ALIGN3D: IMAGE-STRUCTURE CO-ALIGNED 3D SCENE GENERATION WITH OBSERVABILITY-AWARE FIDELITY-PRESERVING GUIDANCE

Anonymous authors
Paper under double-blind review

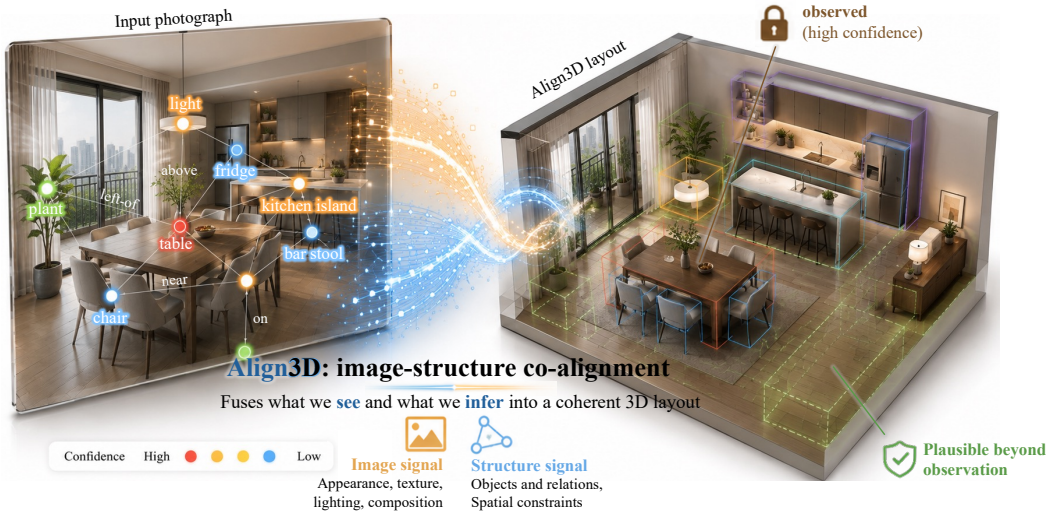

Figure 1: Align3D couples an image signal and a confidence-calibrated structure signal via image-structure co-alignment; Observability-aware fidelity-preserving guidance completes under-observed regions under scene-quality guidance while protecting well-supported layout variables, yielding a coherent 3D scene.

ABSTRACT

We present Align3D, an image-conditioned diffusion model that generates 3D indoor layouts that are both faithful to a single input photograph and physically plausible beyond the observed view. From the photograph we infer a confidence-calibrated scene graph and train a set-diffusion layout denoiser with image-structure co-alignment where confidence-weighted modality fusion balances image and scene-graph alignment using predicted confidences, while observability-aware fidelity-preserving guidance projects conflicting scene-quality gradients and reweights them with an object-level confidence mask so under-observed degrees of freedom absorb physical priors without undoing the image-graph match. On 3D-FRONT, Align3D improves image-alignment and scene-quality metrics over recent baselines.

1 INTRODUCTION

Generating consistent 3D indoor scenes from a single photograph is a fundamental problem in computer graphics and vision, with applications in embodied AI, virtual staging, and interactive design (Nie et al., 2020; Sun et al., 2018). Early methods recover multi-object geometry via shape retrieval or compositional reconstruction (Han et al., 2024; Ardelean et al., 2025), while diffusion models now synthesize entire layouts as unordered object sets (Tang et al., 2024; Yang et al., 2024b). Most layout generators condition on text, category priors, or pre-specified scene graphs (Zhai et al., 2023; Wei et al., 2024; Feng et al., 2026); image-to-scene pipelines (Huang et al., 2025; Meng et al.,

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.

2026; Lin et al., 2025) ground generation in RGB input but seldom enforce *joint* agreement between visual appearance and inferred spatial structure.

The central challenge is reconciling two complementary signals: *image appearance* (materials, lighting, and composition in the reference view) and *scene structure* (objects and relations that should remain coherent across viewpoints) (Johnson et al., 2018; Xu et al., 2017). Visual embeddings anchor layouts to the photograph (Radford et al., 2021), and scene graphs encode editable relational constraints (Zhai et al., 2023; Wei et al., 2024); prior work often optimizes the two sequentially or with fixed loss weights (Ardelean et al., 2025; Yang et al., 2024b), treating every detection and relation as equally reliable even though single-view perception is inherently noisy. However, neither signal alone guarantees a usable room: a layout can match the image from one camera yet still collide, float, or violate relations that were only weakly observed.

Single-view conditioning sharpens this tension: visible silhouettes pin some objects tightly, while depth ordering, occlusions, and hidden clearances remain underdetermined (Nie et al., 2020; Sun et al., 2018). Physics-aware generators improve plausibility (Yang et al., 2024b; Zhai et al., 2023; Wei et al., 2024), but typically combine quality terms with conditioning through a weighted sum that can harm image alignment when objectives conflict — a familiar issue in multi-task learning (Yu et al., 2020; Liu et al., 2021; Sener & Koltun, 2018). Our insight is that observational fidelity and scene quality should not compete along the same parameter directions: strongly observed variables should be protected, while under-constrained ones should absorb physical priors without harming the observation match. In addition, confidence provides the bookkeeping for this split: down-weighting uncertain detections, camera estimates, and relations during alignment, then routing quality updates away from well-supported layout variables.

To this end, we present Align3D: image-structure co-aligned 3D scene generation (Fig. 1), an image-conditioned layout diffusion model that predicts a confidence-calibrated scene graph from a photograph and trains with image-structure co-alignment under observability-aware fidelity-preserving guidance. Our contributions are:

- a state-of-the-art image-to-scene pipeline evaluated with competitive baselines on geometry, layout fidelity, and human preference metrics;
- a confidence-calibrated scene graph inference module from single-view images that attaches detection, depth, camera, and relational confidences to every graph node and edge, enabling supervision without ground-truth graphs; and
- end-to-end confidence propagation through *confidence-weighted modality fusion (CWMF)* and *observability-aware fidelity-preserving guidance (OFG)*: CWMF adaptively balances image-rendering and scene-graph alignment from predicted confidences, while OFG projects conflicting quality gradients and reweights them with an object-level confidence mask derived from the same predicted confidences.

2 Related Work

Scene graph representation and learning. Scene graphs encode objects and their relations as structured, editable constraints (Johnson et al., 2018; Xu et al., 2017). In 3D, early work predicts semantic scene graphs from indoor reconstructions (Wald et al., 2020), while other methods generate and manipulate scenes from given graphs (Dhamo et al., 2021). Subsequent graph-conditioned generators use relational priors for controllable layouts (Zhai et al., 2023; Wei et al., 2024) or couple multimodal graphs with rectified flow (Yang et al., 2026). Related text- or instruction-driven layout systems further show that explicit relational structure improves physical plausibility (Feng et al., 2026), but typically assume language or given objects rather than an observed photograph. We instead *predict* confidence-calibrated scene graphs from images and use them as image-structure co-alignment constraints alongside appearance matching.

3D scene generation from images. Recovering a multi-object 3D scene from a single RGB image has progressed from feed-forward reconstruction and shape retrieval (Nie et al., 2020; Sun et al., 2018) to compositional pipelines that segment, complete, and assemble per-object geometry with layout optimization (Ardelean et al., 2025; Han et al., 2024). Subsequent systems synthesize multi-object meshes or assets jointly from an image (Huang et al., 2025; Lin et al., 2025; Meng et al., 2026), while others use depth-guided arrangement (Zhao et al., 2025), hierarchical 2D-to-3D room

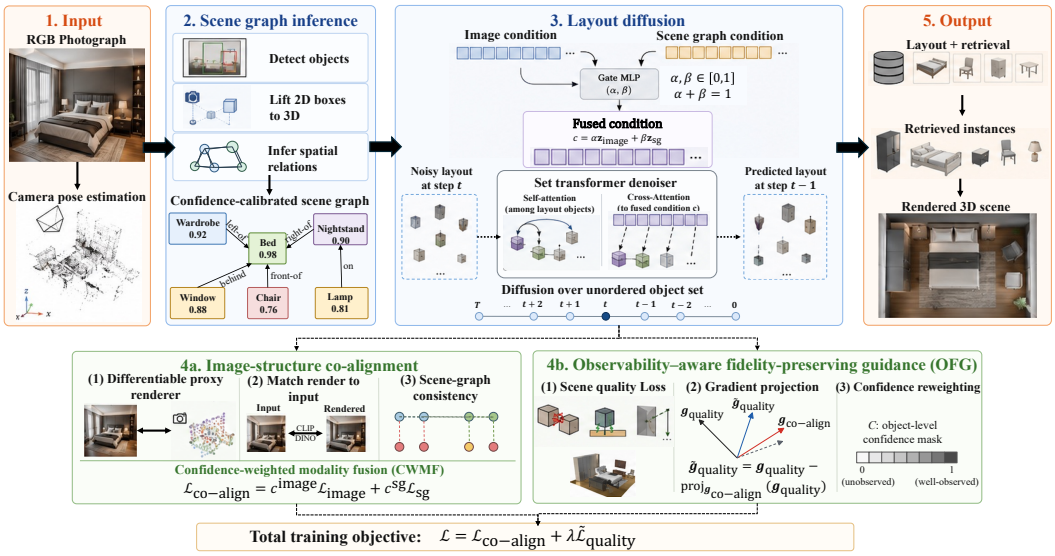

Figure 2: Align3D pipeline, including scene graph inference, fused image-graph conditioning, and set-diffusion layout prediction. Training uses image-structure co-alignment and observability-aware fidelity-preserving guidance, and inference retrieves shapes for rendering.

construction (Wang et al., 2024b), or component-aligned assembly (Yao et al., 2025). Our setting is similar to these baselines, but targets image-structure co-alignment under object-centric attributes.

3D diffusion models for scenes. Denoising diffusion models (Ho et al., 2020; Song et al., 2021) generate samples by gradually removing noise from a corrupted signal and have become a leading generative paradigm; transformer backbones further scale this process in latent space (Peebles & Xie, 2023). Earlier indoor layouts are synthesized autoregressively (Paschalidou et al., 2021) or with diffusion over object sets and relational graphs (Tang et al., 2024; Lin & Mu, 2024; Yang et al., 2024b; Bai et al., 2025), while other diffusion models generate room-scale geometry (Ju et al., 2024), image-conditioned multi-object meshes (Huang et al., 2025; Lin et al., 2025), or vision-guided layouts (Zhu et al., 2025). Most layout-oriented approaches still condition on text, floor plans, or given graphs rather than jointly co-aligning an observed photograph with predicted structure. We instead train an image-conditioned set-diffusion layout model with image-structure co-alignment and observability-aware fidelity-preserving guidance, so quality updates improve under-constrained degrees of freedom without undoing the observation match.

3 METHOD

Align3D generates image-conditioned indoor layouts through a set-diffusion denoiser over unordered object configurations (Fig. 1). Figure 2 summarizes the full system: an input photograph is converted to a confidence-calibrated scene graph, combined with image tokens via fused conditioning, and trained with image-structure co-alignment and observability-aware fidelity-preserving guidance; after training, the model can map a previously unseen photograph to a layout via the same graph-inference and denoising stack.

Training integrates three main modules: scene graph inference (Sec. 3.1), image-structure co-alignment (Sec. 3.2), and observability-aware fidelity-preserving guidance (Sec. 3.3), which folds a scene-quality loss into optimization under projection and object-level confidence weighting. The resulting objective is

$$\mathcal{L} = \mathcal{L}_{\text{co-align}} + \lambda \tilde{\mathcal{L}}_{\text{quality}}, \tag{1}$$

where $\mathcal{L}_{\text{co-align}}$ couples image and graph observations, $\mathcal{L}_{\text{quality}}$ penalizes inter-object overlap, and $\tilde{\mathcal{L}}_{\text{quality}}$ denotes its fidelity-preserving form with object-level confidence weighting in gradient space. We detail each module below. Hyperparameters and implementation details are given in App. A.

Method	CD↓	F-Score↑	IoU↑	CLIP↑	DINO↑	FID↓	HP-Align↑	HP-Scene↑
DiffuScene (Tang et al., 2024)	0.292	36.1	0.289	0.784	0.758	29.8	42.8	44.1
Gen3DSR (Ardelean et al., 2025)	0.058	59.6	0.286	0.808	0.468	52.1	46.2	47.8
PartCrafter (Lin et al., 2025)	0.194	42.0	0.236	0.814	0.552	33.8	46.8	48.1
MIDI (Huang et al., 2025)	0.076	52.4	0.328	0.866	0.698	23.6	51.5	50.8
SceneGen (Meng et al., 2026)	**0.016**	87.8	0.541	0.909	0.824	**20.3**	54.2	54.8
Align3D (ours)	0.019	**88.7**	**0.558**	**0.925**	**0.843**	20.8	**57.2**	**56.1**

Table 1: Main comparison on 3D-FRONT (Fu et al., 2021a). Align3D is best on image-alignment metrics (CLIP, DINO, HP-Align) and is best or second-best across scene-quality metrics.

3.1 Scene Graph Inference

Given a single RGB photograph I, we predict a confidence-calibrated scene graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that supervises image-structure co-alignment. We run open-vocabulary detection with GroundingDINO (Liu et al., 2023): each query names a furniture class from our label vocabulary, and the model returns a category y_i, a 2D bounding box $\mathbf{b}_i^{\text{2D}}$, a detection confidence $c_i^{\text{det}} \in [0,1]$, and an appearance embedding $\mathbf{z}_i^{\text{app}}$. A dense depth map is estimated with Depth Anything (Yang et al., 2024a); we read the estimated depth d_i at the pixel center of each box. Camera intrinsics K, extrinsics $(R, \mathbf{t}^{\text{cam}})$, and a global camera confidence $c^{\text{cam}} \in [0,1]$ are recovered with DUSt3R (Wang et al., 2024a). We cache the bundle $\mathcal{P} \coloneqq \{K, R, \mathbf{t}^{\text{cam}}, c^{\text{cam}}\}$ for 3D lifting and camera-view rendering; c^{cam} also sets the image-modality weight c^{image} in image-structure co-alignment (Eq. (5)).

Each detection is lifted to 3D by perspective back-projection. Let (u_i, v_i) denote the pixel center of $\mathbf{b}_i^{\text{2D}}$, with width w_i^{2D} and height h_i^{2D}; let $(\bar{u}, \bar{v})$ and (f_x, f_y) be the principal point and focal lengths from K, and d_i the Depth Anything depth at (u_i, v_i). We write $\mathbf{p}_i^{\text{cam}}$ for the lifted 3D point in the camera frame and $\mathbf{p}_i^{\text{world}}$ for its center in a shared world frame:

$$\mathbf{p}_i^{\text{cam}} \coloneqq d_i \begin{bmatrix} \frac{u_i - \bar{u}}{f_x} \\ \frac{v_i - \bar{v}}{f_y} \\ 1 \end{bmatrix}, \qquad \mathbf{p}_i^{\text{world}} \coloneqq R^{\mathrm{T}}(\mathbf{p}_i^{\text{cam}} - \mathbf{t}^{\text{cam}}), \tag{2}$$

where $(\cdot)^{\mathrm{T}}$ denotes the transpose. We estimate coarse 3D extents $\mathbf{w}_i = (w_{i,x}, w_{i,y}, w_{i,z})^{\top}$ by back-projecting the 2D box ($w_{i,x} = \frac{w_i^{\text{2D}} d_i}{f_x}$, $w_{i,y} = \frac{h_i^{\text{2D}} d_i}{f_y}$) and setting $w_{i,z} \approx \eta\, w_{i,y}$, where η is a fixed depth-extent ratio because monocular box height does not constrain object thickness along the viewing axis. Node confidence aggregates the detection and depth terms with a fixed calibration,

$$c_i^{\text{node}} \coloneqq \kappa\, c_i^{\text{det}}\, c_i^{\text{depth}}, \qquad c_i^{\text{depth}} \coloneqq \max\left(0,\, 1 - \frac{d_i}{d_{\max}}\right), \tag{3}$$

where c_i^{depth} is a depth confidence that decays with distance, $\kappa \in (0,1]$ conservatively reflects residual uncertainty in monocular depth-based 3D lifting, and $d_{\max}$ sets the depth at which c_i^{depth} reaches zero. Uncertain detections thus contribute less to later fusion and OFG. Spatial relations are inferred from the offset between world-frame centers: letting $\Delta\mathbf{p}_{ij}^{\text{world}} \coloneqq \mathbf{p}_j^{\text{world}} - \mathbf{p}_i^{\text{world}}$, dominant horizontal offsets yield left/right, depth offsets yield front/behind, vertical gaps yield above/on, and nearby contacts yield touching, with proximity radius $r_{\text{close}} = \frac{\|\mathbf{w}_i\| + \|\mathbf{w}_j\|}{2} + \varepsilon_{\text{close}}$. Edge confidence decays with inter-node distance and is limited by the weaker endpoint,

$$c_{ij}^{\text{edge}} \coloneqq \frac{\min(c_i^{\text{node}}, c_j^{\text{node}})}{1 + \frac{\|\Delta\mathbf{p}_{ij}^{\text{world}}\|}{r_{\text{close}}}}. \tag{4}$$

The resulting graph supervises image-structure co-alignment through node and edge confidences c_i^{node} and c_{ij}^{edge}, and node-level signals additionally define OFG's mask C via observability in Sec. 3.3.

3.2 Image-Structure Co-Alignment

image-structure co-alignment trains the denoiser so that generated layouts agree with both the predicted graph from Sec. 3.1 and the input photograph. The module first summarizes per-modality

reliability, then forms a fused conditioning from image and graph tokens, and finally defines the alignment losses that supervise the predicted layout.

Modality confidences. Camera and graph observations can carry unequal reliability across detections. We aggregate this into scalars

$$c^{\text{image}} := c^{\text{cam}}, \qquad c^{\text{sg}} := \frac{1}{|\mathcal{V}| + |\mathcal{E}|} \left(\sum_{i \in \mathcal{V}} c_i^{\text{node}} + \sum_{(i,j) \in \mathcal{E}} c_{ij}^{\text{edge}} \right), \tag{5}$$

where c^{cam}, c_i^{node}, and c_{ij}^{edge} are defined in Sec. 3.1. Both c^{image} and c^{sg} lie in $[0, 1]$ and are reused below for both token fusion and loss weighting.

Fused conditioning. Following prior set-diffusion layout models (Tang et al., 2024), each object is represented by semantic class, oriented bounding box, and a retrievable shape code. The set-diffusion denoiser is dual-conditioned on the input photograph and the predicted scene graph $\mathcal{G}$. Let $\mathbf{z}_{\text{img}}$ denote image tokens obtained by projecting CLIP features (Radford et al., 2021; Zhou et al., 2023) into the denoiser width and $\mathbf{z}_{\text{sg}}$ denote graph tokens (node and relation embeddings of $\mathcal{G}$). A small MLP maps the pooled tokens and modality confidences $(c^{\text{image}}, c^{\text{sg}})$ to fusion weights $\alpha^{\text{image}}, \alpha^{\text{sg}} \in (0, 1)$, which form a fused conditioning

$$\mathbf{h} = \alpha^{\text{image}} \, \mathbf{z}_{\text{img}} + \alpha^{\text{sg}} \, \mathbf{z}_{\text{sg}}. \tag{6}$$

We add a multimodal cross-attention path onto $\mathbf{h}$, so appearance and structure are fused before the alignment losses below are applied. Because the MLP conditions on $(c^{\text{image}}, c^{\text{sg}})$, low graph confidence tends to reduce α^{sg} and low camera confidence tends to reduce α^{image}.

Alignment losses. Generated objects need not be in one-to-one correspondence with predicted graph nodes, so we match them with a soft assignment on node features $\mathbf{f}$ (class embedding concatenated with a linear map of the box). Letting $\mathbf{f}_i^{\text{obs}}$ and $\mathbf{f}_k^{\text{gen}}$ denote features of observed node i and generated object k, the soft matching matrix M is

$$M_{ik} = \mathrm{softmax}_k \left(\frac{\cos(\mathbf{f}_i^{\text{obs}}, \mathbf{f}_k^{\text{gen}})}{\tau} \right), \tag{7}$$

where $\mathrm{softmax}_k$ is the softmax over generated index k, $\tau > 0$ is a temperature that sharpens the assignment, and M_{ik} is the probability of matching i to k. Let $\mathbf{t}_k$ and $\mathbf{s}_k$ denote the translation and size of generated object k. The scene-graph loss then averages differentiable soft relation predicates over observed edges $(i, r, j) \in \mathcal{E}$, where r denotes the relation type (left, right, front, behind, above, etc.). Transporting each predicate onto matched generated pairs gives

$$\mathcal{L}_{\text{sg}} = \frac{1}{|\mathcal{E}|} \sum_{(i,r,j) \in \mathcal{E}} \sum_{k,m} M_{ik} \, M_{jm} \, \ell_r(\mathbf{t}_k, \mathbf{t}_m, \mathbf{s}_k, \mathbf{s}_m), \tag{8}$$

where ℓ_r is a soft hinge for relation r (definitions in App. A). Here M lets gradients flow through uncertain matches rather than forcing a hard permutation.

For appearance we do not back-propagate through the full mesh rasterizer. To keep training efficient we render the current layout as a differentiable 2D Gaussian-splat proxy image I_{proxy} by projecting each generated center into the image under the camera bundle $\mathcal{P}$. Writing $\mathbf{p}_k^{\text{cam}} := R\mathbf{t}_k + \mathbf{t}^{\text{cam}}$ and $z_k := (\mathbf{p}_k^{\text{cam}})_z$, the projected center and image-plane scales are

$$\mathbf{u}_k = \begin{bmatrix} \bar{u} + f_x \, (\mathbf{p}_k^{\text{cam}})_x / z_k \\ \bar{v} + f_y \, (\mathbf{p}_k^{\text{cam}})_y / z_k \end{bmatrix}, \qquad \boldsymbol{\sigma}_k = \begin{bmatrix} f_x \, (\mathbf{s}_k)_x / z_k \\ f_y \, (\mathbf{s}_k)_y / z_k \end{bmatrix}, \tag{9}$$

and the proxy has pixel values

$$I_{\text{proxy}}(u, v) = \sum_k \mathbf{a}_k \exp \left(- \frac{(u - (\mathbf{u}_k)_x)^2}{2 \, ((\boldsymbol{\sigma}_k)_x)^2} - \frac{(v - (\mathbf{u}_k)_y)^2}{2 \, ((\boldsymbol{\sigma}_k)_y)^2} \right), \tag{10}$$

where $\mathbf{a}_k$ is a learnable class color (RGB). Let E_{CLIP} and E_{DINO} denote the image encoders of CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2024). The image alignment loss matches the proxy to the input photograph in both embedding spaces,

$$\mathcal{L}_{\text{image}} = 1 - \cos \left(E_{\text{CLIP}}(I_{\text{proxy}}), E_{\text{CLIP}}(I) \right) + \lambda_{\text{dino}} \left(1 - \cos \left(E_{\text{DINO}}(I_{\text{proxy}}), E_{\text{DINO}}(I) \right) \right), \tag{11}$$

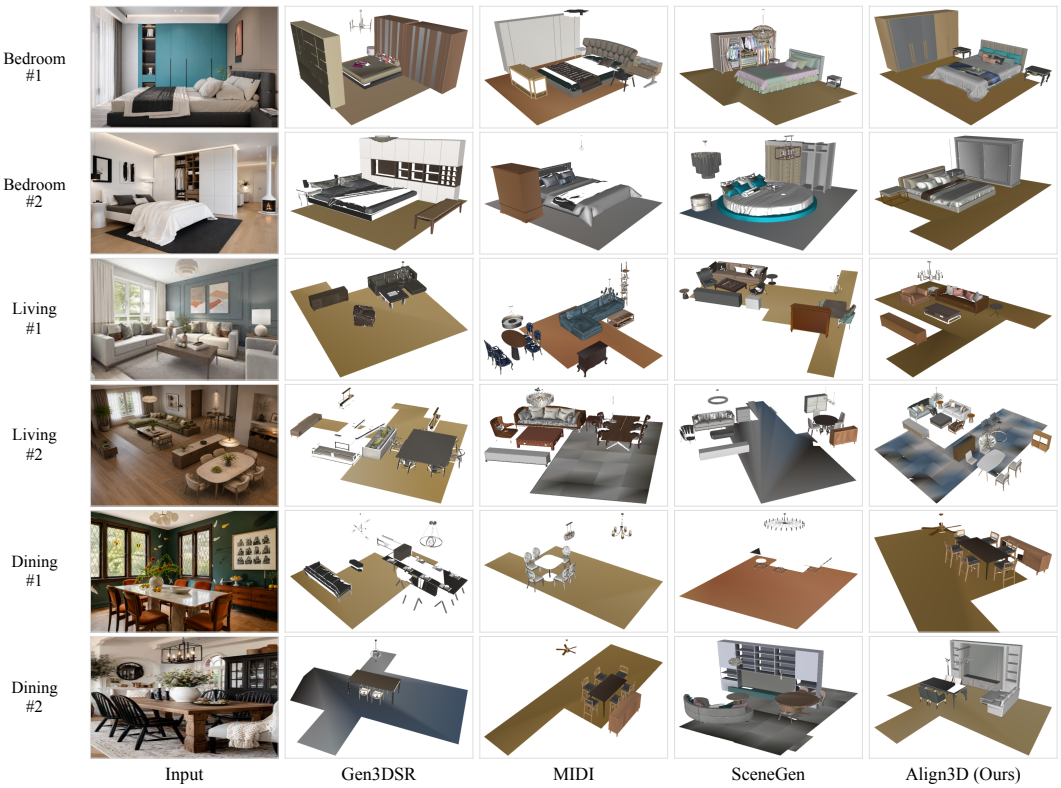

Figure 3: Qualitative image-to-scene comparison on real-world indoor scenes. Each row shows the input photograph and matched-view renderings from different methods. Align3D better matches the observed furniture arrangement and reduces collisions and floating objects relative to the baselines.

where $\lambda_{\text{dino}} > 0$ balances the two embedding terms.

Finally, we propose the *confidence-weighted modality fusion (CWMF)* to combine the two observational terms with the modality confidences from Eq. (5):

$$\mathcal{L}_{\text{co-align}} = c^{\text{image}}\,\mathcal{L}_{\text{image}} + c^{\text{sg}}\,\mathcal{L}_{\text{sg}}. \tag{12}$$

When one modality is unreliable, CWMF automatically down-weights its loss relative to the other.

3.3 Observability-Aware Fidelity-Preserving Guidance

Image-structure co-alignment establishes observational fidelity but does not by itself resolve residual physical ambiguity outside the observed silhouette. Therefore, inspired by DiffuScene (Tang et al., 2024), we introduce a scene-quality loss that penalizes pairwise oriented-box interpenetration,

$$\mathcal{L}_{\text{quality}} = \sum_{i<j} \text{IoU}(i,j), \tag{13}$$

where $\text{IoU}(i,j)$ is the volumetric intersection-over-union of the oriented boxes of objects i and j (Paschalidou et al., 2021).

Nevertheless, simply adding $\mathcal{L}_{\text{quality}}$ to $\mathcal{L}_{\text{co-align}}$ can undo the image-graph match: when the two gradients conflict, the quality term pulls the layout toward a generic collision-free arrangement rather than the observed scene. We therefore introduce *observability-aware fidelity-preserving guidance (OFG)*, which folds $\mathcal{L}_{\text{quality}}$ into training under fidelity-preserving, confidence-aware updates.

Let $\mathbf{g}_{\text{co-align}} := \nabla\mathcal{L}_{\text{co-align}}$ and $\mathbf{g}_{\text{quality}} := \nabla\mathcal{L}_{\text{quality}}$. Adapting gradient projection from multi-task learning (Yu et al., 2020; Liu et al., 2021; Sener & Koltun, 2018), we treat image-structure co-

Variant	CD↓	F-Score↑	IoU↑	CLIP↑	DINO↑	FID↓
Quality loss: $\lambda\mathcal{L}_{\text{quality}}$	0.286	36.8	0.312	0.786	0.754	29.1
Image alignment: $\mathcal{L}_{\text{image}} + \lambda\mathcal{L}_{\text{quality}}$	0.068	76.4	0.438	0.896	0.816	23.9
SG alignment: $\mathcal{L}_{\text{sg}} + \lambda\mathcal{L}_{\text{quality}}$	0.044	81.2	0.516	0.838	0.782	25.4
Co-alignment: $\mathcal{L}_{\text{image}} + \mathcal{L}_{\text{sg}} + \lambda\mathcal{L}_{\text{quality}}$	<u>0.027</u>	<u>86.4</u>	<u>0.542</u>	<u>0.898</u>	<u>0.818</u>	<u>22.1</u>
CWMF: $\mathcal{L}_{\text{co-align}} + \lambda\mathcal{L}_{\text{quality}}$	**0.025**	**86.8**	**0.554**	**0.901**	**0.822**	**21.2**

Table 2: Image-structure co-alignment ablation. Quality loss alone cannot ground layouts in the input; image and scene-graph terms are each necessary for complementary visual and structural gains, and confidence-weighted CWMF introduces further improvements over fixed-weight fusion.

alignment as primary and remove the conflicting component of the quality gradient:

$$\tilde{\mathbf{g}}_{\text{quality}} = \begin{cases} \mathbf{g}_{\text{quality}} - \dfrac{\mathbf{g}_{\text{quality}}^{\mathrm{T}}\mathbf{g}_{\text{co-align}}}{\|\mathbf{g}_{\text{co-align}}\|^2 + \epsilon}\,\mathbf{g}_{\text{co-align}}, & \mathbf{g}_{\text{quality}}^{\mathrm{T}}\mathbf{g}_{\text{co-align}} < -\xi, \\ \mathbf{g}_{\text{quality}}, & \text{otherwise}, \end{cases} \tag{14}$$

and write $\tilde{\mathcal{L}}_{\text{quality}}$ for the quality term under this projected gradient. Here $\xi \geq 0$ is a conflict margin and $\epsilon > 0$ is a numerical stabilizer. By construction, under conflict $\mathbf{g}_{\text{co-align}}^{\mathrm{T}}\tilde{\mathbf{g}}_{\text{quality}} \approx 0$, so a step along $\mathbf{g}_{\text{co-align}} + \lambda\tilde{\mathbf{g}}_{\text{quality}}$ does not increase $\mathcal{L}_{\text{co-align}}$ to first order, while quality still improves in directions orthogonal to alignment.

Moreover, a single view constrains objects unequally: well-observed silhouettes are tightly pinned, while occluded depth and hidden clearances remain free. We summarize this for each observed graph node by an *observability* score $o_i \in [0, 1]$,

$$o_i \coloneqq \text{sigmoid}\left(\alpha^{\text{vis}}\,(2\nu_i - 1) + \alpha^{\text{node}}\left(2c_i^{\text{node}} - 1\right)\right), \tag{15}$$

where $\text{sigmoid}(x) \coloneqq (1 + e^{-x})^{-1}$, $\nu_i \in [0, 1]$ is a visibility score (details in App. A), c_i^{node} is the node confidence from Sec. 3.1, and α^{vis}, α^{node} are fixed observability coefficients. We then transport observability through the matching matrix M (Eq. (7)) to obtain the observability score for each generated object:

$$o_k^{\text{gen}} \coloneqq \sum_i M_{ik}\, o_i, \tag{16}$$

and partition layout parameters into N generated-object blocks with $C \coloneqq \text{diag}(1 - o_1^{\text{gen}}, \ldots, 1 - o_N^{\text{gen}})$ to scale each block of the projected quality gradient, concentrating quality updates on less observable objects:

$$\mathbf{g} = \mathbf{g}_{\text{co-align}} + \lambda\, C\, \tilde{\mathbf{g}}_{\text{quality}}, \tag{17}$$

which implements Eq. (1) with object-level confidence weighting in gradient space. Intuitively, generation freedom is proportional to residual unobservability: high-o_k^{gen} objects should barely move under quality updates, while low-o_k^{gen} ones should absorb the physical prior.

4 Experiments

4.1 Dataset and Baselines

We evaluate on 3D-FRONT (Fu et al., 2021a), a large-scale dataset of synthetic indoor scenes with object layouts and room geometry Following ATISS (Paschalidou et al., 2021) and DiffuScene (Tang et al., 2024), we use three room types—4,041 bedrooms, 900 dining rooms, and 813 living rooms—with an 80/20 train/test split per type. Conditioning inputs are single rendered RGB views of each room; for the quantitative protocol we use the corresponding rendering cameras when available.

We compare Align3D against five recent baselines under a unified protocol: DiffuScene (Tang et al., 2024), Gen3DSR (Ardelean et al., 2025), PartCrafter (Lin et al., 2025), MIDI (Huang et al., 2025), and SceneGen (Meng et al., 2026). For DiffuScene, we use an image-conditioned adaptation of the official denoiser for fair single-image comparison (details in App. A). Baselines use official code and default settings when available; all methods are evaluated on the same conditioning viewpoint.

Variant	CD↓	F-Score↑	IoU↑	CLIP↑	DINO↑	FID↓
Sum: $\mathbf{g}_{\text{co-align}} + \lambda \mathbf{g}_{\text{quality}}$	0.025	86.8	0.554	0.901	0.822	21.2
Projection: $\mathbf{g}_{\text{co-align}} + \lambda \tilde{\mathbf{g}}_{\text{quality}}$	<u>0.020</u>	<u>88.1</u>	<u>0.555</u>	<u>0.921</u>	<u>0.840</u>	21.0
Reversed projection: $\tilde{\mathbf{g}}_{\text{co-align}} + \lambda \mathbf{g}_{\text{quality}}$	0.031	84.6	**0.571**	0.866	0.798	**20.5**
OFG: $\mathbf{g}_{\text{co-align}} + \lambda\, C\tilde{\mathbf{g}}_{\text{quality}}$	**0.019**	**88.7**	<u>0.558</u>	**0.925**	**0.843**	<u>20.8</u>

Table 3: OFG ablation. Unconstrained summation hurts alignment; projection is necessary to avoid gradient conflict, and confidence mask C introduces further improvements.

4.2 METRICS

We report two families of automatic metrics: scene quality and image alignment. Scene-quality metrics are CD↓ (Chamfer distance between point sets sampled from generated and ground-truth scene surfaces) (Fan et al., 2017), F-Score↑ (harmonic mean of precision and recall at a fixed Euclidean distance threshold) (Tatarchenko et al., 2019), IoU↑ (matched object-box intersection-over-union vs. ground truth) (Paschalidou et al., 2021), and FID↓ (Fréchet distance between inception features of matched-view generated renderings and ground-truth test-set renderings) (Heusel et al., 2017). Image-alignment metrics are CLIP↑ and DINO↑ (cosine similarity between global CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2024) embeddings of the conditioning image and a matched-view rendering). In addition, we introduce survey-based metrics, defined next. Unless noted otherwise, the best result in each table is set in **bold** and the second best is <u>underlined</u>.

Human preference study. Automatic metrics do not fully capture perceived image match or placement plausibility, so we run a pairwise preference study. Each trial shows the conditioning image and two anonymized matched-view renderings from two methods drawn uniformly at random from DiffuScene, Gen3DSR, PartCrafter, MIDI, SceneGen, and our Align3D; the participant answers two questions: (i) which layout is closer to the input image, and (ii) which placement is more physically reasonable. We recruit 40 participants; each answers 20 trials (40 binary choices), yielding 800 comparisons per criterion. Order of methods and left/right placement are randomized, and ties are disallowed. We then fit a Bradley-Terry model (Bradley & Terry, 1952) independently for each criterion: each method m receives a latent strength π_m, and the model predicts that m is chosen over m' with probability $P(m \succ m') = \frac{\pi_m}{\pi_m + \pi_{m'}}$. After maximum-likelihood fitting on all pairwise choices, we convert each π_m to a report score on a 0–100 scale where the cohort average maps to 50. We denote the resulting scores for questions (i) and (ii) by HP-Align↑ (human preference for image alignment) and HP-Scene↑ (human preference for scene quality), respectively. These scores indicate how often participants preferred a method relative to the pool average (e.g., 57 means modestly above-average preference).

4.3 RESULTS

Table 1 reports the main quantitative comparison. Align3D is best on image-alignment metrics (CLIP, DINO, HP-Align) and is best or second-best across scene-quality metrics. Figure 3 shows six real-world indoor scenes (two per room type) on held-out photographs (more in App. B). Matched-view renderings use 3D-FUTURE (Fu et al., 2021b) assets, so we advise comparing layout and placement rather than surface color. Across room types, Align3D better recovers the observed arrangement with fewer artifacts such as collisions and floating furniture than the baselines.

4.4 ABLATION STUDIES

Co-alignment components. We study how different alignment terms interact with the scene-quality loss and the impact of confidence weighting. In Tab. 2, each row lists the variant name followed by the training objective it uses. Training with $\mathcal{L}_{\text{quality}}$ alone ("Quality loss") yields weak image-scene agreement (CLIP 0.786, DINO 0.754), confirming that a generic layout prior is insufficient for image-conditioned generation. Adding image alignment ("Image alignment") is essential for visual fidelity: CLIP and DINO increase to 0.896 and 0.816 respectively, but IoU remains low (0.438) because appearance alone does not constrain relational layout. Conversely, adding scene-graph alignment ("SG alignment") is essential for structure: IoU rises to 0.516 and CD drops to 0.044, while CLIP lags behind the image-only variant. Combining both terms ("Co-alignment") is

Method	Training (h)	Inference (s/scene)
DiffuScene (Tang et al., 2024)	120	<u>14</u>
Gen3DSR (Ardelean et al., 2025)	**48**	250
PartCrafter (Lin et al., 2025)	380	34
MIDI (Huang et al., 2025)	120	42
SceneGen (Meng et al., 2026)	600	125
Align3D (ours)	<u>60</u>	**10**

Table 4: Training and inference cost. Our method is fast while producing high quality results.

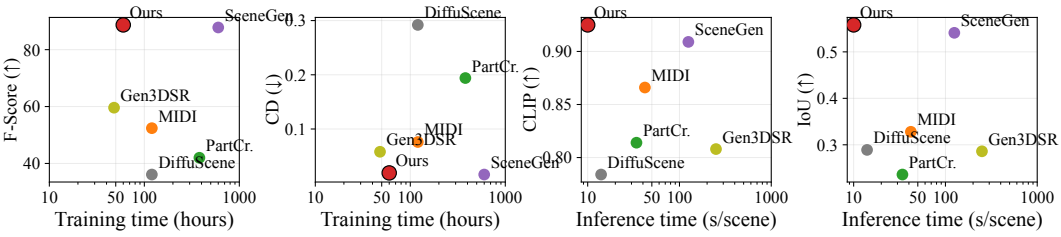

Figure 4: Quality vs. efficiency plots. Our method achieves both high quality and fast runtime.

required to capture both effects, yet fixed equal weighting still underperforms CWMF on every metric, showing that propagating predicted confidences into the alignment objective is necessary when graphs or camera estimates are noisy.

OFG components. We now investigate different ways to combine $\mathcal{L}_{\text{quality}}$ with $\mathcal{L}_{\text{co-align}}$. In Tab. 3, each row lists the variant name followed by the gradient update it uses: "Sum" adds unconstrained quality gradients; "Projection" replaces them by the fidelity-preserving $\tilde{\mathbf{g}}_{\text{quality}}$; "Reversed projection" instead protects quality and projects image-structure co-alignment away from it; "OFG" applies the confidence mask C to the projected quality gradient. Naively summing both losses ("Sum") degrades image alignment relative to projection (CLIP 0.901 vs. 0.921), indicating that gradient projection is necessary whenever quality and image-structure co-alignment objectives conflict. Reversed projection confirms the design choice: protecting $\mathcal{L}_{\text{quality}}$ rather than co-alignment yields higher IoU (0.571) at the cost of substantially worse alignment metrics, i.e. sacrificing observation fidelity for generic plausibility. Full OFG is necessary for the best overall tradeoff—it matches or exceeds projection on alignment metrics while delivering the strongest CD and F-Score, because C routes quality gradients to under-observed objects instead of overwriting image-graph evidence on well-constrained instances.

4.5 Efficiency Analysis

Beyond accuracy, practical systems must also be efficient. In Tab. 4, we compare training and per-scene inference times on a single NVIDIA RTX 4090. At inference, Align3D runs graph prediction ($\sim$3 s), 50-step layout diffusion ($\sim$5 s), and scene composition ($\sim$2 s), totaling $\sim$10 s per scene, the fastest among compared methods. Figure 4 also plots the metrics against training or inference time. Align3D occupies the fast regime on both axes while remaining competitive or best on scene-quality and image-alignment metrics, illustrating a favorable quality-efficiency tradeoff for layout-centric image-to-scene generation.

5 Conclusion

We presented Align3D, which trains an image-conditioned layout generator with image-structure co-alignment and OFG so that generated rooms both explain the observation and remain physically plausible. The training objective protects image-graph evidence while letting scene-quality updates act in under-observed degrees of freedom. Limitations remain: we evaluate only single-image conditioning (App. C sketches a preliminary multi-view extension), proxy rendering does not take colors into account, and errors in predicted scene graphs can propagate into image-structure co-alignment. We leave these limitations for future work.

AI USE STATEMENT

In this work, we used large language models to polish the writing of the paper and to help debug the implementation. We have reviewed all AI-assisted text and code changes and take full responsibility for the final content.

ETHICS STATEMENT

This work studies indoor scene generation on publicly available datasets. The human preference study collects anonymous pairwise judgments about generated layouts and does not collect personally identifying information. Generated layouts are intended for research on graphics and embodied-AI simulation; we discourage misuse for fabricating misleading real-estate or surveillance content.

REPRODUCIBILITY STATEMENT

Code is provided in the supplementary materials. Implementation details and hyperparameters are given in App. A. All quantitative protocols and ablations are described in Sec. 4.

REFERENCES

Andreea Ardelean, Mert Özer, and Bernhard Egger. Gen3DSR: Generalizable 3D scene reconstruction via divide and conquer from a single view. In *International Conference on 3D Vision (3DV)*, 2025.

Tongyuan Bai, Wangyuanfan Bai, Dong Chen, Tieru Wu, Manyi Li, and Rui Ma. FreeScene: Mixed graph diffusion for 3d scene synthesis from free prompts. In *CVPR*, 2025.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Helisa Dhamo, Fabian Manhardt, Nassir Navab, and Federico Tombari. Graph-to-3D: End-to-end generation and manipulation of 3D scenes using scene graphs. In *ICCV*, 2021.

Hao Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3D object reconstruction from a single image. In *CVPR*, 2017.

Haoran Feng, Yifan Niu, Zehuan Huang, Yang-Tian Sun, Chunchao Guo, Yuxin Peng, and Lu Sheng. Repurposing 3d generative model for autoregressive layout generation. In *CVPR*, 2026.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cheng Li, Qixun Zeng, Xun Sun, Rongfei Jia, Binqiang Zhao, et al. 3D-FRONT: 3d furnished rooms with layouts and semantics. In *ICCV*, 2021a.

Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Stephen Maybank, and Dacheng Tao. 3D-FUTURE: 3D furniture shape with textURE. *International Journal of Computer Vision*, 129:3313–3337, 2021b.

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.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

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 *CVPR*, 2025.

Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *CVPR*, 2018.

Xiaoliang Ju, Zhaoyang Huang, Yijin Li, Guofeng Zhang, Yu Qiao, and Hongsheng Li. DiffInD-Scene: Diffusion-based high-quality 3d indoor scene generation. In *CVPR*, 2024.

Chenguo Lin and Yadong Mu. InstructScene: Instruction-driven 3D indoor scene synthesis with semantic graph prior. In *ICLR*, 2024.

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. In *NeurIPS*, 2025.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *NeurIPS*, 2021.

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.

Yanxu Meng, Haoning Wu, Ya Zhang, and Weidi Xie. SceneGen: Single-image 3d scene generation in one feedforward pass. In *International Conference on 3D Vision (3DV)*, 2026.

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 *CVPR*, 2020.

Maxime Oquab, Timothée Darcet, Theo 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.

Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. ATISS: Autoregressive transformers for indoor scene synthesis. In *NeurIPS*, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

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 *ICML*, 2021.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *NeurIPS*, 2018.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3D: Dataset and methods for single-image 3d shape modeling. In *CVPR*, 2018.

Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. Diffuscene: Denoising diffusion models for generative indoor scene synthesis. In *CVPR*, 2024.

Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3D reconstruction networks learn? In *CVPR*, 2019.

Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3D semantic scene graphs from 3D indoor reconstructions. In *CVPR*, 2020.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3d vision made easy. In *CVPR*, 2024a.

Yian Wang, Xiaowen Qiu, Jiageng Liu, Zhehuan Chen, Jiting Cai, Yufei Wang, Tsun-Hsuan Wang, Zhou Xian, and Chuang Gan. Architect: Generating vivid and interactive 3d scenes with hierarchical 2d inpainting. In *NeurIPS*, 2024b.

Yao Wei, Martin Renqiang Min, George Vosselman, Li Erran Li, and Michael Ying Yang. Planner3D: LLM-enhanced graph prior meets 3d indoor scene explicit regularization. *arXiv preprint arXiv:2403.12848*, 2024.

Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In *NeurIPS*, 2024a.

Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. PhyScene: Physically interactable 3d scene synthesis for embodied AI. In *CVPR*, 2024b.

Zhifei Yang, Guangyao Zhai, Keyang Lu, YuYang Yin, Chao Zhang, Zhen Xiao, Jieyi Long, Nassir Navab, and Yikai Wang. FlowScene: Style-consistent indoor scene generation with multimodal graph rectified flow. *arXiv preprint arXiv:2603.19598*, 2026.

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*, 2025.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *NeurIPS*, 2020.

Guangyao Zhai, Evin Pınar Örnek, Shun-Cheng Wu, Yan Di, Federico Tombari, Nassir Navab, and Benjamin Busam. CommonScenes: Generating commonsense 3d indoor scenes with scene graph diffusion. In *NeurIPS*, 2023.

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 *ICCV*, 2025.

Chenliang Zhou, Fangcheng Zhong, and Cengiz Öztireli. CLIP-PAE: Projection-augmentation embedding to extract relevant features for a disentangled, interpretable and controllable text-guided face manipulation. In *ACM SIGGRAPH Conference Proceedings*, 2023.

Xiaoming Zhu, Xu Huang, Qinghongbing Xie, Zhi Deng, Junsheng Yu, Yirui Guan, Zhongyuan Liu, Lin Zhu, Qijun Zhao, Ligang Liu, and Long Zeng. Imaginarium: Vision-guided high-quality 3d scene layout generation. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 2025.

A Implementation Details

Layout denoiser. We instantiate the set-diffusion layout denoiser following DiffuScene (Tang et al., 2024): an image-conditioned denoiser over unordered object tuples (semantics, oriented boxes, shape codes) with shape retrieval from a shape library (3D-FUTURE (Fu et al., 2021b) in our experiments). Training uses Adam with learning rate 10^{-3}, batch size 128, and 60,000 iterations; the diffusion process uses a linear β schedule from 10^{-4} to 2×10^{-2} over $T{=}1000$ steps, with 50-step DDIM sampling at inference. For the DiffuScene baseline, the official unconditional layout denoiser is adapted by encoding the conditioning RGB with CLIP (Radford et al., 2021) and injecting the resulting image embedding as an additional condition into the U-Net (cross-attention over the layout tokens), then training on the same 3D-FRONT splits with matched rendered views while keeping the original network width, diffusion schedule, and retrieval head otherwise unchanged.

Scene graph prediction. Object detection uses GroundingDINO (Liu et al., 2023) with the 23-class vocabulary from 3D-FRONT. Depth estimation uses Depth Anything (Yang et al., 2024a) (base), and camera pose is estimated with DUSt3R (Wang et al., 2024a). For the 3D-FRONT evaluation protocol (Sec. 4), conditioning images are rendered views and we optionally substitute known rendering cameras for K and $(R, \mathbf{t}^{\text{cam}})$ with $c^{\text{cam}}{=}1$. Scene-graph hyperparameters in Sec. 3.1 are $\kappa{=}0.8$, $\eta{=}0.5$, $d_{\max}{=}5\,\text{m}$, and $\varepsilon_{\text{close}}{=}0.2\,\text{m}$. At export, when the generated layout and inferred graph disagree in global yaw, we optionally apply a closed-form floor-plane rotation to align layout translations to graph centers before camera-view rendering; individual layout boxes are not edited.

Training losses. Unless otherwise stated, $\lambda_{\text{dino}}{=}0.3$ and the soft-matching temperature $\tau{=}0.07$. OFG uses $\lambda{=}0.15$ on $\mathcal{L}_{\text{quality}}$, project whenever $\mathbf{g}_{\text{co-align}}^{\mathrm{T}}\mathbf{g}_{\text{quality}} < 0$ ($\xi{=}0$, $\epsilon{=}10^{-8}$), and set $\alpha^{\text{vis}}{=}1.0$, $\alpha^{\text{node}}{=}0.5$ in Eq. (15). Projection always protects image-structure co-alignment except in the reverse-projection ablation. For scene-quality metrics in Sec. 4, F-Score uses a Euclidean distance threshold of $0.05\,\text{m}$.

Relation hinges. Writing $(\mathbf{t}_k)_x$ for the x-component of generated translation $\mathbf{t}_k$, the left hinge is $\ell_{\text{left}}(\mathbf{t}_k, \mathbf{t}_m) \coloneqq \text{softplus}((\mathbf{t}_k)_x - (\mathbf{t}_m)_x + \delta)$ with $\text{softplus}(x) \coloneqq \log(1 + e^x)$ and margin $\delta{=}0.05$; minimizing it encourages object k to lie left of m. Analogous softplus hinges cover right, front, behind, and above.

Visibility. Visibility $\nu_i \in [0, 1]$ in Eq. (15) is the fraction of object i's 2D box area not occluded by other objects closer to the camera after sorting by depth d_i; when depth ordering is unavailable we set $\nu_i{=}1$.

B Additional Qualitative Results

Figure 5 extends Fig. 3 to nine real-world evaluation scenes (three per room type).

C Multi-View Extension

All experiments in the main paper condition on a *single* RGB photograph. This section documents a multi-view generalization that reuses the same confidence-calibrated design: additional views tighten graph inference and image alignment where geometry is ambiguous in any one frame, while unreliable views are down-weighted rather than averaged uniformly. We include it to make the single-view presentation in Sec. 3 explicit about scope and to record the extension path.

Scene graph inference. Given V RGB views $\{I_v\}_{v=1}^{V}$, we run GroundingDINO independently on each view and estimate cameras with DUSt3R (Wang et al., 2024a), yielding per-view intrinsics K_v, extrinsics $(R_v, \mathbf{t}_v^{\text{cam}})$, and camera confidence $c_v^{\text{cam}} \in [0, 1]$. Each 2D detection is lifted to world coordinates with the same back-projection as Sec. 3.1, using $(K_v, R_v, \mathbf{t}_v^{\text{cam}})$ and Depth Anything depth from the view. Node confidence in view v follows the same κ-weighted detection–depth product as in Sec. 3.1.

Cross-view instances of the same physical object must be merged before relation inference. We greedily group same-class detections from different views when their combined similarity exceeds

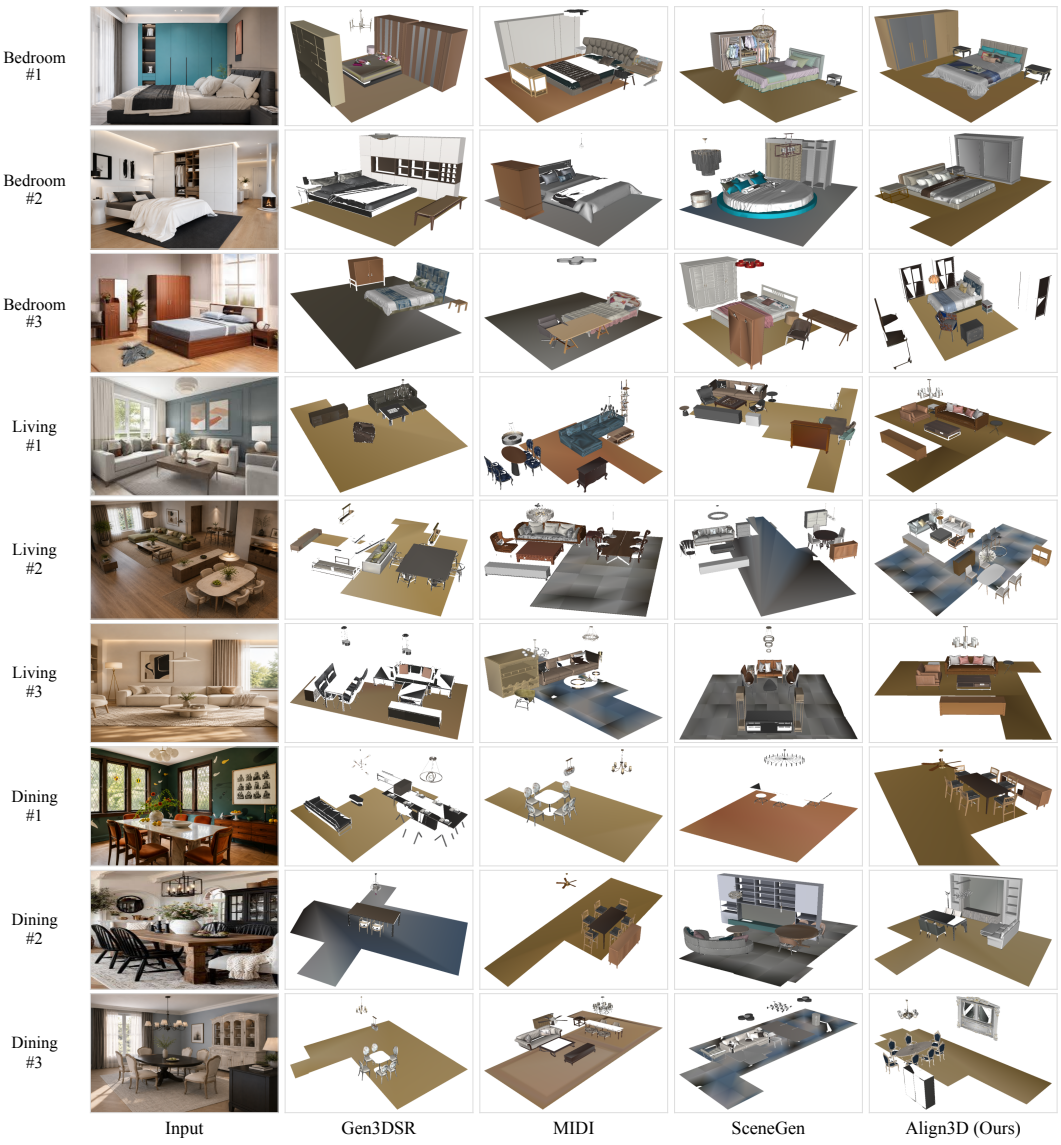

Figure 5: Qualitative comparison on nine real-world indoor scenes.

a threshold τ_{match}=0.5:

$$s_{ij} := \lambda_{\text{app}} \cos(\mathbf{z}_i^{\text{app}}, \mathbf{z}_j^{\text{app}}) + \lambda_{\text{geo}} \frac{1}{1 + \frac{d_{ij}^{\text{3D}}}{r_{\text{merge}}}}, \tag{18}$$

where $d_{ij}^{\text{3D}} := \|\mathbf{p}_i^{\text{world}} - \mathbf{p}_j^{\text{world}}\|$, r_{merge}=0.5 m normalizes the proximity term, and λ_{app}=0.6, λ_{geo}=0.4 with $\lambda_{\text{app}} + \lambda_{\text{geo}}$=1 balance appearance and geometry. Matched detections are fused by confidence-weighted averaging of 3D position and size; fused node confidence is the maximum view confidence (scaled conservatively) so that agreement across views increases structural trust. Spatial relations and edge confidences are then inferred on the fused graph exactly as in Sec. 3.1.

Image-structure co-alignment and fused conditioning. Modality confidences generalize to

$$c^{\text{image}} := \max_{v \in \{1,\ldots,V\}} c_v^{\text{cam}}, \tag{19}$$

so a reliable camera estimate preserves image-modality trust even if other views are uncertain. Image tokens $\mathbf{z}_{\text{img}}$ are extracted per view (CLIP projection as in Sec. 3.2) and mean-pooled before the fused

conditioning in Eq. (6). For appearance alignment, we render the layout proxy from each estimated camera and average view losses with confidence weights:

$$\mathcal{L}_{\text{image}} = \frac{\sum_{v=1}^{V} c_v^{\text{cam}} \mathcal{L}_{\text{image}}^{(v)}}{\sum_{v=1}^{V} c_v^{\text{cam}}}, \tag{20}$$

where $\mathcal{L}_{\text{image}}^{(v)}$ is the per-view CLIP/DINO alignment term defined in Sec. 3.2 for reference image I_v. When $V{=}1$, these expressions reduce to the single-image formulation in Sec. 3.2. Scene-graph alignment, confidence-weighted modality fusion, and OFG are unchanged aside from the richer graph and image terms above.

Inference-time multi-view guidance. At sampling time we additionally support a multi-view Gaussian-splat renderer that projects the current layout into each estimated camera before computing CLIP/DINO guidance, mirroring the training-time proxy but with per-view $K_v, R_v, \mathbf{t}_v^{\text{cam}}$. This path is useful for in-the-wild inputs where DUSt3R provides the camera bundle $\mathcal{P}$ and multiple photographs of the same room are available. Confidence-weighted fusion of per-view guidance gradients follows the same c_v^{cam} weighting as $\mathcal{L}_{\text{image}}$.