DON'T ASK AGENTS TO DO EVERYTHING: SELECTIVE AGENTIC REASONING FOR SINGLE IMAGE TO 3D SCENE GENERATION

Anonymous authors
Paper under double-blind review

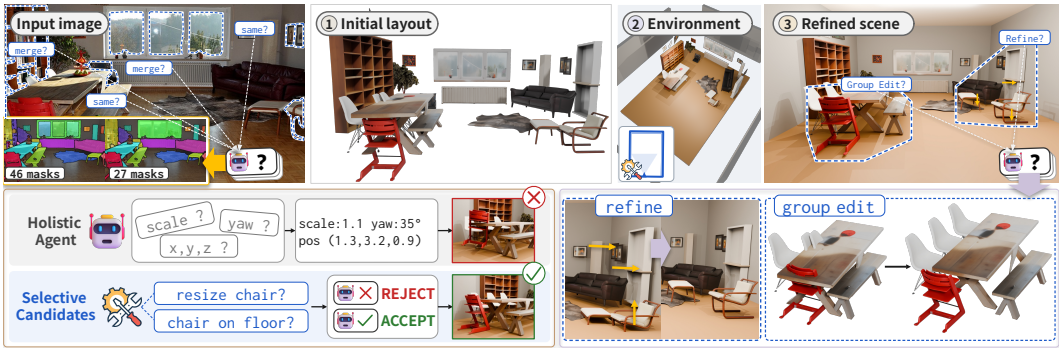

Figure 1: Our framework turns a single image into a complete 3D scene through three stages. By letting agents select among given options, it refines the scene more accurately.

ABSTRACT

Recent agentic scene generation frameworks often rely on holistic pipelines that ask agents to resolve multiple entangled factors at once, which limits their performance in complex real-world environments. We propose a framework that generates a complete 3D scene from a single image in three stages, *i.e.*, scene initialization, environment construction, and scene refinement. Our key idea, termed *selective agentic reasoning*, is to compute what can be measured geometrically and to leave agents mainly the decisions that remain ambiguous, most of which are simple accept-or-reject questions about candidates proposed from geometry. During initialization, an inspector agent segments and describes the objects in the image, and a reviewer agent decides which proposed mask merges and deletions to apply before each object is generated and posed. The scene is then brought to metric scale and enclosed by a room fitted to the observed walls, which anchors every later correction. Finally, the reviewer screens geometry-proposed corrections of the arrangement, and a group editor agent locally rearranges only the groups that remain wrong before they are merged back. The output is a complete 3D scene that stays faithful to the input image. To pose objects accurately, we further introduce a geometry-aware layout predictor supervised by the point map estimated from the input image, so it can also be trained on real-world images without 3D annotation. Experiments on synthetic and real-world benchmarks show that our framework achieves the highest perceptual quality among agent-based and generative approaches while ranking first or second on every physical metric, and that its layout predictor remains competitive in geometric accuracy.

1 INTRODUCTION

3D scene generation serves as a foundation for immersive and interactive applications such as AR/VR, robotics, and digital content creation. Recent generative models have greatly improved object-level 3D asset generation (Mescheder et al., 2019; Xie et al., 2019; Nichol et al., 2022; Vahdat et al., 2022; Gao et al., 2022; Jun & Nichol, 2023). Extending this to an entire scene from a single image, however, remains far more challenging, as it requires jointly inferring object geometry, spatial layout, inter-object relationships, and environmental context from limited visual evidence.

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.

Recent generative methods (Huang et al., 2025b; Lin et al., 2025; Meng et al., 2025; Wang et al., 2026; Team et al., 2025; Yin et al., 2026b; Siddiqui et al., 2026) attempt to produce scene objects together with their layouts, either jointly within a unified generation process or object by object with learned pose estimation. While effective at modeling global structure, training such generative models requires scene-level 3D supervision, which is confined to synthetic or domain-restricted datasets, and thus their scalability is bounded by annotated 3D data rather than by real-world images. On the other hand, agent-based methods use the priors of vision language models (VLMs) to plan and assemble scenes (Noh et al., 2026; Yang et al., 2025; Sun et al., 2025b; Yin et al., 2026a; Yang et al., 2026; Pfaff et al., 2026; Ma et al., 2026), which generalize beyond such supervision and keep arrangements plausible by drawing on commonsense layout knowledge. While VLMs can reason about spatial structure from images and text, obtaining precise metric measurements of scene geometry through vanilla VLM reasoning remains challenging. As a result, they may struggle with accurate object placement when precise geometric reasoning is required, *e.g.*, when aligning a reconstructed scene with the input image, often leading to discrepancies between the generated scene and the photograph.

In this paper, we argue that agents should not be asked to do everything. Many errors of single-image scene generation, *e.g.*, fragmented masks, floating objects, or misaligned walls, can be measured from the image and its point map, and what remains for an agent is a narrow judgment, *e.g.*, whether two adjacent masks belong to one object. We therefore propose a multi-agent framework built on *selective agentic reasoning*, where geometric procedures enumerate a small set of candidate edits from measured evidence and a reviewer agent accepts or rejects each of them against the photograph. Thus, the reviewer never has to output a coordinate, yet its commonsense still rejects the candidates that geometry alone cannot rule out, as shown in Figure 1.

Specifically, the framework consists of three stages. In scene initialization, an inspector agent identifies the objects and reads their attributes, the reviewer rules on the proposed mask candidates, *e.g.*, merges and deletions, and each refined object is reconstructed as a mesh and posed by our layout predictor. Environment construction then scales the scene to real-world units and fits the room to the walls observed in the point map, and this room serves as the reference for the corrections in scene refinement. In scene refinement, the reviewer rules on corrections proposed from the layout, the room, and the object attributes, and only the groups that still fail a geometric re-check are rearranged by a group editor agent before deterministic physical finalization. Our contributions are twofold. First, we propose selective agentic reasoning for single-image 3D scene generation, where agents rule on geometry-enumerated candidates at the two points where the scene is corrected, *i.e.*, mask refinement and scene refinement. Second, we introduce a geometry-aware layout predictor trained with a geometric loss against the estimated point map, which learns from real-world images without 3D layout annotations and expresses every pose in a shared floor-aligned frame. Integrated into a full system that outputs a complete Blender scene, the two produce scenes that are more faithful to the photograph than those of prior agent-based and generative methods, with physical validity comparable to the strongest baseline.

2 Related Work

3D Scene Generation 3D scene generation builds on progress in 3D perception and object-level asset synthesis. Feed-forward perception models (Wang et al., 2024; 2025a;b) efficiently recover geometry from sparse or monocular observations but do not produce complete mesh-based scenes. Object-level 3D generation has advanced rapidly through feed-forward reconstruction (Hong et al., 2024; Xu et al., 2024) and diffusion-based mesh generation (Liu et al., 2024; Xiang et al., 2025; Li et al., 2025; Yang et al., 2024a) on large-scale asset datasets (Deitke et al., 2022), yet it remains focused on individual objects without scene-level reasoning. Compositional methods jointly generate multiple objects or parts in a unified process (Lin et al., 2025; Huang et al., 2025b; Meng et al., 2025; Wang et al., 2026), but such joint models require dense scene-level 3D supervision, which confines them to synthetic or domain-restricted datasets. Object-centric pipelines instead generate individual assets and assemble them via layout prediction or geometric alignment (Team et al., 2025; Yin et al., 2026b; Siddiqui et al., 2026; Yu et al., 2025; Yao et al., 2025); however, their final quality hinges on layout or pose modules trained on synthetic or domain-restricted data, which generalize poorly to in-the-wild images. Modular single-image pipelines (Dogaru et al., 2025; Zhao et al., 2025; Sautter et al., 2025) register each reconstructed object to its estimated depth points, leaving every pose to a

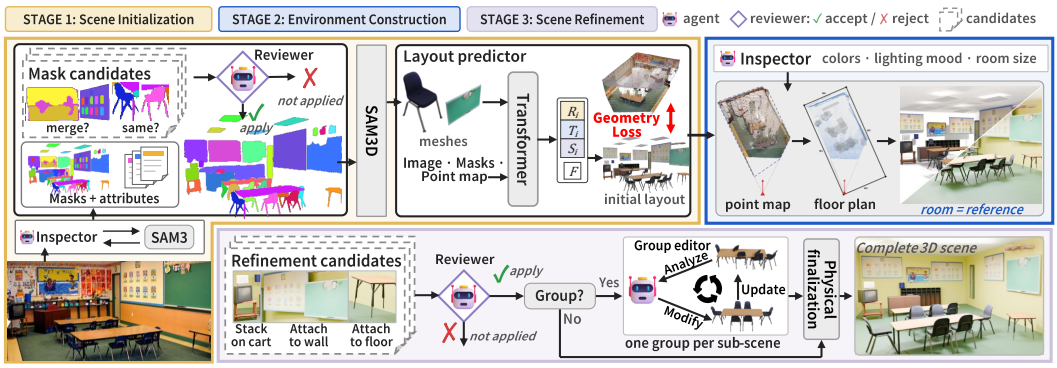

Figure 2: Overview of our framework. From a single image, scene initialization (Stage 1), environment construction (Stage 2), and scene refinement (Stage 3) build a complete 3D scene. In Stage 1, the geometry-aware layout predictor regresses per-object R, T, S, and floor rotation F.

per-object fit against partial and noisy depth, whereas SceneMaker (Shi et al., 2025) decouples de-occlusion and pose estimation from object generation and trains its pose model on synthetic scenes with ground-truth poses. Our layout predictor is likewise decoupled from object generation, but it is also supervised by the point map of the input image, which lets it learn from unannotated real-world images, and it regresses all poses jointly in a shared floor-aligned frame.

Agent-based Methods LLM/VLM agents offer a flexible paradigm for 3D scene generation, covering layout planning, asset retrieval, and scene editing (Feng et al., 2023; Yang et al., 2024b; Huang et al., 2024), but they rarely enforce reference-image constraints such as object composition, spatial arrangement, and room-level appearance. Recent agentic systems add iterative feedback and tool use. VIGA (Yin et al., 2026a) refines scenes through an image-conditioned generator–evaluator loop, SceneWeaver (Yang et al., 2025) repeatedly invokes specialized tools for construction and correction, and Edit-As-Act (Noh et al., 2026) grounds actions semantically to reduce parameter ambiguity. Code-as-Room (Yang et al., 2026), SceneSmith (Pfaff et al., 2026), REST3D (Ma et al., 2026), and SceneOrchestra (He et al., 2026) further explore scene-code synthesis, designer–critic assembly, single-image support-relation trees, and one-pass tool-call trajectories without step-wise review, respectively. Many of these systems rely on broader agent roles or iterative scene-level reasoning. Tool-augmented and propose-then-verify agents compose vision modules via generated programs (Gupta & Kembhavi, 2023; Surís et al., 2023) or pair a foundation model with external proposers and verifiers (Ahn et al., 2022; Yang et al., 2023; Nasiriany et al., 2024; Kambhampati et al., 2024). In 3D, FirePlace (Huang et al., 2025a) prunes constraint-solved placements with a VLM, while LayoutVLM (Sun et al., 2025a) differentiably optimizes VLM-proposed layouts. In contrast, geometry in our framework first measures the input image and enumerates a small set of feasible candidate edits for both the object masks and the scene arrangement, a reviewer agent resolves only the remaining ambiguity by accepting or rejecting each candidate against the photograph, and the accepted edits are checked geometrically again. Unlike FirePlace, which prunes placements for a text instruction, both our candidates and decisions are grounded in the input photograph.

3 Proposed Method

In this section, we describe our three-stage framework that builds a complete 3D scene from a single image, as illustrated in Figure 2. Scene initialization (Stage 1) turns the image into object masks, meshes, and an initial layout, environment construction (Stage 2) builds a metric room around them, and scene refinement (Stage 3) corrects the remaining physical and arrangement errors. Since inferring spatial relations and metric values from rendered views of complex 3D geometry is challenging for an agent, we design *selective agentic reasoning*, where geometric procedures first measure the scene and enumerate candidate edits, and an agent mainly accepts or rejects each of them against the photograph. The candidates are proposed generously so that few correct edits are missed, and the agent only needs to reject the wrong ones. Accordingly, our agents take three roles: an *inspector* reads from the image what cannot be measured, such as object classes and surface colors, a *reviewer* accepts or rejects the candidates, and a *group editor* directly rearranges only the groups that remain unresolved after review. We also leverage off-the-shelf models for segmentation, point-map esti-

mation, and 3D object generation (Carion et al., 2025; Wang et al., 2025b; Team et al., 2025), and introduce a geometry-aware layout predictor, which is trained with point-map geometry and can thus also learn from real-world images without 3D layout annotations. Prompts, candidate rules, and thresholds are in Secs. F, C.2, and E of the supplementary material.

3.1 Stage 1: Scene Initialization

Since every later stage builds on the object masks, the first stage aims at a clean set of objects. The inspector first lists the object classes in the image, which prompt SAM3 to segment their masks, while MoGe estimates the point map of the image. As these masks may be mislabeled, fragmented, or spurious, the inspector then examines each mask from the full image with the mask outlined and a close-up crop, and reports whether the content matches the label, what the object rests on, and how much of it is visible. These reports are also kept for Stage 3, where, *e.g.*, what an object rests on becomes a candidate attachment. From these reports, the masks, and the point map, we enumerate mask candidates, *e.g.*, merging masks that overlap or meet along a seam of continuous depth, deleting small isolated masks or masks whose label the inspector reports as wrong, and letting same-class instances of agreeing size share one object model so that identical objects get the same mesh, each handed over with its measurements. The reviewer accepts or rejects each candidate, judging single masks and pairs from close-ups with a labeled overview and a group from one view of all its members, where it keeps only the members that belong together. A close-up shows whether a single mask is correct, while a view of the whole group lets the agent compare its members without mixing up their IDs. For example, two chairs that touch in the image form a single region by geometry alone, and only the photograph shows that they are two objects, so the reviewer rejects their merge. The accepted candidates form a mask plan, which is validated before execution. From the refined masks, SAM3D generates a canonical mesh for each object, and instead of its per-object poses, our layout predictor (Sec. 3.2) poses all meshes jointly in a floor-aligned frame.

3.2 Geometry-aware Layout Prediction

Since the layout largely affects the final scene even after refinement, we propose a dedicated geometry-aware layout predictor. Unlike methods that generate meshes and poses jointly (Team et al., 2025; Meng et al., 2025; Huang et al., 2025b), it regresses the layout alone, which avoids the cost of mesh generation and lets global attention span more objects. Two design choices make it accurate and trainable on unannotated images, *i.e.*, floor-aligned object poses and a point-map geometry loss. As shown in Figure 2, the image, masks, and point map are encoded into conditioning tokens by DINOv2 (Oquab et al., 2024) and the SAM3D point encoder (Team et al., 2025), and each mesh is embedded into a shape latent by the sparse-structure encoder (Xiang et al., 2025). Each shape latent is concatenated with learnable pose tokens $\{R, T, S, F\}$ and passes through transformer blocks that cross-attend to these conditioning tokens and alternate object-local and scene-global attention, which regress a rotation R_i, a translation T_i, a scale S_i, and a floor rotation F_i for each object i (Sec. C.3 of the supplementary material).

Floor-aligned object poses. All object poses are expressed in a floor-aligned scene frame whose ground plane is the XZ-plane, and a floor rotation F shared by all objects maps this frame to the camera frame, *i.e.*, a mesh vertex $\mathbf{x}_k$ of object i is placed at $\hat{\mathbf{x}}_k = F\left(R_i(S_i \odot \mathbf{x}_k) + T_i\right)$, where $\odot$ is element-wise scaling. Since F belongs to the scene rather than to any object, every object token predicts it in the 6D rotation representation (Zhou et al., 2019), a consistency loss $\mathcal{L}_{\text{cons}}$ penalizes the standard deviation of these predictions, and F is their average at inference, which gives every object and the later stages one consistently aligned floor.

Point-map geometry loss. Since the predictor regresses the layout, the placed meshes can be compared directly against the point map of the input image, which needs no 3D annotation. Let $\hat{\mathcal{M}}_i$ be points sampled from the mesh of object i and placed in the camera frame by this transformation, and $\mathcal{P}_i$ the point-map points inside mask i. Since the point map is sparse, noisy, and observes the visible side of each object alone, we employ a trimmed one-sided Chamfer distance from $\mathcal{P}_i$ to $\hat{\mathcal{M}}_i$ together with a bounding-box term,

$$\mathcal{L}^i_{\text{geo}} = \frac{1}{|\mathcal{P}_i^\tau|} \sum_{\mathbf{p} \in \mathcal{P}_i^\tau} \min_{\hat{\mathbf{x}} \in \hat{\mathcal{M}}_i} \|\mathbf{p} - \hat{\mathbf{x}}\|_2^2 + \lambda \left(\|\hat{\mathbf{b}}_i^{xy} - \mathbf{b}_i^{xy}\|_1 + |\hat{\mathbf{b}}_i^{z_{\min}} - \mathbf{b}_i^{z_{\min}}| \right), \tag{1}$$

where $\mathcal{P}_i^\tau$ keeps the τ closest points to suppress outliers at mask boundaries, and $\mathbf{b}_i$ is a pseudo box computed from robust quantiles of $\mathcal{P}_i$ in the camera frame, *i.e.*, the lateral extent $\mathbf{b}_i^{xy}$ and the near-depth bound $\mathbf{b}_i^{z\min}$, since the far side is unobserved, and $\hat{\mathbf{b}}_i$ is the same quantity from $\hat{\mathcal{M}}_i$. The Chamfer term alone can be minimized by inflating the object along the viewing direction, which the box term prevents by anchoring the lateral extent and nearest depth of the object. The box term is weighted by λ, since pseudo boxes are less reliable under partial occlusion.

3.3 Stage 2: Environment Construction

This stage builds the room that anchors the later refinement, since every correction of Stage 3, *e.g.*, attaching an object to its wall or keeping it inside the room, is expressed against it. As in Stage 1, geometry handles what can be measured, and the inspector only reads what cannot, *i.e.*, the surface colors, the lighting mood, and a coarse room size. Since a scene reconstructed from one image has no real-world unit, we first fix its metric scale from two cues, *i.e.*, the per-class size priors of the reconstructed objects and the camera eye height above the point-map floor, bounded by the room size read by the inspector. We then build a floor plan in a bird's-eye view of the floor-aligned frame (Sec. 3.2). The plan follows the dominant direction of the vertical surfaces in the point map, and each of its sides is placed on a wall plane observed in the point map. Where no wall is visible, *e.g.*, behind the camera, the side is placed just outside the observed floor, the objects, and the camera, so that the room contains everything the image shows. No agent drafts this plan, since the walls are measurable, whereas a drafted polygon may fail to follow them or to enclose every object. Finally, walls are raised along the plan, a ceiling is placed at the height read from the point map, and a preset lighting rig and the surface colors are set from what the inspector reads.

3.4 Stage 3: Scene Refinement

Although the initialized scene follows the photograph globally, it still contains local errors, *e.g.*, floating or interpenetrating objects, identical instances of mismatched scales, or implausible arrangements. Rather than asking an agent to fix them, we repeat the candidate-and-review pattern of Stage 1 on the 3D scene, as illustrated in the last stage of Figure 2. From the initial layout, the room, and the object attributes of Stage 1, we enumerate candidate object groups, such as chairs around a table or frames on one wall, and candidate operations, such as attaching an object to its support or nearest wall and unifying the scale of identical instances. The reviewer rules on the groups and then on the operations against the photograph. Besides accepting or rejecting, it may redirect an attachment to another existing wall or support and describes the photographed arrangement of each confirmed group, but it never outputs a coordinate, so the pose resulting from each of its decisions is computed geometrically. For example, geometry may propose to attach a lamp to the nearest wall, while the photograph shows that it stands on a side table, so the reviewer redirects the attachment to the table. Leveling tilted objects and completing truncated supports are applied without review.

Finally, after the accepted operations are applied, each group that still fails a geometric re-check is handed to the group editor, which sees only that group in a sub-scene around its fixed anchor and iteratively moves its members toward the described arrangement with bounded planar steps, keeping an edit only if it improves a deterministic score. We edit each group apart from the rest of the scene, because objects that interact closely, such as chairs around a table, are hard to fix all at once in the full scene. After the groups are merged back, physical finalization makes the whole scene physically valid, since the reviewed operations and the group edits correct each object or group locally but can still leave objects floating or interpenetrating across groups. It is a fixed sequence of deterministic passes that involves no agent, *i.e.*, remaining interpenetrations are resolved by moving the smaller free object within a small budget, hovering objects are lowered onto their first support, wall-mounted objects are aligned flush to their wall, and wall- or ceiling-hung objects are merged into the room geometry. The refined scene is then rendered with the environment of Sec. 3.3.

4 Experiments & Discussion

Implementation details. All agents in the proposed framework are instantiated from Qwen3.8-27B (Qwen Team, 2026b) with role-specific prompts, and all geometric procedures run in Python and Blender, so that the output is a complete Blender scene. The layout predictor is trained on 3D



Figure 3: Qualitative comparison of complete scene generation on real-world indoor scenes.

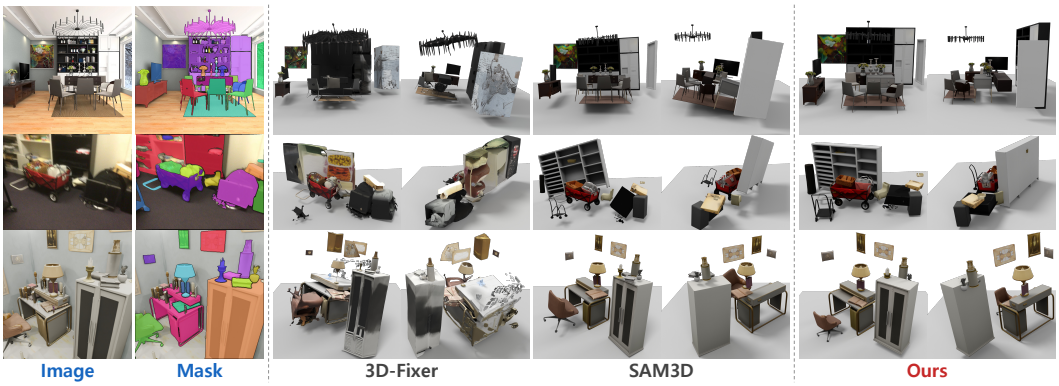

Figure 4: Qualitative comparison of layout prediction with 3D-Fixer and SAM3D.

indoor datasets (Fu et al., 2021; Dai et al., 2017; Xia et al., 2026) with COCO (Lin et al., 2014). We evaluate the proposed framework on two tasks. For complete scene generation, we compare each stage of our framework with agent-based generators (Yin et al., 2026a; Ma et al., 2026; Yang et al., 2026), run on the same Qwen3.8-27B model, and an object-centric generator (Team et al., 2025) on 160 real-world indoor images (Quattoni & Torralba, 2009; Lin et al., 2014). For layout prediction, we compare our predictor with joint (Huang et al., 2025b; Lin et al., 2025; Meng et al., 2025) and object-centric generators (Yin et al., 2026b; Team et al., 2025) on synthetic and real-world benchmarks (Fu et al., 2021; Dai et al., 2017; Xia et al., 2026). Dataset construction, training, alignment, rendering, and evaluation protocols are detailed in Sec. D of the supplementary material.

Evaluation metrics. For complete scene generation, we measure perceptual quality with a VLM score, where Qwen3.5-9B (Qwen Team, 2026a) rates realism, functionality, layout, and image alignment from 0 to 10 given the input photograph and a rendering of the scene, and with CLIP, the cosine similarity between the CLIP (Radford et al., 2021) embeddings of the photograph and the rendering.

Table 1: **Scene-level comparison** on MIT-67 and COCO, where VLM and CLIP measure perceptual quality and Float, Coll., Wall, and Stable measure physical validity.

	MIT-67						COCO					
Method	VLM ↑	CLIP ↑	Float ↓	Coll. ↓	Wall ↓	Stable ↑	VLM ↑	CLIP ↑	Float ↓	Coll. ↓	Wall ↓	Stable ↑
SAM3D	4.3176	0.7262	–	21.3	–	–	4.5267	0.6733	–	15.9	–	–
VIGA	5.7618	0.7028	49.0	38.5	34.6	8.6	5.7733	0.6487	43.8	46.0	40.5	8.5
REST3D	6.4559	0.6962	71.6	5.9	42.1	56.8	7.3480	0.6639	70.2	5.5	37.1	59.7
Code-as-Room	6.9765	0.7155	23.5	18.5	12.6	15.4	7.3233	0.6644	18.8	18.9	14.0	15.3
Ours (Stage 1)	5.6029	0.7509	76.7	20.3	–	6.9	5.9867	0.7011	72.6	16.1	–	12.5
Ours (Stage 2)	6.6176	0.7782	74.2	20.6	13.2	9.3	7.4333	0.7240	72.3	16.4	13.7	12.2
Ours (Stage 3)	7.3471	0.7649	20.2	5.3	2.1	56.4	7.4733	0.7124	23.8	5.7	3.5	62.0

Table 2: **Geometric comparison** of layout prediction on 3D-FUTURE, ScanNet, and SAGE.

	3D-FUTURE			ScanNet			SAGE		
Model	CD ↓	F-Score ↑	mIoU ↑	CD ↓	F-Score ↑	mIoU ↑	CD ↓	F-Score ↑	mIoU ↑
MIDI	0.1180	0.6190	0.2785	0.2088	0.4460	0.2313	0.1879	0.4119	0.0715
PartCrafter	0.2186	0.4641	0.1430	0.3812	0.2904	0.1091	0.2096	0.3932	0.0363
3D-Fixer	0.0406	0.8400	0.4716	0.0638	0.6296	0.3268	0.0239	0.8439	0.2823
SceneGen	0.0337	0.8435	0.4675	0.1532	0.4897	0.1807	0.1229	0.4996	0.0559
SAM3D	0.0122	0.9218	0.5904	0.0499	0.6884	0.3714	0.0038	0.9643	0.4801
Ours	0.0106	0.9246	0.5820	0.0489	0.6944	0.3917	0.0048	0.9511	0.4029

Physical validity is measured on the final meshes as the percentages of objects that float (Float), collide with another object (Coll.), penetrate the room (Wall), and stay in place under rigid-body simulation (Stable) (Ma et al., 2026). For layout prediction, each predicted scene is aligned to the ground truth by one global rigid transform, and we report the Chamfer distance (CD), F-Score, and 3D bounding-box mIoU. Best and second-best results are marked in red and yellow.

4.1 Evaluation

Complete scene generation. As shown in Table 1, our method achieves the best perceptual quality on both datasets and ranks first or second on every metric of physical validity, trailing REST3D in stability on MIT-67 and in interpenetration on COCO and Code-as-Room in floating objects on COCO. The stage-wise rows show how these results build up. Stage 1 already aligns with the input image better than all baselines in CLIP, including SAM3D, which shares our object generator and photo camera, and Stage 2 improves further by placing the objects in a metric room with materials and lighting. Stage 3 then sharply reduces floating objects and increases stable ones, as the refinement candidates accepted by the reviewer and the subsequent deterministic physical finalization, which accounts for about a third of the stability gain (Table S1 of the supplementary material), attach the objects to their supports while keeping the photographed arrangement. Since these corrections are expressed against the room of Stage 2, only 2.1% (MIT-67) and 3.5% (COCO) of our objects end up penetrating the room, compared with at least 12.6% for the baselines. REST3D, which optimizes object poses with a physics simulator, is the closest competitor in collision and stability, yet our framework, which relies on geometric rules alone, is on par with it on both datasets. Our framework also remains far more faithful to the input image, while the simulation of REST3D moves objects away from their observed poses. Code-as-Room, which writes the object coordinates directly, leaves few objects floating, but only about 15% of its objects are stable on both datasets. As shown in Figure 3, the baselines often drift from the observed layout, whereas ours keeps the photographed objects and their arrangement, *e.g.*, the monitors, desk, and shelves of the office scene (fifth row) stand upright on one floor, and the bathroom fixtures (first two rows) rest against the walls without interpenetration, thus leading to stable and richer scenes.

Layout prediction. We further evaluate our layout predictor alone against scene generation methods. As shown in Table 2, it achieves the best results on every metric of ScanNet and on CD and F-Score of 3D-FUTURE, whereas the joint generators fall far behind on every benchmark. The key difference from the object-centric baselines is that our predictor regresses all objects jointly in one floor-aligned frame, so that every object stands upright on a shared floor. As shown in Figure 4, 3D-

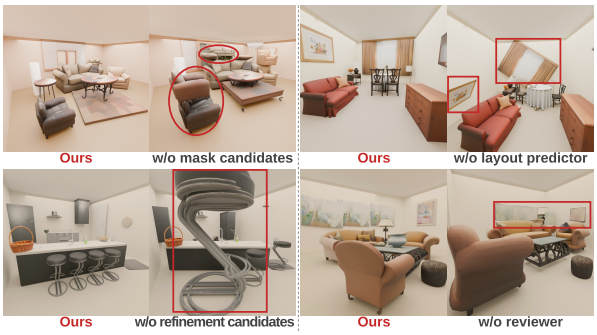

Figure 5: Ablation of the full framework.

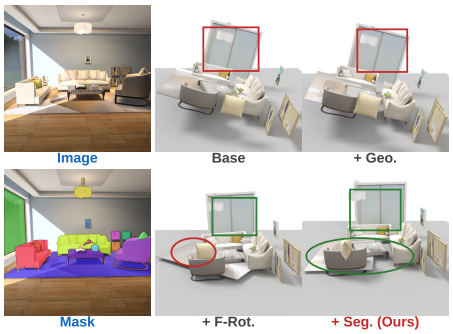

Figure 6: Ablation of the layout predictor.

Table 3: Ablation of the full framework.

	MIT-67		COCO	
Method	VLM ↑	CLIP ↑	VLM ↑	CLIP ↑
w/o mask candidates	6.2971	0.7711	6.8209	0.7132
w/o layout predictor	6.9794	0.7682	6.9867	0.7041
w/o refinement candidates	6.6529	0.7611	7.1933	0.7107
w/o reviewer	6.5824	0.7374	5.7567	0.6865
Ours (full)	7.3471	0.7649	7.4733	0.7124

Table 4: Ablation study of the layout predictor.

Components			ScanNet			SAGE		
Geo.	F-Rot.	Seg.	CD ↓	F-Score ↑	mIoU ↑	CD ↓	F-Score ↑	mIoU ↑
✗	✗	✗	0.0557	0.6689	0.3684	0.0073	0.9131	0.3809
✓	✗	✗	0.0530	0.6711	0.3811	0.0066	0.9225	0.3841
✗	✓	✗	0.0556	0.6472	0.3514	0.0090	0.8866	0.3195
✓	✓	✗	0.0519	0.6595	0.3658	0.0070	0.9174	0.3824
✓	✗	✓	0.0496	0.6965	0.3788	0.0062	0.9261	0.3973
✓	✓	✓	0.0489	0.6944	0.3917	0.0048	0.9511	0.4029

Fixer fuses neighboring objects into distorted meshes and SAM3D leaves the objects on inconsistent planes, whereas our predictor keeps them on a single ground plane with the arrangement of the photograph. This floor-aligned layout also provides the frame on which the room of Stage 2 is fitted and every correction of Stage 3 is expressed. On SAGE, our predictor ranks second after SAM3D. SAGE consists of synthetic scenes rendered from room-overview cameras, where each object appears clean and mostly unoccluded, thus SAM3D, which estimates the pose of each object independently from its own mask, benefits most from this setting.

4.2 Ablation Study & Discussion

Ablation Study for complete scene generation. To validate each component, we compare our full framework with four variants, each of which removes one component, and the results are shown in Table 3 and Figure 5. *w/o mask candidates* runs Stage 1 without the mask candidates and thus without mask refinement, *w/o layout predictor* initializes the layout with the per-object poses of SAM3D, *w/o refinement candidates* lets the agent write the Stage 3 groups and operations instead of ruling on the refinement candidates, and *w/o reviewer* applies every candidate of Stages 1 and 3 whose measurement passes its threshold. Every variant lowers the VLM score on both datasets, and each drop follows the role of the removed component. Without mask candidates, the score drops the most on MIT-67 (by 1.05 points). As noted in Sec. 3.1, every later stage builds on the masks, and indeed no later stage can repair fragmented or spurious objects, which also cause the most inter-penetrations (15.7% vs. 5.3% on MIT-67; Table S1 of the supplementary material). These spurious masks still add photographed appearance, which is why this variant gets the highest CLIP on both datasets, as CLIP rewards appearance regardless of the layout and realism. Without the reviewer, every candidate that passes its threshold is applied, and the score drops the most on COCO (by 1.72 points, with 59 of the 75 scenes worse). Since we propose the candidates generously (Sec. 3), applying all of them merges or deletes many masks that the reviewer would keep, and this variant keeps only about half of the objects of our full framework. This confirms our key idea: geometry narrows the decisions, but the reviewer is still needed to reject the wrong candidates. Without refinement candidates, the agent writes the Stage 3 groups and operations itself, and the score drops on both datasets. As argued in the introduction, an agent struggles with precise geometric reasoning, and since it has to guess each support without measurements, more objects are left floating (27.4% vs. 20.2% on MIT-67; Table S1). Finally, replacing the layout predictor with the per-object poses of SAM3D also lowers the score on both datasets, which agrees with Sec. 3.2 that the initial layout matters even after refinement.

Ablation Study for Layout predictor. To validate each component, we train the layout predictor with and without the geometry loss, the floor-rotation head, and the segmentation-level training data,

and the results are shown in Table 4. The geometry loss improves every metric over the base model on both benchmarks, which verifies that the point-map supervision makes the predictor accurate, as designed in Sec. 3.2, and adding segmentation-level real images on top of it further lowers CD on both benchmarks (ScanNet: 0.0530→0.0496). On top of the geometry loss, the floor-rotation head does not consistently improve accuracy (ScanNet F-Score: 0.6711→0.6595), but it changes how the layout is expressed. As shown in Figure 6, without it (Base, + Geo.), the objects inherit the camera tilt and lean (red boxes), whereas with it (+ F-Rot., + Seg.), they stand upright on one floor plane (green boxes). With all three components, our full predictor achieves the best CD and mIoU on both benchmarks while providing the floor-aligned frame on which the room of Stage 2 and every correction of Stage 3 are built. Training with segmentation-level real images further corrects the misplaced armchair (red and green circles).

Discussion with coding agents. Recent coding agents such as Claude Opus 5.5 (Anthropic, 2026) can generate an entire Blender scene from a single photograph. We therefore compare them with our method to explore their complementary strengths. As shown in Figure 7, when generating from scratch, the coding agent better reproduces the overall room appearance, colors, and lighting, but relies on generic primitives and does not faithfully recover object geometry. We further compare by using the coding agent to edit our results. It improves the layout, room appearance, and lighting, suggesting that the two approaches are complementary: our method provides image-grounded geometry, while coding agents offer stronger appearance editing. A promising direction is to combine these strengths while keeping agent edits grounded in the input photograph.



Figure 7: Comparison with Claude Opus 5.5. From left to right: input image, ours, ours edited by Opus, and the scene Opus writes from scratch.

Limitations and future work. Our framework targets indoor scenes enclosed by a floor and walls, since the room built in Stage 2 is the reference for every later correction, so outdoor or open scenes without walls are out of its scope. Moreover, it builds a single room with flat walls on one floor, so a scene that spans several rooms, *e.g.*, a view through a doorway into the next room, cannot be modeled. Since the agents only select among the candidates that geometry proposes, an edit that no candidate covers cannot be made, and an object missed in Stage 1 cannot be added later. Finally, a single photograph cannot show hidden objects or the back of the room, so these parts are left out or filled in plainly. In future work, we plan to extend the room model to multiple rooms and to open and multi-level spaces, to let agents request new candidates when none of the proposed ones fits, and to complete the unseen parts with generative priors that stay consistent with the photograph.

AI use statement. Generative AI is a component of the proposed framework and of its evaluation. All agents are instantiated from Qwen3.8-27B (Qwen Team, 2026b), Qwen3.5-9B (Qwen Team, 2026a) serves as the VLM judge of perceptual quality, and Qwen3.5-27B (Qwen Team, 2026a) filters the COCO training images of the layout predictor. Claude Opus 5.5 is the coding agent compared above and in Sec. B.1 of the supplementary material. In addition, we used large language models to assist with revising sentences and the agent prompts of our work.

5 Conclusion

We have shown that an agent-based framework builds a complete 3D scene from a single photograph more reliably when its agents are not asked to do everything. With selective agentic reasoning, geometry measures the scene and enumerates candidate edits, and a reviewer agent mainly accepts or rejects them against the photograph in both mask refinement and scene refinement. In between, a layout predictor supervised by point maps provides an accurate floor-aligned initial layout even for real-world images, and a room fitted to the observed walls serves as the reference for every later correction. As a result, our framework generates scenes more faithful to the photograph than prior methods, while ranking first or second on every physical metric. We believe that leaving agents only the decisions that geometry cannot settle is a useful principle for agent-based 3D scene generation.

REFERENCES

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as I can, not as I say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Anthropic. Claude opus 5.5, 2026. URL `https://www.anthropic.com/claude/opus`.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

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. *arXiv preprint arXiv:2212.08051*, 2022.

Andreea Dogaru, Mert Özer, and Bernhard Egger. Gen3dsr: Generalizable 3d scene reconstruction via divide and conquer from a single view. In *International Conference on 3D Vision*, 2025.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. In *Conference on Neural Information Processing Systems*, 2023.

Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129: 3313–3337, 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.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023.

Yun He, Kelin Yu, and Matthias Zwicker. Sceneorchestra: Efficient agentic 3d scene synthesis via full tool-call trajectory generation. *arXiv preprint arXiv:2604.19907*, 2026.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. In *Proceedings of the International Conference on Learning Representations*, 2024.

Ian Huang, Guandao Yang, and Leonidas Guibas. Blenderalchemy: Editing 3d graphics with vision-language models. In *European Conference on Computer Vision*, pp. 297–314. Springer, 2024.

Ian Huang, Yanan Bao, Karen Truong, Howard Zhou, Cordelia Schmid, Leonidas Guibas, and Alireza Fathi. FirePlace: Geometric refinements of LLM common sense reasoning for 3D object placement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2025a.

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

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

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. Position: LLMs can't plan, but can help planning in LLM-Modulo frameworks. In *Proceedings of the International Conference on Machine Learning*, pp. 22895–22907, 2024.

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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

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.

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*, pp. 10072–10083, 2024.

Xiaoxuan Ma, Jiashun Wang, Nicolás Ugrinovic, Yehonathan Litman, and Kris Kitani. Rest3d: Reconstructing physically stable 3d scenes from a single image. *arXiv preprint arXiv:2605.30338*, 2026.

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

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

Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, Annie Xie, Danny Driess, Ayzaan Wahid, Zhuo Xu, Quan Vuong, Tingnan Zhang, Tsang-Wei Edward Lee, Kuang-Huei Lee, Peng Xu, Sean Kirmani, Yuke Zhu, Andy Zeng, Karol Hausman, Nicolas Heess, Chelsea Finn, Sergey Levine, and Brian Ichter. PIVOT: Iterative visual prompting elicits actionable knowledge for VLMs. In *Proceedings of the International Conference on Machine Learning*, pp. 37321–37341, 2024.

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

Seongrae Noh, SeungWon Seo, Gyeong-Moon Park, and HyeongYeop Kang. Edit-as-act: Goal-regressive planning for open-vocabulary 3d indoor scene editing. *arXiv preprint arXiv:2603.17583*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Nicholas Pfaff, Thomas Cohn, Sergey Zakharov, Rick Cory, and Russ Tedrake. Scenesmith: Agentic generation of simulation-ready indoor scenes. *arXiv preprint arXiv:2602.09153*, 2026.

Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 413–420. IEEE, 2009.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026b. URL https://qwen.ai/blog?id=qwen3.8.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021.

Tobias Sautter, Jan-Niklas Dihlmann, and Hendrik Lensch. 3d-re-gen: 3d reconstruction of indoor scenes with a generative framework. *arXiv preprint arXiv:2512.17459*, 2025.

Yukai Shi, Weiyu Li, Zihao Wang, Hongyang Li, Xingyu Chen, Ping Tan, and Lei Zhang. Scenemaker: Open-set 3d scene generation with decoupled de-occlusion and pose estimation model. *arXiv preprint arXiv:2512.10957*, 2025.

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

Fan-Yun Sun, Weiyu Liu, Siyi Gu, Dylan Lim, Goutam Bhat, Federico Tombari, Manling Li, Nick Haber, and Jiajun Wu. Layoutvlm: Differentiable optimization of 3d layout via vision-language models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2025a.

Fan-Yun Sun, Shengguang Wu, Christian Jacobsen, Thomas Yim, Haoming Zou, Alex Zook, Shangru Li, Yu-Hsin Chou, Ethem Can, Xunlei Wu, Clemens Eppner, Valts Blukis, Jonathan Tremblay, Jiajun Wu, Stan Birchfield, and Nick Haber. 3d-generalist: Self-improving vision-language-action models for crafting 3d worlds, 2025b. URL https://arxiv.org/abs/2507.06484.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via Python execution for reasoning. In *Proceedings of the International Conference on Computer Vision*, pp. 11854–11864, 2023.

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

Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. *Advances in Neural Information Processing Systems*, 35:10021–10039, 2022.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2025a.

Ling Wang, Hao-Xiang Guo, Xinzhou Wang, Fuchun Sun, Kai Sun, Pengkun Liu, Hang Xiao, Zhong Wang, Guangyuan Fu, Eric Li, et al. Scenetransporter: Optimal transport-guided compositional latent diffusion for single-image structured 3d scene generation. *arXiv preprint arXiv:2602.22785*, 2026.

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

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

Hongchi Xia, Xuan Li, Zhaoshuo Li, Qianli Ma, Jiashu Xu, Ming-Yu Liu, Yin Cui, Tsung-Yi Lin, Wei-Chiu Ma, Shenlong Wang, Shuran Song, and Fangyin Wei. SAGE: Scalable agentic 3d scene generation for embodied ai. *arXiv preprint arXiv:2602.10116*, 2026.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2025.

Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2690–2698, 2019.

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.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-Mark prompting unleashes extraordinary visual grounding in GPT-4V. *arXiv preprint arXiv:2310.11441*, 2023.

Xianghui Yang, Huiwen Shi, Bowen Zhang, Fan Yang, Jiacheng Wang, Hongxu Zhao, Xinhai Liu, Xinzhou Wang, Qingxiang Lin, Jiaao Yu, et al. Hunyuan3d 1.0: A unified framework for text-to-3d and image-to-3d generation. *arXiv preprint arXiv:2411.02293*, 2024a.

Yandan Yang, Baoxiong Jia, Shujie Zhang, and Siyuan Huang. Sceneweaver: All-in-one 3d scene synthesis with an extensible and self-reflective agent, 2025. URL `https://arxiv.org/abs/2509.20414`.

Yixuan Yang, Zhen Luo, Wanshui Gan, Jinkun Hao, Junru Lu, Jinghao Yan, Zhaoyang Lyu, and Xudong Xu. Code-as-room: Generating 3d rooms from top-down view images via agentic code synthesis. *arXiv preprint arXiv:2605.18451*, 2026.

Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, et al. Holodeck: Language guided generation of 3d embodied ai environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024b.

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. In *ACM SIGGRAPH Conference*, 2025.

Shaofeng Yin, Jiaxin Ge, Zora Zhiruo Wang, Chenyang Wang, Xiuyu Li, Michael J Black, Trevor Darrell, Angjoo Kanazawa, and Haiwen Feng. Vision-as-inverse-graphics agent via interleaved multimodal reasoning. *arXiv preprint arXiv:2601.11109*, 2026a.

Ze-Xin Yin, Liu Liu, Xinjie Wang, Wei Sui, Zhizhong Su, Jian Yang, and Jin Xie. 3d-fixer: Coarse-to-fine in-place completion for 3d scenes from a single image. *arXiv preprint arXiv:2604.04406*, 2026b.

Huangyue Yu, Baoxiong Jia, Yixin Chen, Yandan Yang, Puhao Li, Rongpeng Su, Jiaxin Li, Qing Li, Wei Liang, Zhu Song-Chun, Tengyu Liu, and Siyuan Huang. Metascenes: Towards automated replica creation for real-world 3d scans. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

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. *arXiv preprint arXiv:2507.22825*, 2025.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5745–5753, 2019.