# TOWARDS OCCLUSION-ROBUST AND AMBIGUITY-RESILIENT PANORAMIC ROOM LAYOUT ESTIMATION VIA A LAYOUT RELIABILITY FIELD

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Estimating room layout from a panoramic image is fundamental to downstream tasks such as embodied navigation and scene reasoning, and has attracted extensive research attention. However, existing works are challenged by severe object occlusions and inherent ambiguity of layout annotations. In this paper, we propose an occlusion-robust and ambiguity-resilient layout estimation framework, which follows a paradigm that decomposes the task into horizontal-depth estimation and room height estimation. The horizontal-depth estimation is obtained through the fusion of depth priors and a layout reliability field. The reliability field is constructed under the supervision of geometric depth evidence and semantic textual priors. Specifically, we develop a depth-consistency-driven occlusion representation scheme that leverages depth-scale normalization to stably model object-induced occlusions across diverse scenes. In addition, a candidate-layout reasoning scheme is introduced by designing text-guided prompts and leveraging a multimodal large language model (MLLM) to capture semantic representations of layout ambiguity. Extensive experiments on three widely recognized datasets demonstrate that the proposed method significantly reduces boundary misclassification.

## 1 INTRODUCTION

Obtaining the layout from a single panoramic image input is crucial for indoor scene information perception and downstream tasks such as augmented reality, robot navigation, and 3D reconstruction Lin & Zheng (2026).

Early approaches Zou et al. (2018) recover room layouts by predicting layout boundary probability maps. These methods heavily rely on visible scene boundaries and are vulnerable to furniture-induced occlusions. Later methods formulate layout estimation as a 1D sequence prediction task by compressing panoramic images along the vertical direction. Although this strategy significantly improves computational efficiency Sun et al. (2019; 2021); Jiang et al. (2022); Pintore et al. (2020); Wang et al. (2021), the compression of 2D spatial information inevitably loses the spatial relationships among different geometric planes. To better exploit multi-dimensional information from panoramic images, recent studies Shen et al. (2023b; 2024) attempt to decouple layout estimation into multi-dimensional representations, such as horizon depth estimation, room height estimation, or plane segmentation.

Despite the significant progress, two critical challenges shown in Fig. 1 remain insufficiently addressed: severe object-induced occlusions and layout ambiguities caused by multiple competing layout hypotheses. Existing methods generally achieve accurate predictions in scenes with little or no occlusion. However, as the occlusion level increases, their performance deteriorates significantly. Another challenge arises from layout ambiguities caused by multiple plausible layout hypotheses. For certain special regions, particularly structures behind door openings and transparent or glass regions, it remains difficult to determine whether these areas belong to the interior space of the current room or should be regarded as perspective-visible regions. Consequently, multiple possible layout hypotheses may exist in these regions, making it challenging to identify the true room layout.



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.

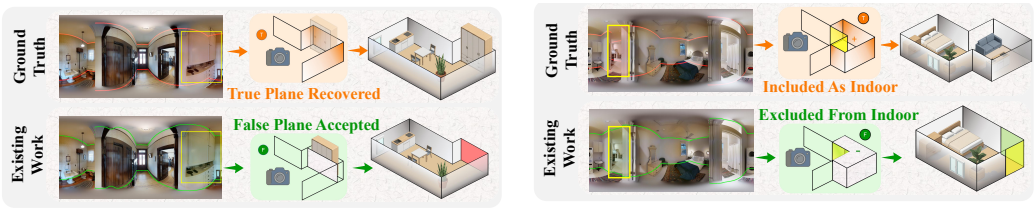


Figure 1: Illustration of two representative failure modes in panoramic room layout estimation. The first example demonstrates erroneous structural-plane identification caused by object-induced occlusion, while the second demonstrates perspective-induced layout ambiguity.

In this work, we jointly address object-induced occlusion and layout ambiguity by explicitly estimating a layout reliability field supervised by geometric depth evidence and semantic textual priors. Specifically, object-induced occlusion is characterized through scale-normalized depth consistency, whereas candidate-layout reasoning is performed via structured textual prompts coupled with ChatGPT-5.4, thereby modeling competing layout hypotheses. During inference, the predicted reliability field guides pixel-wise weighted depth fusion, thereby preserving reliable wall evidence. Extensive experiments on widely adopted datasets demonstrate that the proposed framework reduces the occurrence of low-quality failure cases and improves the robustness of panoramic room layout estimation in challenging indoor environments.

The contributions of this paper are as follows:

- We propose a panoramic room layout estimation framework that leverages a layout reliability field to explicitly suppress unreliable geometric evidence caused by object occlusion and layout ambiguity, reducing layout reconstruction failures in challenging panoramic scenes.
- We construct customized supervision labels for the layout reliability field to characterize occluded regions and ambiguous layout regions in panoramic images.
- We conduct extensive experiments on three widely used datasets. The results demonstrate that our method effectively reduces layout reconstruction failures under severe object occlusion and ambiguous layout hypotheses while maintaining high overall accuracy and robustness.

## 2 Related Work

Indoor room layout estimation from a single RGB panorama has attracted increasing attention in recent years. Mainstream panoramic layout estimation methods are built upon the Manhattan World assumption, which models indoor scenes as box-shaped structures where walls are mutually perpendicular and orthogonal to the floor. Based on this assumption, researchers have explored various layout representations and prediction paradigms, gradually forming several representative directions, including probability map estimation methods, 1D representation-based methods. depth-based geometric Decomposition methods, and multi-hypothesis reasoning methods.

**Probability Map Estimation Methods.** A representative work is LayoutNet Zou et al. (2018), which uses dense prediction to obtain corner boundary probability maps for subsequent layout estimation. In addition, some methods incorporate vanishing point estimation with score-based post-processing strategies to obtain more accurate structural boundary information.

**One-Dimensional Representation-Based Methods.** To reduce the computational cost of dense prediction, Sun *et al.* proposed HorizonNet Sun et al. (2019), which reformulates layout estimation as a 1D representation of the ceiling–wall and floor–wall boundaries and uses an LSTM to model spatial dependencies. HoHoNet Sun et al. (2021) subsequently extends this boundary representation to 1D features for multi-task learning. However, compressing scene information into a 1D representation weakens feature interactions across spatial dimensions, and the resulting information loss cannot be fully compensated for by methods such as Fourier transforms.

**Depth-Based Geometric Decomposition.** To better exploit 3D layout information, Wang *et al.* propose LED2-NetWang et al. (2021), which extends the 1D layout representation to a 2D depth representation, thereby improving the geometric plausibility of layout estimation. Subsequently, Shen

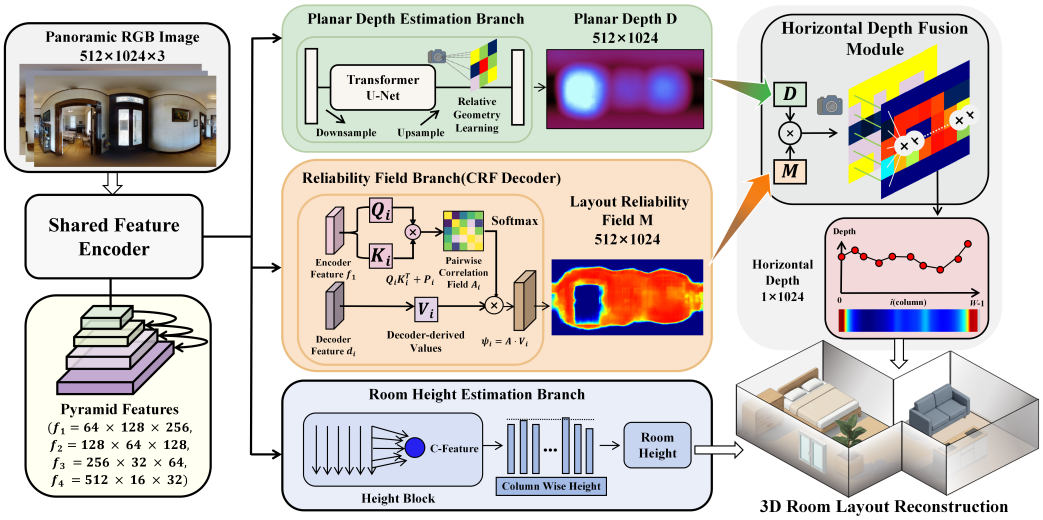


Figure 2: Overview of the proposed framework. A shared encoder predicts $D$, $M$, and $h$. The reliability field $M$ weights $D$ to recover the horizontal depth $\mathbf{d}$ for 3D layout reconstruction.

*et al.* introduce DOPNetShen et al. (2023b), which decomposes layout estimation into room-height and planar-depth estimation and recovers more accurate 3D layouts through multi-dimensional feature interaction. Building upon DOPNet, a follow-up work incorporates multi-view aggregation and camera parameter estimation to improve model generalizationShen et al. (2024).

**Multi-Hypothesis Reasoning Methods.** Seg2Reg Sun et al. (2024) reconstructs room layouts through a differentiable and occlusion-aware process. However, its output remains a one-dimensional representation, which may lead to the loss of spatial information across multiple dimensions. Tsai *et al.* propose the Bi-Layout framework Tsai et al. (2024), which introduces the Extended Layout and Enclosed Layout hypotheses and selects the hypothesis with the higher matching quality as the final result. However, Bi-Layout does not explicitly model the underlying causes of layout ambiguity or address object-induced occlusion.

Although existing methods have made progress in handling object occlusion and layout ambiguity, they do not jointly address these two issues or explicitly model their underlying factors Tabata et al. (2023). On the one hand, foreground furniture occlusion can introduce misleading planar cues and degrade layout reconstruction accuracy Danier et al. (2025); Moon et al. (2024). On the other hand, open regions may correspond to multiple plausible layout hypotheses, making their spatial ownership difficult to determine reliably from visual evidence alone. To address these issues, we propose a framework that explicitly models a layout reliability field to guide the network toward more reliable and structurally consistent layout representations.

## 3 METHOD

The proposed method follows a paradigm that decomposes the layout estimation task into horizontal depth estimation and room-height estimation. As show in Fig. 2, the multi-task framework comprises three branches: a planar depth preprocessing branch, a layout reliability field decoding branch, and a room-height decoding branch, all of which share a common encoder for feature extraction. The planar depth preprocessing branch provides dense geometric cues of the scene, while the layout reliability field decoding branch produces reliability representations under object occlusion and layout ambiguity. Their outputs are fed into a horizontal depth fusion module to obtain the horizontal depth. Meanwhile, the room-height decoding branch estimates the global room height. Finally, the horizontal depth and room height are jointly used to reconstruct the room layout.

### 3.1 PLANAR DEPTH PREPROCESSING BRANCH

We employ a Transformer-based U-Net architecture tailored to panoramic images for pixel-wise planar depth estimation Lin et al. (2026). Moreover, a relative geometry learning strategy are introduced to enhance the stability and generalization of depth representations.

**Relative Geometry Learning.** In this branch, the Transformer self-attention mechanism adopts the PSA module Shen et al. (2023a). To account for the imaging geometry of panoramic projections onto room planes, a positional encoding strategy tailored to planar depth estimation is introduced Lee et al. (2025); Piccinelli et al. (2025), as illustrated in Fig. 3. A residual-centered weighting scheme is introduced based on the local smoothness and consistency of depth within planar regions.

The weight of the central point is fixed to 1 to preserve the original observation, while the weights of its eight neighboring points are adaptively learned, thereby improving robustness to non-planar regions and noise. The formulation is defined as:

$$\hat{D}(\theta,\phi) = \frac{\sum_{i=-1}^{1}\sum_{j=-1}^{1} w_{(i,j)} D(\theta + i\Delta\theta, \phi + j\Delta\phi)}{\sum_{i=-1}^{1}\sum_{j=-1}^{1} w_{(i,j)}}, \tag{1}$$

where $D$ and $\hat{D}$ denote the initial and refined planar depths, $\theta$ and $\phi$ are the horizontal and vertical angles, and $w_{(i,j)}$ denotes the learnable weight, while the central weight is defined as $w_{(0,0)} = 1$. Since each point is constrained by its neighbors, after multi-module training, any two points become implicitly connected, allowing each point to encode both local neighborhood information and global contextual cues Ai & Wang (2024).

**Loss Function.** To emphasize relative geometric relationships rather than absolute metric depth, a median-normalized, scale- and shift-invariant depth loss inspired by Depth Anything V2 Yang et al. (2024); Wang et al. (2025) is adopted. The predicted and ground-truth depths are independently centered by their medians and normalized by their mean absolute deviations. This loss encourages the network to preserve consistent relative depth structures. Further implementation details are provided in the Appendix.

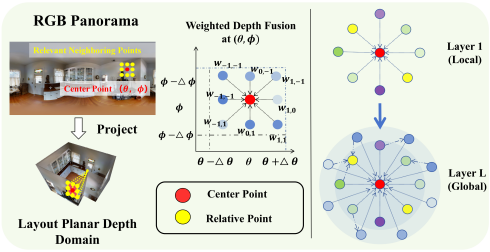


Figure 3: Layout-plane depth refinement through learnable aggregation of neighboring depths.

### 3.2 LAYOUT RELIABILITY FIELD BRANCH

The layout reliability field branch takes the multi-scale features extracted by the shared encoder as input and employs CRF decoder to produce the layout reliability field. To mitigate the susceptibility of implicit learning to shortcut features and appearance-based texture cues, dedicated reliability-field labels are constructed to provide explicit supervision, enabling the branch to learn representations of unreliable regions caused by object occlusion and layout ambiguity. Subsequently, the predicted reliability field is used to guide depth feature fusion for horizontal depth estimation.

**Layout Reliability Field Decoder.** The layout reliability field branch employs the window-based fully connected CRF decoder. Each CRF module takes the encoder feature $f_i$ at the corresponding scale ($i = 1, 2, 3, 4$) as the query $Q$ and key $K$, and the decoder feature $d_i$ as the value $V$. The pairwise correlation weights are obtained by computing $QK^{\mathsf{T}}$, adding the positional encoding $P$, and applying Softmax normalization. The resulting weights are then used to aggregate $V$ and generate the pairwise potential. The pairwise weight matrix is computed as follows:

$$A = \mathrm{SoftMax}(QK^{T} + P), \tag{2}$$

and the feature aggregation is formulated as: $\psi_p = A \cdot V$, which represents the pairwise potential of the neural window-based fully connected CRF. Subsequently, the outputs of the CRF modules at different scales are progressively upsampled and propagated to subsequent layers. Finally, the network generates a layout reliability field with the same spatial resolution as the input image, with values normalized to $[0, 1]$.

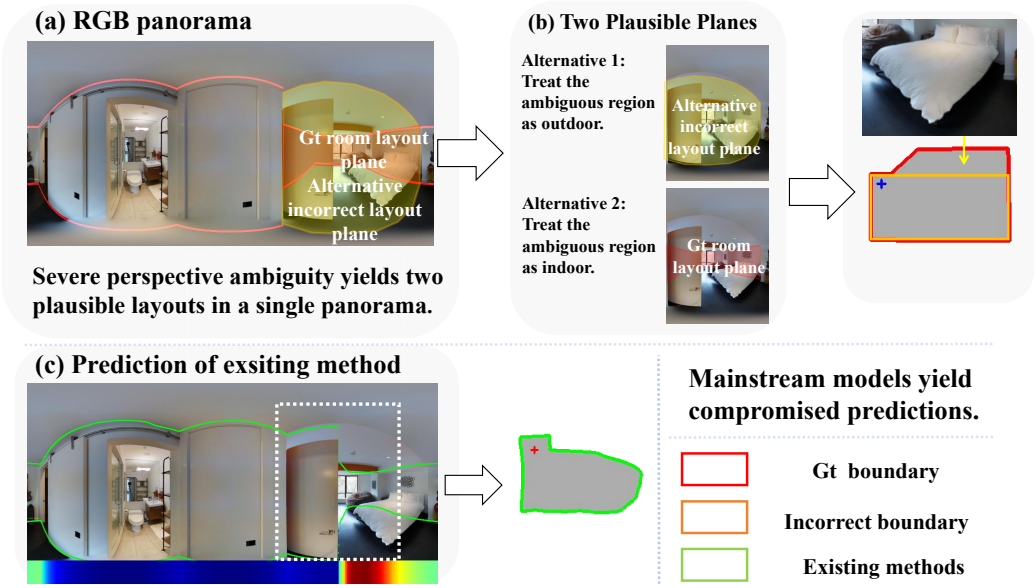


Figure 4: Mainstream models yield compromised predictions.

**Label Construction for the Layout Reliability Field.** Our preliminary experiments show that merely adding a reliability prediction branch without supervising it with labels yields no performance gain. Therefore, we construct explicit layout reliability field labels to supervise the layout reliability field decoder. The occlusion-aware field $M_1$ and the layout-ambiguity-aware field $M_2$ are constructed separately and then combined to form the layout reliability field label $M$.

To construct the occlusion-aware field label $M_1$, the mean relative error (MRE) between the occluding-object depth and the local planar depth is used to measure the occlusion extent and determine pixel-wise reliability. The MRE of each pixel is first used as the initial embedding feature and then mapped from Euclidean space into a unit Poincaré ball using the nonlinear scaling function $\rho = 1 - e^{(-k \cdot mre)}$, where $k$ is a scaling factor and $\rho \in [0, 1)$ denotes the coordinate of the embedded point in the Poincaré ball Cho & An (2026); Sur et al. (2025). In this construction, the origin of the Poincaré ball represents the most reliable geometric state, corresponding to zero depth inconsistency. Since the distance between two points u and v in the Poincaré ball is defined as $d_{(u,v)} = arcosh(1 + \frac{2||u-v||^2}{(1-||u||^2)(1-||v||^2)})$. The hyperbolic distance between the embedded point and the origin is computed as:

$$d_H = ln(\frac{1+\rho}{1-\rho}) = 2 \cdot arctanh(\rho). \tag{3}$$

We interpret the hyperbolic distance $d_H$ as pixel-wise uncertainty and geometric energy, and define the occlusion-aware field label according to the negative exponential relationship in energy-based models:

$$M_1 = e^{-\lambda d_H}, \tag{4}$$

where $\lambda$ controls the decay rate of the reliability field. We adopt the Poincaré-based formulation because its radial hyperbolic distance provides a nonlinear response that tolerates small inconsistencies while imposing increasingly stronger penalties on large geometric deviations. The detailed comparison results of the Poincaré transformation with the exponential and sigmoid methods are provided in the ablation study.

To construct the layout-ambiguity-aware field $M_2$, we introduce textual semantics as an auxiliary constraint Song et al. (2025). Layout ambiguity commonly occurs in open regions, such as doorways, windows, and transparent structures, as illustrated in Fig. 4. For each panorama, we manually construct a structured prompt consisting of a global room description and local landmark-object cues: *This is a [room type]. It contains [present landmarks] and does not contain [absent landmarks].* The global description provides a coarse room-level semantic prior, while the local description provides region-specific evidence for determining whether a perspective-visible region belongs to the target room Liang et al. (2026); Song et al. (2025).

The RGB panorama and the structured prompt are provided to ChatGPT-5.4 in an offline label-construction procedure. For each potentially ambiguous region $R_n$, the MLLM returns a region proposal, an evidence type $c_n \in \{\text{global}, \text{local}\}$, and a modulation score $r_n \in [0, 1]$. $R_n$ directly denotes the proposed ambiguous image region. Each proposal is manually verified against the target-room interpretation. An inconsistent proposal is rejected or regenerated using a more explicit local landmark description. After verification, we compute the region-level reliability value:

$$v_n = \begin{cases} 0.6 + 0.4\phi(r_n), & c_n = \text{global}, \\ 0.05 + 0.55\phi(r_n), & c_n = \text{local}, \end{cases} \qquad \phi(r) = \frac{1 - \cos(\pi r)}{2}. \tag{5}$$

To obtain a pixel-wise smoothly varying reliability field, we compute the distance of each interior pixel to the boundary of its corresponding ambiguous region. For each pixel $(i, j) \in R_n$, let $d_n(i, j) = \min_{(x,y)\in\partial R_n} \sqrt{(i-x)^2 + (j-y)^2}$ denote its distance to the region boundary. We use a cosine transition within an $\delta$-pixel boundary band, where $\delta = 8$:

$$W_n(i,j) = \begin{cases} \dfrac{1 - \cos(\pi d_n(i,j)/\delta)}{2}, & d_n(i,j) < \delta, \\ 1, & d_n(i,j) \geq \delta, \end{cases} \quad M_{2,n}(i,j) = \begin{cases} 1 - (1 - v_n)W_n(i,j), & (i,j) \in R_n, \\ 1, & (i,j) \notin R_n. \end{cases} \tag{6}$$

The final field is obtained by conservative aggregation, $M_2(i, j) = \min_n M_{2,n}(i, j)$. Circular horizontal padding is used when a region crosses the ERP seam. At inference, the network requires only the RGB panorama, without the MLLM or manual verification. Further details of the label-construction procedure are provided in the Appendix.

Finally, the overall layout reliability field $M$ is constructed by performing pixel-wise multiplication of $M_1$ and $M_2$: $M = M_1 \odot M_2$. Through pixel-wise multiplication, only regions less affected by occlusion and layout ambiguity retain high responses. Horizontal depth fusion primarily relies on the distances of wall regions. Therefore, the reliability values in wall regions are preserved, whereas those in the ceiling and floor regions are set to $1 \times 10^{-3}$ to suppress interference caused by blur, low illumination, and missing textures.

**Loss Function.** To accommodate varying reliability distributions caused by different levels of occlusion and ambiguity across panoramic scenes, an image-adaptive weighted binary cross-entropy loss is adopted to dynamically balance positive and negative samples within each image. Given the target $t_i$ and prediction $p_i$ at pixel $i$, with $N = HW$, the mean and variance of the target reliability distribution are first computed as follows:

$$\rho = \frac{1}{N}\sum_{i=1}^{N} t_i, \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(t_i - \rho)^2. \tag{7}$$

Here, $\rho$ represents the mean reliability, while $\sigma^2$ measures the mixture of reliable and unreliable regions. Based on these statistics, the dynamic weights are defined as:

$$w_{\text{pos}} = (1 - \rho)(1 + \sigma^2), \qquad w_{\text{neg}} = \rho(1 + \sigma^2). \tag{8}$$

The reliability loss is:

$$L_2 = -\frac{1}{N}\sum_{i=1}^{N}\Big[w_{\text{pos}}\, t_i \log(p_i) + w_{\text{neg}}\,(1 - t_i)\log(1 - p_i)\Big]. \tag{9}$$

### 3.3 Room Height Estimation Branch

The room-height estimation branch takes the multi-scale features $f_1$, $f_2$, $f_3$, $f_4$ extracted by the shared encoder as input and performs column-wise compression followed by multi-scale feature fusion to predict column-wise room-height estimates $\{h_i\}_{i=0}^{1023}$. Owing to the approximate vertical symmetry of indoor scenes in panoramic images, the regressed room-height information remains reliable even when the visually inferred ceiling and floor positions in individual columns are biased. Finally, the height predictions from all columns are averaged to obtain the global room height.

### 3.4 Horizontal Depth Fusion Module

The horizontal depth fusion module combines the planar depth map $D \in R^{H\times W}$ with the layout reliability field $M \in R^{H\times W}$ to recover the horizontal depth. In an ERP panorama, pixels in the

same column share the same horizontal angle $\theta$, while their depth estimates may be affected by object occlusion and layout ambiguity. Therefore, the reliability field is used as pixel-wise weights to perform soft voting over the depth values in each column, allowing reliable pixels to contribute more while suppressing the influence of unreliable depth estimates. The planar depth $d_{\text{plane}}$ for the i-th column is obtained as follows:

$$d_{\text{planei}} = \frac{1}{\sum_{j=1}^{H} m_{ij}} \sum_{j=1}^{H} d_{\text{ij}} \cdot m_{\text{ij}} \cdot cos(\frac{2j - H - 1}{2H} \cdot \pi), \tag{10}$$

$d_{\text{plane}}$ is the horizontal depth of the i-th column in the panoramic image, $d_{\text{ij}}$ is the depth value at position (i,j) in $D$, and $m_{\text{ij}}$ is the reliability field at position (i,j) in $M$.

**Loss Function.** The horizontal-depth prediction is supervised by a composite loss $L_3$, which combines a distance-aware regression term $L_{dplane}$ and a global shape-consistency term $L_{shape}$:

$$L_3 = L_{dplane} + L_{shape}. \tag{11}$$

The former penalizes column-wise horizontal-depth errors, while the latter preserves the global shape of the predicted depth sequence. Their detailed formulations are provided in Appendix.

# 4 EXPERIMENTS

The proposed method is compared with existing state-of-the-art (SOTA) approaches on three public datasets. All experiments are conducted on a single NVIDIA RTX 4090 GPU with a batch size of 4. The model is implemented in PyTorch and optimized using the Adam optimizer Kingma & Ba (2015), with an initial learning rate of $1 \times 10^{-4}$.

The proposed model adopts a two-stage multi-task training strategy, with implementation details provided in Appendix. Performance is evaluated using 2D IoU, 3D IoU, RMSE, CE, PE and the threshold accuracies $\delta_1$, $\delta_2$, and $\delta_3$ at $1.25$, $1.25^2$, and $1.25^3$.

## 4.1 COMPARISON STUDY

**Results on Structure3d Dataset.** This large-scale CAD-synthetic dataset contains over 30,000 panoramic images. For each scene, panoramas with three occlusion levels Empty, Simple, and Full are provided together with depth, semantic, and layout annotations. We follow the official training, validation, and test splits. To evaluate occlusion robustness, we compare our method with HorizonNet Sun et al. (2019) on the Simple and Full panoramas.

Tables 1 and 2 present the quantitative comparison results under the Simple and Full occlusion settings, respectively, while Fig. 5 provides qualitative comparisons with HorizonNet. The results demonstrate that our method accurately identifies the true structural boundaries even when severe occlusions mislead the network into incorrect plane predictions, while maintaining stable and accurate performance as the occlusion level increases.

**Results on MatterportLayout Dataset.** The MatterportLayout dataset contains 2,295 panoramic images, covering both regular rectangular rooms and complex rooms with multiple corners, as well as numerous cases involving object occlusion and layout ambiguity. We follow the official data splits for comparative experiments, with the quantitative results reported in Table 3. Visualizations of the outputs from individual modules on MatterportLayout are provided in Appendix.

Fig. 6 presents qualitative comparisons with HorizonNet Sun et al. (2019) and DOPNet Shen et al. (2023b). The results demonstrate that our method accurately recovers room layouts in occluded regions, scenes with distant boundaries and complex structures, and regions admitting multiple plausible layout interpretations.

**Results on Stanford2D-3D Dataset.** The Stanford2D-3D dataset contains 552 panoramic images collected from Stanford2D3DS and annotated by Zou *et al.* Zou et al. (2018). Due to its limited size, some methods, such as DOPNet Shen et al. (2023b), jointly train on Stanford2D-3D and PanoContext Zhang et al. (2014). In contrast, our method achieves 87.28% 3DIoU using only the Stanford2D-3D training set, which is competitive with methods trained using additional PanoContext data.

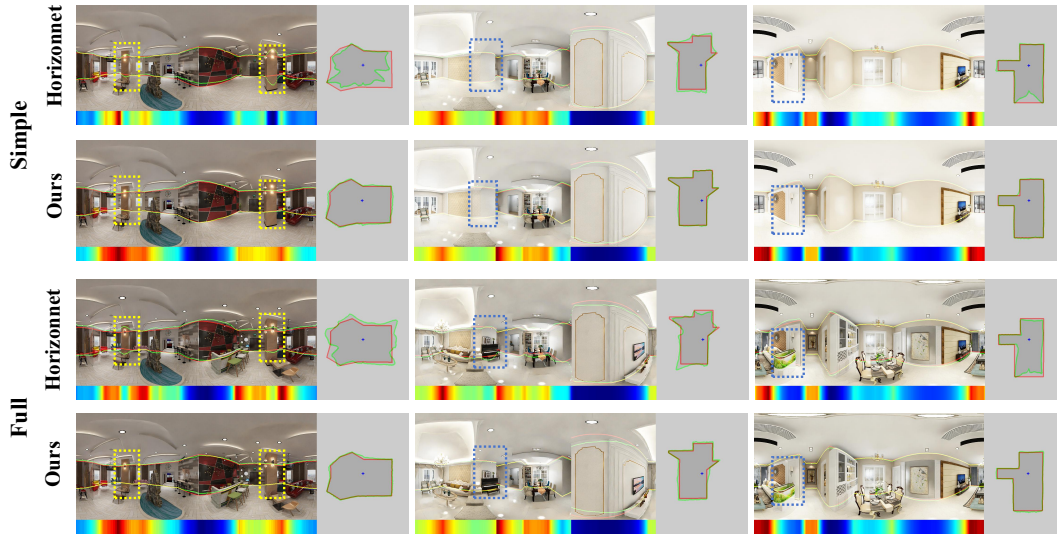


Figure 5: Qualitative comparison with HorizonNet under different occlusion levels on Structured3D. Green and red denote predictions and ground truth, respectively. Yellow and blue boxes highlight boundary recovery and robust reconstruction under severe occlusions, respectively.

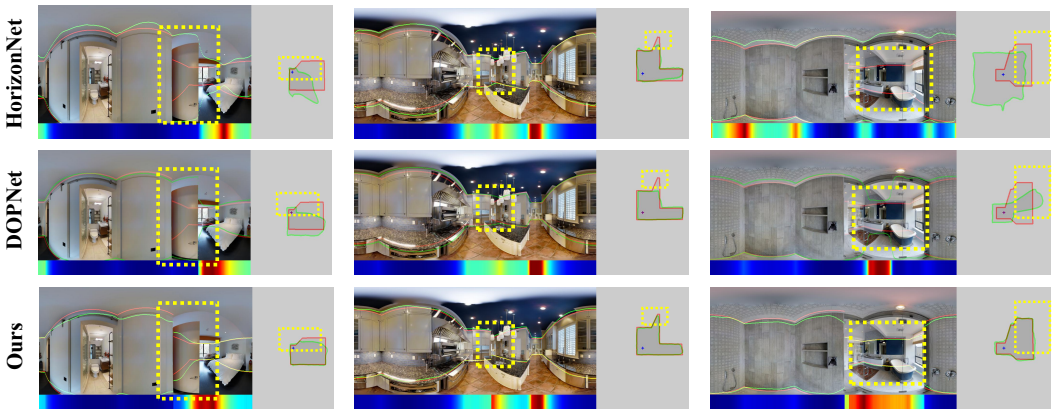


Figure 6: Qualitative comparison of the proposed method with HorizonNet and DOPNet on the MatterportLayout dataset. Green and red denote predictions and ground truth, respectively. Dashed boxes highlight differences caused by severe occlusions or layout ambiguities.

Since constructing our layout reliability-field labels requires raw depth information, we further pretrain the model on Matterport3D, which provides more diverse room layouts and richer scene information. With this additional pretraining, the complete framework achieves 90.58% 3DIoU. Table 4 presents quantitative comparisons between our method and several state-of-the-art approaches.

### 4.2 Ablation Study

To analyze the effects of different construction strategies and hyperparameter settings for the occlusion-aware field $M_1$, we compare the exponential-based mapping, the sigmoid-based mapping, and the Poincaré-embedded mapping. The exponential-based mapping adopts: $M_1 = \exp(-\sigma \cdot \mathrm{mre})$, the sigmoid-based strategy directly maps the MRE to pixel-wise reliability using: $M_1 = \alpha \cdot \frac{1}{1+e^{\beta(mre-\tau)}}$, the Poincaré-embedded strategy instead maps the MRE into hyperbolic space, as described in the Method section. For a fair comparison, the final reliability field is consistently constructed as $M = M_1 \times M_2$ in all experiments. The optimal prompt configuration is fixed for $M_2$, while all other training and evaluation settings remain unchanged. Detailed results are provided in the Appendix.

Table 1: Quantitative comparison on the Structure3D simple dataset.

| Method | 2DIoU (%) ↑ | 3DIoU (%) ↑ | RMSE ↓ | $\delta_1$ ↑ |
|---|---|---|---|---|
| HorizonNet | 86.37 | 84.58 | 0.1781 | 0.9324 |
| LGT-Net | 87.45 | 86.45 | 0.1525 | 0.9226 |
| LED2-Net | 87.81 | 86.63 | 0.1427 | 0.9358 |
| Dop-Net | 88.12 | 86.96 | 0.1346 | 0.9401 |
| **Ours** | **88.60** | **87.57** | **0.1189** | **0.9483** |

Table 2: Quantitative comparison on the Structure3D full dataset.

| Method | 2DIoU (%) ↑ | 3DIoU (%) ↑ | RMSE ↓ | $\delta_1$ ↑ |
|---|---|---|---|---|
| HorizonNet | 85.82 | 83.45 | 0.1929 | 0.9105 |
| LGT-Net | 86.45 | 83.89 | 0.1842 | 0.9187 |
| LED2-Net | 86.63 | 83.26 | 0.1524 | 0.9224 |
| Dop-Net | 87.86 | 84.92 | 0.1221 | 0.9373 |
| **Ours** | **88.23** | **86.41** | **0.1178** | **0.9392** |

Table 3: Quantitative comparison on MatterportLayout dataset.

| Method | 2DIoU (%) ↑ | 3DIoU (%) ↑ | RMSE ↓ | $\delta_1$ ↑ |
|---|---|---|---|---|
| LayoutNetv2 | 78.73 | 75.82 | 0.258 | 0.871 |
| Dula-Netv2 | 78.82 | 75.05 | 0.291 | 0.818 |
| HorizonNet | 81.71 | 79.11 | 0.197 | 0.929 |
| AtlantaNet | 82.09 | 80.02 | - | - |
| LGT-Net | 83.52 | 81.11 | 0.204 | 0.951 |
| LED2-Net | 82.61 | 80.14 | 0.207 | 0.94 |
| Dop-Net | 84.11 | 81.70 | 0.197 | 0.950 |
| Bi-Layout | 85.10 | 82.57 | - | - |
| Seg2Reg | 85.27 | 82.83 | - | - |
| **Ours** | **86.05** | **84.66** | **0.184** | **0.934** |

Table 4: Quantitative comparison on Stanford 2D-3D dataset.

| Method | 3DIoU(%)↑ | CE↓ | PE↓ |
|---|---|---|---|
| LayoutNetv2 | 82.66 | 0.83 | 2.59 |
| Dula-Netv2 | 86.60 | 0.67 | 2.48 |
| HorizonNet | 82.72 | 0.69 | 2.27 |
| AtlantaNet | 83.94 | 0.71 | 2.18 |
| LGT-Net | 86.03 | 0.63 | 2.11 |
| DOPNet | 85.58 | 0.66 | 2.10 |
| Seg2Reg | 87.24 | - | - |
| **Ours** | **87.28** | **0.62** | **2.03** |
| **Ours[w/pre-train]** | **90.58** | **0.58** | **1.99** |

To validate the contribution of $M_2$ to layout estimation, we compare $M = M_1$ with $M = M_1 \times M_2$ on MatterportLayout. As shown in Table 5, incorporating $M_2$ generated from the simple prompt consistently improves all metrics.

We further investigate three schemes Muturi et al. (2024) for constructing the layout-ambiguity-aware field $M_2$: (1) Global Description Only, (2) Local Ambiguous Objects Only, and (3) Global Description + Local Ambiguous Objects. The corresponding $M_2$ fields are quantitatively evaluated in Table 6. For a fair comparison, all experiments use the optimal Poincaré embedding configuration determined by hyperparameter analysis Nickel & Kiela (2017) to construct $M_1$.

The results show that the $M_2$ reliability field constructed by combining global scene descriptions with key object information in ambiguous regions achieves the best performance, demonstrating the complementarity of these semantic cues and the effectiveness of their joint modeling for layout estimation.

Table 5: Quantitative comparison of panoramic layout estimation w/wo layout ambiguity field ($M_2$) supervision on the MatterportLayout dataset.

| $M$ Strategy | 2DIoU(%)↑ | 3DIoU(%)↑ | RMSE↓ | $\delta_1$↑ |
|---|---|---|---|---|
| $M = M_1$ | 81.64 | 79.87 | 0.2301 | 0.8923 |
| $M = M_1 \times M_2$ | 83.25 | 81.82 | 0.2038 | 0.9135 |

Table 6: Quantitative comparison of panoramic layout estimation using different layout ambiguity field ($M_2$) construction strategies on the MatterportLayout dataset.

| $M_2$ Strategy | 2DIoU(%)↑ | 3DIoU(%)↑ | RMSE↓ | $\delta_1$↑ |
|---|---|---|---|---|
| Ours[G] | 83.25 | 81.82 | 0.2038 | 0.9135 |
| Ours[LAO] | 84.89 | 82.47 | 0.1959 | 0.9246 |
| Ours[G + LAO] | 86.05 | 84.66 | 0.1839 | 0.9339 |

## 5 CONCLUSION

This paper addresses the challenges of object occlusion and layout ambiguity in panoramic layout estimation and proposes an end-to-end multi-task learning framework for indoor room layout recovery. In addition, we manually construct layout reliability field supervision labels that explicitly model both occlusion and layout ambiguity, aiming to alleviate the low-quality reconstruction problem caused by these factors. Experimental results show that the proposed supervised reliability learning mechanism significantly improves the network's capability in complex scenes. Particularly, for panoramic images with severe occlusions and strong layout ambiguities, our method can accurately identify occlusion information and distinguish the correct spatial structure from multiple plausible layout interpretations.

## 6 AI USE STATEMENT

AI tools were used only to assist with and polish the manuscript's writing. The authors remain fully responsible for the content, analysis, and conclusions.

## REFERENCES

Hao Ai and Lin Wang. Elite360d: Towards efficient 360 depth estimation via semantic- and distance-aware bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9926–9935, 2024.

J. Cho and J. An. OASIS-DC: Generalizable depth completion via output-level alignment of sparse-integrated monocular pseudo depth. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2026.

Duolikun Danier, Mehmet Aygün, Changjian Li, Hakan Bilen, and Oisin Mac Aodha. Depthcues: Evaluating monocular depth perception in large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20049–20059, 2025.

Zhengyu Jiang, Zhiyi Xiang, J. Xu, and M. Zhao. LGT-Net: Indoor panoramic room layout estimation with geometry-aware transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

J. Lee, B. Solarte, Chia-Hung Wu, Jui-Cheng Jhang, Fu-En Wang, Yi-Hsuan Tsai, and Min Sun. uLayout: Unified room layout estimation for perspective and panoramic images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025.

Huizhi Liang, Yichao Shen, Yu Deng, Sicheng Xu, ZhiYuan Feng, Tong Zhang, Yaobo Liang, and Jiaolong Yang. HiSpatial: Taming hierarchical 3d spatial understanding in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2502–2514, 2026.

Xin Lin, Meixi Song, Dizhe Zhang, Wenxuan Lu, Haodong Li, Bo Du, Ming-Hsuan Yang, Truong Nguyen, and Lu Qi. Depth any panoramas: A foundation model for panoramic depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26835–26844, 2026.

Zekai Lin and Xu Zheng. PanoEnv: Exploring 3d spatial intelligence in panoramic environments with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9647–9657, 2026.

Jaeho Moon, Juan Luis Gonzalez Bello, Byeongjun Kwon, and Munchurl Kim. From-ground-to-objects: Coarse-to-fine self-supervised monocular depth estimation of dynamic objects with ground contact prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10519–10529, 2024.

T. W. Muturi, B. A. Kyem, J. K. Asamoah, N. J. Owor, R. Dyzinela, A. Danyo, Y. Adu-Gyamfi, and A. Aboah. Prompt-guided spatial understanding with RGB-D transformers for fine-grained object relation reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, 2017.

Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unik3d: Universal camera monocular 3d estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1028–1039, 2025.

Giovanni Pintore, Marco Agus, and Enrico Gobbetti. AtlantaNet: Inferring the 3D indoor layout from a single 360 image beyond the manhattan world assumption. In *Proceedings of the European Conference on Computer Vision*, 2020.

Alessio Quercia, Erenus Yildiz, Zhuo Cao, Kai Krajsek, Abigail Morrison, Ira Assent, and Hanno Scharr. Enhancing monocular depth estimation with multi-source auxiliary tasks. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 6435–6445, 2025.

Zhijie Shen, C. Lin, L. Nie, K. Liao, and Y. Zhao. PanoFormer: Panorama transformer for indoor 360-degree depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023a.

Zhijie Shen, Z. Zheng, C. Lin, L. Nie, K. Liao, S. Zheng, and Y. Zhao. Disentangling orthogonal planes for indoor panoramic room layout estimation with cross-scale distortion awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023b.

Zhijie Shen, C. Lin, J. Zhang, L. Nie, and K. Liao. 360 layout estimation via orthogonal planes disentanglement and multi-view geometric consistency perception. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15768–15780, 2025.

Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. HorizonNet: Learning room layout with 1D representation and pano stretch data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Cheng Sun, Min Sun, and Hwann-Tzong Chen. HoHoNet: 360 indoor holistic understanding with latent horizontal features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Cheng Sun, Wei-En Tai, Yu-Lin Shih, Kuan-Wei Chen, Yong-Jing Syu, Kent Selwyn The, Yu-Chiang Frank Wang, and Hwann-Tzong Chen. Seg2reg: Differentiable 2d segmentation to 1d regression rendering for 360 room layout reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10435–10445, 2024.

Tanuj Sur, Samrat Mukherjee, Kaizer Rahaman, Subhasis Chaudhuri, Muhammad Haris Khan, and Biplab Banerjee. Hyperbolic uncertainty-aware few-shot incremental point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11810–11821, 2025.

M. Tabata, K. Kurata, and J. Tamamatsu. Shape-Net: Room layout estimation from panoramic images robust to occlusion using knowledge distillation with 3D shapes as additional inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

Yi-Hsuan Tsai, Chia-Hao Chen, Yu-Cheng Lin, and Yung-Yu Chuang. No more ambiguity in 360-degree room layout via bi-layout estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. LED2-Net: Monocular 360-degree layout estimation via differentiable depth rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

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

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Zhao, Jiashi Xu, Hengshuang Zhao, et al. Depth anything V2. In *Advances in Neural Information Processing Systems*, 2024.

Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. PanoContext: A whole-room 3D context model for panoramic scene understanding. In *Proceedings of the European Conference on Computer Vision*, 2014.

Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. LayoutNet: Reconstructing the 3D room layout from a single RGB image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

# A Appendix

## A.1 Loss for the Planar Depth Preprocessing Branch

This section describes the loss function used for the planar depth preprocessing branch. Since layout estimation relies more on relative depth relationships than on precise metric depth, we focus on preserving global shape consistency, thereby providing reliable geometric cues for the subsequent fusion of depth information and the layout reliability field to recover horizontal depth. Considering the varying spatial scales of indoor scenes, we adopt a median-based normalization strategy inspired by Depth Anything V2 Yang et al. (2024). The loss function is defined as follows:

$$\begin{aligned} L_1 &= \frac{1}{HW}\sum_{i=1}^{HW}\rho(\hat{d}_i^*, \hat{d}_i) \\ \rho(\hat{d}_i^*, \hat{d}_i) &= |\hat{d}_i^* - \hat{d}_i| \\ \hat{d}_i &= \frac{d_i - t(d)}{s(d)} \\ t(d) = \mathrm{median}(d), \quad s(d) &= \frac{1}{HW}\sum_{i=1}^{HW}|d_i - t(d)| \end{aligned} \tag{12}$$

Here, $d_i^*$ and $d_i$ denote the predicted and ground-truth depth values, respectively. The predicted and ground-truth depth maps are independently processed using the same median-based normalization defined above, yielding the normalized depths $\hat{d}_i^*$ and $\hat{d}_i$, respectively.

## A.2 Loss for the Horizontal-depth

The horizontal-depth loss consists of a distance-aware regression loss $L_{dplane}$ and a global shape-consistency loss $L_{shape}$. For the output $d_{\text{planei}}$ ($i = 0, 1, 2, \ldots, 1023$), the distance-aware regression loss is computed as:

$$L_{dplane} = \frac{|d_{plane}^{pred} - d_{plane}^{gt}| \cdot d_{plane}^{gt}}{\sum (d_{plane}^{gt})^2}. \tag{13}$$

The global shape-consistency loss constrains the overall shape of the predicted horizontal-depth sequence by measuring its cosine similarity with the ground truth:

$$L_{shape} = 1 - \frac{d_{pred} \cdot d_{gt}}{||d_{pred}||||d_{gt}||}. \tag{14}$$

The final horizontal-depth loss is obtained by combining the two terms:

$$L_3 = L_{dplane} + L_{shape}. \tag{15}$$

## A.3 Training strategy

The proposed model employs a two-stage multi-task training strategy. The planar-depth estimation branch is first pretrained to initialize the shared encoder, after which all branches are jointly optimized with task weights $\alpha_1 = 0.05$, $\alpha_2 = 1$, and $\alpha_3 = 1$ for planar-depth estimation, layout reliability field prediction, and room height estimation, respectively. The fused output of the

first two branches is further supervised by an auxiliary horizontal-depth regression loss with weight $\lambda_{aux} = 1$ Quercia et al. (2025).

### A.4 Implementation Details of the Layout-Ambiguity-Aware Field $M_2$

The layout-ambiguity-aware field $M_2$ is constructed through an offline, MLLM-assisted label-generation procedure. The central objective is to determine whether the regions visible through doorways, windows, transparent surfaces, or other open structures should be incorporated into the target room layout. Unlike ordinary wall regions, these regions may admit two plausible interpretations: they may either belong to the current room or correspond to an adjacent space that should be excluded.

The construction procedure consists of four stages: reliability-field initialization, global and local semantic reasoning, ambiguous-region identification and verification., and continuous reliability assignment. The MLLM is first informed of the semantic interpretation rules and reliability-mapping functions. The RGB panorama and its structured prompt are then jointly provided to the MLLM to identify regions whose reliability should be reduced. After the identified regions and semantic judgments are manually verified, their inconsistency scores are converted into pixel-wise reliability values according to predefined continuous functions.

**Reliability-Field Initialization.** For an input panorama $I \in \mathbb{R}^{H\times W\times 3}$, the layout-ambiguity-aware field is initialized as:

$$M_2^{(0)}(i,j) = 1, \qquad 1 \le i \le H, \quad 1 \le j \le W. \tag{16}$$

This initialization assumes that every pixel provides reliable layout evidence.

**Global and Local Semantic Prompts.** The structured prompt contains two complementary components: a global scene description and a local landmark-object description. The global description specifies the semantic category of the current room, whereas the local description provides targeted evidence for deciding whether a particular open or perspective-visible region should be included in the target layout.

The global prompt follows the template:

> *This is a [room type].*

For example, the prompt *This is a bedroom* provides a global semantic prior regarding the objects and functional regions expected to appear in the current room. If a visible region contains visual content that is clearly incompatible with a bedroom, the region is considered less likely to belong to the target room and is therefore assigned a moderately reduced reliability value.

The local prompt follows the template:

> *It contains [present landmark objects] and does not contain [absent landmark objects].*

The local landmark-object cues are constructed according to whether a visible region should be incorporated into the target layout. If an extended region should be included in the target room, a representative object located in that region is added to the positive description, e.g., *It contains a bed.* This informs the MLLM that the region containing the object should be incorporated into the ground-truth layout and should therefore retain a high reliability value.

Conversely, if an extended region should be excluded from the target layout, a representative object located in that region is added to the negative description, e.g., *It does not contain a washing machine.* The MLLM then regards the region containing that object as not belonging to the ground-truth layout and assigns it a lower reliability value.

In this manner, the global description performs coarse scene-level semantic screening, while the local landmark-object description provides a strong region-specific constraint for resolving layout ambiguity.

**Prior Instruction for MLLM-Based Reasoning.** Before processing the RGB panorama and its structured prompt, the MLLM is provided with a fixed instruction that specifies the objective and the reliability-assignment rules. The instruction informs the MLLM of the following principles:

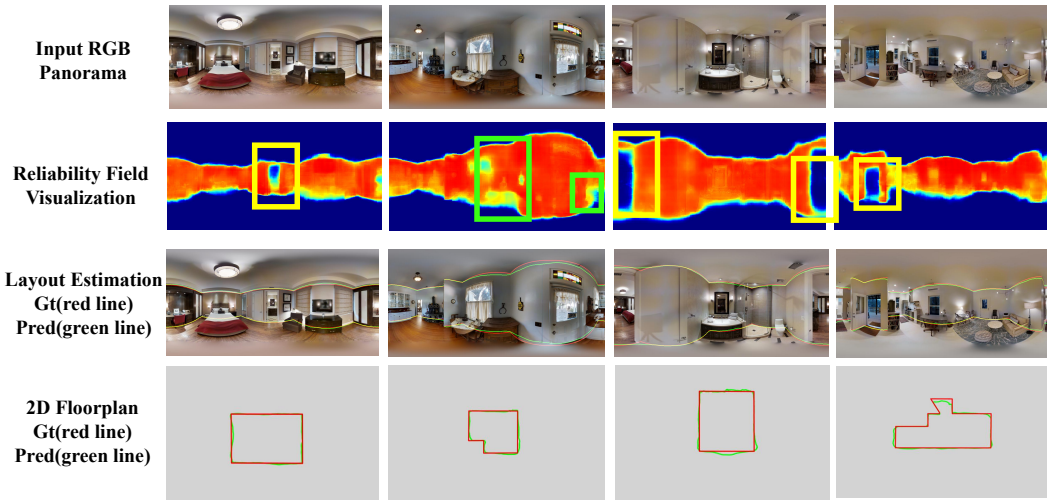


Figure 7: Qualitative visualization of the proposed method on representative MatterportLayout samples. From top to bottom, the rows show the input RGB panoramas, the predicted layout reliability fields, the estimated ceiling–wall and floor–wall boundaries in the panoramic view, and the reconstructed 2D floorplans in the top-down view. Green and yellow boxes in the predicted reliability fields highlight representative regions affected by object-induced occlusion and layout ambiguity, respectively. In the panoramic layout-estimation and top-down floorplan views, red and green lines denote the ground-truth and predicted layouts, respectively.

1. The initial reliability value of every pixel is one.
2. The MLLM outputs its assessment of the layout-ambiguous regions, which is subsequently manually reviewed to determine whether the assessment is reasonable. that are potentially inconsistent with the semantic category of the current room.
3. The local landmark-object description is used to determine whether a specific ambiguous region should be incorporated into the target room layout.
4. Regions identified only through global semantic inconsistency receive a moderate reliability reduction.
5. Regions explicitly rejected by local landmark-object evidence receive a stronger reliability reduction.
6. Spatially abrupt binary assignments should be avoided, and the output should preserve gradual reliability transitions.

Under the above instructions, the RGB panorama and its corresponding structured prompt are jointly provided to ChatGPT-5.4. The MLLM first visualizes, in the image plane, the regions $R_n$ whose reliability values should be reduced. We then manually inspect the visualization of $R_n$ generated by the MLLM to verify whether the model has correctly identified the layout-ambiguous regions.

**Ambiguous-Region Identification and Verification.** For each input panorama, the MLLM first identifies regions that may involve layout ambiguity. For each identified region $R_n$, the MLLM returns a structured region proposal containing a visualization of the region, a suggested spatial mask or boundary coordinates, the type of semantic evidence, and a region-level modulation score.

Let

$$c_n \in \{\text{global}, \text{local}\} \tag{17}$$

denote the type of semantic evidence associated with region $R_n$, indicating whether the region is identified according to the global scene description or the local landmark-object description introduced above. The MLLM also assigns a region-level modulation score:

$$r_n \in [0, 1], \tag{18}$$

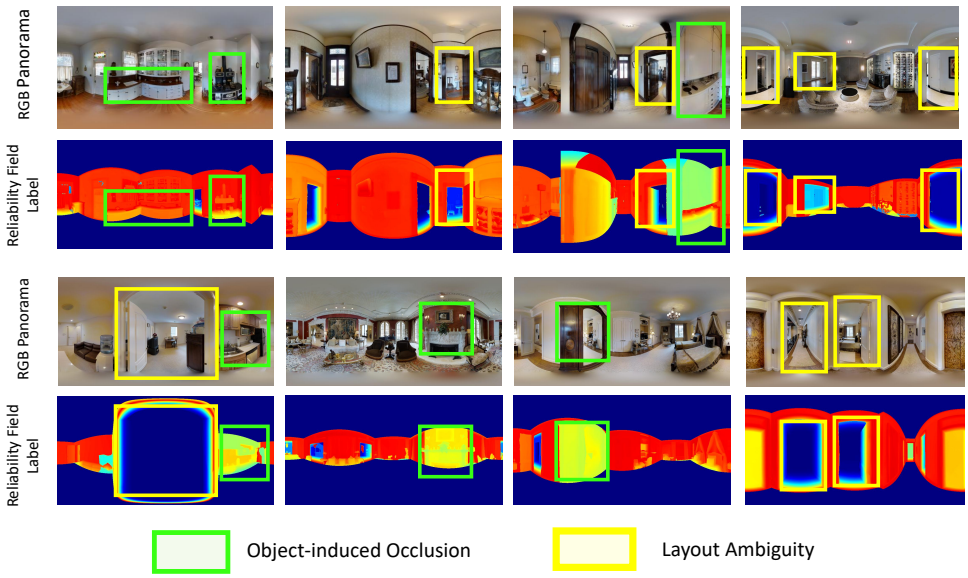


Figure 8: Visualization of layout reliability field supervision labels on representative Matterport-Layout samples. Each RGB panorama is paired with its corresponding layout reliability field label. Green boxes highlight regions affected by object-induced occlusion, which are assigned lower reliability values. Yellow boxes indicate ambiguous regions that admit multiple plausible layout interpretations, where candidate regions inconsistent with the ground-truth room layout are assigned lower reliability values.

The score $r_n$ is not interpreted as the final absolute reliability of $R_n$. Instead, it modulates the reliability-reduction strength within the range prescribed by the corresponding semantic-evidence type $c_n$. A smaller $r_n$ indicates a stronger reliability reduction, whereas a larger $r_n$ indicates a weaker reduction within the type-specific range. Since $R_n$ has already been identified as a layout-ambiguous region, its final reliability is further constrained by whether it is supported by global or local semantic evidence.

Before generating the final reliability field $M_2$, we manually inspect the regions $R_n$ returned by the MLLM. The verification focuses on whether the model's decision to include or exclude each region is consistent with the target room annotation. If the MLLM output is inconsistent with the scene content or the target layout, the response is rejected, and the MLLM is queried again using a more explicit local description while retaining the same reliability-assignment rules. This verification process is performed only during offline label generation and is not required during network training or inference.

**Global-Local Reliability Mapping.** After the region $R_n$ predicted by the MLLM has been manually verified as reasonable, we use the MLLM output and its corresponding modulation score $r_n$ as the initial reliability assignment. However, directly thresholding these continuous values into binary assignments may introduce abrupt discontinuities into the reliability field, making the resulting representation unsuitable for the decoder adopted in this work. We therefore define the following smooth transition function:

$$\phi(r) = \frac{1 - \cos(\pi r)}{2}, \qquad r \in [0, 1]. \tag{19}$$

Since $\phi(0) = 0$ and $\phi(1) = 1$, this function provides a smooth transition of the modulation scores between the two endpoints.

For a region identified through global scene semantics, the mapped reliability value is defined as:

$$v_{\text{global}}(r_n) = 0.6 + 0.4\phi(r_n). \tag{20}$$

This mapping introduces a moderate reliability reduction because the global room category provides only coarse semantic evidence and may not be sufficient to determine the spatial ownership of a specific open region.

For a region explicitly identified for reliability reduction through local landmark-object evidence, the mapped reliability value is defined as:

$$v_{\mathrm{local}}(r_n) = 0.05 + 0.55\phi(r_n). \tag{21}$$

This mapping produces a stronger reliability reduction because the presence or absence of a distinctive landmark object provides more direct evidence for determining whether the corresponding ambiguous region belongs to the target room.

The target reliability value of region $R_n$ is therefore computed as:

$$v_n = \begin{cases} v_{\mathrm{global}}(r_n), & c_n = \mathrm{global}, \\ v_{\mathrm{local}}(r_n), & c_n = \mathrm{local}. \end{cases} \tag{22}$$

Regions supported by positive local landmark-object evidence are considered part of the target room and are not subjected to the local reliability-reduction mapping. Their reliability values remain at the initial value of one.

**Pixel-Wise Field Generation.** The layout-ambiguity-aware field is initialized as an all-one field. For each verified ambiguous region $R_n$, its interior is assigned the mapped region-level reliability value $v_n$, while pixels outside the region retain the initial value of one. To avoid an abrupt transition from one to $v_n$ at the region boundary, we introduce a narrow boundary band and apply local smoothing only within this band.

For an interior pixel $(i, j) \in R_n$, its distance to the region boundary is defined as:

$$d_n(i, j) = \min_{(x,y)\in\partial R_n} \sqrt{(i - x)^2 + (j - y)^2}, \tag{23}$$

where $\partial R_n$ denotes the boundary of $R_n$. For regions crossing the horizontal ERP seam, we apply circular padding along the horizontal direction before computing the boundary distance. The central ERP interval is retained after smoothing.

For each interior pixel $(i, j) \in R_n$, we define the local transition weight as:

$$W_n(i, j) = \begin{cases} \dfrac{1 - \cos\left(\pi d_n(i, j)/\delta\right)}{2}, & d_n(i, j) < \delta, \\ 1, & d_n(i, j) \geq \delta. \end{cases} \tag{24}$$

Here, $\delta$ controls the width of the locally smoothed boundary band and is fixed to 8 pixels in all experiments. The transition weight is zero at the region boundary and gradually increases to one toward the region interior. Pixels located more than $\delta$ pixels from the boundary retain a weight of one. Pixels outside $R_n$ are not modified and retain their initial reliability value of one.

The reliability field associated with region $R_n$ is extended to the complete ERP image domain as:

$$M_{2,n}(i, j) = \begin{cases} 1 - (1 - v_n)\, W_n(i, j), & (i, j) \in R_n, \\ 1, & (i, j) \notin R_n. \end{cases} \tag{25}$$

Accordingly, pixels outside $R_n$ retain their initial reliability value of one. Within the boundary transition band, the reliability value gradually decreases from one to $v_n$, whereas pixels in the interior of $R_n$ are assigned the region-level reliability value $v_n$. This operation modifies the reliability values within the verified ambiguous region while restricting the smooth transition to a narrow neighborhood inside its boundary.

When multiple verified regions overlap, the final layout-ambiguity-aware field is obtained by retaining the minimum reliability response:

$$M_2(i, j) = \min_{1\leq n\leq N} M_{2,n}(i, j), \tag{26}$$

where $N$ denotes the number of verified layout-ambiguous regions in the panorama. This conservative aggregation ensures that a stronger local reliability reduction is not overridden by a weaker global semantic assessment. Pixels not covered by any verified region retain their initial reliability value of one.

Table 7: Quantitative comparison of panoramic layout estimation using different parameter settings of the exponential-based $M_1$ on the MatterportLayout dataset.

| **exponential-Based** | **2DIoU(%)↑** | **3DIoU(%)↑** | **RMSE↓** | $\delta_1$↑ |
|---|---|---|---|---|
| $\sigma = 0.5$ | 82.18 | 79.92 | 0.2326 | 0.8993 |
| $\sigma = 1.0$ | 82.35 | 80.42 | 0.2305 | 0.9028 |
| $\sigma = 1.5$ | 82.73 | 80.56 | 0.2258 | 0.9056 |
| $\sigma = 2.0$ | 81.92 | 79.41 | 0.2351 | 0.8965 |

Table 8: Quantitative comparison of panoramic layout estimation using different parameter settings of the sigmoid-based $M_1$ on the MatterportLayout dataset.

| **Sigmoid-Based** | **2DIoU(%)↑** | **3DIoU(%)↑** | **RMSE↓** | $\delta_1$↑ |
|---|---|---|---|---|
| $\alpha = 1.5, \beta = 2, \tau = 0.2$ | 83.25 | 81.05 | 0.2251 | 0.9083 |
| $\alpha = 1.8, \beta = 1, \tau = 0.2$ | 82.93 | 81.23 | 0.2245 | 0.9097 |
| $\alpha = 1.5, \beta = 2, \tau = 0.1$ | 84.24 | 82.85 | 0.2128 | 0.9102 |
| $\alpha = 1.8, \beta = 1, \tau = 0.1$ | 84.46 | 82.73 | 0.2158 | 0.9136 |

The layout-ambiguity-aware field $M_2$ is thereby obtained and is subsequently used to construct the complete supervision label. The complete label is generated by multiplying the occlusion-aware field $M_1$ and the layout-ambiguity-aware field $M_2$ pixel-wise:

$$M = M_1 \odot M_2, \tag{27}$$

where $\odot$ denotes pixel-wise multiplication. Consequently, only regions that are reliable with respect to both object-induced occlusion and layout ambiguity retain high response values. Since the layout reliability field $M$ is subsequently used in the soft-voting strategy for horizontal-depth regression, we preserve the reliability values in wall regions while setting those in the ceiling and floor regions to $1 \times 10^{-3}$. This suppresses interference caused by image blur, low illumination, and missing textures.

### A.5 Additional Qualitative Results on MatterportLayout

The visualization results of our method on the MatterportLayout dataset are presented in Fig. 7. As shown, our method produces layout estimates that are highly consistent with the ground truth for both regular cuboid rooms and non-cuboid rooms with multiple corners. Notably, the layout reliability fields shown in the second row demonstrate the effectiveness of our design: the network assigns lower reliability values to regions affected by object-induced occlusion and layout ambiguity.

### A.6 Visualization for Layout Reliability Field Label

To provide an intuitive understanding of the constructed layout reliability field supervision labels, Fig. 8 presents representative examples from the MatterportLayout dataset. Each RGB panorama is paired with its corresponding layout reliability field label. The green boxes highlight regions affected by object-induced occlusion. Since foreground objects in these regions may introduce misleading geometric evidence, they are assigned lower reliability values. The yellow boxes highlight layout-ambiguous regions that admit multiple plausible layout interpretations. Within these regions, candidate areas that are inconsistent with the ground-truth room layout are assigned lower reliability values, whereas areas consistent with the target layout retain higher responses. These visualizations demonstrate that the constructed labels explicitly characterize unreliable regions caused by both object-induced occlusion and layout ambiguity.

### A.7 Additional Ablation Studies on the Construction of $M_1$

As a supplement to the comparison of the $M_1$ construction strategies and parameter settings presented in the main paper, we further investigate the effects of different occlusion-aware field construction schemes and hyperparameter configurations on model performance.

Table 9: Quantitative comparison of panoramic layout estimation using different parameter settings of the Poincaré-embedded $M_1$ on the MatterportLayout dataset.

| **Poincaré-embedded** | **2DIoU(%)↑** | **3DIoU(%)↑** | **RMSE↓** | $\delta_1$↑ |
|---|---|---|---|---|
| $k = 0.5, \lambda = 1$ | 82.58 | 80.24 | 0.2247 | 0.9105 |
| $k = 0.6, \lambda = 1$ | 83.36 | 81.61 | 0.2251 | 0.9086 |
| $k = 0.7, \lambda = 1$ | 83.75 | 81.82 | 0.2204 | 0.9057 |
| $k = 0.8, \lambda = 1$ | 83.89 | 82.03 | 0.2187 | 0.9035 |
| $k = 0.9, \lambda = 0.5$ | 85.24 | 83.75 | 0.1992 | 0.9237 |
| $k = 0.9, \lambda = 1$ | 85.72 | 84.23 | 0.1983 | 0.9213 |
| $k = 0.9, \lambda = 1.5$ | 86.05 | 84.66 | 0.1839 | 0.9339 |
| $k = 0.9, \lambda = 2$ | 85.12 | 83.94 | 0.2025 | 0.9182 |

The first strategy uses a simple exponential decay function:

$$M_1 = \exp(-\sigma \cdot \mathrm{MRE}), \tag{28}$$

where $\sigma > 0$ controls the decay rate. This mapping assigns the maximum reliability to pixels with zero depth inconsistency and monotonically decreases the reliability as the MRE increases.

The second strategy employs a Sigmoid-like empirical function, It is formulated as follows:

$$M_1 = \alpha \cdot \frac{1}{1 + e^{\beta(mre-\tau)}} \tag{29}$$

The third category is based on the recently introduced Poincaré ball embedding framework. The detailed implementation has been described in the Method section.

To ensure a fair comparison between the three $M_1$ construction strategies, the final reliability field is constructed as $M = M_1 \odot M_2$ in all experiments. The layout ambiguity reliability $M_2$ is fixed to the optimal prompt construction strategy identified in the previous ablation study. Subsequently, hyperparameter sensitivity analyses are conducted for both schemes to quantitatively evaluate the influence of different hyperparameter settings on the final layout estimation performance. The corresponding results are presented in Tables 7–9.

The experimental results demonstrate that the Poincaré-embedded strategy achieves better overall performance than the Exponential-based and Sigmoid-based mapping. Accordingly, we adopt $k = 0.9$ and $\lambda = 1.5$ as the default configuration for constructing $M_1$ in all experiments.