InLayDiffusion: Indoor Layout Estimation from a Single Panorama via Structural Point Cloud Lifting and Guided Diffusion

Giovanni Pintore[1][0000−0001−8944−1045], Uzair Shah[2][0000−0002−6729−5654], Marco Agus[2][0000−0003−2752−3525], and Enrico Gobbetti[1][0000−0003−0831−2458]

[1] CRS4, Cagliari, Italy
[2] HBKU, Doha, Qatar

Abstract. We introduce a novel end-to-end deep-learning method that combines structural geometric lifting with prior-guided diffusion to recover a 2D floorplan and a 3D room layout from a 360° image. Structural lifting predicts gravity-aligned, structure-aware depth and projects the resulting colored 3D point cloud onto the floor plane through a differentiable operation, directly linking panoramic image features to geometric reasoning and improving robustness to clutter while encoding cues such as room height and coarse footprint geometry. Footprint reconstruction is then formulated as polygon denoising and completion within a diffusion framework. The layout is represented as a fixed-vertex polygon jointly encoded with floor-projected RGB and density features, enabling refinement of noisy predictions and completion of occluded boundaries using structural priors. A variational guidance network further regularizes initialization by encoding indoor layout priors. Panoramic benchmarks show that our diffusion approach provides a strong foundation for general room reconstruction, even in cluttered, non-Manhattan environments.

Keywords: 3D Reconstruction · Scene Analysis and Understanding · Panoramic and 360 images · Layout from Single Image.

1 Introduction

Single-image indoor layout estimation infers the room's structural boundaries (walls, ceiling, and floor) from a photograph of a furnished space [34]. Recently, 360° panoramic imagery has gained attention for its ease of use, broad field of view, and affordability [27,23,12]. However, reconstruction remains challenging because furniture occlusions, non-convex geometries, and untextured surfaces obscure or distort structural cues. Overcoming these challenges requires deep contextual understanding and strong geometric priors [34,14]. Furthermore, the equirectangular representation provided by the vast majority of capture setups [12] does not preserve areas or angles and introduces severe horizontal stretching near the poles, complicating shape analysis.

Because single-image layout estimation is under-constrained, it fundamentally relies on priors. The current trend is shifting away from pure image analysis

paired with rigid geometric constraints [34] toward a hybrid mix of geometry and learning from large-scale examples, with Transformers emerging as a well-adapted solution for capturing long-range panoramic dependencies (see section 2). Most frameworks assume gravity-aligned panoramas to project 3D space onto a 2D floor plan, since this optimizes implementations and is easy to obtain even from casual captures [25,4]. Typically, these methods predict structural cues like boundaries or horizon depth [3]. However, handling furniture occlusions and non-convex geometries often forces reliance on heuristic, Manhattan-world post-processing [34]. Furthermore, height is often treated as an isolated auxiliary task, meaning that robustly completing occluded regions and ensuring metric consistency remain open challenges [22].

In this work, we investigate the use of deep generative *guided diffusion* models to reconstruct and complete indoor layouts from a single panorama, introducing, to the best of our knowledge, the first diffusion-based formulation for room layout estimation from a single panoramic image.

Existing approaches typically rely on direct regression of structural boundaries, horizon-aligned signals, or transformer-based reasoning in equirectangular image space. Rather than reasoning directly in distorted equirectangular space [12], we predict a structure-oriented depth using gravity-aligned features, lift it into a colored 3D point cloud, and differentially project it onto the floor plane. While this projection captures only visible architectural surfaces, leaving out occluded regions, it transfers visual evidence into a floor-aligned domain where structural regularities become explicit. In this space, walls and layout elements form coherent geometric patterns, while clutter and non-structural content appear more fragmented or radially distributed. This makes wall boundaries and room height explicit while reducing panoramic distortions and clutter.

Building on this representation, we cast layout estimation as a guided denoising problem. A diffusion network refines a coarse polygonal footprint into a fixed-vertex 2D floorplan, correcting noisy visible structures and completing occluded or weakly observed boundaries. This is well-suited to single-panorama layout inference, which is inherently under-constrained due to furniture, occlusions, and concave geometries. The denoising process enforces global polygon consistency while integrating floor-projected appearance cues. This formulation uses diffusion to infer layout geometry from visual evidence rather than synthesizing images or scenes from a given layout, as in common diffusion-based generative models. The process is guided by floor-aligned appearance features and a learned variational prior over indoor structures, enabling coherent layout recovery even under strong occlusions.

Our main contributions are summarized as follows:

– **Structural Lifting from Panoramas to Floorplan Space.** We introduce a geometry-aware representation that maps panoramic evidence into planar layout space. A *structure-oriented depth* predicted with gravity-aligned features [10] is lifted into a colored 3D point cloud and differentiably projected onto the floor plane. Unlike full-scene depth, this representation focuses on stable architectural surfaces, suppresses clutter and equirectangular distor-

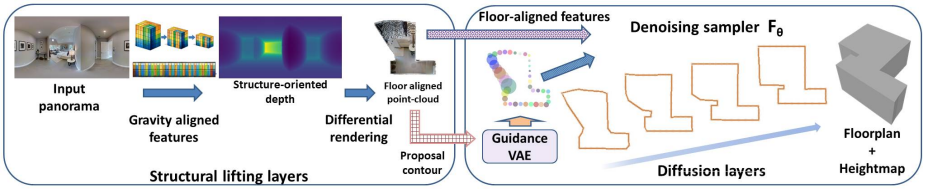


Fig. 1: Overview. From a single equirectangular panorama, we predict a structure-oriented depth and lift it into a colored point cloud, which is projected onto the floor plane using a differentiable operator. This floor-aligned representation organizes visible structural evidence into coherent geometric patterns, yielding a height map and a coarse footprint proposal. The final layout is obtained via diffusion-based polygonal denoising, which refines the footprint and completes occluded regions under the guidance of learned indoor layout priors.

tions, preserves room height, and provides a structured basis for polygonal reasoning, beyond prior panorama- or ceiling-based formulations [26,11].

- **Diffusion-Based Polygonal Layout Reconstruction.** We formulate layout estimation as denoising over polygon vertices, where a diffusion network refines a noisy proposal into a fixed-vertex floorplan conditioned on geometric and appearance evidence. Unlike generative methods that synthesize scenes from layouts, we use diffusion to infer layout geometry from visual observations, completing occluded regions and improving structural coherence over direct regression approaches [18,3].
- **Variational Guidance with Learned Indoor Layout Priors.** We introduce a variational guidance module that learns a latent prior over plausible indoor room geometries. Injected during denoising, this prior constrains refinement toward metrically consistent layouts, reduces geometric drift, and supports both free-form and Manhattan-regularized initializations.

To validate the effectiveness of the proposed approach, we conduct an extensive experimental evaluation on widely used panoramic layout benchmarks, including MatterportLayout [33] and the Zillow Indoor Dataset (ZInD) [2]. Beyond standard quantitative comparisons, we provide a comprehensive analysis of the proposed pipeline, including detailed computational statistics and component-level ablation studies. These experiments highlight both the robustness of the diffusion-based formulation and its good behavior across diverse indoor scenes, showing that our diffusion approach provides a strong foundation for solving the general room reconstruction problem, even in cluttered, non-Manhattan environments.

2 Related work

As a fundamental task in indoor reconstruction, single-image panoramic layout estimation has shifted from traditional geometric reasoning to deep learning frameworks. We refer the reader to established surveys [34,**?**,12] for a comprehensive review of this evolution, limiting our discussion here to the approaches most closely related to ours.

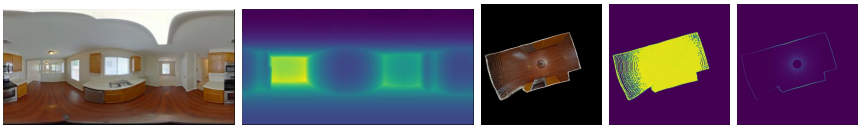
To cope with the under-constrained layout reconstruction problem, the current trend is to employ a hybrid mix of geometry priors derived from architectural construction considerations (e.g., wall alignment) and learning from large-scale examples [12]. Zou et al. [34] observed that many data-driven methods retain a three-stage pipeline: pre-alignment, 2D prediction, and post-processing. LayoutNet [32] predicted corners and boundaries, while HorizonNet [18] used 1D floor–wall, ceiling–wall, and wall–wall junction signals, later regularized with MWM. DuLaNet [26] fused ERP and ceiling projections, Led^{2}Net [24] added horizon-depth supervision, and Zeng et al. [29] combined semantic and depth cues. However, these methods often still require Manhattan post-processing [34].

Other works relaxed restrictive geometric assumptions. Vanishing-line and structural post-processing constraints [34,30,7,32,18,26] were addressed by AtlantaNet [11], which used the Atlanta World Model (AWM) and dual horizontal projections. Deep3DLayout [14] predicted watertight 3D meshes beyond predefined priors, while SSLayout360 [21] showed that semi-supervised learning can approach fully supervised accuracy with fewer labels. These methods broaden representational capacity, but often decouple 2D layout recovery from full 3D reasoning.

Transformers have recently become prominent. LGT-Net [3] introduced horizon-sequence attention; DMH-Net [31] used a learnable Manhattan Hough space; DOP-Net [16] disentangled orthogonal plane features; and PanelNet [28] modeled panoramas as continuous horizontal panels. Bi-Layout [22] handled annotation ambiguities with dual layouts, while Seg2Reg [20] combined 2D segmentation and 1D regression via differentiable rendering. More recently, HUSH [8] used spherical harmonics as structured queries for multi-task geometry prediction, achieving state-of-the-art results on depth, normals, and layout. Transformers have also been applied to self-supervised indoor point-cloud completion from a single panorama [15,9].

Despite these advances, existing methods still rely on assumptions or modular pipelines that limit geometric completeness. Reprojection-based approaches [26,11] mainly detect visible ceiling or floor boundaries in equirectangular or perspective space, leaving height estimation and occluded regions to heuristics. Horizon-line formulations [18,3,19] reduce layout inference to 1D signals, discarding the full 2D footprint structure. Recent transformer models [31,16,20] are more expressive, but often treat completion and height reasoning as auxiliary tasks requiring constraints or post-hoc refinement.

In contrast, we decouple structural evidence extraction from layout completion. We lift panoramic cues into a floor-aligned representation through structure-oriented depth, then formulate layout estimation as guided denoising in floorplan space. A diffusion model completes the polygonal footprint beyond visible evidence, while a variational indoor prior regularizes the process toward plausible room geometries, enabling recovery of wall height and occluded boundaries without heuristic post-processing.



(a) Input RGB. **(b)** Struct. depth **(c)** F. PC **(d)** H map **(e)** O map

Fig. 2: Structural lifting from panorama to floor-aligned representations. The input panorama is converted into structure-oriented depth, lifted to a colored 3D point cloud, and differentially projected onto the floor plane. This yields a projected point cloud, a height map H, and an occupancy map O. The upper 4% of the depth image is removed to discard ceiling regions. These outputs provide the geometric basis for diffusion-based polygon refinement.

3 Overview

Figure 1 summarizes our end-to-end pipeline. Given an input equirectangular panorama, we first lift visual evidence into a floor-aligned geometric domain. A structure-oriented depth map, predicted from gravity-aligned features, is converted into a colored 3D point cloud and differentially projected onto the floor plane. This projection provides a height map for recovering the final 3D layout and ceiling geometry, and a coarse footprint proposal given by the contour of the projected point cloud. Although partial under occlusions, this proposal initializes the subsequent inference. We then perform diffusion-based layout estimation (section 5). The footprint is represented as a closed fixed-vertex polygon encoded with floor-projected RGB features. Starting from the proposal, a denoising network refines noisy boundaries and completes unobserved structures, guided by a variational prior over plausible indoor layouts. The guidance supports both free-form and Manhattan-regularized coarse proposals. Training is staged for stability. We first train structural lifting and variational guidance while freezing the denoiser (section 4); then we freeze these modules and train the diffusion network (section 5). At inference, all components jointly produce the final 2D footprint and corresponding 3D room layout (section 6). Additional implementation details and qualitative results are provided in the Supplementary Materials.

4 Structural Lifting

The structural lifting stage maps an input equirectangular panorama into a metric, floor-aligned representation for layout reasoning. As shown in Figure 2, it does not directly predict the final room layout, but reorganizes image evidence so that room extent, wall structures, and architectural regularities become easier to infer.

Structure-Oriented Depth and Indoor Priors. Given an input panorama $\mathcal{I}_{\text{erp}} \in \mathbb{R}^{H\times W\times 3}$, the lifting module predicts a structure-oriented depth map

$$D_s = f_\theta(\mathcal{I}_{\text{erp}}), \tag{1}$$

encoding only visible architectural surfaces, i.e., walls, floor, and ceiling. Occluded regions and complete room geometry are not inferred at this stage.

Indoor priors are embedded through gravity-aligned features, assuming vertical walls and horizontal floors/ceilings. Intermediate features are compressed along the vertical image dimension,

$$\Phi_{\text{gaf}} = \text{Compress}_v(\Phi_{\text{enc}}(\mathcal{I}_{\text{erp}}))\,, \tag{2}$$

to emphasize vertical continuity and wall-boundary discontinuities while suppressing clutter.

Lifting RGB Evidence into 3D. The predicted depth lifts both geometry and RGB evidence into 3D. For each pixel (u, v), equirectangular coordinates are

$$\lambda(u,v) = 2\pi\frac{u}{W} - \pi, \qquad \varphi(u,v) = \pi\left(\frac{v}{H} - \frac{1}{2}\right), \tag{3}$$

with unit ray

$$\mathbf{r}(u,v) = \left[\cos\varphi \sin\lambda \sin\varphi \cos\varphi \cos\lambda\right]^\top. \tag{4}$$

The 3D point and colored point cloud are

$$\mathbf{p}(u,v) = D_s(u,v)\,\mathbf{r}(u,v), \qquad \mathcal{P} = \{(\mathbf{p}_i, \mathbf{c}_i)\}_{i=1}^{N}. \tag{5}$$

This lifting highlights structural organization rather than fine geometry. Wall points accumulate along room boundaries after floor projection, while clutter spreads across the interior, producing contextual patterns useful for layout reasoning.

Differentiable Floor-Plane Projection. Assuming vertical walls, each point $\mathbf{p}_i = (x_i, y_i, z_i)$ is projected onto the floor as

$$\tilde{\mathbf{p}}_i = (x_i, z_i). \tag{6}$$

A differentiable splatting operator accumulates projected points on a 2D grid, producing a floor-projected point cloud, an occupancy map

$$O(\mathbf{u}) = \sum_i \exp\left(-\frac{\|\tilde{\mathbf{p}}_i - \mathbf{u}\|^2}{2\sigma^2}\right), \tag{7}$$

and a height map

$$H(\mathbf{u}) = \frac{\sum_i y_i \exp\left(-\frac{\|\tilde{\mathbf{p}}_i - \mathbf{u}\|^2}{2\sigma^2}\right)}{\sum_i \exp\left(-\frac{\|\tilde{\mathbf{p}}_i - \mathbf{u}\|^2}{2\sigma^2}\right)}. \tag{8}$$

Before projection, the upper 4% of the depth image is discarded to avoid ceiling artifacts inside the footprint, while the height map H preserves vertical extent for 3D layout recovery.

The occupancy map O is used only for training supervision. Since it is sparse in single-room cases, it is not provided to the diffusion stage, which instead uses floor-projected RGB and geometric cues.

Structural lifting outputs floor-projected RGB features, a height map, and auxiliary density information. A coarse footprint is extracted by thresholding and contour detection from visible structures and used only to initialize the subsequent fixed-vertex diffusion refinement.

Training losses and optimization. The lifting network is trained to predict structure-oriented depth for layout reconstruction using depth, regularization, floor-projection, height-map, and 3D consistency losses. Given predicted depth $\hat{D}$ and ground-truth layout depth D, we use an inverse Huber depth loss

$$\mathcal{L}_{\text{depth}} = \text{Huber}^{-1}(\hat{D}, D), \tag{9}$$

and horizon-based gradient/normal regularization [3],

$$\mathcal{L}_{\text{reg}} = \mathcal{L}_{\text{norm}} + \mathcal{L}_{\text{grad}}. \tag{10}$$

Floorplan-space consistency is encouraged through an occupancy loss after projection [13],

$$\mathcal{L}_{\text{occ}} = \mathcal{L}_{\text{occ}}(\hat{D}, D). \tag{11}$$

For 3D consistency, predicted and ground-truth depths are lifted into point clouds $\hat{\mathcal{P}}$ and $\mathcal{P}$. Local PCA provides planarity π and sharpness s scores, yielding

$$\mathcal{L}_{\text{pc}} = \|\hat{\pi} - \pi\|_1 + \|\hat{s} - s\|_1. \tag{12}$$

The height map is supervised with a shape-aware Dice loss,

$$\mathcal{L}_{\text{hmap}} = \mathcal{L}_{\text{Dice}}(\hat{H}, H). \tag{13}$$

The total objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{reg}} + \lambda_{\text{occ}}\mathcal{L}_{\text{occ}} + \lambda_{\text{hmap}}\mathcal{L}_{\text{hmap}} + \lambda_{\text{pc}}\mathcal{L}_{\text{pc}}, \tag{14}$$

with $\lambda_{\text{occ}} = 0.01$, $\lambda_{\text{hmap}} = 0.01$, and $\lambda_{\text{pc}} = 0.1$.

5 Diffusion-based polygon reconstruction

Given the floor-aligned representation from structural lifting (section 4), we reconstruct the complete 2D room footprint as a closed fixed-vertex polygon. The input consists of floor-projected RGB features from the lifted point cloud and a coarse polygonal proposal extracted from the projected point cloud. The target is the full room contour, including occluded boundaries, represented as $P \in [-1, 1]^{M\times 2}$, with $M = 40$ in all experiments. The fixed-vertex representation simplifies learning and enables structured attention while preserving the dominant layout shape. The final 3D layout is obtained by extruding the predicted footprint

using the room height estimated during structural lifting. We formulate footprint reconstruction as guided denoising, combining: (i) a variational guidance network that predicts a Gaussian prior over polygon vertices from the proposal polygon; and (ii) a diffusion denoiser that refines a noisy polygon into a complete footprint conditioned on this prior and floor-aligned visual features.

Variational guidance network. The polygon P is vectorized as $\mathbf{y} \in \mathbb{R}^{2M}$ and modeled with a variational autoencoder with Gaussian encoder $q_\psi(\mathbf{z} \mid \mathbf{y})$ and decoder $p_\psi(\mathbf{y} \mid \mathbf{z})$. It is trained with the β-VAE objective [6]:

$$\mathcal{L}_{\mathrm{VAE}} = -\mathbb{E}_{q_\psi}[\log p_\psi(\mathbf{y} \mid \mathbf{z})] + \beta\, \mathrm{KL}(q_\psi(\mathbf{z} \mid \mathbf{y}) \,\|\, \mathcal{N}(0, I)). \tag{15}$$

At inference, the decoder predicts per-vertex Gaussian parameters

$$(\boldsymbol{\mu}_{\mathrm{guide}}, \boldsymbol{\sigma}_{\mathrm{guide}}) = \mathrm{Dec}_\psi(\mathbf{z}), \tag{16}$$

defining a distribution over plausible room footprints and a structured initialization for diffusion.

To avoid leakage, the guidance network is trained only on training-split layout polygons. Validation and test polygons are never used to train the VAE prior, tune the denoiser, or define the initialization distribution. At inference, the guidance network receives only the coarse proposal extracted from the projected point cloud.

Diffusion-style polygon denoising. Following EDM [5], we define a noisy polygon state $\mathbf{x}_\sigma \in \mathbb{R}^{2M}$ at noise level σ, initialized as

$$\mathbf{x}_{\sigma_{\max}} = \boldsymbol{\mu}_{\mathrm{guide}} + \sigma_{\max}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{17}$$

Using a power-law schedule with N steps, the denoiser predicts the clean state

$$\widehat{\mathbf{x}}_0 = F_\theta(\mathbf{x}_\sigma, \sigma, \boldsymbol{\mu}_{\mathrm{guide}}, \phi(\mathcal{F}_{\mathrm{floor}})), \tag{18}$$

where $\phi(\mathcal{F}_{\mathrm{floor}})$ are features from the floor-projected colored point cloud. The polygon is updated with the EDM solver, optionally with Heun correction, and the final state $\mathbf{x}_0$ is reshaped into P.

Denoiser architecture. The denoiser F_θ combines VAE guidance with floor-aligned visual evidence. Each vertex $\mathbf{v}_k$ is mapped to normalized sampling coordinates

$$\mathbf{r}_k = \tfrac{1}{2}(\mathbf{v}_k + \mathbf{1}) \in [0, 1]^2, \tag{19}$$

used as query points over the floor-projected feature maps. A convolutional backbone extracts multi-scale features, encoded by a deformable transformer. Decoder layers update polygon tokens through multi-scale deformable cross-attention, self-attention, and feed-forward layers, with diffusion time injected by FiLM modulation.

Diffusion polygon module

Architecture		Complexity	
Component	**Config**	**Params**	**GFLOPs**
ResNet50 Backbone	–	25.60M	5.36
Input Projections	4 levels	1.51M	0.23
Embeddings	dim=256	0.56M	–
Encoder	6 layers	4.54M	6.42
Decoder	6 layers	10.46M	0.89
Output Heads	6×	0.79M	0.03
Other	–	0.27M	–
PolyModel	total	**43.73M**	**12.93**
Guidance Network	2 layers	2.37M	–
System Total	–	**46.10M**	**12.93**
RTX 4090	10 steps	2 ms / 1 ms cached	
RTX 3090	10 steps	4 ms / 2.3 ms cached	

Structural lifting

Backbone / Metric	Encoder	Slicing	MHSA	Decoder
ResNet18				
Params	11.18M	2.16M	2.10M	6.29M
Latency	6.34 ms	1.68 ms	0.95 ms	4.06 ms
ResNet34				
Params	21.28M	2.16M	2.10M	6.29M
Latency	10.20 ms	1.41 ms	0.92 ms	4.19 ms
ResNet50				
Params	23.51M	34.47M	33.58M	100.67M
Latency	14.31 ms	6.10 ms	10.31 ms	42.61 ms

Table 1: Computational statistics. Left: diffusion-based polygon reconstruction, evaluated at 256×256 with $M = 40$ vertices. Right: structural lifting statistics for 512×1024 RGB input. Structural lifting dominates the cost, while diffusion refinement adds only millisecond-level overhead with 10 denoising steps.

Training objective. Let $\mathbf{y} \in \mathbb{R}^{2M}$ be the ground-truth polygon. Following EDM [5], we sample σ from a log-normal distribution and define

$$\mathbf{x}_\sigma = \mathbf{y} + \sigma\,(\boldsymbol{\sigma}_{\text{guide}} \odot \epsilon), \qquad \epsilon \sim \mathcal{N}(0, I). \tag{20}$$

The denoiser predicts $\widehat{\mathbf{x}}_0$ and is trained with

$$\mathcal{L} = \mathcal{L}_{\text{dm}} + \lambda_{\text{edge}}\,\mathcal{L}_{\text{edge}} + \lambda_{\text{angle}}\,\mathcal{L}_{\text{angle}}, \tag{21}$$

where $\mathcal{L}_{\text{edge}}$ enforces consistent edge lengths and $\mathcal{L}_{\text{angle}}$ preserves corner angles.

6 Results

We implemented our approach with *PyTorch* and evaluated it on widely used benchmarks for panoramic indoor layout estimation, covering a wide range of scene layouts, clutter levels, and architectural variability. All experiments are conducted using a single panoramic image as input, following standard evaluation protocols.

Datasets. Our evaluation is conducted on standard benchmarks for panoramic layout estimation, enabling direct comparison with recent state-of-the-art methods. While datasets such as *PanoContext* [30] and *Stanford2D3D* [17] mainly contain small-scale near-cuboid layouts, our focus is on more complex geometries and layout completion under occlusions. For this reason, we report results on the more challenging *MatterportLayout* [33] and *Zillow Indoor Dataset (ZInD)* [2]. *MatterportLayout* provides annotated panoramic images for training, validation, and testing in cluttered indoor scenes under Manhattan constraints, while *ZInD* is the largest real-world benchmark and includes both simple and raw annotations covering diverse non-Manhattan layouts.

Computational stats. Table 1 summarizes the cost of the two main pipeline components. The budget is dominated by structural lifting, whose encoder and decoder scale with backbone capacity and panoramic input resolution. In contrast, diffusion-based polygon reconstruction operates on a compact fixed-vertex representation and adds only a small inference overhead.

Fig. 3: Qualitative comparison. Input panoramas and comparison views on MatterportLayout and ZInD, with *InLayDiffusion* in blue, LGT-Net [3], DOPNet [16], and ground truth in green.

Training details. Structural lifting, variational guidance, and diffusion denoising—are trained in separate stages using Adam with learning rate 2×10^{-4}. Further implementation details, hyperparameters, and training schedules are provided in the supplementary materials.

6.1 Quantitative results

We evaluate layout estimation using the standard Intersection over Union (IoU) metric, reporting both 2D IoU for footprint accuracy and 3D IoU, which also accounts for wall height estimation. Table 2 compares our method with representative panoramic layout approaches on MatterportLayout and ZInD. The considered baselines differ in their internal layout representations, including ERP corner or boundary maps [34], ceiling or floor projections [26,11], 1D horizon-depth signals [24], and horizon-aligned vectors or density representations [18,3,20]. As these approaches represent mature solutions, their quantitative results often converge to similar ranges, with differences partly influenced by annotation protocols and by the form of the predicted output (e.g., corner sequences, polygonal approximations, or heuristically post-processed layouts). Moreover, as discussed later, ground-truth annotations themselves may contain approximations and inconsistencies, which can further affect numerical comparisons. For these reasons, we focus our analysis on methods for which inference code and data format details are publicly available. Rather than emphasizing small differences in absolute performance, we highlight the architectural characteristics of the compared pipelines and the formulation proposed in this work. In particular, *InLayDiffusion* approaches panoramic layout estimation by first mapping visual evidence into a floorplan-aligned geometric space and subsequently refining the footprint through guided polygonal denoising. Operating in a geometric domain allows the method to integrate floor-aligned visual cues with learned structural priors during refinement, providing a flexible mechanism to handle partially observed boundaries and irregular layouts while maintaining competitive performance across different datasets (see subsection 6.2).

6.2 Qualitative results

Figure 3 illustrates representative qualitative comparisons across multiple benchmarks. For each example, the input panorama in ERP format is shown together

Method	Model	2D IoU	3D IoU
(a) MatterportLayout results (%)			
LayoutNetV2 [34]	ERP corners	78.73	75.82
DuLaNetV2 [34]	2D map	78.82	75.05
HorizonNet [18]	ERP vectors	81.71	79.11
AtlantaNet [11]	2D maps	82.09	80.02
LED2-Net [24]	ERP 1D depth	82.61	80.14
LGT-Net [3]	ERP vectors	83.52	81.11
DOPNet [16]	ERP density	84.11	81.70
InLayDiffusion	Polygon	**84.24**	**81.82**
(b) ZInD results (%)			
HorizonNet [18]	ERP vectors	90.44	88.59
LED2-Net [24]	ERP 1D depth	90.36	88.49
LGT-Net [3]	ERP vectors	91.77	89.95
DOPNet [16]	ERP density	91.94	90.13
InLayDiffusion	Polygon	**92.10**	**90.15**

Table 2: Benchmark results of representative panoramic layout estimation methods on common datasets. MatterportLayout and ZInD. The table reports 2D and 3D IoU scores whenever available, including only methods for which directly comparable results were provided in the original papers. Rows marked as *Ours* correspond to the proposed InLayDiffusion, with results to be reported.

Variant	2D IoU (%)	3D IoU (%)
Baseline	77.68	75.12
+ Layout depth	78.56	76.67
+ Denoising	82.22	79.99
+ Guided diffusion **(our full)**	**84.24**	**81.82**

Table 3: Ablation study on MatterportLayout. Effect of the main components of the proposed pipeline. The baseline derives footprint and height directly from structure-oriented depth.

with a square visualization that juxtaposes the predicted layout produced by InLayDiffusion (blue), the result of LGT-Net [3] (orange) and DOPNet [16] (red), and the ground-truth annotation (filled green). We report qualitative comparisons against LGT-Net and DOPNet as they provide publicly available code and pretrained models, enabling direct and fair visual evaluation.

Across MatterportLayout and ZInD, our method consistently produces sharper and more metrically accurate contours, particularly in regions affected by clutter or partial occlusions. In such scenarios, LGT-Net often produces incomplete or locally distorted boundaries, reflecting the limitations of horizon-aligned representations when large portions of the room structure are not directly visible. In contrast, the diffusion-based refinement in InLayDiffusion leverages floor-aligned structural cues to complete missing boundaries in a coherent and globally consistent manner.

6.3 Ablation study

We perform an ablation study on MatterportLayout to assess the contribution of the main pipeline components. Table 3 reports increasingly complete variants evaluated with 2D and 3D IoU. We analyze three design choices. First, layout-oriented depth is compared with full-scene depth from a foundation model [1]. By suppressing non-architectural elements such as doors, windows, and clutter, layout-oriented depth produces cleaner floor projections and fewer spurious

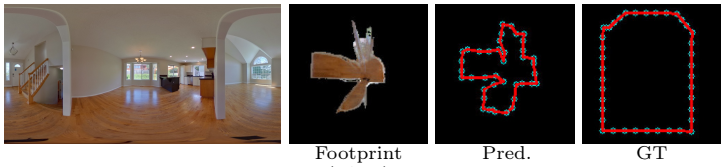


Fig. 4: Limitation example. Top: input equirectangular panorama. Bottom: lifted floor-aligned footprint (left), predicted polygon (middle), and ground truth (right). Weak structural cues and coarse annotations, including simplified boundaries and slanted-ceiling approximations, can penalize visually plausible predictions.

boundaries. Second, we compare direct footprint extraction with diffusion-based reconstruction. Diffusion denoising enables iterative refinement and completion of missing layout regions beyond visible evidence. Third, we evaluate guided diffusion by contrasting random initialization with a regularized footprint proposal. Structured initialization stabilizes denoising and improves geometric coherence. Overall, the results support the combination of layout-specific geometry and guided diffusion in floorplan space.

6.4 Limitations

Our method depends on the quality of structural lifting. Although diffusion can denoise and complete missing structures, it still requires consistent wall evidence in the floor-aligned representation. When clutter dominates or boundary accumulation is weak, the model may produce an incorrect but plausible footprint, as shown in Figure 4. A further limitation is annotation quality. Dataset layouts are sometimes coarse or misaligned with the panorama, affecting training and evaluation. This can penalize predictions that better match the observed scene, especially for complex geometries such as slanted ceilings, which are often simplified in the ground truth.

7 Conclusions

We presented *InLayDiffusion*, a geometry-aware approach for indoor layout estimation from a single panorama that combines structural lifting with diffusion-based polygon reconstruction. By operating in a floor-aligned geometric space, the method separates structural evidence extraction from layout completion. Structural lifting organizes visual cues into geometric patterns, while diffusion refinement denoises and completes the footprint under learned indoor layout priors. Experiments on multiple benchmarks show robustness to clutter and occlusions and performance on par with the current state-of-the-art on both regular and irregular layouts. As for all diffusion models, fully exploiting our approach's potential requires a large number of training examples. The results obtained using restricted training sets already show that our method provides a strong foundation for general room reconstruction in challenging, non-Manhattan environments. We plan to explore this path by improving the training scheme, in

particular to achieve generalization with limited supervision, including pseudo-labeling, teacher–student or distillation strategies to reduce reliance on manual annotations. We also plan to improve the depth estimation phase by exploring the usage of pre-trained foundation models.

Acknowledgments. This publication was supported by NPRP-S 14th Cycle grant 0403-210132 AIN2 from the Qatar National Research Fund (a member of Qatar Foundation). GP and EG also acknowledge support from Sardinian Regional Authorities under the XDATA project (Art9 LR 20/2015), while US and MA acknowledge support from HBKU Vice President of Research under the Thematic Grant HBKU-INT-VPR-TG-03-03 *3G: Gaze, Guide, and Grasp: LLM-Augmented Immersive Experiences for Qatar's Heritage Sites*. The findings herein reflect the work and are solely the responsibility of the authors.

References

1. Cao, Z., Zhu, J., Zhang, W., Ai, H., Bai, H., Zhao, H., Wang, L.: PanDA: Towards panoramic depth anything with unlabeled panoramas and mobius spatial augmentation. In: Proc. CVPR. pp. 982–992 (2025). https://doi.org/10.1109/CVPR52734.2025.00100
2. Cruz, S., Hutchcroft, W., Li, Y., Khosravan, N., Boyadzhiev, I., Kang, S.B.: Zillow indoor dataset: Annotated floor plans with 360° panoramas and 3D room layouts. In: Proc. CVPR. pp. 2133–2143 (2021). https://doi.org/10.1109/CVPR46437.2021.00217
3. Jiang, Z., Xiang, Z., Xu, J., Zhao, M.: LGT-Net: Indoor panoramic room layout estimation with geometry-aware transformer network. In: Proc. CVPR. pp. 1654–1663 (2022). https://doi.org/10.1109/CVPR52688.2022.00170
4. Jung, R., Lee, A.S.J., Ashtari, A., Bazin, J.: Deep360Up: A deep learning-based approach for automatic VR image upright adjustment. In: Proc. IEEE VR. pp. 1–8 (2019). https://doi.org/10.1109/VR.2019.8798326
5. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. In: Proc. NeurIPS. pp. 1926:1–1926:13 (2022). https://doi.org/10.48550/arXiv.2206.00364
6. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: Proc. ICLR (2014). https://doi.org/10.48550/arXiv.1312.6114
7. Lee, D.C., Hebert, M., Kanade, T.: Geometric reasoning for single image structure recovery. In: Proc. CVPR. pp. 2136–2143 (2009). https://doi.org/10.1109/CVPR.2009.5206872
8. Lee, J., Park, H., Lee, B.U., Joo, K.: Hush: Holistic panoramic 3d scene understanding using spherical harmonics. In: Proc. CVPR. pp. 16599–16608 (June 2025). https://doi.org/10.1109/CVPR52734.2025.01547
9. Li, T., Zhang, Z., Wang, Y., Cui, Y., Li, Y., Zhou, D., Yin, B., Yang, X.: Self-supervised indoor scene point cloud completion from a single panorama. The Visual Computer **41**(3), 1891–1905 (2025). https://doi.org/10.1007/s00371-024-03509-w
10. Pintore, G., Agus, M., Almansa, E., Schneider, J., Gobbetti, E.: SliceNet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In: Proc. CVPR. pp. 11536–11545 (2021). https://doi.org/10.1109/CVPR46437.2021.01137

11. Pintore, G., Agus, M., Gobbetti, E.: AtlantaNet: Inferring the 3D indoor layout from a single 360 image beyond the Manhattan World assumption. In: Proc. ECCV. pp. 432–448 (2020). https://doi.org/10.1007/978-3-030-58598-3_26
12. Pintore, G., Agus, M., Schneider, J., Gobbetti, E.: State-of-the-art in deep learning approaches for automatic single-panorama indoor modeling and exploration. Computer Graphics Forum **45**(2), e70396 (2026). https://doi.org/10.1111/cgf.70396
13. Pintore, G., Agus, M., Signoroni, A., Gobbetti, E.: DDD++: Exploiting density map consistency for deep depth estimation in indoor environments. Graphical Models **140**, 101281 (August 2025). https://doi.org/10.1016/j.gmod.2025.101281
14. Pintore, G., Almansa, E., Agus, M., Gobbetti, E.: Deep3DLayout: 3D reconstruction of an indoor layout from a spherical panoramic image. ACM TOG **40**(6), 250:1–250:12 (2021). https://doi.org/10.1145/3478513.3480480
15. Shah, U., Tukur, M., Alzubaidi, M., Pintore, G., Gobbetti, E., Househ, M., Schneider, J., Agus, M.: MultiPanoWise: holistic deep architecture for multi-task dense prediction from a single panoramic image. In: Proc. CVPRW - OmniCV. pp. 1311–1321 (2024). https://doi.org/10.1109/CVPRW63382.2024.00138
16. Shen, Z., Zheng, Z., Lin, C., Nie, L., Liao, K., Zheng, S., Zhao, Y.: Disentangling orthogonal planes for indoor panoramic room layout estimation with cross-scale distortion awareness. In: Proc. CVPR. pp. 17337–17345 (2023). https://doi.org/10.1109/CVPR52729.2023.01663
17. Stanford University: BuildingParser Dataset (2017), https://sdss.redivis.com/datasets/9q3m-9w5pa1a2h, [Accessed: 2025-09-16]
18. Sun, C., Hsiao, C.W., Sun, M., Chen, H.T.: HorizonNet: Learning room layout with 1D representation and pano stretch data augmentation. In: Proc. CVPR. pp. 1047–1056 (2019). https://doi.org/10.1109/CVPR.2019.00114
19. Sun, C., Sun, M., Chen, H.T.: HoHoNet: 360° indoor holistic understanding with latent horizontal features. In: Proc. CVPR. pp. 2573–2582 (2021). https://doi.org/10.1109/CVPR46437.2021.00260
20. Sun, C., Tai, W.E., Shih, Y.L., Chen, K.W., Syu, Y.J., Wang, Y.C.F., Chen, H.T.: Seg2Reg: Differentiable 2D segmentation to 1D regression rendering for 360 room layout reconstruction. In: Proc. CVPR. pp. 10435–10445 (2024). https://doi.org/10.1109/CVPR52733.2024.00993
21. Tran, P.V.: Sslayout360: Semi-supervised indoor layout estimation from 360° panorama. In: Proc. CVPR. pp. 15353–15362 (2021). https://doi.org/10.1109/CVPR46437.2021.01510
22. Tsai, Y.J., Jhang, J.C., Zheng, J., Wang, W., Chen, A.Y., Sun, M., Kuo, C.H., Yang, M.H.: No more ambiguity in 360
23. Tukur, M., Ur Rehman, A., Pintore, G., Gobbetti, E., Schneider, J., Agus, M.: PanoStyle: Semantic, geometry-aware and shading independent photorealistic style transfer for indoor panoramic scenes. In: Proc. ICCVW. pp. 1553–1564 (2023). https://doi.org/10.1109/ICCVW60793.2023.00170
24. Wang, F.E., Yeh, Y.H., Sun, M., Chiu, W.C., Tsai, Y.H.: LED2-Net: Monocular 360 layout estimation via differentiable depth rendering. In: Proc. CVPR. pp. 12956–12965 (2021). https://doi.org/10.1109/CVPR46437.2021.01276
25. Xian, W., Li, Z., Fisher, M., Eisenmann, J., Shechtman, E., Snavely, N.: UprightNet: geometry-aware camera orientation estimation from single images. In: Proc. ICCV. pp. 9974–9983 (2019). https://doi.org/10.1109/ICCV.2019.01007
26. Yang, S.T., Wang, F.E., Peng, C.H., Wonka, P., Sun, M., Chu, H.K.: DuLa-Net: A dual-projection network for estimating room layouts from a single RGB panorama. In: Proc. CVPR. pp. 3363–3372 (2019). https://doi.org/10.1109/CVPR.2019.00348

27. Yang, S., Li, B., Cao, Y.P., Fu, H., Lai, Y.K., Kobbelt, L., Hu, S.M.: Noise-resilient reconstruction of panoramas and 3D scenes using robot-mounted unsynchronized commodity RGB-D cameras. ACM TOG **39**(5), 152:1–152:15 (2020). https://doi.org/10.1145/3389412
28. Yu, H., He, L., Jian, B., Feng, W., Liu, S.: Panelnet: Understanding 360 indoor environment via panel representation. In: Proc. CVPR. pp. 878–887 (2023). https://doi.org/10.1109/CVPR52729.2023.00091
29. Zeng, W., Karaoglu, S., Gevers, T.: Joint 3D layout and depth prediction from a single indoor panorama image. In: Proc. ECCV. pp. 666–682 (2020). https://doi.org/10.1007/978-3-030-58517-4_39
30. Zhang, Y., Song, S., Tan, P., Xiao, J.: PanoContext: A whole-room 3D context model for panoramic scene understanding. In: Proc. ECCV. pp. 668–686 (2014). https://doi.org/10.1007/978-3-319-10599-4_43
31. Zhao, Y., Wen, C., Xue, Z., Gao, Y.: 3D room layout estimation from a cubemap of panorama image via deep Manhattan hough transform. In: Proc. ECCV. pp. 637–654. Springer (2022). https://doi.org/10.1007/978-3-031-19769-7_37
32. Zou, C., Colburn, A., Shan, Q., Hoiem, D.: LayoutNet: Reconstructing the 3D room layout from a single RGB image. In: Proc. CVPR. pp. 2051–2059 (2018). https://doi.org/10.1109/CVPR.2018.00219
33. Zou, C., Su, J.W., Peng, C.H., Colburn, A., Shan, Q., Wonka, P., Chu, H.K., Hoiem, D.: MatterportLayout (2017), https://github.com/ericsujw/Matterport3DLayoutAnnotation, [Accessed: 2025-09-16]
34. Zou, C., Su, J., Peng, C., Colburn, A., Shan, Q., Wonka, P., Chu, H., Hoiem, D.: Manhattan room layout reconstruction from a single 360 image: A comparative study of state-of-the-art methods. International Journal of Computer Vision **129**, 1410–1431 (2021). https://doi.org/10.1007/s11263-020-01426-8