SPOON: TOWARDS COHERENT COMPOSITIONAL 3D SCENE GENERATION FROM UNCALIBRATED MULTI-VIEW IMAGES

Anonymous authors
Paper under double-blind review



Figure 1: Given multi-view images, current 3D generation pipelines often yield a soup of objects. SPOON can recover a coherent compositional 3D scene with accurate object geometry and layout.

ABSTRACT

Compositional 3D scene generation aims to recover complete 3D object shapes and their spatial arrangement from visual observations. Recent image-conditioned 3D generators provide strong priors for producing high-quality object geometry, making the generation of complex scenes increasingly practical. A central challenge is therefore to spatially organize these generated assets into a globally coherent scene while remaining consistent with multi-view observations. Existing approaches either entangle scene layout with object generation or separately estimate spatial placement from view-specific observations, where pose hypotheses may remain ambiguous and inconsistent across views, often resulting in an incoherent *object-camera soup*. We introduce SPOON, a framework that reformulates multi-view compositional 3D generation as scene-level, geometry-grounded pose reasoning. Rather than treating view-specific object pose hypotheses independently, SPOON coordinates them using reconstruction-derived multi-view geometry through a *Guide-Route-Reconcile* paradigm. This progressively organizes object poses and camera configurations into a coherent scene-level spatial arrangement. Extensive experiments on ARSG-110K and MIDI-3D-Front demonstrate consistent improvements in object placement and scene composition across varying numbers of input views. On ARSG-110K, SPOON reduces scene-level and object-level Chamfer distances by 12.7% and 17.7%, respectively, compared with a strong baseline.

1 INTRODUCTION

Multi-view compositional 3D scene generation can construct coherent scenes by recovering complete object assets together with their spatial arrangement from visual observations, supporting applications in embodied AI, simulation, and 3D content creation. Recent advances in image-conditioned 3D generation Xiang et al. (2025b;a); Huang et al. (2026a); Chen et al. (2026) provide strong priors for recovering high-quality object geometry, making compositional 3D scene generation increasingly practical. A challenge is therefore how to incorporate multi-view observations and recover the spatial layout that organizes these generated assets into a coherent scene.

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.

One strategy is to incorporate layout directly into the generation process. For example, 3D-Fixer Yin et al. (2026) conditions object completion on partial geometry in the scene coordinate frame, using the observed geometry as a spatial anchor to generate objects in place. While avoiding explicit post-generation pose recovery, this formulation entangles shape generation with scene-space placement, requiring object generative priors learned in canonical space to additionally accommodate incomplete observations under arbitrary spatial transformations.

An alternative strategy preserves object generation in its canonical space and models spatial placement separately. SAM 3D Chen et al. (2026), for example, generates canonical object geometry while recovering its transformation with respect to the observing camera. Such a decomposition preserves the advantages of object-centric generative priors, while making accurate pose recovery critical to coherent scene composition. Pose estimates inferred from individual object-view observations are inherently scale-ambiguous and can become inaccurate under occlusion, image-boundary truncation, and limited visibility. Multiple views provide complementary evidence, yet do not automatically resolve this ambiguity: different observations may support pose estimation with substantially different reliability, and independently plausible hypotheses can become mutually inconsistent when expressed in a shared scene frame. Thus, these methods often provide an incoherent *object–camera soup*.

Scene-level 3D reconstruction provides complementary geometric cues for resolving this ambiguity. It jointly recovers camera geometry and scene structure in a shared coordinate system, establishing globally consistent relations across observations. This shared geometry provides a common reference for coordinating object hypotheses across cameras and reasoning about their support within the overall scene. This raises a natural question: *Can scene-level geometry organize locally plausible object–camera hypotheses into a globally coherent 3D composition?*

We answer this question with SPOON, standing for **S**cene-level **P**ose and **O**bject layout **O**ptimizatio**N**. We cast multi-view object placement as a *scene-level pose reasoning problem*, where view-conditioned hypotheses are coordinated through shared scene geometry rather than considered in isolation. In this way, SPOON *stirs the object-camera soup into a coherent scene*, combining strong object-centric generative priors with scene-level geometric reasoning for coherent spatial placement. SPOON follows a *Guide–Route–Reconcile* paradigm. First, we **Guide** pose generation with reconstruction-derived scene geometry, grounding view-conditioned predictions in globally consistent spatial cues. Second, we **Route** the resulting pose hypotheses by assessing the geometric reliability of each view-specific observation against the reconstructed object geometry. Finally, we **Reconcile** the object layouts and camera poses through joint differentiable optimization, reducing residual object–camera inconsistencies while preserving the generated object geometry. Together, these stages progressively coordinate local pose hypotheses into a coherent scene-level spatial configuration without altering the underlying object generative prior. Our main contributions are:

- We identify the *object–camera soup* problem in compositional 3D scene generation, where locally plausible object-camera hypotheses may not constitute a coherent scene-level configuration.
- We propose SPOON, a scene-level pose reasoning framework that follows a *Guide–Route–Reconcile* paradigm to guide pose generation using shared geometry, route view-specific hypotheses by geometric reliability, and jointly reconcile object layouts and camera poses.
- Extensive experiments on ARSG-110K and MIDI-3D-Front show consistent improvements in scene composition and object placement across varying numbers of input views.

2 Related work

2.1 3D Shape Generation

Recent advances in 3D generation have substantially improved the quality and scalability of object-level asset creation from textual or visual inputs. Early methods were largely developed on relatively small-scale 3D datasets Wang et al. (2018); Gkioxari et al. (2019); Chang et al. (2015), whereas the emergence of large-scale collections such as Objaverse Deitke et al. (2023b) and Objaverse-XL Deitke et al. (2023a) has enabled increasingly powerful generative priors. Modern native 3D generators explore a wide range of representations, including VecSets Zhang et al. (2023), tri-planes

Gao et al. (2022); Wu et al. (2024), geometry images Yan et al. (2025); Elizarov et al. (2024), 3D primitives Chen et al. (2025d), structured latents Xiang et al. (2025b); Wu et al. (2025); Chen et al. (2025b), and autoregressive mesh representations Fang et al. (2025); Chen et al. (2024); Tang et al. (2025); Siddiqui et al. (2024); Chen et al. (2025a;c). Together with advances in scalable architectures and optimized 3D operators Wu et al. (2025); Chen et al. (2025b); Xiang et al. (2025a), these developments have significantly improved the geometric fidelity and structural complexity of generated 3D assets. The increasing availability of high-quality object generation makes it increasingly feasible to extend object-level generative models toward compositional 3D scene generation. Our work builds upon this progress and focuses on reliable object placement and spatial organization in complex multi-view scenes.

2.2 Compositional 3D Scene Generation

Building upon strong object-level generative priors, recent studies have extended 3D generation toward compositional scenes with multiple objects and observations. Two design questions are particularly relevant in this setting: how to integrate complementary evidence across views, and how to recover the spatial layout of generated objects within a shared scene frame.

For multi-view object generation, recent approaches increasingly combine reconstruction and generative priors to improve geometric completeness and input consistency. ReconViaGen Chang et al. (2026), UniRecGen Huang et al. (2026b), and Mix3R Lin et al. (2026) transfer knowledge from 3D reconstruction foundation models Wang et al. (2025; 2026b) to condition 3D generators on multi-view observations. These approaches demonstrate the benefit of integrating reconstruction priors or multi-view consistency into generative models, but primarily focus on aggregating information across observations for object shape enhancement.

Accurately modeling object layouts further requires 3D generation models to be explicitly pose-aware. Recent approaches decouple shape from camera or object pose using representations such as UV volumes Huang et al. (2026a) or camera tokens Chen et al. (2026). For example, ShapeR Siddiqui et al. (2026) first predicts object-to-world transformations from detected bounding boxes and subsequently generates their 3D shapes under the estimated placements. Another line of work directly applies layout transformations to object representations and learns to model the resulting SE(3)-transformed shapes within a scene volume. While this formulation enables pixel-aligned geometry Li et al. (2026b), entangling arbitrary object transformations with shape representation broadens the distribution that the generative model must capture and can degrade the geometry of generated meshes Ling et al. (2026); Yin et al. (2026). This motivates compositional formulations that preserve object geometry in canonical space while modeling spatial transformations separately. SceneGen Meng et al. (2026) predicts object positions before assembling independently generated assets, while CAST Yao et al. (2025) further refines the resulting scene composition using photometric and physics-aware objectives. MV-SAM3D Li et al. (2026a) extends to multi-view observations through Multi-Diffusion Bar-Tal et al. (2023) and confidence-aware latent velocity fusion. These formulations retain strong object-level generative priors while exposing spatial layout as an explicit variable for scene composition.

3 Preliminaries

3.1 Joint Shape and Pose Modeling

Given an image prompt, SAM3D Chen et al. (2026) can recover the object shape and the camera pose. It adopts a two-stage flow-matching architecture: the first stage jointly generates canonical object geometry (as low-resolution occupancy) and the relative camera pose; the second stage generates the object's texture and fine geometry. Given an object observation $\mathcal{I}$, the first-stage shape and pose branches predict the corresponding velocity fields:

$$\mathbf{v}_\theta^S, \mathbf{v}_\theta^P = f_\theta(\mathbf{z}_t^S, \mathbf{z}_t^P, \mathcal{I}, t), \tag{1}$$

where $\mathbf{z}_t^S$ denotes the shape latent and $\mathbf{z}_t^P = (\mathbf{z}_t^R, \mathbf{z}_t^T, \mathbf{z}_t^s)$ is the camera-relative pose latent, parameterizing rotation, translation, and scale at flow time t. Under rectified flow, both states evolve according to their predicted velocities:

$$\mathbf{z}_{t+\Delta t}^S = \mathbf{z}_t^S + \Delta t\, \mathbf{v}_\theta^S, \qquad \mathbf{z}_{t+\Delta t}^P = \mathbf{z}_t^P + \Delta t\, \mathbf{v}_\theta^P. \tag{2}$$

At the endpoint, the shape latent is decoded into a canonical object representation $\mathcal{G}_o$, while the pose state yields a camera-frame transformation $T^c_{o,v}$ that places the generated object relative to view v. Instead of entangling object shape with its layout, separately modeling geometry and layout keeps the data distribution simple and easy to learn, which is beneficial in generating high-fidelity 3D shapes. However, this might sacrifice the accuracy of the layout estimation.

3.2 Multi-View Flow Fusion

Multi-Diffusion Bar-Tal et al. (2023) fuses prompt-conditioned velocity predictions into a global coherent direction at each generation step, which makes a flow matching model accept multiple conditions. MV-SAM3D Li et al. (2026a) extends SAM3D to multi-view generation to exploit complementary cues across observations. Given V views, the corresponding shape velocities $\{\mathbf{v}^S_{\theta,v}\}^V_{v=1}$ are aggregated into a shared update:

$$\mathbf{v}^S_\theta = \mathcal{A}\big(\{\mathbf{v}^S_{\theta,v}\}^V_{v=1}\big), \qquad \mathbf{z}^S_{t^+} = \mathbf{z}^S_t + \Delta t\, \mathbf{v}^S_\theta, \tag{3}$$

where $\mathcal{A}$ denotes the multi-view velocity fusion operator, $t^+ = t + \Delta t$. This yields a shared shape trajectory that integrates complementary evidence across observations. In contrast, the pose trajectories remain view-conditioned and evolve independently. Consequently, multi-view fusion produces a shared canonical shape while retaining view-specific camera-relative pose hypotheses.

4 Method

Given V unposed RGB images $\{I_v\}^V_{v=1}$ and the corresponding instance masks $\{M_{o,v}\}$, our goal is to reconstruct a compositional 3D scene containing O observed objects. We leverage feedforward 3D reconstruction models Wang et al. (2026a) to recover per-view point maps X_v and camera intrinsics K_v and world-to-camera extrinsics E_v in a shared world coordinate system.

For each object, we follow MV-SAM3D Li et al. (2026a) to obtain a shared canonical representation $\mathcal{G}_o$ together with view-specific camera-frame pose hypotheses $\{T^c_{o,v}\}^V_{v=1}$. We seek an object-to-world transformation T^{ow}_o for each canonical object, yielding the compositional scene:

$$\mathcal{S} = \{(\mathcal{G}_o, T^{\text{ow}}_o)\}^O_{o=1}, \tag{4}$$

where T^{ow}_o places $\mathcal{G}_o$ into the shared world frame.

While MV-SAM3D produces plausible camera-relative object placements, these view-specific hypotheses may collectively lack a scene-coherent spatial configuration, giving rise to an *object–camera soup*. To address this, SPOON uses multi-view geometric cues as a scene-level reference and organizes these hypotheses through a *Guide–Route–Reconcile* paradigm, as described below.

4.1 Geometry-Guided Object Pose Generation

While multi-view flow fusion (Sec. 3.2) aggregates complementary cues for shape generation, pose generation in Li et al. (2026a) remains view-specific. Such independently inferred poses may be ambiguous and unreliable from a single image, whereas feed-forward 3D reconstruction models Wang et al. (2026a) recover camera poses from the full multi-view input with stronger geometric consistency. Thus, we propose to use these reconstruction-derived poses as an inference-time geometric prior to guide pose generation.

A straightforward solution is to align reconstruction-derived camera poses in the world frame with generated poses in the object canonical frame and use the aligned poses as guidance. However, this requires decoding intermediate pose states from the flow trajectory, which are often unstable and can introduce substantial errors into the cross-frame alignment. In contrast, the canonical object surface remains substantially more stable than the decoded pose states and therefore provides a more reliable anchor for cross-frame alignment. Therefore, we establish the cross-frame correspondence through object geometry, aligning the aggregated reconstruction with the canonical object surface to bridge the two coordinate systems.

Surface coordinate bridge. Concretely, we construct object surfaces in both coordinate frames. Using the point maps and camera extrinsics from 3D reconstruction methods Wang et al. (2026a),

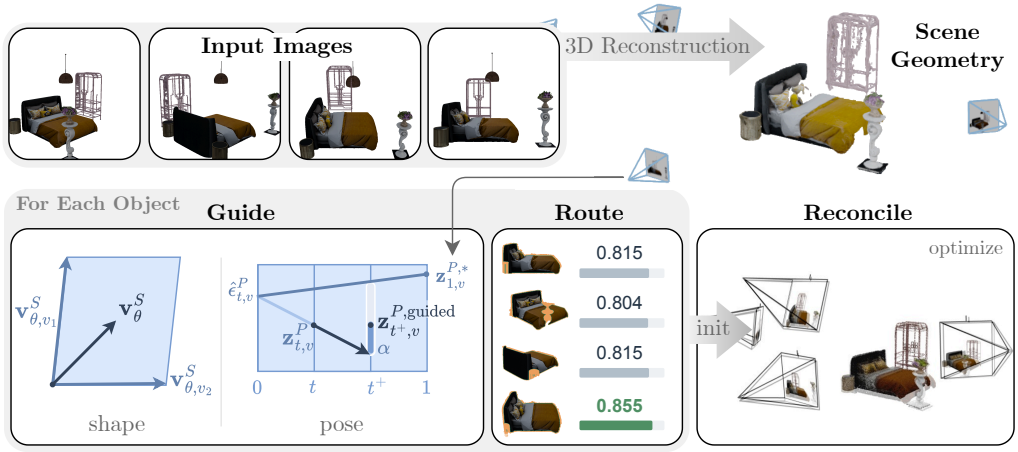

Figure 2: **Overview of SPOON.** Given multiple uncalibrated images, 3D reconstruction methods first recover the scene geometry as camera poses and depth maps. Then, for each object, we prompt the mesh generation process with multi-view visual cues and *guide* the pose recovery with reconstruction guidance, which is then placed using an optimal layout hypothesis through *routing*. Finally, the object layouts and camera poses are jointly optimized and *reconciled* to provide a coherent compositional 3D scene.

we transform the masked object points into the shared world frame and aggregate them into a world-space point cloud $\mathcal{P}_W$. To establish a correspondence with the canonical frame where each shape is generated, we recover a clean occupancy from the noisy shape latent and the fused velocity:

$$\widehat{\mathbf{z}}_1^S = \mathbf{z}_t^S + (1-t)\mathbf{v}_\theta^S, \tag{5}$$

where $\mathbf{v}_\theta^S$ denotes the fused shape velocity. We decode $\widehat{\mathbf{z}}_1^S$ and sample its surface to obtain the canonical point cloud $\mathcal{P}_C$. Since $\mathcal{P}_W$ and $\mathcal{P}_C$ lie in different coordinate systems, solving the transformation between them can robustly transfer the reconstructed camera pose into the canonical space in shape generation. After centering and normalizing both point cloud sets, we search for an optimal rotation that best aligns them, yielding the bridge-aligned pose $\boldsymbol{\pi}_v^*$. We then encode this aligned pose using the SAM3D pose encoder:

$$\mathbf{z}_{1,v}^{P,*} = \mathrm{Enc}_P\left(\boldsymbol{\pi}_v^*\right), \tag{6}$$

which serves as the geometry-derived clean endpoint for the pose rectified flow.

Flow-consistent endpoint injection. Directly replacing the intermediate pose state $\mathbf{z}_{t,v}^P$ with the geometry-derived endpoint $\mathbf{z}_{1,v}^{P,*}$ would ignore the current flow time and deviate from the learned trajectory. We instead construct a reference state at the same flow time. From the current state and predicted velocity, we recover the corresponding noise endpoint as $\widehat{\boldsymbol{\epsilon}}_{t,v}^P = \mathbf{z}_{t,v}^P - t\mathbf{v}_{\theta,v}^P$. We then keep this noise realization fixed and replace only the clean endpoint with $\mathbf{z}_{1,v}^{P,*}$ to define a time-matched reference state, which is then used to guide the Euler update:

$$\mathbf{z}_{t^+,v}^{P,\mathrm{ref}} = (1-t^+)\widehat{\boldsymbol{\epsilon}}_{t,v}^P + t^+\mathbf{z}_{1,v}^{P,*}, \tag{7}$$

$$\mathbf{z}_{t^+,v}^{P,\mathrm{guided}} = (1-\alpha_P)\mathbf{z}_{t^+,v}^{P,\mathrm{base}} + \alpha_P\mathbf{z}_{t^+,v}^{P,\mathrm{ref}}. \tag{8}$$

where $\mathbf{z}_{t^+,v}^{P,\mathrm{base}}$ denotes the standard Euler proposal. This update retains the noise endpoint inferred from the current linear flow approximation while steering the clean endpoint toward the geometry-derived pose.

4.2 Geometry-Aware Pose Routing

Geometry guidance improves object pose generation, yet the resulting hypotheses can still vary in reliability across views. Differences in visibility and boundary truncation may affect the geometric

evidence available for pose estimation. Since pose accuracy cannot be directly assessed at inference time, we use geometric observability as a proxy for reliability. Specifically, we derive a per-view reliability score from its support under the aggregated multi-view geometry and use it to select the pose hypothesis for world-frame placement. For clarity, we omit the object index o below.

Multi-view geometric support. We first lift the masked point-map observations from all views into the shared world frame using the estimated camera extrinsics. Their union is voxel-downsampled into an aggregated object point cloud $\mathcal{P}_W$, which combines complementary observations and serves as a geometric proxy for the object extent. For each view v, we reproject $\mathcal{P}_W$ onto the image plane: $\widehat{M}_v = \mathrm{Rasterize}\left(\{\Pi(K_v\,\mathcal{T}(E_v, \mathbf{P})) \mid \mathbf{P} \in \mathcal{P}_W\}\right)$, where $\Pi(\cdot)$ denotes perspective projection. We evaluate $\widehat{M}_v$ before restricting it to the valid image domain Ω_v, so that it also captures geometry projected beyond the image boundary. Together with the observed mask M_v, this provides complementary cues for boundary truncation and visible coverage.

Geometry-aware reliability. We characterize geometric observability using frame completeness F_v and visible coverage C_v:

$$F_v = \frac{|\widehat{M}_v \cap \Omega_v|}{|\widehat{M}_v|}, \quad C_v = \frac{|M_v \cap \widehat{M}_v|}{|\widehat{M}_v \cap \Omega_v|}, \quad S_v = [F_v \geq \tau_f]\, F_v^{\alpha} C_v, \tag{9}$$

where F_v measures how much of the projected object extent remains inside the image, while C_v measures how much of the in-frame geometric support is covered by the observation. Here, τ_f is the completeness threshold used to filter severely truncated observations. Their combination S_v serves as a geometry-aware proxy for pose reliability.

Reliability-aware pose routing. We select the most reliable view and transfer its pose hypothesis to the shared world frame:

$$v^\star = \arg\max_v S_v, \qquad T_o^{\mathrm{ow}} = E_{v^\star}^{-1} T_{o,v^\star}^{c}, \tag{10}$$

where $T_{o,v^\star}^{c}$ is the object transformation predicted in the selected camera frame. This converts heterogeneous multi-view pose hypotheses into a single placement supported by the most geometrically reliable observation.

4.3 Multi-View Pose Reconciliation

The strategy introduced in Sec. 4.2 improves the object layout by selecting the most plausible layout among the candidates estimated from different input views. However, the quality of the resulting layout remains bounded by the accuracy of the initial per-view layout estimations. To further improve the consistency of the reconstructed scene, we leverage the multi-view observations and jointly refine the object layouts through differentiable Gaussian splatting.

Specifically, for the o-th object, we introduce an additional $\mathrm{Sim}(3)$ transformation parameterized by (s', R', t') to refine its initial layout $(s, R, t) = T_o^{\mathrm{ow}}$. The resulting transformation is given by:

$$s \leftarrow s's, \qquad R \leftarrow R'R, \qquad t \leftarrow s'R't + t'. \tag{11}$$

We initialize the additional transformation as the identity, i.e., $s' = 1$, $R' = I_{3\times3}$, and $t' = 0$. The transformed object Gaussians are then composited and rendered from the input viewpoints, whose camera poses are obtained from a 3D reconstruction foundation model Wang et al. (2026a). In this way, the multi-view observations provide image-space supervision for refining the object layout without requiring explicit cross-view point or feature correspondences.

Although this optimization can effectively correct moderate layout errors, its performance remains coupled with the estimated camera geometry. Errors in the estimated camera poses perturb multi-view consistency, limiting the effectiveness of object-level refinement. To address this issue, we refine the camera poses jointly with the object layouts. We employ a Gaussian rasterization kernel with explicit camera Jacobians Matsuki et al. (2024), which enables efficient gradient-based optimization of the camera parameters through the rendering process. Conceptually, this optimization is analogous to Bundle Adjustment (BA) in Structure-from-Motion, where camera poses and scene geometry are jointly optimized to achieve multi-view consistency. Rather than relying on explicit point or feature correspondences, our formulation directly minimizes the discrepancy between

Table 1: **Main results on the ARSG-110K and MIDI-3D-Front test sets under varying numbers of input views.** We report scene- and object-level metrics together with 3D bounding-box IoU for spatial layout evaluation. Best results within each input-view setting are highlighted in bold.

Method	Views	ARSG-110K					MIDI-3D-Front				
		$\mathrm{CD}_S \downarrow$	$\mathrm{FS}_S \uparrow$	$\mathrm{CD}_O \downarrow$	$\mathrm{FS}_O \uparrow$	IoU $\uparrow$	$\mathrm{CD}_S \downarrow$	$\mathrm{FS}_S \uparrow$	$\mathrm{CD}_O \downarrow$	$\mathrm{FS}_O \uparrow$	IoU $\uparrow$
Gen3DSR	1	0.265	46.72	0.546	31.95	0.304	0.123	40.07	0.157	38.11	0.363
MIDI	1	0.801	15.35	**0.179**	35.99	0.033	0.080	50.19	0.103	53.58	**0.518**
I-Scene	1	0.277	34.69	0.563	21.25	0.228	0.311	10.76	0.187	69.90	0.025
3D-Fixer	1	0.159	**68.82**	0.197	**57.85**	**0.519**	**0.069**	**78.67**	**0.032**	**94.39**	0.492
SAM3D	1	**0.146**	65.32	0.257	48.48	0.445	0.084	77.06	0.053	92.62	0.430
ShapeR	4	0.073	78.10	0.148	61.83	0.502	0.082	87.28	0.024	96.31	0.506
MV-SAM3D	4	0.063	82.37	0.113	69.48	0.585	0.083	87.84	0.020	96.76	0.597
Ours	4	**0.055**	**87.08**	**0.093**	**78.02**	**0.662**	**0.069**	**91.73**	**0.016**	**97.66**	**0.684**
MV-SAM3D	6	0.056	84.56	0.108	72.98	0.606	0.086	85.62	0.030	95.40	0.568
Ours	6	**0.047**	**89.68**	**0.087**	**81.42**	**0.695**	**0.070**	**91.13**	**0.023**	**97.09**	**0.684**
MV-SAM3D	8	0.051	86.90	0.101	74.33	0.615	0.083	86.23	0.029	95.54	0.572
Ours	8	**0.038**	**92.84**	**0.078**	**84.54**	**0.723**	**0.069**	**91.21**	**0.026**	**96.62**	**0.684**

the rendered and observed images. Moreover, the scene geometry is represented by the generated object-level Gaussian representations, while the object motion is parameterized by a compact $\mathrm{Sim}(3)$ transformation.

Formally, given the set of input images, we jointly optimize the camera poses $\{E_v\}$ and the per-object similarity transformations $\{(s'_o, R'_o, t'_o)\}$ by minimizing the multi-view rendering objective:

$$\min_{\{E_v\},\{s'_o,R'_o,t'_o\}} \sum_v \mathcal{L}_{\text{render}} \left(\mathcal{R} \left(\mathcal{G}_o(s'_o, R'_o, t'_o)_o, E_v \right), I_v \right), \tag{12}$$

where $\mathcal{G}_o$ denotes the Gaussian representation of the o-th object, $\mathcal{R}$ denotes differentiable Gaussian rasterization, and I_v is the v-th input image. The rendering loss can incorporate both photometric and depth supervision, where depth is from 3D reconstruction model as a regularization term. By jointly optimizing the camera and object parameters, the method allows their relative configurations to adapt under shared multi-view supervision, leading to a more globally consistent reconstruction.

5 Experiments

5.1 Experimental Setup

Implementation Details. We use VGGT-Omega Wang et al. (2026a) to estimate camera poses and scene point cloud from the input views, and adopt SAM3D with its default settings as the underlying object generator. All experiments are conducted on a single NVIDIA H100 GPU. We evaluate our method on two benchmarks for compositional 3D scene generation: the ARSG-110K test set Yin et al. (2026) and the MIDI-3D-Front Fu et al. (2021). Both benchmarks provide object-level geometry together with ground-truth scene configurations, enabling evaluation of object placement and overall scene composition.

Baselines. We compare with representative single-view and multi-view scene generation methods. Single-view baselines include Gen3DSR Ardelean et al. (2025), MIDI Huang et al. (2025), I-Scene Ling et al. (2026), 3D-Fixer Yin et al. (2026), and SAM3D Chen et al. (2026). We further compare with ShapeR Siddiqui et al. (2026) and MV-SAM3D Li et al. (2026a), which leverage multiple observations for 3D generation and provide the most relevant comparisons to our multi-view setting.

Metrics. Following 3D-Fixer Yin et al. (2026), we evaluate scene composition at both the scene and object levels. We report Chamfer Distance and F-Score for the complete scene (CD_S, FS_S) and for individual objects (CD_O, FS_O), with the predicted object poses retained during evaluation. These metrics therefore reflect both geometric fidelity and errors in object placement. The F-Score is computed with a distance threshold of 0.1. We further report the volumetric IoU between predicted and ground-truth 3D object bounding boxes as a more direct measure of spatial layout accuracy. Lower CD and higher F-Score and IoU indicate better performance.

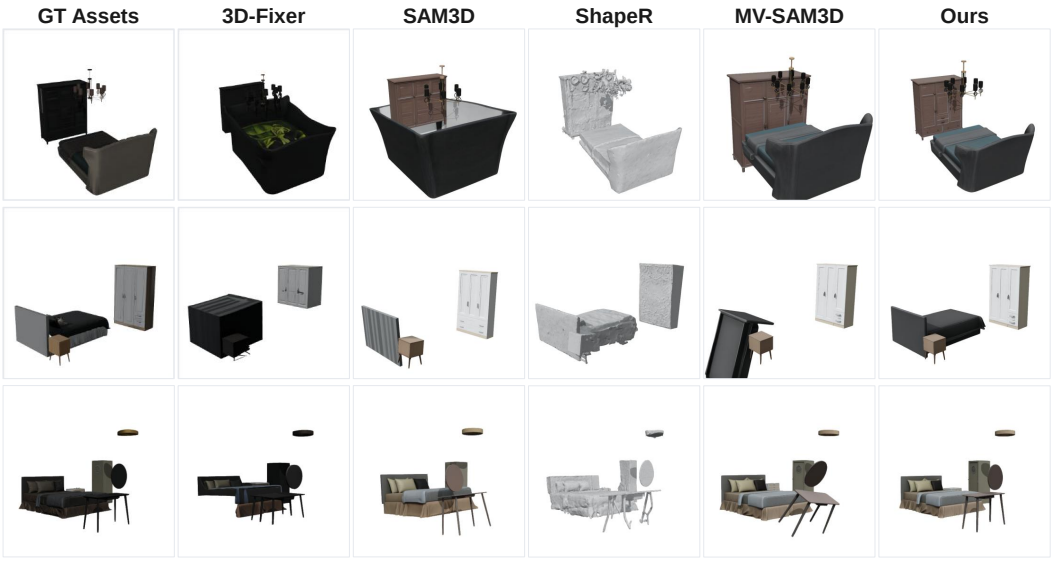

Figure 3: **Qualitative comparison on the MIDI-3D-Front.** Compared with existing single-view and multi-view methods, our method produces more accurate object placements and scene layouts.

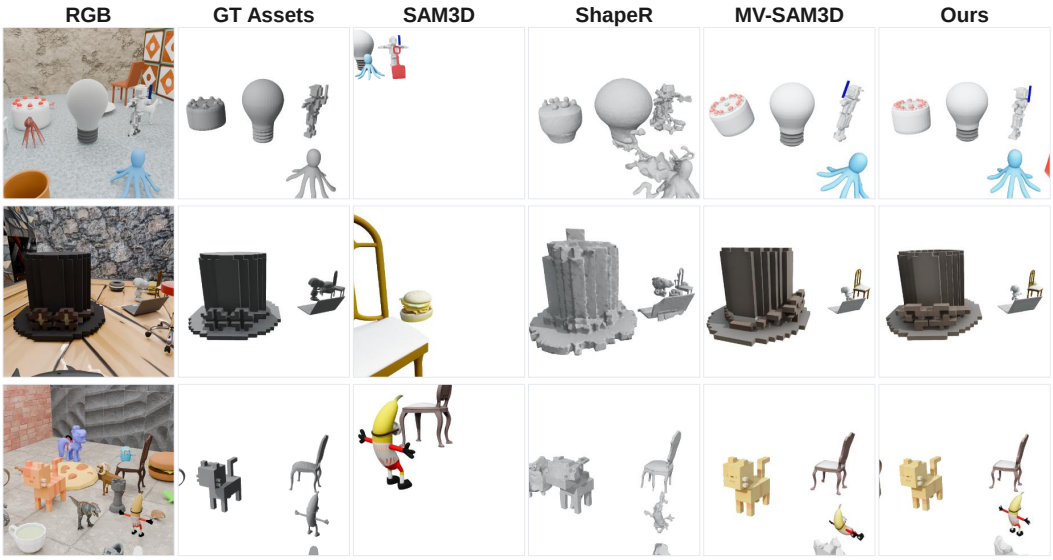

Figure 4: **Qualitative comparison on the ARSG-110K test set.**

5.2 Comparison with State-of-the-Art Methods

ARSG-110K. In Table 1, under the 4-view setting, multi-view methods substantially outperform single-view approaches, confirming the benefit of complementary observations. Compared with MV-SAM3D, our method reduces CD_S and CD_O by 12.7% and 17.7%, while improving FS_S and FS_O by 5.7% and 12.3%, respectively. More notably, object-layout IoU improves by 13.2%, showing that geometry-grounded pose reasoning more effectively converts complementary multi-view evidence into accurate object placement and coherent scene composition.

MIDI-3D-Front. A similar trend is observed on MIDI-3D-Front. Under the 4-view setting, our method reduces CD_S and CD_O by 16.9% and 20.0% over MV-SAM3D, respectively, while improving object-layout IoU substantially by 14.6%. This suggests that the main advantage of our method lies in more reliable spatial placement and scene composition rather than merely improving object-level agreement.

Table 2: **Ablation of pose routing.** Pose guidance and multi-view reconciliation are disabled.

Routing Strategy	$CD_S \downarrow$	$FS_S \uparrow$	IoU↑
Reference View	0.083	87.84	0.597
Largest Mask	0.075	89.64	0.611
Max Frame Completeness F_v	0.079	88.12	0.598
Max Visible Coverage C_v	0.075	89.94	0.614
Max Reliability S_v (Ours)	**0.072**	**90.08**	**0.617**

Table 3: **Ablation of reconciliation.**

Variant	$CD_S \downarrow$	$FS_S \uparrow$	IoU↑
Object Only	0.062	86.21	0.651
Object & Camera	**0.055**	**87.08**	**0.662**

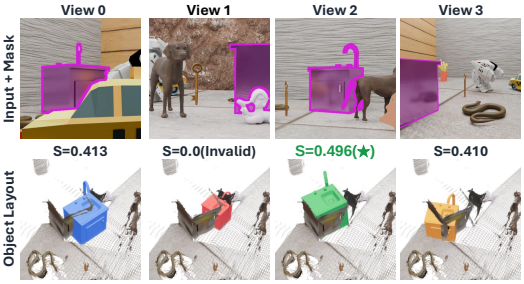

Figure 5: **Analysis of reliability-aware pose routing.** Boundary truncation in Views 1 and 3 yields unreliable pose hypotheses. Our reliability score favors View 2, whose placement is more consistent with the scene layout.

Scaling with input views. Our method remains robust as the number of input views increases. On ARSG-110K, the relative IoU gain over MV-SAM3D increases from 13.2% with 4 views to 14.7% with 6 views and 17.6% with 8 views, while the reduction in CD_S grows from 12.7% to 16.1% and 25.5%, respectively. On MIDI-3D-Front, SPOON maintains stable scene-level CD and layout IoU as the number of views increases. These results show that our method is more stable in exploiting increasing view coverage via reliability-aware pose reasoning.

Qualitative comparison. Fig. 3 and 4 show qualitative comparisons on MIDI and ARSG, respectively. Our method produces more accurate object orientations and spatial arrangements, resulting in fewer misplaced objects and more coherent scene layouts.

5.3 Ablation Studies

Table 4 shows that the full model achieves superior results, while removing any stage consistently degrades scene-level reconstruction and layout accuracy, confirming the complementary roles of the three components. Removing Guide indicates that geometry-grounded generation provides better pose initialization, while the drop without Route highlights the benefit of selecting geometrically reliable pose hypotheses. Moreover, the IoU gain shows the effectiveness of reconciliation in enforcing coherent scene-level placement. Table 2 shows that pose reliability cannot be adequately captured by mask size or a single geometric cue alone. Frame completeness and visible coverage capture complementary aspects of geometric observability, and their combination in S_v enables more reliable pose selection across views. Fig. 5 further illustrates that boundary-truncated observations yield less reliable hypotheses, while S_v favors a placement more consistent with the scene layout. Table 3 shows that jointly refining object placements and camera poses improves scene-level reconstruction and layout accuracy.

Table 4: **Component ablation of our method.**

Variant	$CD_S \downarrow$	$FS_S \uparrow$	$CD_O \downarrow$	$FS_O \uparrow$	IoU↑
Baseline	0.083	87.84	0.020	96.76	0.597
w/o Guide	0.073	91.47	0.018	97.58	0.677
w/o Route	0.071	90.63	0.019	97.10	0.669
w/o Reconcile	0.072	90.20	**0.014**	97.50	0.617
Ours	**0.069**	**91.73**	0.016	**97.66**	**0.684**

6 Conclusion

We presented SPOON, a scene-level pose reasoning framework for compositional 3D scene generation. Rather than treating object poses as independent predictions from local observations, SPOON uses reconstructed multi-view geometric cues as a shared scene reference to guide pose generation, route reliable pose hypotheses, and reconcile object–camera configurations through the proposed *Guide–Route–Reconcile* paradigm. This formulation turns independently inferred object–camera pose hypotheses into a coherent scene configuration while preserving the strong object priors of 3D generators. Experiments on ARSG-110K and MIDI-3D-Front demonstrate consistent improvements in object placement and scene composition quality across different numbers of input views.

REFERENCES

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

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *Proceedings of the International Conference on Machine Learning*, pp. 1737–1752, 2023.

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.

Jiahao Chang, Chongjie Ye, Yushuang Wu, Yuantao Chen, Yidan Zhang, Zhongjin Luo, Chenghong Li, Yihao Zhi, and Xiaoguang Han. Reconviagen: Towards accurate multi-view 3d object reconstruction via generation. In *International Conference on Learning Representations*, volume 2026, pp. 90387–90401, 2026.

Sijin Chen, Xin Chen, Anqi Pang, Xianfang Zeng, Wei Cheng, Yijun Fu, Fukun Yin, Billzb Wang, Jingyi Yu, Gang Yu, et al. Meshxl: Neural coordinate field for generative 3d foundation models. *Advances in Neural Information Processing Systems*, 37:97141–97166, 2024.

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, 2026.

Yiwen Chen, Tong He, Di Huang, Weicai Ye, Sijin Chen, Jiaxiang Tang, Zhongang Cai, Lei Yang, Gang Yu, Guosheng Lin, et al. Meshanything: Artist-created mesh generation with autoregressive transformers. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Yiwen Chen, Zhihao Li, Yikai Wang, Hu Zhang, Qin Li, Chi Zhang, and Guosheng Lin. Ultra3d: Efficient and high-fidelity 3d generation with part attention. *arXiv preprint arXiv:2507.17745*, 2025b.

Yiwen Chen, Yikai Wang, Yihao Luo, Zhengyi Wang, Zilong Chen, Jun Zhu, Chi Zhang, and Guosheng Lin. Meshanything v2: Artist-created mesh generation with adjacent mesh tokenization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13922–13931, 2025c.

Zhaoxi Chen, Jiaxiang Tang, Yuhao Dong, Ziang Cao, Fangzhou Hong, Yushi Lan, Tengfei Wang, Haozhe Xie, Tong Wu, Shunsuke Saito, et al. 3dtopia-xl: Scaling high-quality 3d asset generation via primitive diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26576–26586, 2025d.

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.

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.

Slava Elizarov, Ciara Rowles, and Simon Donné. Geometry image diffusion: Fast and data-efficient text-to-3d with image-based surface representation. *arXiv preprint arXiv:2409.03718*, 2024.

Shuangkang Fang, I Shen, Yufeng Wang, Yi-Hsuan Tsai, Yi Yang, Shuchang Zhou, Wenrui Ding, Takeo Igarashi, Ming-Hsuan Yang, et al. Meshllm: Empowering large language models to progressively understand and generate 3d mesh. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14061–14072, 2025.

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 *Proceedings of IEEE/CVF International Conference on Computer Vision*, pp. 10913–10922. IEEE, 2021.

Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances in neural information processing systems*, 35:31841–31854, 2022.

Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9785–9795, 2019.

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, 2026a.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23646–23657. IEEE, 2025.

Zhisheng Huang, Jiahao Chen, Cheng Lin, Chenyu Hu, Hanzhuo Huang, Zhengming Yu, Mengfei Li, Yuheng Liu, Zekai Gu, Zibo Zhao, et al. Unirecgen: unifying multi-view 3d reconstruction and generation. *arXiv preprint arXiv:2604.01479*, 2026b.

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.

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.

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, 2026.

Lu Ling, Yunhao Ge, Yichen Sheng, and Aniket Bera. I-scene: 3d instance models are implicit generalizable spatial learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26974–26983, 2026.

Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. Gaussian splatting slam. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18039–18048. IEEE, 2024.

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.

Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, and Matthias Nießner. Meshgpt: Generating triangle meshes with decoder-only transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19615–19625, 2024.

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.

Jiaxiang Tang, Zhaoshuo Li, Zekun Hao, Xian Liu, Gang Zeng, Ming-Yu Liu, and Qinsheng Zhang. Edgerunner: Auto-regressive auto-encoder for artistic mesh generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. Vggt-omega. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21486–21499, 2026a.

Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *European conference on computer vision*, pp. 55–71. Springer, 2018.

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

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *Advances in Neural Information Processing Systems*, 37:121859–121881, 2024.

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.

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. *arXiv preprint arXiv:2512.14692*, 2025a.

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*, pp. 21469–21480, 2025b.

Xingguang Yan, Han-Hung Lee, Ziyu Wan, and Angel X Chang. An object is worth $64\times$ 64 pixels: Generating 3d object via image diffusion. In *2025 International Conference on 3D Vision (3DV)*, pp. 123–133. IEEE, 2025.

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.

Ze-Xin Yin, Liu Liu, Xinjie Wang, Wei Sui, Zhizhong Su, Jian Yang, and Jin Xie. 3d-fixer: Coarse-to-fine in-place completion for 3d scenes from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12753–12763, 2026.

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.

A APPENDIX

In this appendix, we first provide further methodological (A.2) and implementation details (A.3). We then present additional analyses and qualitative results (A.4), followed by further discussion of limitations and future directions (A.5).

A.1 THE USE OF LARGE LANGUAGE MODELS

We use Large Language Models (LLMs) only for minor language editing to improve grammar and readability. All methodological design, experiments, equations, and results are developed by the authors.

A.2 ADDITIONAL METHOD DETAILS

This section provides additional details of the SPOON inference procedure. We first summarize the complete *Guide–Route–Reconcile* pipeline in Algorithm 1, and then detail the surface-based coordinate bridge used to transfer reconstruction-derived pose into the object canonical frame.

Algorithm 1 SPOON Inference

Require: Multi-view images $\{I_v\}_{v=1}^{V}$ and object masks $\{M_{o,v}\}$
Ensure: Compositional scene $\mathcal{S} = \{(\mathcal{G}_o, T_o^{\mathrm{ow}})\}_{o=1}^{O}$

Recover point maps $\{X_v\}$, camera intrinsics $\{K_v\}$, and world-to-camera extrinsics $\{E_v\}$ using the feed-forward reconstruction model.
for each object o **do**
  Aggregate the masked reconstruction points into the world-space object surface $\mathcal{P}_W^o$.
  Initialize the shared shape latent $\mathbf{z}^S$ and the view-specific pose latents $\{\mathbf{z}_v^P\}_{v=1}^{V}$.
  Initialize the surface bridge as invalid.
  for each flow step t **do**
    Predict per-view shape and pose velocities $\{\mathbf{v}_{\theta,v}^S, \mathbf{v}_{\theta,v}^P\}_{v=1}^{V}$.
    Fuse $\{\mathbf{v}_{\theta,v}^S\}_{v=1}^{V}$ and update the shared shape trajectory.
    if geometry guidance is activated **then**
      if the surface bridge has not been initialized **then**
        Recover the predicted clean shape endpoint and sample the canonical surface $\mathcal{P}_C^o$.
        Estimate the surface-based coordinate bridge $\mathcal{B}_{W\to C}^o$ between $\mathcal{P}_W^o$ and $\mathcal{P}_C^o$.
        Validate and cache $\mathcal{B}_{W\to C}^o$ if the geometric alignment is reliable.
      end if
      if $\mathcal{B}_{W\to C}^o$ is valid **then**
        Transfer the reconstruction-derived camera poses into the object canonical frame and construct geometry-derived pose endpoints $\{\mathbf{z}_{1,v}^{P,*}\}_{v=1}^{V}$.
        Apply flow-consistent endpoint guidance to the view-specific pose trajectories.
      else
        Apply the original pose-flow update.
      end if
    else
      Apply the original pose-flow update.
    end if
  end for
  Decode the canonical object representation $\mathcal{G}_o$ and the view-specific pose hypotheses $\{T_{o,v}^c\}_{v=1}^{V}$.
  Reproject $\mathcal{P}_W^o$ into each view and compute the geometric reliability scores $\{S_{o,v}\}_{v=1}^{V}$.
  Select $v_o^\star = \arg\max_v S_{o,v}$ and initialize $T_o^{\mathrm{ow}} = E_{v_o^\star}^{-1} T_{o,v_o^\star}^c$.
end for
Jointly refine the object transformations $\{T_o^{\mathrm{ow}}\}$ and camera poses $\{E_v\}$ using differentiable multi-view rendering.
return $\mathcal{S} = \{(\mathcal{G}_o, T_o^{\mathrm{ow}})\}_{o=1}^{O}$.

Overall Inference Procedure. Given multi-view images and their object instance masks, SPOON first recovers a shared geometric reconstruction, including point maps and camera parameters. For each object, during flow matching sampling, multi-view shape velocities are fused into a shared canonical shape trajectory, while the view-specific pose trajectories are geometrically guided once a sufficiently stable canonical surface becomes available. After generation, the view-specific pose

hypotheses are routed according to their geometric reliability to initialize the object placements in the shared world frame. Finally, object transformations and camera poses are jointly reconciled through differentiable multi-view rendering.

Surface-Based Coordinate Bridge. We provide additional details on the surface alignment used to estimate the coordinate bridge in Sec. 4.1. Given the world-space and canonical object surfaces $\mathcal{P}_W$ and $\mathcal{P}_C$ defined in the main paper, we independently center and isotropically normalize the two point sets to remove translation and global scale. After aligning the up axes of the reconstruction and canonical frames, we denote the resulting normalized surfaces by $\bar{\mathcal{P}}_W$ and $\bar{\mathcal{P}}_C$. The remaining rotational ambiguity is reduced to a yaw rotation, which we estimate by:

$$\theta^* = \arg\min_{\theta} \mathrm{TrimRMS}_{\bar{\mathbf{q}}\in\bar{\mathcal{P}}_C} \left[\min_{\bar{\mathbf{p}}\in\bar{\mathcal{P}}_W} \left\| \mathbf{R}_{\mathrm{yaw}}(\theta)^\top \bar{\mathbf{q}} - \bar{\mathbf{p}} \right\|_2 \right]. \tag{13}$$

We solve this objective using a global coarse-to-fine yaw search over the aggregated multi-view geometry. The resulting rotation is shared across all views of the same object, thereby preserving their relative camera configuration. The estimated yaw defines the surface bridge $\mathbf{B}^o_{W\to C}$ from the reconstruction world frame to the object canonical frame. Given the reconstruction-derived world-to-camera transformation $\mathbf{E}_v$, the corresponding geometry-derived object-to-camera transformation is obtained as:

$$\mathbf{T}^{c,*}_{o,v} = \mathbf{E}_v \left(\mathbf{B}^o_{W\to C}\right)^{-1}. \tag{14}$$

where $\mathbf{T}^{c,*}_{o,v}$ is converted to the pose parameterization of SAM3D, yielding $\boldsymbol{\pi}^*_v$ used in Eq. 6. To improve robustness against partial observations and reconstruction outliers, we compute $\mathrm{TrimRMS}$ using the lowest-error fraction τ_{trim} of nearest-neighbor residuals, with $\tau_{\mathrm{trim}} = 0.85$ across all experiments. Although the reconstructed surface may only partially cover the complete canonical object, aggregating observations across multiple views provides broader geometric support than any individual view. The normalized alignment is used only to estimate the remaining rotational correspondence between the two frames, while TrimRMS reduces the influence of unmatched regions and reconstruction outliers. The estimated bridge is accepted only when the normalized fitting RMSE is below 0.15 and at least four valid views provide sufficient geometric support. Otherwise, geometry guidance is disabled, and the original pose flow is used. Once accepted, the bridge is cached and reused for the remaining guided sampling steps.

A.3 Implementation Details

For geometric pose guidance, we set the guidance strength to $\alpha_P = 0.8$ and activate guidance from $t = 0.6$ onward, applying it at every subsequent sampling step. These two settings are selected based on preliminary experiments. For reliability-aware pose routing, we set $\alpha = 2$ in Eq. equation 9 and apply a completeness threshold $\tau_f = 0.5$ to exclude severely truncated observations. For the reconciliation stage, we jointly optimize the object transformations and camera poses using AdamW, with learning rates of 3×10^{-3} and 1×10^{-4}, respectively. The loss is a combination of photometric loss and depth loss. Since we only use depth loss as a regularization term, the depth loss is formulated as $\left\| M \odot \left(\exp(-d) - \exp(-\hat{d}) \right) \right\|_2^2$, where d is the depth map estimated by 3D reconstruction models, $\hat{d}$ is the rendered depth map, M is the object mask.

We evaluate computational efficiency on a single NVIDIA H100 GPU using scenes containing five objects under the same input setting. With four-view input, MV-SAM3D requires 195s per five-object scene, while the complete SPOON pipeline takes 326.5s. The proposed geometry guidance and pose routing introduce only modest additional costs of 10s and 1.5s, respectively, while the iterative reconciliation accounts for the remaining 120s.

We use the object instance masks provided by the benchmark datasets for all methods requiring object-level observations. For MV-SAM3D, we follow its official inference configuration, with both layout injection during generation and post-generation object-pose refinement enabled. Both MV-SAM3D and our method place generated objects in the world coordinate system recovered by VGGT-Ω, whose global gauge generally differs from that of the ground-truth scene. For evaluation, we estimate a single global $\mathrm{Sim}(3)$ transformation by aligning the reconstructed camera centers with the ground-truth camera centers and apply it uniformly to the entire predicted scene. For MV-SAM3D, this transformation is estimated from the initial camera poses predicted by VGGT-Ω, whereas for our method it is estimated from the camera poses after multi-view reconciliation.

No additional per-object alignment is applied for either scene-level or object-level evaluation. For single-view baselines, where multi-view camera-center alignment is unavailable, we follow the coordinate and scale calibration protocols of their official implementations. For example, 3D-Fixer performs in-place completion in the scene frame defined by its fragmented geometric input, with the scene scale calibrated according to its released evaluation pipeline.

A.4 Additional Results

We provide additional analyses and qualitative results to further validate the proposed method. We first examine the effect of geometry-guided pose generation and the sensitivity of the reliability-aware routing strategy to the exponent α. We then present additional qualitative comparisons on the MIDI-3D-Front and ARSG-110K test sets to demonstrate the consistency of the improvements across diverse scenes.

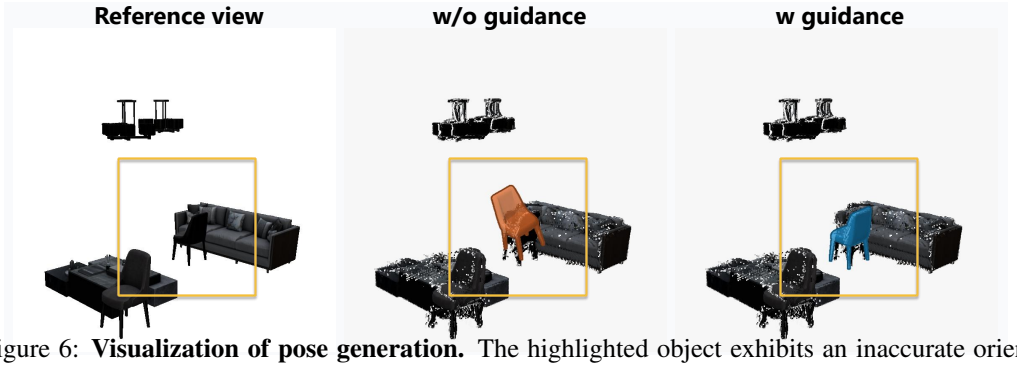

Figure 6: **Visualization of pose generation.** The highlighted object exhibits an inaccurate orientation without guidance, while pose guidance steers the generated pose toward an orientation more consistent with the object layout.

Effect of pose guidance. Fig. 6 qualitatively demonstrates the effect of geometry-guided pose generation. Without guidance, the generated objects can exhibit noticeable orientation errors, whereas geometry guidance steers the pose toward a configuration more consistent with the observed scene layout. This qualitatively demonstrates that the reconstructed geometric cues provide effective guidance for improving pose generation.

Table 5: Sensitivity analysis of the routing exponent α on MIDI with four input views, with multi-view reconciliation disabled.

α	$CD_S \downarrow$	$FS_S \uparrow$	IoU↑
1	0.072	90.08	0.619
2	0.072	90.20	0.617
3	0.072	90.07	0.618

Sensitivity to the routing exponent. Table 5 shows that the routing strategy is insensitive to the choice of α. Across $\alpha \in \{1, 2, 3\}$, all three metrics remain nearly unchanged, indicating that the proposed reliability score is robust to moderate variations in the completeness weighting.

Additional qualitative comparisons. Figs. 7 and 8 provide additional qualitative comparisons on the MIDI-3D-Front and ARSG-110K test sets, respectively. Across diverse scenes, our method produces object configurations with more consistent orientations and relative placements, resulting in scene layouts that better agree with the reference observations. These results further demonstrate the benefit of geometry-grounded pose reasoning for improving scene-level spatial coherence under multi-view observations.

A.5 Limitation and Future Work

Although SPOON improves scene-level coherence, individual objects are still generated independently rather than through an explicit multi-instance generative process. Extending the current object-wise formulation toward joint multi-instance generation could enable stronger coupling between object generation and scene-level spatial organization. In addition, the effectiveness of



Figure 7: **Qualitative comparison on the MIDI-3D-Front test set.**

geometry-grounded reasoning depends on the quality of the underlying multi-view reconstruction and may be affected by inaccurate or incomplete geometry.



Figure 8: **Qualitative comparison on the ARSG-110K test set.**