CC-FMO: Camera-Conditioned Zero-Shot Single Image to 3D Scene Generation with Foundation Model Orchestration

Boshi Tang
Tsinghua University
Beijing, China
tbs16@mails.tsinghua.edu.cn

Henry Zheng
Tsinghua University
Beijing, China
jh-zheng22@mails.tsinghua.edu.cn

Rui Huang
Tsinghua University
Beijing, China
hr20@mails.tsinghua.edu.cn

Gao Huang*
Tsinghua University
Beijing, China
gaohuang@tsinghua.edu.cn

Abstract

High-quality 3D scene generation from a single image is crucial for AR/VR and embodied AI applications. Early approaches struggle to generalize due to reliance on specialized models trained on curated small datasets. While recent advancements in large-scale 3D foundation models have significantly enhanced instance-level generation, coherent scene generation remains a challenge, where performance is limited by inaccurate per-object pose estimations and spatial inconsistency. To this end, this paper introduces CC-FMO, a zero-shot, camera-conditioned pipeline for single-image to 3D scene generation that jointly conforms to the object layout in input image and preserves instance fidelity. CC-FMO employs a hybrid instance generator that combines semantics-aware vector-set representation with detail-rich structured latent representation, yielding object geometries that are both semantically plausible and high-quality. Furthermore, CC-FMO enables the application of foundational pose estimation models in the scene generation task via a simple yet effective camera-conditioned scale-solving algorithm, to enforce scene-level coherence. Extensive experiments demonstrate that CC-FMO consistently generates high-fidelity camera-aligned compositional scenes, outperforming all state-of-the-art methods.

1. Introduction

The rapid progress of foundation models has revolutionized numerous research domains, demonstrating impressive generalization capabilities in text-to-image synthesis[20, 61], 3D generation[21, 45, 46, 63], embodied intelligence[2, 26, 82], and many others. Despite these successes, to achieve generalizability in single image-to-scene generation methods remains a challenge. Previous methods[24, 37] often depend heavily on training with carefully curated, task-specific datasets, which inherently constrains their ability to generalize beyond the training distributions. Moreover, unlike large-scale 3D instance-level datasets[9, 10], curating diverse 3D scene-level datasets is significantly more difficult, resulting in limited data availability and reduced cross-dataset generalization.

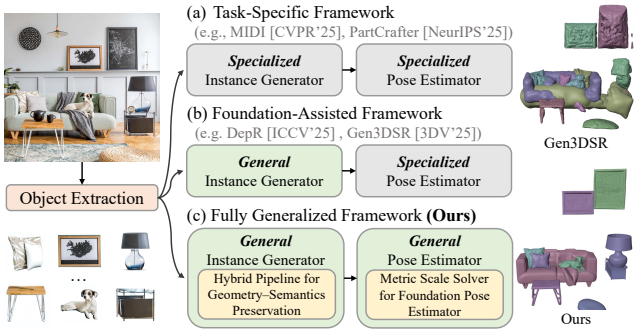

Figure 1. CC-FMO generates high-fidelity 3D scenes from single images in a zero-shot manner by addressing key challenges in applying 3D foundation models for scene generation. It combines (1) a hybrid instance generator to preserve instance semantics and geometric fidelity and (2) a foundation pose estimation model augmented with a novel metric-scale solving algorithm.

Note that single-image to 3D scene generation is commonly formulated as a two-stage process comprising (1) instance generation and (2) pose estimation, and existing methods primarily differ in how these two stages are conducted, as illustrated in Fig. 1. Given the aforementioned

*Corresponding author.

limitation, several recent approaches [12, 89, 92] integrate foundation models into single-image 3D scene generation pipelines, but only at selected stages rather than throughout the pipeline. They mostly utilize foundation models for instance generation, but employ lightweight, task-specific pose estimators or classical point-cloud registration techniques for subsequent pose estimation. These pose estimators frequently become the dominant source of error, and as a consequence, limiting the overall robustness and generalization capability of such methods, preventing them from generalizing reliably across diverse scenes and camera setups. To the best of our knowledge, none of existing methods has realized a fully foundation-model-driven pipeline, particularly for the pose estimation stage. This motivates the central research question of our work:

Can we achieve generalizable 3D scene generation from a 2D image with 3D foundation models only?

In this research, we aim to answer this question by introducing a fully foundation model–powered pipeline for generalizable 3D scene generation. Specifically, we identify two key issues that impact the performance of foundation models in compositional scene generation: occlusion-induced errors during instance generation, where the generated instances deviate from the correct object semantics; and scale mismatch, where the object scales of generated meshes deviate from real-world scales.

To address occlusion-induced errors, we propose a hybrid instance generator that preserves object semantics using the vector latent set representation (VecSet) [84] while improving geometric fidelity with structured latents (SLAT)[75]. Specifically, we first leverage a massively pretrained VecSet-based object generator to produce coarse 3D assets. VecSet's low-dimensional latent formulation captures semantic structure robustly—even under occlusion—thus preventing semantics from shifting toward incorrect object types, as shown in Fig. 4. The coarse geometry is then refined by a generic detail-enhancing model built on SLAT representation and normal estimation[80], which produces detailed geometry without altering the underlying semantics. To resolve the scale mismatch problem in pose estimation, we propose a simple yet effective camera-conditioned scale solving algorithm that admits a closed-form linear solution. This algorithm is conditioned on camera parameters and produces metric scales across components, enabling foundation-scale pose estimation models to be used directly for compositional scene generation, without requiring specialized, task-specific pose estimation modules.

In summary, our main contributions are:

- We propose CC-FMO, a fully generalized framework for zero-shot 3D scene generation from single images.
- We introduce a novel hybrid instance generation paradigm for both semantics-preserving and high-fidelity 3D object generation, together with a simple yet effective camera-conditioned algorithm with closed-form solution, to enable the application of foundation pose estimation models to the compositional scene generation task.
- We evaluate CC-FMO for generating diverse open-world 3D scenes that align with camera parameters, and show its superiority over the state-of-the-art zero-shot or training-dependent methods.

2. Related Works

2.1. Single Image to 3D Scene Generation.

Generating a set of posed 3D scenes given a single image remains a core and long-standing problem in computer vision. Depending on how to process the scene, current approaches in the literature can be broadly classified into two categories: feed-forward reconstruction techniques [4, 7, 8, 17, 39, 50, 51, 85, 87], and compositional generation frameworks [5, 12, 18, 24, 62, 79, 92].

Feed-Forward Scene Reconstruction. Feed-forward reconstruction techniques [4, 7, 8, 17, 39, 50, 51, 85, 87] typically formulate scene reconstruction as a regression problem, exploiting spatial regularities of indoor environments and relying on 3D data to train networks in an end-to-end manner. They commonly adopt encoder-decoder architectures to predict scene geometry directly from a single input image. Although jointly inferring scene layout and object poses simplifies the training objective, the reconstruction quality is often limited by the scarcity and bias of annotated 3D scene datasets, leading to poor generalization on out-of-distribution objects. In the single-image setting considered in this work, scene reconstruction is conditioned on a single 2D view, so the network is strongly constrained only on visible pixels, while occluded and back-facing regions cannot be reliably reconstructed. As a result, such methods often produce incomplete or biased object geometries that do not faithfully capture the underlying scene.

Compositional Scene Generation. Recent compositional generation frameworks [5, 12, 18, 92] integrate foundation models for both image [28, 43, 49, 55–57, 59, 60] and 3D object [14, 25, 86] domains to enhance scene reconstruction. These approaches generally follow a multi-stage procedure involving image segmentation [57], object completion [59], per-object generation [25, 86], and pose estimation [69]. Although they improve generalization by utilizing pre-trained models, none of them is able to apply foundational pose estimation abilities for scene generation, limiting the applicability and generalization of the overall pipelines. Our method tackles the problem by employing a foundation pose estimation model, which is made possible with a simple yet effective scale solving algorithm with

closed-form solutions.

2.2. Single Image to 3D Object Generation

Recent advancements of 3D generation [13, 21, 23, 31, 38, 41, 42, 45–47, 58, 63, 64, 67, 70–72, 74, 76, 86, 90] has been largely driven by progress in diffusion models [20, 61] and the creation of large-scale datasets [9, 10]. A number of image-to-3D object generation works [22, 45, 46, 63, 64, 67, 70, 76] generate 3D assets in two stages that first generates multi-view images and then reconstructs the 3D geometry from these views. These methods typically involve fine-tuning pre-trained image [55, 59] or video [1] diffusion models to produce coherent multi-view images, and subsequently recover the 3D shape with either large-scale reconstruction models [21, 63, 77, 93] or optimization-based techniques [66]. There are also 3D-native techniques [31, 33, 72, 86, 90] that concentrate on generating 3D-native geometries by training large-scale generative models, which often train a latent diffusion transformer (DiT) [53] over the latent space of a variational autoencoder [27]. Owing to the diversity and scale of datasets they are trained on, these models are capable of producing high-quality 3D geometries with remarkable generalizability. Our work is built upon these works, integrating both a low-dimensional VecSet-based object generation model for semantics preservation, and a high-dimensional SLAT-based one for geometry fidelity.

2.3. Pose Estimation

Depending on the level of generalization, pose estimation techniques can be roughly divided into three categories: instance-level[19, 34, 54, 65, 83], category-level[3, 11, 36, 40, 91] and unseen object methods[30, 35, 44, 48, 68], in which a model is able to either: a) only estimate the pose of a specific object on which it is trained, b) estimate intra-class unseen instances if the class has been seen during training, or c) handle unseen object categories during training. In our setting we take advantage of the unseen object methods owing to the richness of objects in the input scenes. However, the foundation models for unseen object pose estimation are still deficient, where FoundationPose is a representative. In this work we bridge the gap between FoundationPose and the compositional scene generation task to achieve accurate and generalizable pose estimation.

3. Methodology

The goal of CC-FMO is to develop a generalizable method that generates 3D scenes directly from RGB images, performing robustly across diverse real-world environments. To this end, we design a zero-shot pipeline built entirely on foundation models to achieve both semantics preservation together with enhanced geometry fidelity in instance generation, and improved accuracy in per-object pose estimation. We first carefully segment and inpaint 2D object instances with occlusion awareness using 2D foundation models, providing clean inputs for subsequent 3D instance generation (Sec. 3.1). To produce high-quality geometry and semantically accurate instances, we integrate VecSet and SLAT 3D foundation generation models for instance synthesis (Sec. 3.2). Lastly, for coherent scene composition, we incorporate a foundational pose estimation model with a camera conditioned scale solver that resolves scale ambiguity inherent to 2D images, to align the generated objects with indicated spatial layout (Sec. 3.3). Together, these components form a coherent and generalizable framework capable of efficiently generating complex 3D scenes (see Fig. 2).

3.1. 2D Preprocessing

Object Extraction. To prepare a 2D image I for 3D instance generation, we first employ an open-set image tagging model, Recognize Anything [88], to obtain object labels $\{l_i\}_{i=1}^N$, where N denotes the number of objects in I. For each label l_i, we extract the corresponding 2D bounding box b_i and segmentation mask m_i using the open-vocabulary object segmentation model Grounded-SAM [57]. We then remove background categories such as "floor," "room," and "wall". Using the remaining object masks, we extract the associated pixels to produce instance-level 2D object images $\{O_i\}_{i=1}^N$ for subsequent processing.

Occlusion-Aware 2D Inpainting. Given that $\{O_i\}_{i=1}^N$ tends to suffer from object occlusion, before sending them to the 3D generation models we inpaint those occluded O_i, to prevent lose of information. We first employ the widely adopted Depth-Anything-v2[78], together with previously estimated masks, to estimate per-object depths $\{D_i\}_{i=1}^N$. We determine the occlusion of each object instance i, by jointly considering its adjacent objects in the image and their depths. For each pair of adjacent objects, the instance with the smaller depth value is considered unoccluded, while the other is considered occluded and processed by our inpainting model to restore the occluded regions. We utilize Stable-Diffusion-XL-1.0-Inpainting-0.1 [55] due to its wide usage. The corresponding instance masks are updated with the inpainted results accordingly. Our depth-guided inpainting procedure distinguishes itself from previous approaches [12, 18] in that it avoids unnecessary inpainting, thereby preserving the fidelity of the original object images as much as possible.

3.2. Object Generation

Recent works[73, 75, 81] based on structured latent representation (SLAT) have shown excellent generation performance on real-world image-based 3D instance generation.

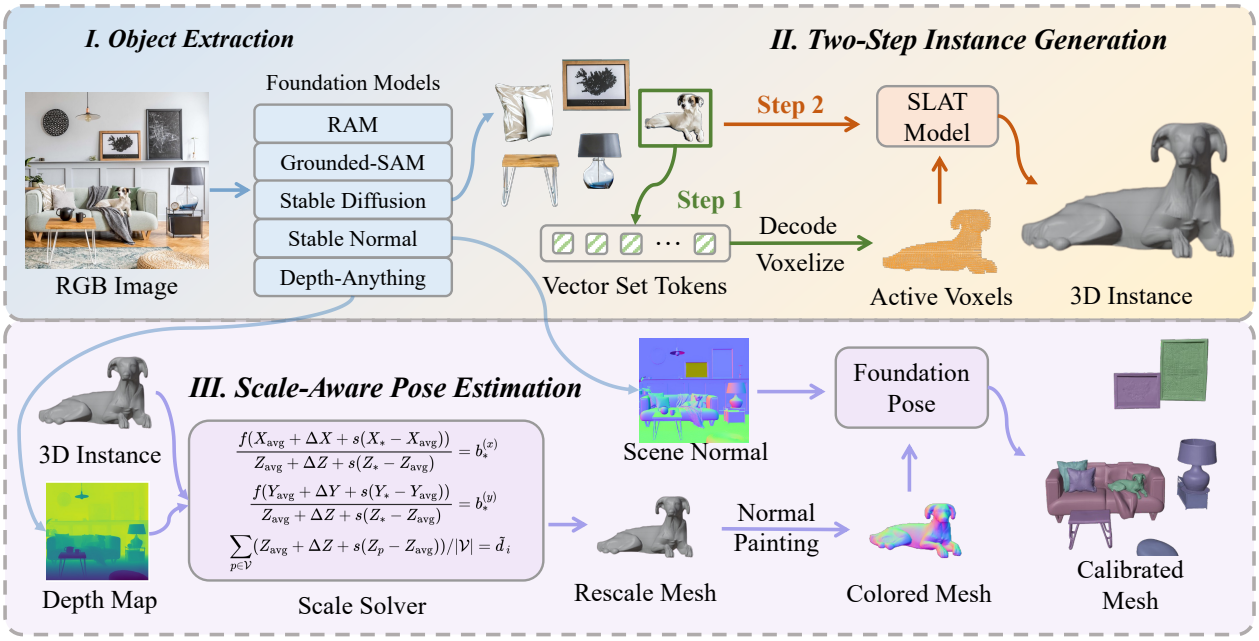

Figure 2. Illustration of CC-FMO, a zero-shot framework built entirely on foundation models for single-image to 3D scene generation. Our method delivers semantically accurate, high-fidelity geometry alongside precise per-object pose estimation. We first extract clean object instances through occlusion-aware segmentation and inpainting, providing reliable inputs for 3D generation. Next, we synergize a hybrid 3D foundation generation approach that melds vector latent-set (VecSet) and structured-latent (SLAT) models to synthesize high-quality, semantically consistent instances. Finally, to ensure coherent scene composition, we orchestrate a scale-solving algorithm with a foundational pose estimation model to resolve scale ambiguity and align generated objects with the input's spatial layout in a camera-conditioned manner.

Unfortunately, in our experiments (see Fig. 4), SLAT-based models alone exhibit suboptimal performance, as they tend to focus on local geometry details, instead of overall object semantics. For instance, given an image of a framed painting, SLAT-based methods such as Stable3DGen[81] incorrectly interpret it as the side of a delicate aquarium with aquatic plants. Therefore, inspired by Ultra3D[6], we employ a two-stage object generation pipeline.

First, to preserve object semantics, we generate a coarse 3D object using Step1X3D[32], a vector latent-set (VecSet)–based object generator trained on large-scale instance data. Due to its low-dimentional nature, the VecSet representation provides high-level semantic guidance, encouraging the model to capture the global semantics of the object. However, in contrast to SLAT-based methods, VecSet-based ones alone generally fall short of geometry fidelity, as shown in Fig. 4. To leverage the strengths of both, we voxelize the output of Step1X3D into indices of active voxels, and then employ the widely used Stable3DGen[81] to predict the corresponding structured latents, which are then populated into the decoder for fine-grained object meshes $\{M_i\}_{i=1}^N$. Although our two-stage generation procedure is reminiscent of that of Ultra3D, our approach distinguishes itself in that it employs foundation models, instead of specifically trained ones, in the pipeline. Furthermore, our design purpose differs, as we use VecSet models primarily for semantics extraction rather than resolution enhancement, thus adopting no voxel super-resolution techniques.

3.3. Scale-Aware Pose Estimation

Given the scene image I, camera parameters c, and generated meshes $\{M_i\}_{i=1}^N$, a central challenge is to recover per-object transformations (s_i, R_i, t_i) to align the meshes with input spatial layouts. Noticeably, no prior work has successfully applied a foundation pose estimation model to scene generation. Existing methods either train a scale-aware pose estimation module from scratch [24], or resort to time-consuming optimization on 3D discrepancy[92], which, in practice (see Tab. 1), often degrade on out-of-domain images and produce unstable and inaccurate pose estimations.

A naive approach is to apply open-set foundation pose estimation models, such as FoundationPose [69], to our problem. However, this cannot be done directly due to scale mismatch between the generated meshes and the target scene. In particular, FoundationPose only estimates the rotation R and translation t of an object while assuming a known scale s. Should this assumption be violated, pose estimation accuracy degrades substantially (see Tab. 3). In

our setting, the instance-generation foundation models are trained on objects with diverse and unconstrained physical sizes, thus, the scales of the generated meshes M_i are generally inconsistent with the metric scale implied in the input image. This mismatch makes naive application of foundation pose estimators infeasible for accurate 3D scene generation.

To incorporate FoundationPose into our pose estimation procedure, we propose a scale estimation algorithm, which together with FoundationPose constitutes the overall pose estimation module of our scene generation pipeline, as illustrated in Fig. 2.

Solving for Scale. We denote the translation and scale to be solved by $t := (\Delta X, \Delta Y, \Delta Z)$ and s, respectively. Our derivation is based on the standard pinhole camera projection model, $\frac{f}{Z}X = x$, where X is the x-coordinate of a 3D point in camera space, Z is its depth, f is the focal length, and x is the corresponding x-coordinate on the image plane. We formulate the constraints for the algorithm in both 2D and 3D spaces.

- **Constraint 1**: After rescaling and translation, the 2D bounding box of M_i should align with b_i.

To enforce the constraint, we first rasterize M_i, to get the uppermost, lowermost, leftmost and rightmost 2D points (x_*, y_*), and their 3D counterparts (X_*, Y_*, Z_*), where $*$ can be one of "upper", "lower", "left" or "right".

Next, we compute the 2D bounding box of the transformed mesh. When a point (X, Y, Z) is transformed by translation $(\Delta X, \Delta Y, \Delta Z)$ and scaling s, its new location becomes:

$$\begin{aligned}&\mathcal{T}_p((X,Y,Z),(\Delta X,\Delta Y,\Delta Z,s)) = \\ &(X_{\text{avg}} + \Delta X + s(X - X_{\text{avg}}),\\ &Y_{\text{avg}} + \Delta Y + s(Y - Y_{\text{avg}}),\\ &Z_{\text{avg}} + \Delta Z + s(Z - Z_{\text{avg}})),\end{aligned} \tag{1}$$

where X_{avg}, Y_{avg}, and Z_{avg} are the means of the X, Y, and Z coordinates of mesh M_i, respectively, and $\mathcal{T}_p$ denotes the pointwise linear transformation. To make the computation of the bounding boxes of the transformed meshes tractable, we approximate that the leftmost, rightmost, uppermost, and lowermost visible 3D points of a mesh remain to be the extremal visible points after transformation.

With this approximation, substituting Eq. 1 into the pinhole camera projection model yields the 2D coordinate of the leftmost visible point of the transformed mesh, which should match the x-coordinate of the leftmost edge of b_i, namely:

$$\frac{f(X_{\text{avg}} + \Delta X + s(X_{\text{left}} - X_{\text{avg}}))}{Z_{\text{avg}} + \Delta Z + s(Z_{\text{left}} - Z_{\text{avg}})} = b^{(x)}_{\text{left}} \tag{2}$$

Algorithm 1 Solving for object-wise scale and translation with closed-form solution. Omit superscript i for brevity

Require: X_{avg}, Y_{avg}, Z_{avg}, $\{X_*\}$,$\{Y_*\}$,$\{Z_*\}$,$\{b^{(x)}_*\}$, $\{b^{(y)}_*\}$, $\forall * \in \{\text{left}, \text{right}, \text{upper}, \text{lower}\}$, $\tilde{d}$, f, Z ▷ Z is the object-wise depth map

$dX_*, dY_*, dZ_* \leftarrow X_* - X_{\text{avg}}, Y_* - Y_{\text{avg}}, Z_* - Z_{\text{avg}}$

$dZ \leftarrow Z - Z_{\text{avg}}$

$$A \leftarrow \begin{bmatrix} f & 0 & -b^{(x)}_{\text{left}} & f \cdot dX_{\text{left}} - b^{(x)}_{\text{left}} \cdot dZ_{\text{left}} \\ f & 0 & -b^{(x)}_{\text{right}} & f \cdot dX_{\text{right}} - b^{(x)}_{\text{right}} \cdot dZ_{\text{right}} \\ 0 & f & -b^{(y)}_{\text{upper}} & f \cdot dY_{\text{upper}} - b^{(y)}_{\text{upper}} \cdot dZ_{\text{upper}} \\ 0 & f & -b^{(y)}_{\text{lower}} & f \cdot dY_{\text{lower}} - b^{(y)}_{\text{lower}} \cdot dZ_{\text{lower}} \\ 0 & 0 & 1 & \frac{\sum dZ}{\sum \mathbf{1}[Z \neq 0]} \end{bmatrix}$$

$$B \leftarrow \begin{bmatrix} b^{(x)}_{\text{left}} \cdot Z_{\text{avg}} - f \cdot X_{\text{avg}} \\ b^{(x)}_{\text{right}} \cdot Z_{\text{avg}} - f \cdot X_{\text{avg}} \\ b^{(y)}_{\text{upper}} \cdot Z_{\text{avg}} - f \cdot Y_{\text{avg}} \\ b^{(y)}_{\text{lower}} \cdot Z_{\text{avg}} - f \cdot Y_{\text{avg}} \\ \tilde{d} - Z_{\text{avg}} \end{bmatrix}$$

Solving $AX = B$, where $X = \begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \\ s \end{bmatrix}$

where the superscript (i) is omitted for simplicity, and $b^{(x)}_{left}$ represents the x-coordinate of the leftmost edge in b_i. For brevity, we only show the equation for the leftmost visible points. The equations for the other visible points follow analogously, yielding four equations in total. Note that these equations are linear with respect to t_i and s_i.

Due to depth–scale ambiguity, the equations defined in Eq. 2 admit an infinite number of solutions for t and s. To constrain the solution space, we define a second constraint:

- **Constraint 2**: After transformation, the average depth of the visible region of an object matches that of the same object in I.

To obtain the depth of I, we run Depth-Anything-v2 [78] to estimate a depth map. We then compute the average depth of the i-th object's visible region, denoted by $\tilde{d}_i$. This yields another equation:

$$\sum_{p \in \mathcal{V}} \left(Z_{\text{avg}} + \Delta Z + s(Z_p - Z_{\text{avg}})\right) / |\mathcal{V}| = \tilde{d}_i \tag{3}$$

where the numerator is the depth of all visible points of the transformed mesh, and the denominator $\mathcal{V}$ denotes the set of all visible points.

We jointly solve the five linear equations defined in Eq. 2 and Eq. 3 for s_i and t_i. Since these equations are linear with respect to s and t, they can be solved efficiently using standard linear equation solvers within milliseconds. Due to the approximation introduced in *Constraint 1*, the resulting

t_i and s_i are not perfectly accurate. We therefore iteratively solve the equations multiple times to refine s_i and t_i. In our experiments, we find that four iterations are sufficient. The algorithm for one step of scale solving is shown in Alg. 1.

Mesh Texturing for Accurate Pose Estimation. With $\{s_i\}_{i=1}^N$ and $\{t_i\}_{i=1}^N$, applying FoundationPose becomes feasible. However, a key challenge remains: FoundationPose assumes textured meshes as input, whereas generating consistent object textures from a single RGB image is inherently difficult. In our preliminary experiments, we employed a state-of-the-art mesh inpainter [22] and observed that the inpainting quality was limited, which in turn severely degraded pose estimation accuracy. To address this issue, we texture the generated meshes using vertex normals and estimate the normal map of I, denoted by I_{normal}, using StableNormal [80]. We then invoke FoundationPose on I_{normal} together with the normal-textured meshes to obtain estimates of R_i. Given R_i, we solve the five linear equations to obtain refined s_i and t_i, yielding the final transformed meshes. By encoding geometric information into the texture in the form of normal maps, this mechanism encourages FoundationPose to focus on aligning geometric details rather than appearance textures.

Camera-Conditioned Generation. Unlike recent training-dependent scene generation methods [24, 37, 89], CC-FMO is camera-conditioned: the focal length and other camera parameters naturally enter the formulation of our constraints and are preserved throughout. This is an important feature because monocular images suffer from well-known scale ambiguity: different combinations of object scale, depth, and focal length can produce identical 2D projections. By enforcing consistency with camera parameters in our constraints, we effectively restrict the solution space and recover object poses and scales that are geometrically meaningful rather than only image-consistent.

Moreover, as demonstrated in our experiments (see Tab. 1), our model can zero-shot adapt to scenes with different camera parameters at test time simply by updating the intrinsics used in the projection equations. This feature stands in contrast to prior works [24, 37], which typically assume fixed intrinsics and absorb such prior information into learned parameters, limiting the ability to generalize across cameras or to produce metrically calibrated 3D scenes.

4. Experiments

4.1. Experimental Setup

For fair and consistent comparisons, we follow the evaluation protocol of MIDI [24]. All random seeds are fixed to 0 for reproducibility. The details are as follows:

Table 1. Quantitative results on 3D-FRONT [15] measured by scene-level Chamfer Distance (CD-S) and F-Score (F-Score-S), object-level Chamfer Distance (CD-O) and F-Score (F-Score-O), and object bounding box Volume IoU (IoU-B). Colors indicate whether methods are training-dependent or zero-shot. Best and second-best results are in bold and underlined, respectively. CD and FScore are shown in percentage.

Method	CD-S↓	FScore-S↑	CD-O↓	FScore-O↑	IoU-B↑
DepR [89]	10.71	39.81	22.24	39.51	0.142
PartCrafter [37]	10.82	47.19	3.201	80.10	0.294
MIDI [24]	<u>2.241</u>	<u>75.81</u>	<u>0.802</u>	<u>89.42</u>	<u>0.527</u>
Gen3DSR [12]	22.81	26.78	10.78	54.95	0.113
DPA* [92]	12.47	37.71	4.054	70.29	0.137
Ours	**1.805**	**89.60**	**0.738**	**91.00**	**0.670**

Implementation Details. Recall that our pipeline operates in a zero-shot setting. Object detection and segmentation are conducted with Grounded-SAM [57], where Recognize Anything [88] is first used to obtain object category labels. We adopt Stable-Diffusion-XL-1.0-Inpainting-0.1 [55] as our 2D inpainting model, followed by SAM [29] to retrieve segmentation masks on the inpainted images.

Object generation is performed using Step1X-3D [32] and Stable3DGen [81], where the latter relies on StableNormal [80] for normal estimation. FoundationPose [69] estimates per-object poses given the estimated scene-level normal map and per-object normal maps from StableNormal, together with a scene depth map estimated by Depth-Anything-v2 [78]. The linear systems defined in Eq. 2 and Eq. 3 are solved using the `linalg.lstsq` solver in PyTorch, and rasterization is implemented with PyTorch3D.

Datasets. For quantitative evaluation, we adopt the MIDI test partition of 3D-FRONT [24]. However, we observe that this partition contains samples that are ill-posed for scene generation evaluation. For example, some views place the cameras so far from the scene that the rendered content occupies only tiny, indistinguishable regions of the image, while others produce almost entirely black images due to severe occlusion by large foreground objects. We manually filter out such degenerate cases and evaluate all models on the remaining samples, which comprise more than 500 scenes. In contrast to MIDI, our rendering setup randomizes the camera field of view between 20 and 60 degrees, and the camera position is adjusted to ensure that the scene occupies a substantial portion of the rendered image. The camera coordinate axes are aligned with the world coordinate axes, like MIDI. Following MIDI, we also compute ground-truth depth maps and normal maps during rendering, which are used in place of the outputs of Depth-Anything-v2 and StableNormal during quantitative evaluation.

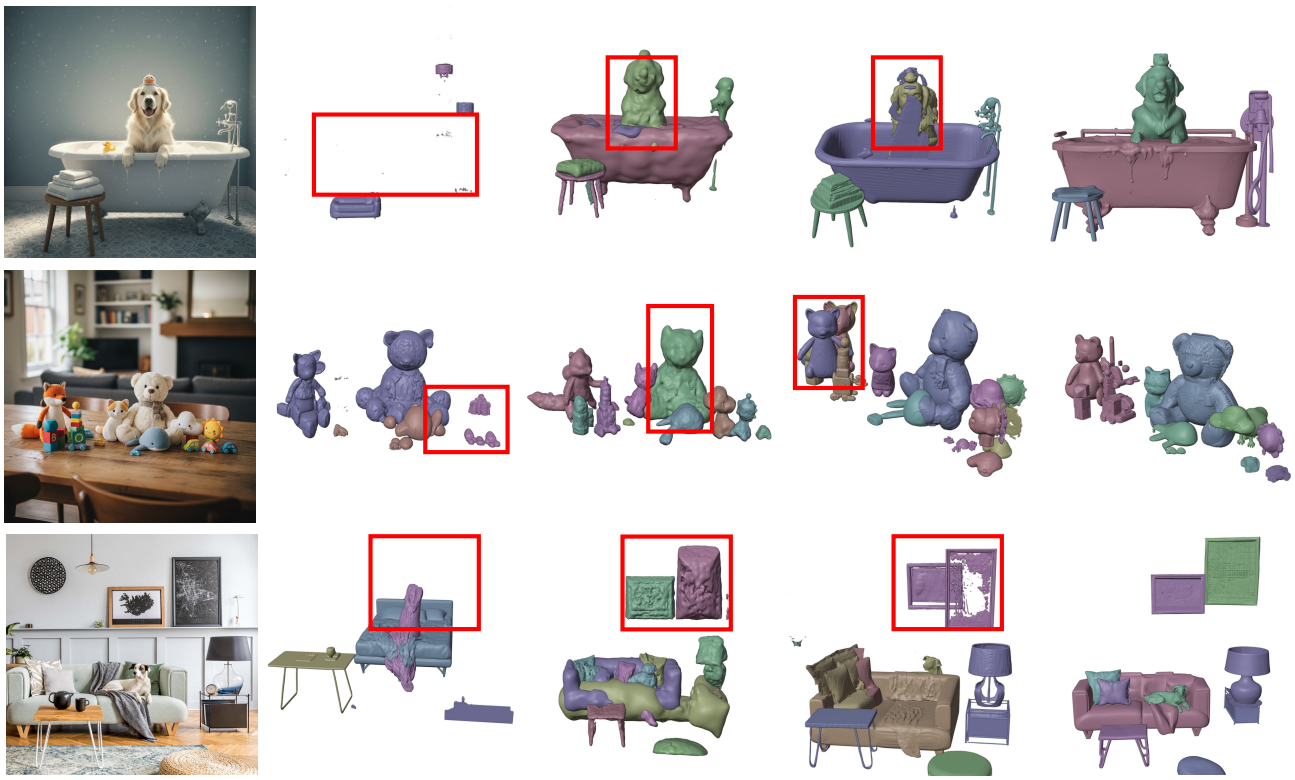

Figure 3. Qualitative results. CC-FMO demonstrates strong performance in this challenging scenario. Compared to baseline methods, our model produces meshes with superior geometric fidelity while maintaining object layouts that closely align with the input images. Refer to our supplementary materials for more qualitative comparisons.

It is worth noting that the real-world images in our evaluation are substantially more complex than those used in prior work [24, 37, 89]. For example, Fig. 3 illustrates that our model can handle complex indoor scenes.

Baselines. We compare CC-FMO with single image-based compositional scene generation models, including MIDI[24], DPA[92], Gen3DSR[12], PartCrafter[37] and DepR[89]. As DPA is also a zero-shot pipeline with foundation instance generator more outdated than ours, we replace its instance generator with ours to make fair comparisons. Thus evaluation difference between DPA and CC-FMO results from our effectiveness in pose estimation. For methods which are dependent on segmentations, like MIDI, we provide them with the segmentation masks same to ours to eliminate the effects from these irrelevant factors. Also the camera parameters are set as the same across all the methods. These baselines adopt various recipes for generation and training. The comparison between our model and them thouroughly testifies the effectiveness of our model design.

Metrics. Following MIDI [24], we adopt the widely used Chamfer Distance (CD) and F-Score (with the default threshold of 0.1) to evaluate scene-level performance. In addition, we compute *object-level* Chamfer Distances and F-Scores between each generated object and its ground-truth counterpart to assess geometric fidelity at the object level.

We further calculate the Volumetric Intersection over Union (Volume IoU) between the bounding boxes of generated objects and those of the ground truth to evaluate the accuracy of the estimated object layout and spatial consistency.

4.2. Quantitative Experiments

As shown in Tab. 1, our zero-shot pipeline outperforms all baselines on all scene-level metrics, which is particularly noteworthy given that MIDI is trained directly on 3D-FRONT. Since our evaluation dataset is rendered with different camera parameters and offsets, we attribute this superior performance over dataset-specific models to the camera-conditioned design of CC-FMO. In particular, training-dependent methods such as MIDI implicitly encode the training-time camera configuration in their model

Table 2. Ablation study of our hybrid instance generator. Quantitatively evaluated on 3D-FRONT [15] using object-level Chamfer Distance (CD-O) and F-Score (F-Score-O), shown in percentage.

VecSet	SLAT	CD-O↓	FScore-O↑
	✓	1.369	81.06
✓		1.166	78.66
✓	✓	**0.756**	**89.11**

Table 3. Ablation study of our pose estimation method. Quantitatively evaluated on 3D-FRONT [15] using scene-level Chamfer Distance (CD-S), F-Score (F-Score-S), and object bounding box Volume IoU (IoU-B). CD and FScore are shown in percentage.

Scale Solving	Foundation-Pose	Normal Painting	CD-S↓	FScore-S↑	IoU-B↑
	✓	✓	16.40	32.56	0.351
✓		✓	3.159	81.26	0.555
✓	✓		2.714	83.22	0.569
✓	✓	✓	**1.923**	**87.20**	**0.629**

parameters. When the camera parameters at test time differ from those seen during training, their performance degrades substantially, as they tend to produce erroneous inter-object distances due to compression or expansion of depth of view across camera setups.

Moreover, CC-FMO achieves the best performance on all object-level metrics, namely CD-O and FScore-O. We attribute this advantage to the generalizable instance generation capability and occlusion robustness afforded by our carefully designed instance generation pipeline.

4.3. Qualitative Experiments

We qualitatively compare CC-FMO with baseline models on real-world images in Fig. 3. The evaluated scenes can contain around 10 objects. As shown in the figure, pipelines powered by foundation models produce scenes with smoother and more coherent geometry, highlighting the effectiveness of foundation models for this task. In contrast, methods trained on specific datasets, such as MIDI, tend to generate meshes with noticeable geometric discontinuities, as illustrated in the bottom example. We hypothesize that this degradation arises from distributional shifts between real-world objects and the curated training datasets. These observations support the use of foundation models as the core of our pipeline. Compared to baselines, CC-FMO produces meshes with the better geometric quality, and the resulting object layouts aligns with the input images.

4.4. Ablation Study

Hybrid Instance Generation. As shown in Tab. 2 and Fig. 4, ablating the proposed hybrid instance generation

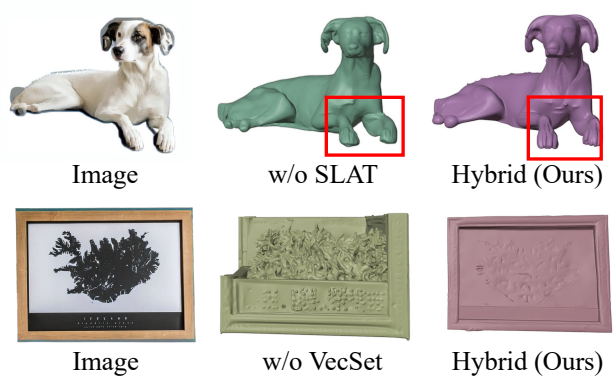

Figure 4. Ablation visualization. Removing the structured latent component (SLAT) leads to degraded geometric quality, while omitting the vector set (VecSet) results in reduced semantic accuracy. Our instance generator effectively preserves semantic accuracy and simultaneously produces detailed geometry.

scheme yields objects with either incorrect semantics or noticeably degraded geometric details. In contrast, our instance generator strikes a balance between semantics preservation and geometric fidelity. Interestingly, although SLAT models produce instances with higher geometric fidelity when provided with unoccluded image inputs, removing SLAT from our instance generation pipeline leads to a smaller performance drop than removing the VecSet model. This observation highlights the crucial role of semantics preservation, particularly in occlusion-heavy scenarios.

Scale-Aware Pose Estimation. In Tab. 3, we compare pose estimation performance with and without the proposed closed-form scale solving algorithm and FoundationPose. Due to the scale priors implicitly learned during the training of instance generation models, FoundationPose cannot be directly applied without explicit scale estimation.

We further evaluate CC-FMO with the proposed normal-painting technique. The results demonstrate that this component is essential for accurate pose estimation. By encoding geometric details into the texture as normal maps, the method encourages FoundationPose to prioritize alignment of geometric structure over appearance-based textures.

5. Future Work and Conclusion

Building upon our work, future works can extend the input modalities for instance generation, instead of considering only images as inputs, which may fall short in fidelity-sensitive scenarios, due to scale-depth-focal ambiguity. We propose CC-FMO, a compositional scene generation framework with end-to-end foundation models. With proposed CC-FMO, we demonstrate that scene-level finetuning is not

strictly necessary to achieve high-fidelity compositional scene generation. We identify key challenges specific to compositional scene generation that hinder the direct application of foundation models, and then address them through carefully designed model integration and a novel camera-aware, metric pose estimation algorithm. Our experiments show that CC-FMO generates high-fidelity scenes in a camera-conditioned manner. Despite being training-free, CC-FMO attains fidelity comparable to, or even surpassing, that of methods requiring task-specific training, highlighting the effectiveness of our design. We believe that our approach can benefit downstream applications that demand precise 3D scene generation.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[2] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023. 1

[3] Kai Chen and Qi Dou. Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In *ICCV*, 2021. 3

[4] Yixin Chen, Junfeng Ni, Nan Jiang, Yaowei Zhang, Yixin Zhu, and Siyuan Huang. Single-view 3d scene reconstruction with high-fidelity shape and texture. In *2024 International Conference on 3D Vision (3DV)*, pages 1456–1467. IEEE, 2024. 2

[5] Yongwei Chen, Tengfei Wang, Tong Wu, Xingang Pan, Kui Jia, and Ziwei Liu. Comboverse: Compositional 3d assets creation using spatially-aware diffusion guidance. *arXiv preprint arXiv:2403.12409*, 2024. 2

[6] 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*, 2025. 4

[7] Tao Chu, Pan Zhang, Qiong Liu, and Jiaqi Wang. Buol: A bottom-up framework with occupancy-aware lifting for panoptic 3d scene reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4937–4946, 2023. 2

[8] 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. 2

[9] 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, 3

[10] 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, 2024. 1, 3

[11] Yan Di and Ruida Zhang. Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. In *CVPR*, 2022. 3

[12] Andreea Dogaru, Mert Özer, and Bernhard Egger. Generalizable 3d scene reconstruction via divide and conquer from a single view. *arXiv preprint arXiv:2404.03421*, 2024. 2, 3, 6, 7

[13] Junting Dong, Qi Fang, Zehuan Huang, Xudong Xu, Jingbo Wang, Sida Peng, and Bo Dai. Tela: Text to layer-wise 3d clothed human generation. In *European Conference on Computer Vision*, pages 19–36. Springer, 2025. 3

[14] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021. 2

[15] 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 the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021. 6, 8

[16] Wei Gao and Russ Tedrake. Filterreg: Robust and efficient probabilistic point-set registration using gaussian filter and twist parameterization, 2019. 1

[17] Georgia Gkioxari, Nikhila Ravi, and Justin Johnson. Learning 3d object shape and layout without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1695–1704, 2022. 2

[18] Haonan Han, Rui Yang, Huan Liao, Jiankai Xing, Zunnan Xu, Xiaoming Yu, Junwei Zha, Xiu Li, and Wanhua Li. Reparo: Compositional 3d assets generation with differentiable 3d layout alignment. *arXiv preprint arXiv:2405.18525*, 2024. 2, 3

[19] Yisheng He and Wei Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *CVPR*, 2020. 3

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3

[21] 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. 1, 3

[22] Zehuan Huang, Yuanchen Guo, Haoran Wang, Ran Yi, Lizhuang Ma, Yan-Pei Cao, and Lu Sheng. Mv-adapter: Multi-view consistent image generation made easy. *arXiv preprint arXiv:2412.03632*, 2024. 3, 6

[23] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, et al. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9784–9794, 2024. 3

[24] 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, 2, 4, 6, 7

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

[26] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model, 2024. 1

[27] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 2

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 6

[30] Yann Labbé and Lucas Manuelli. Megapose: 6d pose estimation of novel objects via render & compare. In *CoRL*, 2022. 3

[31] Weiyu Li, Jiarui Liu, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*, 2024. 3

[32] Weiyu Li, Xuanyang Zhang, Zheng Sun, Di Qi, Hao Li, Wei Cheng, Weiwei Cai, Shihao Wu, Jiarui Liu, Zihao Wang, et al. Step1x-3d: Towards high-fidelity and controllable generation of textured 3d assets. *arXiv preprint arXiv:2505.07747*, 2025. 4, 6

[33] 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. 3

[34] Zhigang Li and Gu Wang. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *ICCV*, 2019. 3

[35] Jiehong Lin and Lihua Liu. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. In *CVPR*, 2024. 3

[36] Jiehong Lin and Zewei Wei. Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks. In *ECCV*, 2022. 3

[37] Yuchen Lin, Chenguo Lin, Panwang Pan, Honglei Yan, Yiqiang Feng, Yadong Mu, and Katerina Fragkiadaki. Partcrafter: Structured 3d mesh generation via compositional latent diffusion transformers. *arXiv preprint arXiv:2506.05573*, 2025. 1, 6, 7

[38] Anran Liu, Cheng Lin, Yuan Liu, Xiaoxiao Long, Zhiyang Dou, Hao-Xiang Guo, Ping Luo, and Wenping Wang. Part123: part-aware 3d reconstruction from a single-view image. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 3

[39] 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. 2

[40] Jianhui Liu and Yukang Chen. Ist-net: Prior-free category-level pose estimation with implicit space transformation. In *ICCV*, 2023. 3

[41] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10072–10083, 2024. 3

[42] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[43] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2

[44] Yuan Liu and Yilin Wen. Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. In *ECCV*, 2022. 3

[45] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 1, 3

[46] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 1, 3

[47] Quan Meng, Lei Li, Matthias Nießner, and Angela Dai. Lt3sd: Latent trees for 3d scene diffusion. *arXiv preprint arXiv:2409.08215*, 2024. 3

[48] Van Nguyen Nguyen and Thibault Groueix. Gigapose: Fast and robust novel object pose estimation via one correspondence. In *CVPR*, 2024. 3

[49] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 16784–16804, 2022. 2

[50] 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. 2

[51] Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. Atiss: Autoregressive transformers for indoor scene synthesis. *Advances in Neural Information Processing Systems*, 34:12013–12026, 2021. 2

[52] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 1

[53] 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. 3

[54] Sida Peng and Yuan Liu. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *CVPR*, 2019. 3

[55] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 3, 6

[56] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022.

[57] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 2, 3, 6

[58] Barbara Roessle, Norman Müller, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, Angela Dai, and Matthias Nießner. L3dg: Latent 3d gaussian diffusion. *arXiv preprint arXiv:2410.13530*, 2024. 3

[59] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[60] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[61] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 3

[62] Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. Diffuscene: Denoising diffusion models for generative indoor scene synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20507–20518, 2024. 2

[63] 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, 2025. 1, 3

[64] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2025. 3

[65] Chen Wang and Danfei Xu. Densefusion: 6d object pose estimation by iterative dense fusion. In *CVPR*, 2019. 3

[66] 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. *Advances in Neural Information Processing Systems*, 34:27171–27183, 2021. 3

[67] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. *arXiv preprint arXiv:2403.05034*, 2024. 3

[68] Bowen Wen and Wei Yang. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *CVPR*, 2024. 3

[69] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868–17879, 2024. 2, 4, 6

[70] Hao Wen, Zehuan Huang, Yaohui Wang, Xinyuan Chen, Yu Qiao, and Lu Sheng. Ouroboros3d: Image-to-3d generation via 3d-aware recursive diffusion. *arXiv preprint arXiv:2406.03184*, 2024. 3

[71] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *arXiv preprint arXiv:2405.20343*, 2024.

[72] 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. *arXiv preprint arXiv:2405.14832*, 2024. 3

[73] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Yajie Bao, Jiachen Qian, Siyu Zhu, Philip Torr, Xun

Cao, and Yao Yao. Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention. *arXiv preprint arXiv:2505.17412*, 2025. 3

[74] Zhennan Wu, Yang Li, Han Yan, Taizhang Shang, Weixuan Sun, Senbo Wang, Ruikai Cui, Weizhe Liu, Hiroyuki Sato, Hongdong Li, et al. Blockfusion: Expandable 3d scene generation using latent tri-plane extrapolation. *ACM Transactions on Graphics (TOG)*, 43(4):1–17, 2024. 3

[75] 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, 2025. 2, 3

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

[77] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*, 2024. 3

[78] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 3, 5, 6

[79] Kaixin Yao, Longwen Zhang, Xinhao Yan, Yan Zeng, Qixuan Zhang, Wei Yang, Lan Xu, Jiayuan Gu, and Jingyi Yu. Cast: Component-aligned 3d scene reconstruction from an rgb image, 2025. 2

[80] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (TOG)*, 43(6):1–18, 2024. 2, 6

[81] Chongjie Ye, Yushuang Wu, Ziteng Lu, Jiahao Chang, Xiaoyang Guo, Jiaqing Zhou, Hao Zhao, and Xiaoguang Han. Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. *arXiv preprint arXiv:2503.22236*, 2025. 3, 4, 6

[82] Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montse Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, Brian Ichter, Ted Xiao, Peng Xu, Andy Zeng, Tingnan Zhang, Nicolas Heess, Dorsa Sadigh, Jie Tan, Yuval Tassa, and Fei Xia. Language to rewards for robotic skill synthesis, 2023. 1

[83] Sergey Zakharov and Ivan Shugurov. Dpod: 6d pose object detector and refiner. In *ICCV*, 2019. 3

[84] 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. 2

[85] Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, and Shuaicheng Liu. Holistic 3d scene understanding from a single image with implicit representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8833–8842, 2021. 2

[86] 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

[87] Xiang Zhang, Zeyuan Chen, Fangyin Wei, and Zhuowen Tu. Uni-3d: A universal model for panoptic 3d scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9256–9266, 2023. 2

[88] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1724–1732, 2024. 3, 6

[89] Qingcheng Zhao, Xiang Zhang, Haiyang Xu, Zeyuan Chen, Jianwen Xie, Yuan Gao, and Zhuowen Tu. Depr: Depth guided single-view scene reconstruction with instance-level diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5722–5733, 2025. 2, 6, 7

[90] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[91] Linfang Zheng and Chen Wang. Hs-pose: Hybrid scope feature extraction for category-level object pose estimation. In *CVPR*, 2023. 3

[92] Junsheng Zhou, Yu-Shen Liu, and Zhizhong Han. Zero-shot scene reconstruction from single images with deep prior assembly. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2, 4, 6, 7

[93] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10324–10335, 2024. 3

CC-FMO: Camera-Conditioned Zero-Shot Single Image to 3D Scene Generation with Foundation Model Orchestration

Supplementary Material

6. More Implementation Details

It is worth mentioning that $\{b_i\}_{i=1}^N$, $\{m_i\}_{i=1}^N$ and $\{l_i\}_{i=1}^N$ are paired, which enables us to filter out the object segmentations that do not stand for object instances at all, such as those with labels "floor", "room", or "wall". To facilitate the subsequent object generation procedure, we enlarge the object regions in $\{I_i := I * m_i\}_{i=1}^N$ by centering and normalizing, such that the non-zero region occupies $\alpha = 60\%$ of the max width or height of the image, where the value is determined from small-scale experiments. The inpainting is done with prompt "a complete model of l_i, white background". For all the models we align their generations with ground-truth ones by FilterReg[16] during testing such that their orientations or offsets won't affect the scores.

7. Rationale for Baseline Choice

The selected baselines span the spectrum between specifically-trained and foundation-assisted pipelines:

- DepR is specifically trained. It utilizes depth information throught both training and inference, where the depth is provided by a foundation model.
- PartCrafter is specifically trained, which generates 3D instances in part-level with a 3D-native DiT[52]. It is also able to be applied on scene generation by taking object instances as "parts".
- MIDI-3D is specifically trained, featured by compatibility with occluded object image inputs and a cross-instance attention mechanism to incorporate global context during scene generation.
- DPA is zero-shot. Whereas it generates 3D objects with foundation models, it calibrates generated objects by minimizing Chamfer Distances between point clouds with RANSAC-like algorithm.
- Gen3DSR is zero-shot. It decomposes scene generation as subtasks such as entity segmentation, depth estimation, etc. Each subtask is tackled via a different foundation model.

These baselines adopt various recipes for generation and training. The comparison between our model and them thouroughly testifies the effectiveness of our model design.

8. More Qualitative Comparisons

In Fig. 5, we provide more qualitative comparisons between our model and MIDI, the strongest baseline. The cases contain examples with various lights, backgrounds, numbers of objects, and scenes. Especially, the cameras are not aligned with the coordinates of objects, which makes the cases challenging.

Figure 5. Qualitative comparison between CC-FMO and MIDI. Note that the segmentation masks are the same.