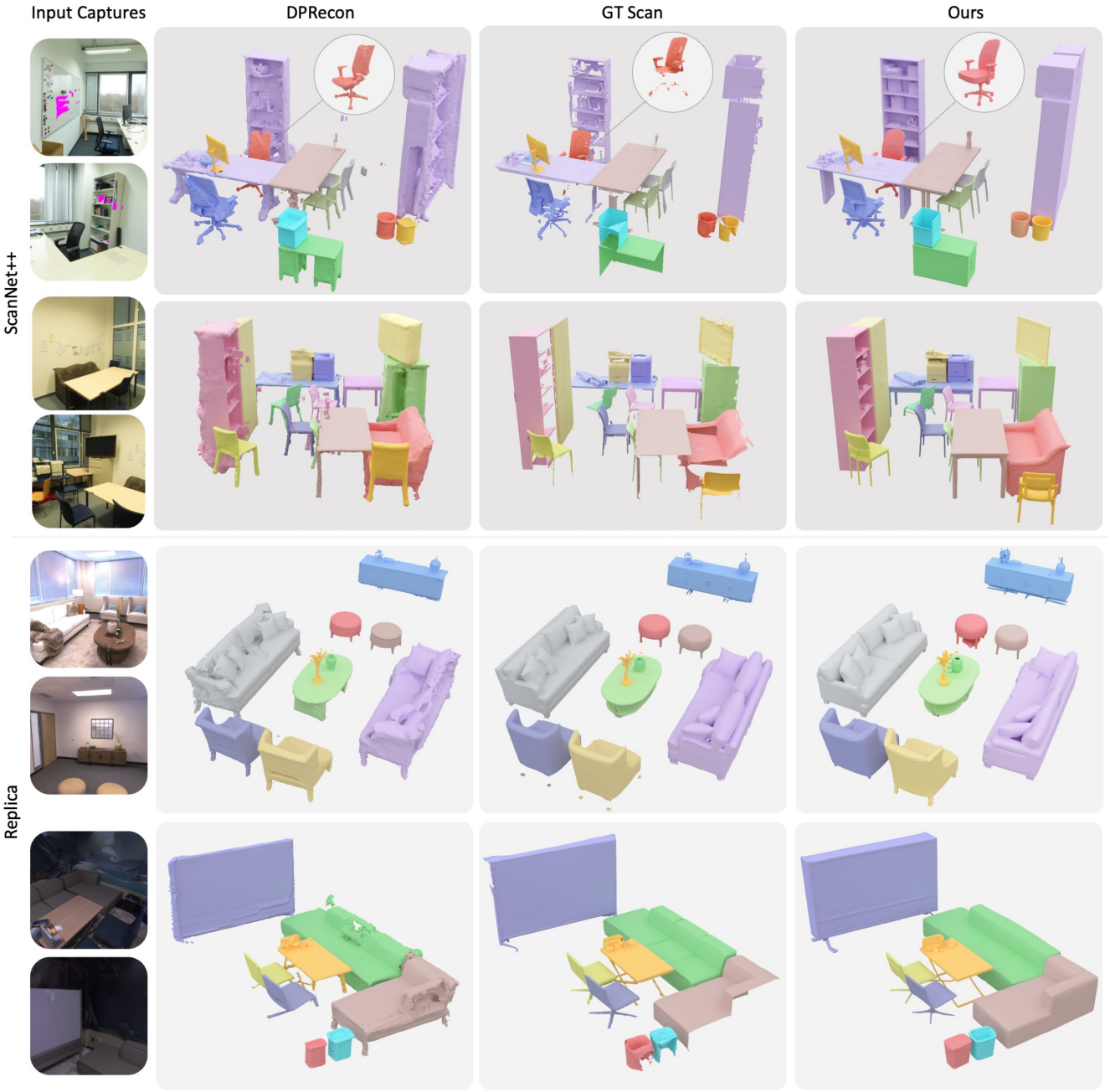

Figure 18. Reconstruction results on ScanNet++ [87] and Replica [71] scenes, compared to DPRecon [56]. ShapeR produces complete reconstructions, often surpassing the ground-truth scans in completeness, as the latter lack geometry in occluded regions.

GPUs. The rectified flow transformer comprises 16 dual-stream and 32 single-stream blocks, each with 16 attention heads and a hidden width of 1024. Training is performed for 550K steps using 128 H100 GPUs, progressively increasing the latent sequence length. The effective batch size is 512. Both networks are optimized using Adam with a learning rate of 5×10^{-5}.

D. Limitations

While ShapeR advances 3D shape generation under casual capture scenarios, several limitations remain. First, for objects captured with low image fidelity or observed in very few views, reconstructions can be incomplete or lack fine detail due to insufficient geometric and visual evidence. Second, when objects have other items stacked or closely attached (for example, tables supporting other objects), the

Table 4. Reconstruction results on the DTC [23] Active and Passive datasets, each with approximately 100 sequences, compared against LIRM [44]. ShapeR achieves comparable performance to LIRM on the highly controlled Active sequences, and surpasses LIRM on the more challenging Passive sequences.

Methods	DTC Active			DTC Passive		
	CD$\times 10^2 \downarrow$	NC$\uparrow$	F1$\uparrow$	CD$\times 10^2 \downarrow$	NC$\uparrow$	F1$\uparrow$
LIRM [44]	**0.90**	**0.94**	0.92	1.37	0.91	0.88
ShapeR	0.94	0.91	**0.94**	**0.95**	0.91	**0.95**

reconstructed meshes sometimes include remnants of these adjacent structures instead of cleanly isolating the target object. Finally, ShapeR depends on upstream 3D instance detection; thus, missed detections or inaccurate bounding boxes directly propagate to the reconstruction stage, where missed objects cannot be recovered.