arXiv:2611.30219v1 [cs.CV] 23 Nov 2026

ROOMTOKENS：面向室内场景合成的结构化令牌掩码扩散

Xiaohao Sun[1] Angel X. Chang[1,2]
[1]西蒙菲莎大学[2]阿尔伯塔机器智能研究所
(A
mii)https://3dlg-hcvc.github.io/RoomToken/

摘要

生成三维室内场景需要一种能够捕捉物体语义、实例身份、高度结构和朝向，同时遵循房间几何结构的表示方法。将场景建模为物体向量序列的方法能够隐式学习空间约束，因而将碰撞和边界违规留给损失函数或后处理处理；语义地图方法虽然明确表示占用情况，却会丢失实例和方向结构。我们提出 RoomTokens，这是一种切片式结构化令牌表示，其中每个令牌都存储语义类别、实例颜色标识和朝向字段。该表示使场景能够以离散网格的形式生成，同时保留实例分组、朝向估计和三维范围预测所需的线索。为学习 RoomToken 分布，我们提出 MDIRT（RoomTokens掩码扩散），一种掩码离散扩散模型，通过共享的跨切片主干网络联合对语义／实例颜色／朝向通道进行去噪。建筑标记被钳制为明确的平面图约束，因此模型生成的家具布局既遵循房间几何结构，又保留实例和物体属性结构。在 3D-FRONT 上的实验表明，与基于序列／向量和基于地图的方法相比，结合 MDIRT 的 RoomTokens 提升了布局保真度和物理一致性，说明结构化空间令牌是离散三维场景合成的一种实用选择。

1 引言

自动生成三维场景日益重要，因为它能够扩展用于三维场景理解和具身人工智能的训练数据，并为增强现实、游戏和娱乐等应用提供内容。组合式三维室内场景生成（Tam 等人，2026b）将这一问题表述为物体布局问题：我们必须在遵循房间几何形状的同时，确定物体类别、各个物体的身份、位置摆放和朝向。为确保生成场景合理可信，同时建模空间结构和物体级语义至关重要。

大多数现有方法将场景表示为物体向量集合或序列（Wang 和 Liu，2026；Lin 和 Mu，2024），其中每个物体向量都编码了该物体的类别和包围框。虽然这种表示很紧凑，但它没有明确表示占用约束，也无法确保生成的两个包围框不会占据同一空间。因此，生成场景中可能出现物体间碰撞、墙体穿透，以及物体和房间区域周围的通行间隙不足等问题。因此，先前的系统会加入辅助机制，例如在训练目标中加入相交惩罚（Tang 等人，2024）、对无效插入进行拒绝式重采样（Ritchie 等人，2019），或在初步预测后进行基于优化的布局细化（Luo 等人，2020）。另一类工作（Feng 等人，2023；Yang 等人，2025；Xia 等人，2026）利用大语言模型/视觉语言模型中编码的广泛知识，以获得更丰富的语义先验，但其坐标输出仍需经过几何清理，并可能需要借助视觉评估器进行多轮处理，才能确保场景在物理上合理可信。

基于语义地图生成场景的方法（Yang 等人，2021；Sun 等人，2026）通过为密集网格中的单元分配语义标签，明确捕捉空间布局。近期工作 SemLayoutDiff（Sun 等人，2026）展示了在平面图条件下进行生成时，碰撞和越界物体更少。然而，单一的自上而下语义

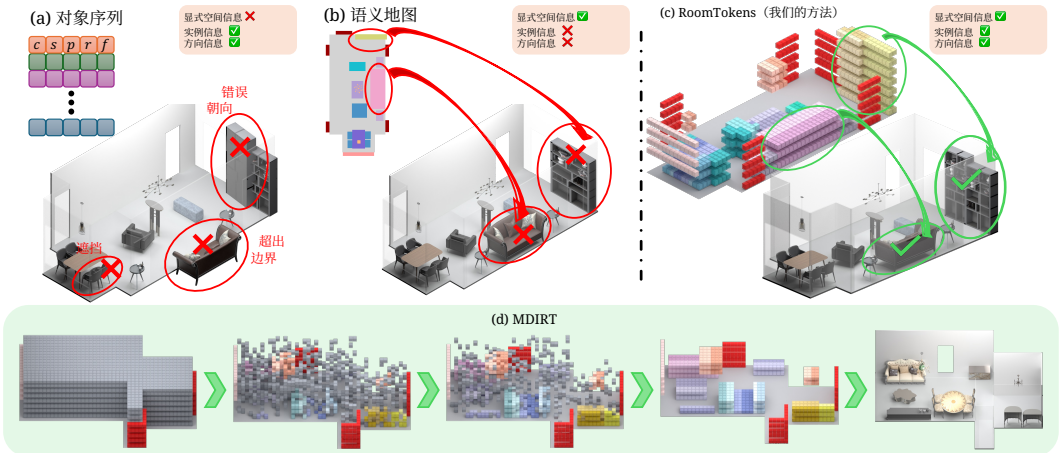

图1：上图。室内场景生成中场景表示的比较。对象序列（a）编码了对象属性，但缺少明确的空间关系信息。语义地图（b）明确呈现了空间结构，但缺少垂直、实例和方向信息。我们提出 RoomTokens （ c ），通过在切片空间网格中编码语义占用、实例身份和朝向线索，弥补这些缺失的信息。下图。我们提出 MDIRT（d），这是一种基于 RoomTokens 的掩码离散扩散模型。该模型在固定建筑标记的同时，对缺失的令牌字段进行去噪，使生成场景符合房间边界。

地图无法捕捉三维场景生成所需的重要信息。它会压缩垂直信息，在连通分量解码过程中合并类别相同且相邻的物体，也无法提供用于估计对象方向的方向线索。在这项工作中，我们着手解决如何将实例级和垂直信息引入这些空间地图这一难题。

我们提出**RoomTokens**，这是一种用于室内场景的分片式结构化令牌表示。我们将房间表示为由多个高度感知网格堆叠而成的结构，每个网格单元都存储一个 RoomToken，其中包含三个字段，分别表示语义类别、实例身份和朝向线索。实例身份的编码颇具挑战，因为为对象分配特定 ID 是任意的（即可以置换或重新分配），而且场景中的对象数量并不固定。由于我们的目标是在平面图上区分空间上相邻的对象实例（而非提供实例 ID），因此可以使用四种“颜色”（遵循四色定理（Robertson 等人，1997））来区分不同实例。对于方向，朝向字段编码家具标记占据对象主体部分还是前部；非家具标记则不使用该字段。我们发现，这两种家具标签足以解码朝向。这些属性共同提供了充足的信息，使模型能够生成离散的空间布局，同时保留实例分组、方向估计和三维范围预测所需的线索。

为学习 RoomToken 分布，我们提出 **MDiRT** (RoomTokens掩码扩散)，这是一种用于室内场景生成的单阶段掩码离散扩散模型。每个词元属性都被嵌入为语义、实例颜色或朝向通道。这些通道由共享的跨切片 UNet 处理，并联合解码。去噪过程将掩蔽的 RoomTokens 映射为结构化的三维场景布局，同时以平面图中的建筑标记为条件并对其进行钳制（图 1 (底部)），因此家具标记会在建筑条件所定义的有效内部区域内生成。不同于以往工作（Sun 等人，2026）先预测语义、再预测实例级属性的做法，这种多通道参数化通过单一反向过程中的共享特征，学习语义、实例颜色和朝向之间的依赖关系。在 3D-FRONT（Fu etal.，2021a）上的实验表明，与 SemLayoutDiff (Sun 等人，2026）相比，MDIRT 生成的布局更合理。生成布局与参考房间更加吻合（FID 为 50.14，对比 71.18），并且能够捕捉到更合理的物体关系（成对关系为 12.09，对比 20.65）。此外，生成布局中的碰撞更少（COL 为 12.64，对比 16.24），门和窗户也更少受到遮挡（物体覆盖率为 82.31，对比 56.02）。

总而言之，我们的贡献如下：1)**一种切片式结构化 RoomToken 表示**，在每个考虑高度的网格词元中存储语义类别、图着色实例标识和朝向信息。2)**一种单阶段掩码离散扩散模型 MDiRT，适用于 RoomTokens** ，该模型在整个去噪过程中固定建筑条件，从而确保生成的家具位于有效的房间

内部，同时利用共享的跨切片主干网络联合建模语义／实例颜色／朝向通道。3)一**项实证分析**表明，显式的实例颜色和朝向场能够提升实例恢复与朝向估计效果，并生成符合物理合理性的场景布局。

2 相关工作

对象向量与关系场景合成。许多室内场景合成方法将房间表示为一组或一系列物体，并通过属性指定物体的语义类别、位置和朝向。早期的神经网络和基于图的方法学习对象排列的空间先验（Wang 等，2018；Ritchie 等，2019；Li 等，2019；Wang 等，2019；Zhang 等，2020），而近期的自回归方法和扩散方法则直接预测物体属性（Wang 等，2021；Paschalidou 等，2021；Sun 等，2024b；Tang 等，2024；Hu 等，2026；Maillard 等，2024；Sun 等，2025b）。其他近期工作利用场景图、关系三元组或混合图扩散，以更好地建模对象布局（Lin 和 Mu，2024；Zhai 等，2023；2024；Bai 等，2025；Ruiz 等，2025），并采用物理约束或人体感知约束目标及优化方法（Yi 等，2023；Yang 等，2024a；Sun 等，2024a；Su 等，2025；Lin 等，2026）。这些方法在条件化和先验方面各不相同，但最终生成的布局都是一组对象边界框。因此，生成场景中可能出现对象碰撞和边界重叠，还需要额外的后处理，因为这种表示本身无法避免这些问题。MDIRT 则生成离散占用网格，其中每个单元格仅对应一个语义分配，因此网格级重叠是显式的。通过在高度感知切片上显式建模物体占用情况，该表示有助于减少物体间碰撞。

大语言模型与基于智能体的场景生成。借助大语言模型和视觉语言模型，可以根据语言输入进行条件控制，并利用这些预训练模型中存储的关于物体摆放方式的常识知识。LayoutGPT（Feng 等，2023）利用语言模型预测布局；Holodeck（Yang 等，2024b）、AnyHome（Wen 等，2024）、SceneTeller（Ocal 等，2024）以及开放世界系统（Aguina-Kang 等，2024）则将文本提示与物体检索、约束求解或场景组装相结合。近期的大语言模型/视觉语言模型及智能体方法采用程序合成、视觉语言反馈或多智能体优化，以提升整体可控性（Hu 等人，2024；Celen 等，2024；Sun 等人，2025a；Yang 等，2025；Pfaff 等，2026；Gao 等，2025）。这些系统提供了灵活的交互界面，但对于密集的室内布局，坐标级或程序级输出仍需经过几何清理。我们的研究重点与之互补：我们从带标注的 3D-FRONT 房间中学习布局分布，并使表示本身包含局部占用情况、实例和朝向结构。

用于场景合成的离散空间表示。 离散空间表示通过直接为二维或三维晶格上的位置分配标签来编码场景，使占用约束具有局部性且明确。 Yang 等人 (2021) 使用体素标签，Lee 等人 (2023) 对场景尺度的类别化三维数据进行建模，而 SemLayoutDiff（Sun 等人，2026）将多项离散扩散应用于自上而下语义图。这一系列方法与我们的方法最为接近，因为它将生成过程从无序的边界框转向了明确的空间场。然而，纯语义空间网格无法区分相互接触的同类别实例，俯视图会压缩垂直结构，并且每个单元格中的单一类别标签无法直接提供方向线索。我们的表示保留了网格的优势，同时将每个切片单元升级为结构化的 RoomToken，其中包含语义、实例颜色和朝向字段。预处理过程中使用图着色来分配实例颜色，而四色定理（Appel 和 Haken，1977；Robertson 等人，1997）给出了分隔相邻同类别实例所需的最少颜色数。类似的着色思想已在判别式实例分割中得到研究（Kulikov 等人，2018；Zhang 等人，2025），但作为生成式建模的紧凑实例通道，仍基本未得到探索。

离散扩散与掩码令牌建模。掩码离散扩散通过将词元替换为吸收掩码状态来破坏类别数据，并训练去噪器逆转这一过程。多项式与掩码变体（Hoogeboom 等人，2021；Shi 等，2024；Lou 等人，2024；Sahoo 等人，2024；Nie 等人，2025）已在离散序列和令牌网格上展现出良好效果。布局专用扩散模型也会使用离散变量或离散-连续混合变量来表示结构化二维布局和室内场景属性（Chai 等人，2023；Inoue 等人，2023；Levi 等人，2023；Hui 等人，2023；Choi 等人，2026），但这些变量通常被组织为布局令牌或对象属性，而非逐单元场景占用场。我们将掩码离散扩散从

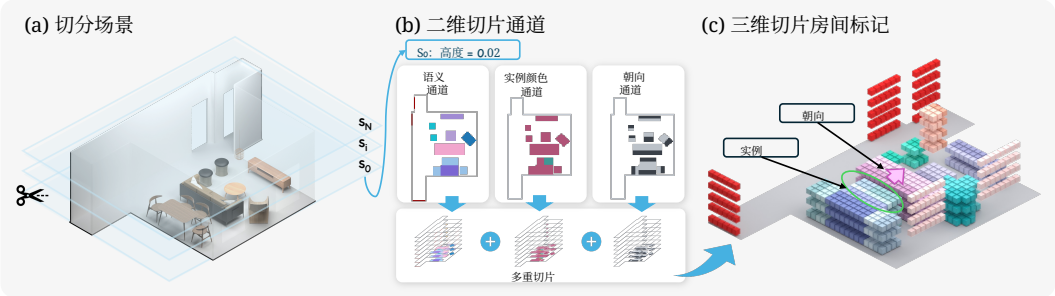

图 2：RoomToken 数据表示。(a) 三维场景表示为在选定垂直高度处截取的水平切片。(b) 每个切片编码三个对齐的二维通道：语义类别、实例颜色标识和朝向字段。将所有高度上的这些通道堆叠起来，便形成切片标记体。(c) 由堆叠切片构成的三维表示（RoomTokens）存储语义、实例和方向信息，高度则由切片索引确定。

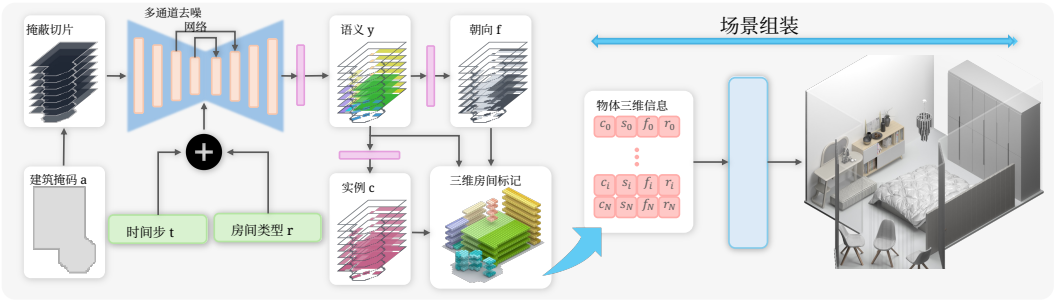

图 3：MDIRT 概览。给定选定的高度切片和建筑掩码，我们固定房间结构，并使用带有跨切片注意力的共享 UNet 对掩码室内房间标记进行去噪。去噪器在语义门控有效性约束下预测语义切片、实例颜色切片和朝向切片：建筑/背景标记不包含实例颜色或朝向线索，而家具标记包含这些信息。随后，将有效房间标记网格归组为对象记录，并利用其类别、尺寸和姿态检索并放置匹配的 3D资产。

单一语义地图或对象属性词元化扩展至结构化 RoomTokens。MDIRT 在训练中采用多通道参数化，在有效性约束和平面图条件下，同时对语义、实例颜色和朝向值进行去噪。

3 方法

给定房间类型 r 和输入的建筑掩码（建筑掩码），MDIRT 会生成完整的对象级 3D 布局（图 3）。借助我们的 RoomToken 表示，不同高度的切片存储带有实例和朝向线索的语义占用情况，使得我们能够从网格词元中恢复物体（第 3.1 节）。我们使用掩码扩散来生成室内 RoomTokens，同时保持建筑标记不变（第 3.2 节）。在每个反向步骤中，多通道去噪网络 g_θ会预测各高度切片中的掩蔽字段（第 3.3 节）。最后，在场景组装过程中，词元网格会被转换为对象记录，并检索匹配资产以构建最终场景（第 3.4 节）。

3.1 切片式 RoomTokens

为了从离散占用图中恢复对象实例及其姿态，我们将每个房间编码为由切片式 RoomTokens 组成的堆栈。我们将地面平面离散化为一个 $H \times W$ 网格，并使用 S 选定的高度切片 $\mathcal{Z} = \{\zeta_1, \ldots, \zeta_S\}$ 来捕捉垂直范围（图 2）。我们通过对训练集中的家具高度区间进行贪心覆盖来选择这些高度：如果对象的垂直区间与候选高度 h 相交 $[h-\tau, h+\tau]$，则该对象覆盖该候选高度，其中 τ=0.10m 与栅格化容差相匹配。从固定的地板和天花板锚点开始，我们反复添加能够覆盖最多此前未覆盖对象的候选高度。我们在第 4.1节中给出具体的网格尺寸、切片高度及其他参数。

我们通过 $i=(u,\ v,\ s)$ 对词元进行索引，其中 $u\in\{1,\dots,H\}$、 $v\in\{1,\dots,W\}$ 和 $s\in\{1,\dots,S\}$。每个网格位置存储一个词元 $\mathbf{x}_i=(y_i,\ c_i,\ f_i)$，其包含语义类别 y_i、实例颜色标签 c_i 和朝向标签 f_i。语义标签使用 K_{sem} 类别。实例颜色标签要么是 C实例颜色之一，要么是无值 $\varnothing_c$；朝向标签可以是 body 、 front 或无值 $\varnothing_f$。这三个值必须与语义类别一致。令 Y_{arch} 表示建筑 / 背景类别，令 Y_{furn} 表示家具类别。只有满足以下条件时，RoomToken $\mathbf{x}_i=(y_i,\ c_i,\ f_i)$ 才有效：

$$\text{valid}(y_i,c_i,f_i)\iff\begin{cases}c_i=\varnothing_c \text{ and } f_i=\varnothing_f, & y_i\in Y_{\text{arch}},\\ c_i\in\{0,\dots,C-1\} \text{ and } f_i\in\{\text{body},\text{front}\}, & y_i\in Y_{\text{furn}}.\end{cases}\tag{1}$$

This 该规则将实例和朝向线索与语义占用关联起来：建筑/背景标记携带 y no a属性，而家具标记同时携带实例颜色和朝向标签。

3.2 房间词元上的掩码扩散

由于建筑掩码已指定房间边界，扩散过程应仅生成未知的内部区域，同时保持已观测到的建筑结构不变。为此，我们采用掩码扩散（Shi 等人，2024）。与逐渐向输入添加噪声的常规扩散模型不同，掩码扩散会在前向过程中逐步用掩码词元替换并隐藏内部的 RoomToken 标签。反向过程见图1 下方。

前向过程。我们将前向掩码过程定义为：从干净的 RoomToken 单元$\mathbf{x}$ 到部分被掩蔽的 $\mathbf{z}_t$ ，在时间步 t 反复采样，以决定保留原始词元还是将其替换为掩码词元。对于我们的三通道 RoomToken $\mathbf{x}_i=(y_i,\ c_i,\ f_i)$，我们将使用三个掩码词元 $[\text{M}]_y$，$[\text{M}]_c$，$[\text{M}]_f$。我们从保留概率为$\alpha(t)$的伯努利分布中，分别独立抽取语义标签、实例颜色标签和朝向标签的保留指示变量 m_i、 n_i： $m_i,\ n_i\sim\text{Bernoulli}(\alpha(t))$。仅当语义标签可见时，实例颜色和朝向标签才会显示。指示变量 $a_i=1$ 标记固定的建筑单元，而 $a_i=0$ 标记待生成的单元。对于 $\rho\in\{c,f\}$，令 $x_i^c=c_i$ 且 $x_i^f=f_i$。前向掩码过程 $q_t(\mathbf{z}_t\mid\mathbf{x},\ \mathbf{a})$ 为

$$z_{t,i}^{y}=\begin{cases}y_i, & \text{if } a_i=1 \text{ or } m_i=1,\\ [\text{M}]_y, & \text{otherwise,}\end{cases}\qquad z_{t,i}^{\rho}=\begin{cases}\varnothing_\rho, & \text{if } a_i=1,\\ x_i^\rho, & \text{if } a_i=0 \text{ and } m_i=n_i=1,\\ [\text{M}]_\rho, & \text{otherwise.}\end{cases}\tag{2}$$

公式（2）会在未固定单元格的语义标签被掩蔽时，同时掩蔽这两个属性，无论其真实类别为何。采样时，所有未固定单元格的三个字段均初始化为掩蔽状态，而给定的建筑结构保持固定。

反向过程。反向过程在相同约束下，从时间 t 逐步推进至更早的去噪时间 $s<t$ ，以填充被掩蔽的词元。我们以 $\kappa=(r,\ \mathbf{a})$为条件，其中 r 是房间类型，$\mathbf{a}$是建筑掩码指示变量。给定部分掩蔽的网格 $\mathbf{z}_t$、时间 t和条件κ，网络 $g_\theta(\mathbf{z}_t,\ t,\ \kappa)$会预测语义、实例颜色和朝向场的标签概率。我们利用这些概率对候选样本 $\hat{}\mathbf{z}_s$进行采样。随后，我们恢复固定的建筑掩码词元，并将生成的词元投影至有效房间词元支持集：

$$\mathbf{z}_{s,i}=\begin{cases}\mathbf{x}_i, & a_i=1,\\ \Pi_{\text{valid}}(\hat{\mathbf{z}}_{s,i}), & a_i=0.\end{cases}\tag{3}$$

此处$\hat{}\mathbf{z}_s$ 表示投影前的模型提案，而 Π_{valid} 则在保留未解析掩码的同时，确保满足公式（1）中的有效性规则。该投影仅移除无效的语义—属性组合；不会合并实例，也不会改变位置摆放。我们在附录 B.中分析了去噪过程中投影前后的词元有效性。

在去噪过程中，每个反向步骤内都会先更新语义标签，然后根据更新后的语义状态重新计算实例颜色和朝向预测。已经揭示的标签会保留。如果语义标签处于掩蔽状态，则实例颜色和朝向属性也保持掩蔽状态。当建筑/背景被揭示时，这些属性会设为空值。总之，掩码值表示未知值，而空值表示不适用的属性。这种语义优先采样发生在一个反向过程中。详情请参阅附录 F 。

3.3 多通道去噪网络

我们训练了一个多通道去噪网络 g_θ，用于预测被掩蔽词元的标签概率。该模型 g_θ 由一个共享的 UNet 组成 U_θ，它会处理各高度切片上求和后的场嵌入；此外还有三个预测头（D_y、D_c、D_f），分别对应语义、实例颜色和朝向场。在每个单元格 i处，三个输出头会生成逻辑值 $\boldsymbol{\ell}_i^y$、$\boldsymbol{\ell}_i^c$、$\boldsymbol{\ell}_i^f$，这些逻辑值定义了第 3.2 节中使用的标签概率。

联合 RoomToken 去噪。主要的去噪步骤基于单个共享隐藏表示 $\mathbf{h}_\theta$预测 RoomToken 值，并以房间类型 r 和建筑掩码 $\mathbf{a}$为条件。对于通道嵌入 E_y, E_c, E_f、共享 UNet U_θ和输出头 D_y, D_c, D_f，去噪图在时间 t 、位置 i 处的嵌入词元 $\mathbf{e}_{t,i}$ 定义如下：

$$\begin{aligned} \mathbf{e}_{t,i} &= E_y(z^y_{t,i}) + E_c(z^c_{t,i}) + E_f(z^f_{t,i}), \quad \text{for each grid location } i, \\ \mathbf{h}_\theta &= U_\theta(\mathbf{e}_t; t, \kappa), \quad \boldsymbol{\ell}^\rho_i = D_\rho(\mathbf{h}_{\theta,i}), \quad \rho \in \{y, c, f\}. \end{aligned} \tag{4}$$

这里，$\mathbf{h}_{\theta,i}$ 表示隐藏特征，$\boldsymbol{\ell}^\rho_i$ 表示通道 ρ的逻辑值，而 $\kappa = (r,\mathbf{a})$表示条件输入。干净的 $\mathbf{x}_i$ 和带噪的 $\mathbf{z}_{t,i}$ 变量是离散的分类状态；$\mathbf{e}_{t,i}$ 和 $\mathbf{h}_{\theta,i}$ 是连续向量，三个解码器将共享向量映射为分类通道逻辑值。求和后的嵌入和共享隐藏状态共同参数化一个 RoomToken 分布，而非采用先语义后属性的流程。

跨切片一致性。跨切片注意力在同一楼层位置的高度切片之间交换信息。UNet 将每个位置的S 切片嵌入拼接后，把 $H \times W$ 网格作为空间域。在注意力分辨率下，跨切片模块将每个 (u, v) 特征重塑为 S 个词元，沿切片维度应用注意力，并通过残差投影将结果写回。因此，$\{(u, v, s)\colon\ s = 1, \ldots, S\}$ 的预测能够交换有关物体范围和主体／正面朝向对齐的信息。

跨通道一致性。 跨通道一致性使实例颜色和朝向预测依赖于语义预测。实例颜色和朝向的逻辑值共享语义主干特征，而属性分支则通过学习得到的门控和偏置，由语义特征分支调制。这使属性在同一去噪步骤中具备语义感知能力，而第 3.2节则确保采样期间通道组合有效。

训练目标。 训练目标不仅监督词元去噪，也监督场景组装所需的几何量。主要项对掩蔽的 RoomToken 通道进行去噪，而辅助项则校准占用情况以及下游使用的几何预测头。我们优化

$$\mathcal{L} = \lambda_y \mathcal{L}_y + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} + \lambda_c \mathcal{L}_c + \lambda_f \mathcal{L}_f + \lambda_{\text{occ}} \mathcal{L}_{\text{occ}} + \lambda_{\text{attr}} \mathcal{L}_{\text{attr}}. \tag{5}$$

此处，每个 λ都是一个标量权重。 $\mathcal{L}_y$ 是针对语义类别的掩码交叉熵，而 $\mathcal{L}_{\text{cls}}$ 则是针对长尾家具分布、按物体数量均衡的语义类别损失。该损失采用根据训练物体数量计算的固定平方根倒数权重，建筑结构／背景权重保持为 1；先对每个真实标注物体所占单元的语义交叉熵取平均，再对所有物体取平均。这样一来，稀有类别和小型实例便能获得与常见或大型家具相当的影响力。 $\mathcal{L}_c$ 和 $\mathcal{L}_f$ 是针对可见家具单元上实例颜色和朝向标签的掩码交叉熵损失。 $\mathcal{L}_{\text{occ}}$ 是根据语义逻辑值计算的前景占用情况的带逻辑值二元交叉熵。 $\mathcal{L}_{\text{attr}}$ 汇总了基于实例池化特征，对偏航角 $(\cos\theta,\ \sin\theta)$、高度和垂直偏移进行的实例级均方回归损失。详情请参见第 D.2 节。

3.4 推理与场景组装

推理时，输入建筑掩码中的建筑单元会在整个反向过程中保持固定，而 MDIRT 则为带家具的室内空间采样语义、实例颜色和朝向切片。由于这些建筑词元永远不会被覆盖，掩码离散扩散设计能在整个生成过程中保留建筑条件，使模型能够学习在有效的房间内部布局家具。在我们的实验中（第4.2节），我们表明，在最终资产摆放后，越界错误为零。

我们将具有相同语义和实例颜色标签的相连RoomToken分组为对象实例。它们的词元决定类别、水平中心和水平尺寸；主体/前向

标签表示朝向。MDIRT中的MLP头使用对每个实例掩码求平均的去噪特征来预测连续高度和垂直偏移；当主体/前向标签无法确定朝向时，则由偏航角头提供朝向。我们通过从3D-FUTURE（Fu等，2021b）中检索适当的3D资产来实例化最终场景。对于每个对象，我们使用物体类别查询3D数据库，并根据3D尺寸范围选择最接近的资产。将检索到的资产沿各轴缩放，以匹配估计范围；旋转至预测偏航角，并平移至恢复的对象位置。我们采用与SemLayoutDiff相同的基于类别和尺寸的资产检索与放置流程。资产放置后，不进行裁剪、额外优化或任何手工设计的场景修正。附录F详细介绍了字段移除消融实验中的组装以及分组/朝向流程。

4 实验

We 评估架构条件生成与近期扩散方法的表现。消融实验和 a
user 研究评估设计选择和感知质量。

4.1 实验设置

数据集。 我们在 3D-FRONT（Fu 等人，2021a）上训练和评估，采用先前工作（Paschalidou 等人，2021）所确立的数据划分，涵盖三种房间类型（卧室、客厅和餐厅），包含 4,496 个训练房间、274 个验证房间和 390 个测试房间。
模型细节。遵循 SemLayoutDiff（Sun 等，2026），我们将所有方法训练为跨房间类型的统一模型，而非为每种房间类型单独训练一个模型。每个场景都被栅格化到 128×128 地面平面网格上，其中一个像素／词元在水平面上的边长为 0.10 m。网格沿垂直方向在 $S{=}8$ 个固定高度切片处重复，切片高度分别为 (0.02, 0.45, 0.80, 1.20, 1.60, 2.00, 2.40, 2.70) m；这些高度通过垂直容差为 0.10 m 的贪心覆盖分析一次性选定。对于 MDIRT，RoomTokens 使用 $K_{\text{sem}}{=}39$ 语义类别、 $C{=}3$ 实例颜色以及 $\varnothing_c$，并采用身体／正面／$\varnothing_f$朝向标签；有效的语义–实例颜色–朝向三元组被展平为包含 209 个 ID 的调色板，用于可视化和单通道消融实验。模型细节请参见第 D.1 节，数据处理细节请参见附录 E。

基线方法。 我们比较了 DiffuScene (Tang 等人，2024）（连续扩散）、MIDIFFU- SION（Hu 等人，2026）（混合离散-连续扩散）和 SEMLAYOUTDIFF（Sun 等人，2026）（语义图扩散），它们均在相同的 3D-FRONT数据划分上作为统一模型进行训练。我们还重新训练了一个 SEMLAYOUTDIFF 变体：保留多项式扩散，同时在训练和采样期间固定外部区域、门和窗户像素，并将家具限制在有效的地面区域内。我们遵循 SemLayoutDiff 的生成和评估流程，并复现其报告的基线方法结果。

指标。 遵循以往工作（Sun 等人，2026），我们使用 FID 和 KID 评估渲染布局的保真度，并使用 CKL 将物体类别分布与 3D-FRONT 测试集进行比较。我们引入 Pair-Rel，用于比较物体对类别和空间关系。对于物理合理性，我们报告场景的越界率（OOB_S）和物体的越界率（OOB_O）、对象碰撞（COL）、可导航性（NAV）以及我们的门窗开口净空指标（OC）。我们还报告每个场景的平均物体数（对象数量）。指标定义和计算细节见附录 A。所有场景质量指标均在资产检索、缩放、旋转和位置摆放之后计算，使用与 SemLayoutDiff 相同的 3D-FUTURE 资产、SceneToolkit 渲染方式和评估脚本。不进行后续裁剪、优化或人工设计的场景修正。

4.2 结果与分析

物理有效性与有竞争力的布局质量。我们使用五个随机采样的不同随机种子生成场景，将 MDIRT 的输出与基线方法进行比较。表 1 显示，MDIRT 的 FID 优于 SemLayoutDiff（50.14 对 71.18），KID/CKL 则与之相当（10.35/6.81 对 8.52/4.85），展现出有竞争力的分布学习能力。此外，MDIRT 在所有空间合理性指标上均表现最佳：较低的成对关系指标（12.09）意味着物体之间的关系更合理；资产放置后零越界，表明其边界合规性良好；较低的 COL（12.64）反映相交更少；较高的 NAV（99.25）表示自由空间连通性更好；较高的

表 1：在 3D-FRONT 上的定量比较。结果为推理期间随机采样的 5 个随机种子的均值 ± 标准差，每个训练好的模型均保持固定。KID 报告为 KID$\times 10^3$，CKL 报告为 CKL$\times 10^2$。最佳均值以粗体标示，次佳均值以斜体标示。

	分布／构成					空间／物理合理性				
方法	FID↓	KID↓	CKL↓	成对关系↓	对象数量	越界$_S$↓	越界$_O$↓	COL↓	NAV↑	OC↑
DIFFUS场景 Tang 等人(2024)	88.52 ± 0.35	30.15 ± 0.42	20.49 ± 0.28	32.38 ± 0.95	8.71 ± 0.15	66.35 ± 0.48	20.15 ± 0.31	41.06 ± 0.44	94.22 ± 0.25	59.58 ± 1.05
MID扩散 (Hu 等人,2026)	93.63 ± 0.41	30.95 ± 0.37	37.05 ± 0.22	40.52 ± 0.46	7.42 ± 0.18	60.55 ± 0.39	34.78 ± 0.98	51.51 ± 0.29	96.28 ± 0.33	*75.82 ± 0.47*
SEMLAYOUTDIFF (Sun 等人，2026)	*71.18 ± 0.34*	**8.52 ± 0.26**	*4.85 ± 0.19*	*20.65 ± 0.43*	7.96 ± 0.12	16.71 ± 0.38	6.55 ± 0.21	*16。24 ± 0.92*	*96。45 ± 0.27*	56.02 ± 0.45
SEMLAYOUTDIFF（架构钳制）	73.45 ± 0.36	10.63 ± 0.24	**4.51 ± 0.17**	22.15 ± 0.41	7.32 ± 0.14	*0.41 ± 0.08*	*0.35 ± 0.05*	40.89 ± 0.49	95.25 ± 0.31	65.48 ± 1.02
Ours	**50.14 ± 0.27**	*10.35 ± 0.21*	6.81 ± 0.66	**12.09 ± 0.07**	8.25 ± 0.18	**0.00 ± 0.00**	**0.00 ± 0.00**	**12.64 ± 0.34**	**99.25 ± 0.24**	**82.31 ± 0.34**

表 2：MDIRT 的消融实验。最佳值以粗体显示。

	分布／构成					空间／物理合理性				
配置	FID↓	KID↓	CKL↓	成对关系↓	对象数量	越界$_S$↓	越界$_O$↓	COL↓	NAV↑	OC↑
多通道训练集 (MDIRT)	**49.84**	**10.41**	**6.43**	**12.13**	8.36	**0.00**	**0.00**	**12.75**	**99.30**	**82.04**
不进行类别平衡	74.77	34.18	21.31	20.03	6.31	**0.00**	**0.00**	37.73	96.58	76.20
不使用多通道（单通道）	95.35	34.91	30.05	48.32	7.12	**0.00**	**0.00**	49.30	95.74	66.40
不使用语义优先采样	88.21	28.74	32.40	34.54	7.84	**0.00**	**0.00**	58.81	73.07	54.90
不使用实例颜色字段	98.73	36.42	51.96	37.32	9.41	**0.00**	**0.00**	88.77	83.52	48.30
不使用朝向字段	84.92	26.18	28.62	36.63	8.96	**0.00**	**0.00**	89.07	75.69	50.70
不使用跨切片注意力	91.05	31.27	37.82	34.78	9.18	**0.00**	**0.00**	89.01	87.16	55.60

物体覆盖率（82.31）表示开口周围的通行间隙更好。MDIRT 的每个场景平均物体数为 8.25，相比之下，SemLayoutDiff 为 7.96，真实标注为 9.39，这证实了性能提升并非源于布局稀疏。

FID 和 KID 使用不同统计量比较渲染特征分布；CKL 比较类别频率。它们都无法直接衡量三维碰撞、开口净空或场景质量。因此，我们还评估物体关系、物理合理性和人类偏好。

采用建筑约束钳制的 SemLayoutDiff 降低了越界率，但 COL 从 16.24 升至 40.89，因为其反向过程对随机语义值进行去噪，而非在固定建筑结构下对掩蔽家具进行去噪。钳制也无法考虑分组、检索和旋转过程中新增的几何形状，因此部分物体仍有一部分位于房间之外。MDIRT 联合建模以建筑结构为条件的语义、实例颜色和朝向切片，在实现零越界的同时保持较低的碰撞率，且不同运行之间的 COL 和物体覆盖率变化较小。

MDIRT 可减少明显的布局失败。图 4 显示，定量改进也体现在生成布局中：基线方法可能将物体放置在边界两侧、使近距离家具发生碰撞，或堵住门窗等建筑开口；而 MDIRT 能更好地保留通畅的开口、房间几何结构和物体结构。更多结果请参见附录 G。

消融分析与失败分析。我们对模型进行了消融实验，以确定不同组件的影响（表 2 和 3）。所有消融模型均从头开始训练，采用相同的数据划分和训练设置，仅更改指定组件。所有模型均使用建筑条件，并采用相同的场景组装、渲染和评估流程。

类别平衡与多通道预测。未进行类别平衡（$\mathcal{L}_{\text{cls}}$）时，物体数量从 8.36 降至 6.31，CKL 从 6.43 升至 21.31，生成的场景更加稀疏，类别频率也不够准确。将三个字段合并为一个类别标记，同时保留全部八个切片后，FID 从 49.84 升至 95.35，Pair-Rel 从 12.13 升至 48.32，这支持了多通道预测的有效性。

实例颜色与朝向。 移除实例颜色字段后，COL升至88.77，表明对象实例分离的可靠性降低。移除朝向字段后，COL升至89.07，NAV则从99.30降至75.69，表明场景中的重叠增加，可导航性降低。若没有实例颜色，相连的同类别词元会被视为同一个实例。若没有朝向，辅助头会根据实例特征预测朝向（附录 F）。这些对比结果支持保留实例分离和朝向线索。

跨切片注意力与采样。 移除跨切片注意力会使 COL 从 12.75 升至 89.01，表明它对于避免物体相交至关重要。若不采用语义优先采样，

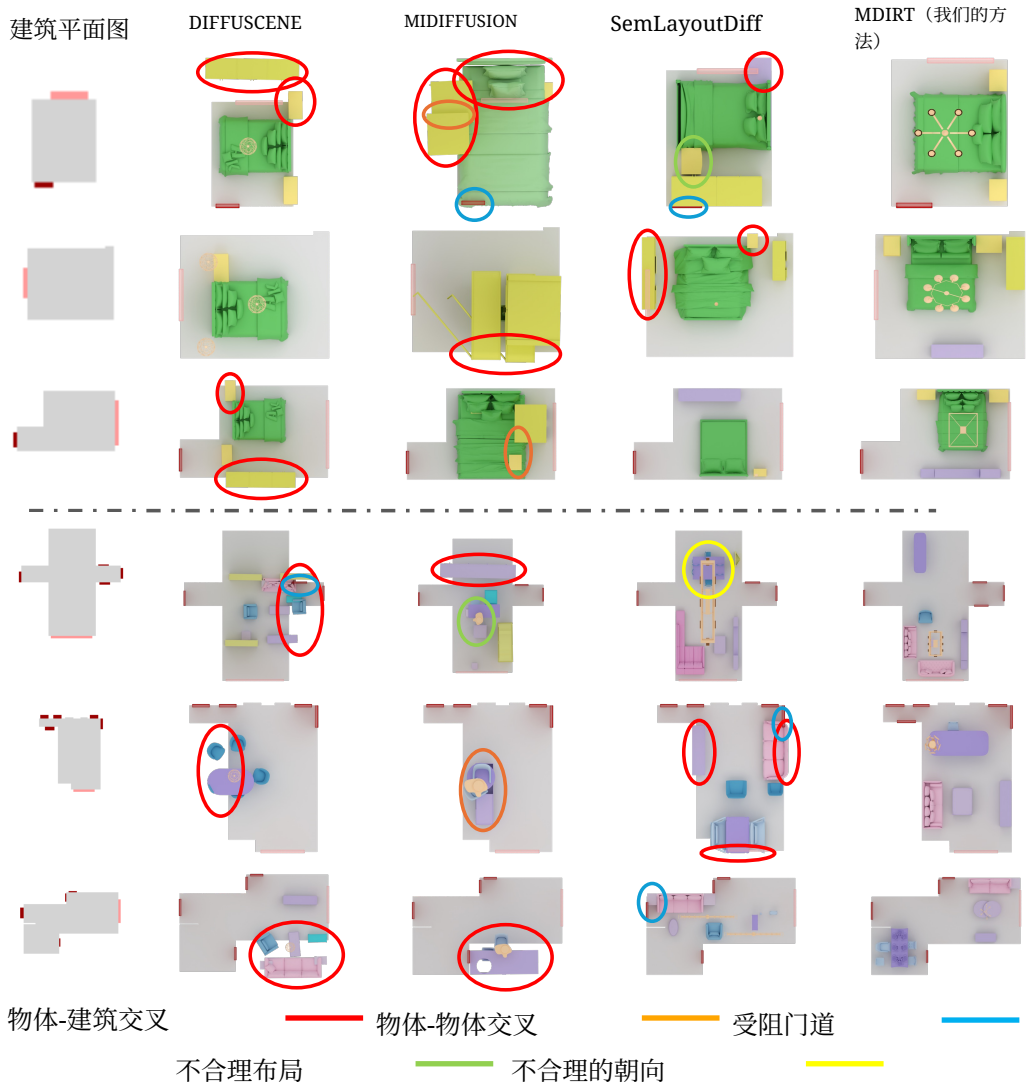

图4：在 3D-FRONT 上进行的拱形掩码条件定性比较。我们使用语义正交视图，重点比较布局评估结果。基线方法存在边界违规、碰撞或门窗开口受阻等问题，而 MDIRT 保留了畅通的开口、房间几何结构以及结构化物体线索。

表3：MDIRT 在表示、实例颜色、扩散和建筑约束方面的消融实验。KID 和 CKL 分别按 10^3 和 10^2 缩放。最佳值以粗体显示。

变体设置				分布／构成					空间／物理合理性			
表示	实例颜色	扩散	钳制	FID↓	KID↓	CKL↓	成对关系↓	对象数量	越界$_S$↓	越界$_O$↓	COL↓	NAV↑	OC↑
多切片（完整）	3	掩蔽	On	**49.84**	10.41	6.43	12.13	8.36	**0.00**	**0.00**	**12.75**	**99.30**	**82.04**
单切片（自上而下）	3	掩蔽	On	62.57	18.64	11.11	9.02	6.08	**0.00**	**0.00**	13.75	97.35	74.67
自顶向下语义图	–	掩蔽	On	54.69	**8.89**	10.15	13.97	7.03	**0.00**	**0.00**	19.50	97.59	74.97
多切片	4	掩蔽	On	58.03	14.06	9.74	**8.40**	6.67	**0.00**	**0.00**	17.49	97.26	67.74
多切片	3	掩蔽	Off	57.61	14.02	8.71	8.98	6.59	1.17	0.18	17.64	97.44	70.94
多切片	3	多项式	On	70.57	16.14	**5.32**	11.02	8.00	0.03	0.0042	32.16	93.10	62.53
多切片	3	多项式	Off	70.14	18.59	12.92	16.78	8.67	34.60	15.88	17.74	96.09	69.65

COL 会升至 58.81，而物体覆盖率会降至 54.90，这支持了先预测类别、再预测实例的判断 - color 和朝向场。

多切片与单切片 RoomTokens 对比。 单切片变体将场景投影到一张俯视图中。多切片 RoomTokens 生成的物体更多（8.36 对比 6.08），且物体覆盖率更高

（82.04 对比 74.67），COL 更低（12.75 对比 13.75）。因此，保留垂直信息可以提升物理合理性，而无需依赖更稀疏的场景。

RoomTokens 与语义地图。 在固定掩码扩散和钳制后，与使用 SemLayoutDiff 属性预测模型（APM）的自上而下语义地图相比，多切片 RoomTokens 的 COL 更低（12.75 对 19.50），物体覆盖率（OC）更高（82.04 对 74.97），尽管生成的物体更多（8.36 对 7.03）。这表明，对实例颜色、朝向和垂直信息进行编码有助于提升物理合理性。

三种与四种实例颜色。 四种实例颜色将有效的语义–实例颜色–朝向组合从 209 种扩展到 277 种，并增加了嵌入条目和输出通道，而这些组合在三种颜色已足够时并未使用。四种颜色改善了成对关系（Pair-Rel），但 COL 变差（从 12.75 升至 17.49），物体覆盖率（OC）也降低（从 82.04 降至 67.74），这支持采用三种颜色的选择。附录 C 提供了实例分离诊断。

扩散与建筑约束。 在采用钳制时，掩码扩散的 COL 低于多项式扩散（12.75 对 32.16），物体覆盖率（OC）则更高（82.04 对 62.53）。移除钳制后，掩码扩散的场景越界率从 0.00% 升至 1.17%，多项式扩散则从 0.03% 升至 34.60%。钳制能减少两种模型的越界，但会增加多项式扩散中的碰撞；采用钳制的掩码扩散实现了零越界和最低的 COL。

失败案例。 图 5展示了具有代表性的失败情况：不合理的家具布局、物体之间相交以及受阻门道。有效的 RoomTokens 和边界内放置并不能保证生成场景中的布局具有实用性或留有足够的通行间隙。

用户研究。 我们请参与者根据建筑保真度和物体布局的可行性对场景进行排序。每位参与者会看到 20 组场景，每组包含由不同方法生成的四个场景。场景顺序随机排列，参与者需要从优到劣进行排序。用户研究共有 12 位参与者。总体而言，参与者平均将 MDIRT 排在首位（表 4）。经 Holm 校正后，配对威尔科克森检验表明，该偏好相较于所有基线方法均具有显著性（p=0.0015），这说明参与者认为 MDIRT 的感知建筑保真度和物体布局优于基线方法。配对威尔科克森检验考虑了同一参与者的重复评分。详见附录 B 有关用户研究的详细信息。

表 4：由 12 名参与者对生成场景进行排名的用户研究。MDIRT 在近 80% 的情况下排名第 1。

方法	平均排名↓	第一名↑	前-2↑
MDIRT（我们的方法）	**1.25**	**79.6%**	**96.3%**
SEMLAYOUTDIFF	2.15	18.3%	73.3%
DIFFUSCENE	3.15	2.1%	17.9%
MIDIFFUSION	3.45	0.0%	12.5%

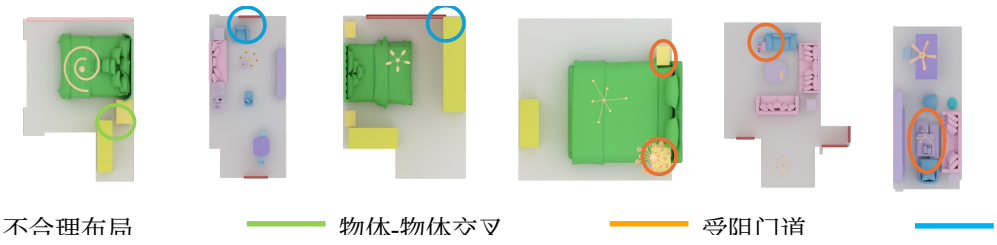

图 5：失败案例：不合理布局（绿色）、物体交叉（橙色）和受阻门道（蓝色）。

5 结论

我们提出了 MDIRT，这是一种针对切片式 RoomTokens 的掩码离散扩散框架，用于生成室内场景布局。RoomTokens 为语义占用图添加了高度切片、实例颜色线索和朝向线索。在去噪过程中钳制建筑标记，可保留给定的平面图，使房间几何结构保持明确，同时支持物体分组、朝向估计和场景组装。

在 3D-FRONT 上，MDIRT 达到了具有竞争力的分布质量，并在评估的方法中展现出最强的整体空间合理性。受控对比表明，其性能提升不仅来自建筑约束

钳制和稀疏生成；盲测人工排名研究更倾向于我们的场景。表征检查显示信息损失很小；消融实验支持因子化的 RoomToken 令牌字段、语义优先采样以及跨切片注意力。结构化占用词元还可支持室外场景、带部件的物体以及其他语义占用任务。

局限性。 固定网格和高度切片可能会遗漏细节和狭窄间隙。小规模的 3D-FRONT 训练集在物体类别和房间类型方面分布不均。RoomTokens 和有效性投影无法防止分组、连续姿态估计、缩放或资产检索造成的网格碰撞。更大规模、更多样化的数据以及更强的场景组装能力，或可改善分布匹配和检索；自适应分辨率切片和几何感知细化则有助于捕捉细节并减少碰撞。

致谢

本研究部分由 CIFAR AI Chair 和 NSERC 探索基金资助，并得到了以下机构的支持：加拿大数字研究联盟 以及 CFI/BCKDF JELF。感谢 Jiayi Liu 和 Yiming Zhang 提供建设性意见并参与讨论。

AI 使用声明

我们在以下任务中使用了 AI 工具：

- 协助构思概念框架：编码代理帮助集思广益，为场景表示和相关方法设计提出备选方案。作者选定并完善了最终方案。
- 设计或提供有关研究方法或实验的反馈：编码代理对评估协议、指标和消融实验提供了反馈。
- 实现方法：编码代理协助实现和调试模型、训练集、场景组装及评估代码。
- 支持定性和主题数据分析：编码代理帮助检查实验结果和消融实验，识别不一致之处，并解读报告指标中的模式。

其他披露内容包括使用人工智能进行头脑风暴、创建和编辑软件代码、搜索信息和相关文献，以及起草和编辑论文文本以提高清晰度。人工智能还帮助澄清了我们对训练目标和评估指标的文字及数学描述。

我们已审查并验证所有人工智能辅助完成的工作，并对工作的最终内容负责，包括借助人工智能生成的文本、主张、代码及其他成果。

参考文献

Rio Aguina-Kang、Maxim Gumin、Do Heon Han、Stewart Morris、Seung Jean Yoo、Aditya Gane- shan、R KennyJones、Qiuhong Anna Wei、Kailiang Fu 和 Daniel Ritchie。使用大语言模型程序合成和未经整理的物体数据库进行开放宇宙室内场景生成。*arXiv* 预印本 *arXiv:2403.09675*，2024。

Kenneth Appel 和 Wolfgang Haken。每个平面图都是四色可着色的。第一部分：放电法。《伊利诺伊数学杂志》，21(3)：429–490，1977。

TongyuanBai、WangyuanfanBai、Dong Chen、Tieru Wu、Manyi Li 和 Rui Ma。FreeScene：通过自由提示进行混合图扩散的三维场景合成。载于*IEEE*计算机视觉与模式识别会议，第5893–5903页，2025。

Ata Celen、Guo Han、Konrad Schindler、Luc Van Gool、Iro Armeni、Anton Obukhov 和 Xi Wang。I-Design：个性化大语言模型室内设计师。*arXiv*预印本 *arXiv:2404.02838*，2024。doi：10.48550/arXiv.2404.02838。

Shang Chai、Liansheng Zhuang 和 Fengying Yan。LayoutDM：用于布局生成的基于Transformer的扩散模型。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第18349–18358页，2023。Jeongjun Choi、Yeonsoo Park 和 H. Jin Kim。SceneNAT：用于语言引导的室内场景合成的掩码生成建模。*arXiv*预印本 *arXiv:2601.07218*，2026。Weixi Feng、Wanrong Zhu、Tsu-jui Fu、Varun Jampani、Arjun Akula、Xuehai He、Sugato Basu、Xin Eric Wang 和 William Yang Wang。LayoutGPT：使用大语言模型进行组合式视觉规划与生成。神经信息处理系统进展，2023。Huan Fu、Bowen Cai、Lin Gao、Ling-Xiao Zhang、Jiaming Wang、Cao Li、Qixun Zeng、Chengyue Sun、Rongfei Jia、Binqiang Zhao 等。3D-FRONT：具有布局和语义的带家具的三维房间。载于*IEEE/CVF*国际计算机视觉会议，第10933–10942页，2021a。Huan Fu、Rongfei Jia、Lin Gao、Mingming Gong、Binqiang Zhao、Steve Maybank 和 Dacheng Tao。3D-FUTURE：具有纹理的三维家具形状。国际计算机视觉期刊，129(12)：3313–3337，2021b。Jialin Gao、Donghao Zhou、Mingjian Liang、Lihao Liu、Chi-Wing Fu、Xiaowei Hu 和 Pheng-Ann Heng。DisCo-Layout：在多智能体框架中解耦并协调语义与物理细化，以实现三维室内布局合成。*arXiv*预印本 *arXiv:2510.02178*，2025。Emiel Hoogeboom、Didrik Nielsen、Priyank Jaini、Patrick Forr´e 和 Max Welling。Argmax流与多项式扩散：学习类别分布。神经信息处理系统进展，34：12454–12465，2021。Siyi Hu、Diego Martin Arroyo、Stephanie Debats、Fabian Manhardt、Luca Carlone 和 Federico Tombari。用于三维室内场景合成的混合扩散。载于冬季计算机视觉应用会议，2026。Ziniu Hu、Ahmet Iscen、Aashi Jain、Thomas Kipf、Yisong Yue、David A Ross、Cordelia Schmid 和 Alireza Fathi。SceneCraft：用于将三维场景合成为Blender代码的大语言模型智能体。载于*ICLR 2024*大语言模型（*LLM*）智能体研讨会，2024。Mude Hui、Zhizheng Zhang、Xiaoyi Zhang、Wenxuan Xie、Yuwang Wang 和 Yan Lu。利用解耦扩散模型统一布局生成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第1942–1951页，2023。Naoto Inoue、Kotaro Kikuchi、Edgar Simo-Serra、Mayu Otani 和 Kota Yamaguchi。LayoutDM：用于可控布局生成的离散扩散模型。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第10167–10176页，2023。Victor Kulikov、Victor Yurchenko 和 Victor Lempitsky。通过深度着色进行实例分割。*arXiv*预印本 *arXiv:1807.10007*，2018。Jumin Lee、Woobin Im、Sebin Lee 和 Sung-Eui Yoon。用于场景尺度三维类别数据的扩散概率模型。*arXiv*预印本 *arXiv:2301.00527*，2023。Elad Levi、Eli Brosh、Mykola Mykhailych 和 Meir Perez。DLT：结合离散-连续扩散布局Transformer的条件布局生成。载于国际计算机视觉会议，2023。Manyi Li、Akshay Gadi Patil、Kai Xu、Siddhartha Chaudhuri、Owais Khan、Ariel Shamir、Changhe Tu、Baoquan Chen、Daniel Cohen-Or 和 Hao Zhang。GRAINS：用于室内场景的生成式递归自编码器。*ACM*图形学汇刊（*TOG*），38(2)：1–16，2019。Chenguo Lin 和 Yadong Mu。InstructScene：基于语义图先验的指令驱动三维室内场景合成。载于学习表征国际会议，2024。Guying Lin、Kemeng Huang、Michael Liu、Ruihan Gao、Hanke Chen、Lyuhao Chen、Beijia Lu、Taku Komura、Yuan Liu、Jun-Yan Zhu 和 Minchen Li。PAT3D：物理增强的文本到三维场景生成。载于学习表征国际会议，2026。

Aaron Lou、Chenlin Meng 和 Stefano Ermon。通过估计数据分布的比率进行离散扩散建模。发表于国际机器学习会议，2024年。Andrew Luo、Zhoutong Zhang、Jiajun Wu 和 Joshua B Tenenbaum。场景布局的端到端优化。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第3754–3763页，2020年。L´eopold Maillard、Nicolas Sereyjol-Garros、Tom Durand 和 Maks Ovsjanikov。DeBaRA：基于去噪的三维房间布置生成。神经信息处理系统进展，2024年。doi：10.48550/arXiv.2409.18336。Shen Nie、Fengqi Zhu、Zebin You、Xiaolu Zhang、Jingyang Ou、Jun Hu、Jun Zhou、Yankai Lin、Ji-Rong Wen 和 Chongxuan Li。大语言扩散模型。*arXiv*预印本 *arXiv:2502.09992*，2025年。¨Bas¸ak Melis Ocal、Maxim Tatarchenko、Sezer Karaoglu 和 Theo Gevers。SceneTeller：从语言生成三维场景。发表于欧洲计算机视觉会议，2024年。Despoina Paschalidou、Amlan Kar、Maria Shugrina、Karsten Kreis、Andreas Geiger 和 Sanja Fi- dler。ATISS：用于室内场景合成的自回归变换器。神经信息处理系统进展，34：12013–12026，2021年。Nicholas Pfaff、Thomas Cohn、Sergey Zakharov、Rick Cory 和 Russ Tedrake。SceneSmith：用于生成可用于仿真的室内场景的智能体生成。*arXiv*预印本 *arXiv:2602.09153*，2026年。Daniel Ritchie、Kai Wang 和 Yu-an Lin。通过深度卷积生成模型实现快速、灵活的室内场景合成。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第6182–6190页，2019年。Neil Robertson、Daniel Sanders、Paul Seymour 和 Robin Thomas。四色定理。组合理论杂志*B*辑，70(1)：2–44，1997年。Antonio Ruiz、Tao Wu、Andrew Melnik、Qing Cheng、Xuqin Wang、Lu Liu、Yongliang Wang、Yan- feng Zhang 和 Helge Ritter。GeoSceneGraph：用于文本引导的三维室内场景合成的几何场景图扩散模型。*arXiv*预印本 *arXiv:2511.14884*，2025年。Subham Sekhar Sahoo、Marianne Arriola、Yair Schiff、Aaron Gokaslan、Edgar Marroquin、Justin T. Chiu、Alexander Rush 和 Volodymyr Kuleshov。简单而有效的掩码扩散语言模型。发表于神经信息处理系统进展，2024年。Jiaxin Shi、Kehang Han、Zhe Wang、Arnaud Doucet 和 Michalis K. Titsias。面向离散数据的简化且泛化的掩码扩散。发表于神经信息处理系统进展，2024年。Chong Su、Yingbin Fu、Zheyuan Hu、Jing Yang、Param Hanji、Shaojun Wang、Xuan Zhao、Cengiz ¨Oztireli 和 Fangcheng Zhong。CHOrD：为三维室内场景生成无碰撞、房屋尺度且布局有序的数字孪生，并支持可控平面图和最优布局。*arXiv*预印本 *arXiv:2503.11958*，2025年。Fan-Yun Sun、Weiyu Liu、Siyi Gu、Dylan Lim、Goutam Bhat、Federico Tombari、Manling Li、Nick Haber 和 Jiajun Wu。LayoutVLM：通过视觉语言模型对三维布局进行可微优化。发表于*IEEE*计算机视觉与模式识别会议，2025a。Jia-Mu Sun、Jie Yang、Kaichun Mo、Yu-Kun Lai、Leonidas Guibas 和 Lin Gao。Haisor：通过深度强化学习实现具有人类感知能力的室内场景优化。*ACM*图形学汇刊，43(2)：1–17，2024a。Kaifan Sun、Bingchen Yang、Peter Wonka、Jun Xiao 和 Haiyong Jiang。RelTriple：将关系三元组融入扩散过程，学习合理的室内布局。*arXiv*预印本 *arXiv:2503.20289*，2025b。doi：10.48550/arXiv.2503.20289。Qi Sun、Hang Zhou、Wengang Zhou、Li Li 和 Houqiang Li。Forest2Seq：重振序列室内场景合成的顺序先验。发表于欧洲计算机视觉会议，2024b。doi：10.48550/arXiv.2407.05388。

Xiaohao Sun、Divyam Goel 和 Angel X. Chang。SemLayoutDiff：使用扩散模型进行室内场景合成的语义布局生成。发表于国际三维视觉会议，2026年。Hou In Ivan Tam、Hou In Derek Pun、Austin T Wang、Angel X Chang 和 Manolis Savva。SceneEval：评估文本条件三维室内场景合成中的语义连贯性。发表于冬季计算机视觉应用会议，2026a。Hou In Ivan Tam、Hou In Derek Pun、Austin T Wang、Xiaohao Sun、Qirui Wu、Han-Hung Lee、Angel X Chang 和 Manolis Savva。组合式三维室内场景生成综述。计算机图形论坛，第e70591页，2026b。Jiapeng Tang、Yinyu Nie、Lev Markhasin、Angela Dai、Justus Thies 和 Matthias Nießner。Diffuscene：用于生成室内场景合成的去噪扩散模型。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），2024年。Kai Wang、Manolis Savva、Angel X Chang 和 Daniel Ritchie。用于室内场景合成的深度卷积先验。*ACM*图形学汇刊（*TOG*），37(4)：1–14，2018年。

KaiWang、Yu-AnLin、BenWeissmann、Manolis Savva、AngelX Chang 和 DanielRitchie。PlanIT：利用关系图和空间先验网络规划并实例化室内场景。*ACM*图形学汇刊 *(TOG)*，38(4)：1–15，2019。

Xinpeng Wang、Chandan Yeshwanth 和 Matthias Nießner。Sceneformer：使用 Transformer 生成室内场景。发表于国际三维视觉会议，第 106–115 页。IEEE，2021。

Zehao Wen、Zichen Liu、Srinath Sridhar 和 Rao Fu。AnyHome：开放词汇生成结构化和纹理化三维住宅。发表于欧洲计算机视觉会议，2024。

Hongchi Xia、Xuan Li、Zhaoshuo Li、Qianli Ma、Jiashu Xu、Ming-Yu Liu、Yin Cui、Tsung-Yi Lin、Wei-Chiu Ma、Shenlong Wang 等。SAGE：面向具身人工智能的可扩展智能体三维场景生成。*arXiv* 预印本 *arXiv:2602.10116*，2026。

Ming-Jia Yang、Yu-Xiao Guo、Bin Zhou 和 Xin Tong。基于语义分割深度图像集合的室内场景生成。在 *IEEE/CVF*国际计算机视觉会议论文集，第 15203–15212 页，2021 年。

Yandan Yang、Baoxiong Jia、Peiyuan Zhi 和 Siyuan Huang。PhyScene：面向具身人工智能的物理可交互三维场景合成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第16262–16272页，2024a。Yandan Yang、Baoxiong Jia、Shujie Zhang 和 Siyuan Huang。SceneWeaver：借助可扩展的自反思智能体实现一体化三维场景合成。载于神经信息处理系统进展，2025。Yue Yang、Fan-Yun Sun、Luca Weihs、Eli VanderBilt、Alvaro Herrasti、Winson Han、Jiajun Wu、Nick Haber、Ranjay Krishna、Lingjie Liu 等。Holodeck：语言引导的三维具身人工智能环境生成。载于*IEEE*计算机视觉与模式识别会议，第30卷，第20–25页。IEEE/CVF，2024b。Hongwei Yi、Chun-Hao P Huang、Shashank Tripathi、Lea Hering、Justus Thies 和 Michael J Black。MIME：人类感知三维场景生成。载于*IEEE*计算机视觉与模式识别会议，第12965–12976页，2023年。¨Guangyao Zhai、Evin Pinar Ornek、Shun-Cheng Wu、Yan Di、Federico Tombari、Nassir Navab 和 Benjamin Busam。CommonScenes：利用场景图生成常识三维室内场景。神经信息处理系统进展，2023年。¨Guangyao Zhai、Evin Pınar Ornek、Dave Zhenyu Chen、Ruotong Liao、Yan Di、Nassir Navab、Federico Tombari 和 Benjamin Busam。Echoscene：通过场景图扩散中的信息回声实现室内场景生成。载于欧洲计算机视觉会议，第167–184页。Springer，2024年。

Ye Zhang、Yu Zhou、Yifeng Wang、Jun Xiao、Ziyue Wang、Yongbing Zhang 和 Jianxu Chen。细胞实例分割的四色定理。载于第四十二届国际机器学习会议，2025年。

Zaiwei Zhang、Zhenpei Yang、Chongyang Ma、Linjie Luo、Alexander Huth、Etienne Vouga 和 Qixing Huang。基于混合表示的场景合成深度生成建模。*ACM*图形学汇刊 *(TOG)*，39(2):1–21，2020年。

附录

在本附录中，我们提供了评估指标定义（附录 A）、补充实验（附录 B）、数据集诊断（附录 C）、模型实现细节（附录 D）、RoomTokens 构建细节（附录 E）、最终场景组装细节（附录 F）以及补充定性结果（附录 G）。

A 评估指标

我们遵循先前的工作（Sun 等人，2026），并使用 SceneEval（Tam 等人，2026a）中介绍的指标，评估生成场景与参考场景的分布相似度，以及物理布局合理性。

分布与语义质量。对于 FID 和 KID，我们将每个生成场景和每个 3D-FRONT 测试场景渲染为 256×256 正交俯视图。物体按语义类别着色，生成场景和测试场景均使用相同的固定房间类型调色板。我们将渲染图转换为 RGB，并在白色背景上补边至 256×256。我们提取 clean-fid 的“clean” Inception 特征，并据此计算 FID 和 KID。对于每次比较，我们使用图像数量相等的集合，从较大的集合中无放回抽样，并对十次抽样的分数取平均。总体 MDIRT 分数将全部 390 张测试渲染图与每次抽样时从 3,000 个生成场景中抽取的 390 张渲染图进行比较；各房间分数则单独计算。报告的 KID 数值已乘以 10^3。

CKL 是测试场景与生成场景中汇总的家具类别计数归一化后分布之间的 $D_{\mathrm{KL}}(p_{\mathrm{gt}}\|p_{\mathrm{gen}})$，不包括建筑元素。CKL 报告时乘以 10^2。

成对布局关系。我们的 Pair-Rel 指标比较物体对类别与相对几何关系的联合分布。对于每一对具有有效边界框的不同家具物体，按顺序排列为 $(i,\ j)$，我们构造一个关系元组 $r_{ij} = (k_i,\ k_j,\ d_{ij},\ b_{ij})$，其中 k_i 和 k_j 是物体类别，d_{ij} 是物体 j 相对于物体 i的方向，b_{ij} 是距离分箱。$(i,\ j)$ 和 $(j,\ i)$ 均计入。

为计算 d_{ij}，我们在物体 i的朝向坐标系中表示物体中心之间的位移。位移幅度较大的轴决定左、右、前或后；中心重合时标记为相同。我们将距离测量为物体有向边界框之间的最小水平间隔；若边界框相交，则赋值为零。距离分箱为接触（≤ 0.05 米）、近距离（$(0.05,\ 0.5]$ 米）、中距离（$(0.5,\ 1.5]$ 米）和远距离（> 1.5米）。

我们汇总每个集合中所有场景的关系元组，并对其计数进行归一化，以获得生成样本和真实值直方图 P_{gen} 及 P_{gt}。令 $M = \frac{1}{2}(P_{\mathrm{gen}} + P_{\mathrm{gt}})$。Pair-Rel 是

$$\begin{aligned}\mathrm{PairRel} &= 100\ \mathrm{JSD}(P_{\mathrm{gen}}, P_{\mathrm{gt}}) \\ &= 50\left[D_{\mathrm{KL}}(P_{\mathrm{gen}}\|M) + D_{\mathrm{KL}}(P_{\mathrm{gt}}\|M)\right].\end{aligned} \tag{6}$$

数值越低，表示一致性越高。由于每个元组都包含物体类别，Pair-Rel 同时反映配对组成和相对几何形状。

物理合理性。OOB_S 表示包含至少一个物体位于房间边界之外的生成场景比例。OOB_O 表示位于房间边界之外的生成对象比例。COL 在进行连通分量分组、方向估计和物体放置后，衡量生成对象之间的碰撞情况。NAV 评估最终场景中的自由空间可导航性。开口净空（OC）是经过评估的朝向房间的门窗净空区域中不与家具网格相交的区域百分比。平均物体数量与物理指标一并报告，以揭示空置或异常稀疏的生成结果。所有指标均使用最终场景，并遵循第 4.1 节所述的评估协议。

B 附加实验与详细信息

我们提供无条件生成和逐房间结果、去噪诊断以及用户研究分析。

表6：3D-FRONT 上按房间类型评估。部分客厅/餐厅源房间同时出现在多个类别细分中，因此这些行是诊断性子集，而非对 390 个不同测试房间的划分。

房间类型	方法	分布／构成					空间／物理合理性				
		FID↓	KID↓	CKL↓	成对关系↓	对象数量	越界$_S$↓	越界$_O$↓	COL↓	NAV↑	OC↑
卧室	DIFFUSCENE	103.16	47.46	40.89	13.54	7.80	55.10	20.77	37.69	91.98	56.80
	MIDIFFUSION	103.74	32.06	61.72	52.26	6.65	65.25	37.81	33.60	95.25	73.30
	SEMLAYOUTDIFF	75.42	**5.93**	**5.27**	30.92	7.20	13.80	4.00	26.21	93.80	52.80
	Ours	**51.06**	9.86	7.92	**13.21**	7.55	**0.00**	**0.00**	**22.29**	**98.41**	**78.60**
餐厅	DIFFUSCENE	87.45	21.14	8.92	44.15	8.35	71.50	12.79	42.65	96.07	60.80
	MIDIFFUSION	95.93	36.47	43.71	47.79	7.10	46.63	30.74	64.62	97.22	77.10
	SEMLAYOUTDIFF	70.87	9.97	**4.97**	16.80	7.75	15.80	7.53	**5.37**	99.36	58.20
	Ours	**49.78**	**7.81**	5.38	**10.47**	8.05	**0.00**	**0.00**	11.22	**99.59**	**83.50**
客厅	DIFFUSCENE	74.79	22.06	11.47	39.78	9.80	72.80	26.53	42.57	94.98	60.78
	MIDIFFUSION	80.87	24.66	5.48	21.90	8.30	70.15	35.44	56.64	96.10	77.45
	SEMLAYOUTDIFF	66.90	**10.05**	**4.10**	14.65	8.72	20.30	8.32	16.82	95.92	57.45
	Ours	**48.68**	13.56	5.99	**12.71**	9.48	**0.00**	**0.00**	**4.73**	**99.90**	**84.02**

表5：在 3D-FRONT 上进行房间类型条件下的无条件生成。我们的两阶段流水线首先生成一张建筑布局图，并将其作为输入条件传递给冻结的 MDiRT 布局生成器。

方法	分布／构成					空间／物理合理性				
	FID↓	KID↓	CKL↓	成对关系↓	对象数量	越界$_S$↓	越界$_O$↓	COL↓	NAV↑	OC↑
SEMLAYOUTDIFF（无条件）	93.93	**10.72**	**17.21**	29.80	7.34	0.20	0.02	14.32	96.87	57.40
ArchGen → MDiRT（我们的方法）	**69.96**	12.86	18.35	**21.70**	8.27	**0.00**	**0.00**	**12.59**	**99.55**	**81.10**

无条件生成。 对于无条件生成，我们使用一个独立的房间类型条件多项式扩散模型，首先生成一张 128×128 建筑布局图。该建筑布局图作为输入提供给 MDiRT，由其将其提升为八个 RoomToken切片，并传递给冻结的布局生成器。表 5 显示，该两阶段流水线（**ArchGen → MDiRT**）在保持零越界的同时，相较于无条件 SemLayoutDiff 提升了 FID、COL 和 NAV。

按房间类型划分的结果。我们在表6中报告了按房间类型评估的结果。MDIRT 在所有房间类型中均保持零越界，并且在卧室和客厅中的碰撞率最低，而 SemLayoutDiff 在餐厅中的碰撞较少。MDIRT 在开口周围的通行间隙最佳（物体覆盖率最高）。与其他方法相比，它在每种房间类型中实现了相当的物体比例，同时取得了合理的合理性指标。分布表现则较为复杂：MDIRT 的 FID 和成对关系指标最低，而 SemLayoutDiff 在大多数按房间划分的 KID/CKL 结果上仍略占优势。

投影前的词元有效性。 我们在去噪过程中 Π_{valid} 之前和之后测量词元有效性，以确定模型无需校正即可预测出有效字段组合的频率。对于完整的 390 条件测试集，我们记录了全部 256 个反向步骤，每个条件使用一个随机种子。在有效性投影之前，最终步骤中所有经评估的词元有 99.49% 有效，而仅包含家具的词元有效率为 100%。投影后，有效率为 100%。语义显现的核心区间约为第 74–240 步；到第 248 步时，最终占用情况的 95% 已生成；到第 256 步时，完整语义–实例颜色–朝向三元组的 99% 已生成。因此，在最终投影之前，去噪器已能为几乎所有词元生成兼容的字段组合，包括所有家具标记。

模型是否在记忆训练数据? 我们还检查模型能否通过从训练集中检索最相近的样本来生成新的布局。实验从三种房间类型中各选取 100 个条件，并为每个条件使用 10 个随机种子。表 7 中的结果表明模型并非直接记忆训练数据，但不能证明生成内容具有新颖性：在 300 个条件中，有 12 个条件至少触发了一个样本的严格精确重复启发式规则。

用户研究详情。 该用户研究由12名参与者组成，共提供240个排名。每位参与者评估20组场景：10组卧室和10组客厅/餐厅。参与者查看随机选取的不同场景组，涵盖三种房间类型的50组场景。对于每组场景，我们以随机顺序展示四个匿名化场景，每种方法在每个

表 7：对 300 个条件分别使用 10 个随机种子时的条件多样性与最近训练样本分析（3,000 个生成结果）。

指标	结果
不同随机种子下家具的 1 – 交并比	0.3227
完全相同条件下的重复样本对	0.5111%
与配对真实值不一致	0.6689
生成样本 →训练集中位距离	5.229
留出集真实值 →训练集中位距离	4.938
生成样本与训练集的完全重复项	1.3333%
留出集真实值与训练集的完全重复项	4.0000%

表8：按房间类型划分的用户研究结果。平均排名越低越好；每种房间类型中的最佳平均排名以粗体显示。

房间类型	平均排名↓ MDIRT	SemLayoutDiff	DIFFUSCENE	MIDIFFUSION	MDIRT 第一名（%）↑
卧室	**1.24**	2.16	3.19	3.41	81.7
客厅	**1.42**	1.92	3.17	3.50	66.7
餐厅	**1.20**	2.23	3.08	3.49	82.1

显示位置上出现的次数相同。参与者根据建筑保真度和物体布局，在无并列的情况下将场景从最好到最差进行排名。

为考虑重复评分，我们对每位参与者针对每种方法给出的排名取平均，并对这些参与者平均值进行统计检验。弗里德曼检验结果为 $\chi^2(3) = 34.0\,(p < 2 \times 10^{-7})$。对参与者平均值进行的配对威尔科克森检验经 Holm 校正后，与每个基线方法相比仍具有显著性（$p = 0.0015$）。全部 12 位参与者平均而言都更偏好 MDIRT。表 8按房间类型报告了描述性结果：MDIRT 在每种类型中的平均排名都是最佳的。

C 数据集与表示诊断

表示合理性检查。 我们进行了一系列表示检查，以验证在模型评估之前，Room- Tokens 是否保留了物体可见性、实例分离和朝向。处理后的数据中，99.47% 的家具框可见，无效令牌占 0.00%，同类别颜色冲突仅占 0.92%，直接朝向平均绝对误差为 1.13°，表明表示损失很小。

RoomToken诊断。 这些指标用于评估 RoomTokens 表示，不用于比较方法。

表9 报告了模型训练前处理后的3D-FRONT RoomToken表示的数据集级统计。物体覆盖率衡量至少出现在一个栅格化高度切片中的家具框所占比例。无效令牌统计违反表示规则的语义–实例颜色–朝向三元组数量。颜色冲突统计被分配相同图着色的相邻同类别、不同实例令牌对数量。实例数量平均绝对误差将从$(y_i,\ c_i)$ 恢复的连通分量与每个场景的真实实例数量进行比较。B/F 配对是同时具有主体和前部单元的已恢复分量所占比例，朝向平均绝对误差衡量通过直接解码主体/前部几何结构得到的偏航角误差。

表9：处理后的3D-FRONT RoomToken表示的数据分析。

Data	场景	物体覆盖率	无效令牌	颜色冲突	实例数量 MAE	实例完全匹配率	B/F 配对	方向 MAE
3D-FRONT（所有划分）	5,160	99.47%	0.00%	0.92%	0.18	91.53%	94.88%	1.13°

实例颜色诊断。 使用三种实例颜色时，属于同一类别中不同实例的相邻词元对，有 99.08% 的颜色不同。按语义和实例颜色标签对词元进行分组后，在 91.53% 的场景中能够正确还原实例数量，实例

数量平均绝对误差 (MAE) 为 0.18。这些数据分析结果表明，处理后的表示中颜色冲突较少；但这并不能证明三色在所有情况下都足够。在表 3中，我们比较了使用三种和四种实例颜色进行生成的结果，并通过实验表明，虽然使用四种颜色可获得更低的 Pair-Rel 分数，但使用三种颜色的碰撞率和开口净空更好。

训练数据中的碰撞来源。我们审查了原始 3D- FRONT 训练集和测试集房间记录中的家具间碰撞。对于每件家具，我们加载其 3D-FUTURE 资产网格，并应用记录的场景变换。我们通过 trimesh 和 python-fcl，使用 Flexible Collision Library (FCL) 检测每对网格是否相交；不包括建筑元素。如果有可用的接触点，我们便从平均接触点向某个网格的质心方向移动该网格 5mm，并且只有在碰撞仍然存在时才计入该碰撞对。否则，我们保留初始的相交检测结果。包含至少一个计入的碰撞对的房间记录被视为发生碰撞；属于至少一个碰撞对的物体被视为发生碰撞。

在训练集记录中，4,882 个房间中有 2,112 个（43.26%）发生碰撞，31,256 件家具物体中有 6,279 件（20.09%）参与碰撞。4,096 个碰撞对表示房间记录中不同的无序对；同一物体可能属于多个碰撞对。在测试集记录中，390 个房间中有 177 个（45.38%）发生碰撞，4,965 件物体中有 840 件（16.92%）发生碰撞，共涉及 592 个碰撞对。这些原始记录计数与用于训练和评估模型的筛选后的数据划分不同。它们表明源数据中存在家具相交；重建并放置生成对象时，还可能产生额外的相交。

D 实现细节

D.1 模型细节

模型仅在训练划分上进行训练；验证划分用于超参数和检查点选择，留出测试集仅用于最终评估。MDIRT 通过多通道训练参数化，学习结构化 RoomTokens 上的分布。去噪器使用与扁平词元模型相同的 UNet 主干网络，通道倍增系数为（1、1、2、2），并在 32×32 和 16×16 处使用自注意力与跨切片注意力。除非另有说明，模型在 2 个 NVIDIA L40S GPU 上训练 3000 个训练周期，每个 GPU 的批量大小为 24，使用 AdamW、学习率 3×10^{-4} 和权重衰减 0.01，并采用 256 个反向扩散步数进行采样。结构化标记损失采用掩码语义损失（使用按对象数量平衡的类别加权）、实例颜色/朝向损失、占用率校准以及偏航角/高度/偏移头。

D.2 训练损失

令 $\mathcal{M}_b^y$ 表示场景 b 中的掩码语义单元。集合 $\mathcal{M}_b^c$ 和 $\mathcal{M}_b^f$ 包含家具语义标签可见的掩码实例颜色单元和朝向单元。用 CE 表示分类交叉熵，三个词元损失为

$$\mathcal{L}_\rho = \frac{1}{B}\sum_{b=1}^{B}\sum_{i\in\mathcal{M}_b^\rho} \mathrm{CE}(\boldsymbol{\ell}_{b,i}^\rho, x_{b,i}^\rho), \qquad \rho\in\{y,c,f\}. \tag{7}$$

对于按物体数量平衡的语义项，令 $\mathcal{O}_b$ 表示具有掩码语义单元的真实家具物体，令属于物体 o 的单元为 $\mathcal{S}_{b,\,o}$。我们使用

$$\mathcal{L}_{\mathrm{cls}} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{|\mathcal{O}_b|}\sum_{o\in\mathcal{O}_b}\frac{w_{k(o)}}{|\mathcal{S}_{b,o}|}\sum_{i\in\mathcal{S}_{b,o}}\mathrm{CE}(\boldsymbol{\ell}_{b,i}^y, y_{b,i}), \tag{8}$$

其中，含有 $\mathcal{O}_b=\varnothing$ 的场景贡献为零，$k(o)$ 是物体 o的类别，固定的家具类别权重满足 $w_k\propto(n_k+\epsilon)^{-1/2}$，对应训练集物体数量 n_k。先对每个物体内部求平均，再对所有物体求平均，可以避免较大的实例在这一项中占据主导。

设 Y_{furn} 和 Y_{arch} 分别为家具类别和建筑/背景类别。前景对数几率和目标为

$$g_{b,i} = \text{LSE}_{k\in Y_{\text{furn}}}\, \ell^{y}_{b,i,k} - \text{LSE}_{k\in Y_{\text{arch}}}\, \ell^{y}_{b,i,k}, \qquad q_{b,i} = \mathbf{1}[y_{b,i} \in Y_{\text{furn}}], \tag{9}$$

其中 LSE 表示对数和指数。占用损失为

$$\mathcal{L}_{\text{occ}} = \frac{1}{B}\sum_{b=1}^{B}\sum_{i\in\mathcal{M}_b^{y}} \text{BCEWithLogits}(g_{b,i}, q_{b,i}). \tag{10}$$

最后，实例池化特征用于预测偏航角、高度和垂直偏移。记 $\mathbf{v}_o$ =(cos θ_o, sin θ_o)，我们将它们的回归损失归纳为

$$\mathcal{L}_{\text{attr}} = \beta_\theta\, \text{MSE}(\hat{\mathbf{v}}, \mathbf{v}) + \beta_h\, \text{MSE}(\hat{h}, h) + \beta_\delta\, \text{MSE}(\hat{\delta}, \delta), \tag{11}$$

其中，每个均方误差均在有监督家具对象上取平均， β 值用于对三个属性头加权。

E RoomTokens 构建细节

贪心切片高度选择。切片高度的选择旨在通过固定数量的水平面尽可能覆盖更多家具垂直区间。对于每个物体 n，设 $I_n = [z_n^-, z_n^+]$为其垂直范围。候选高度 h若满足以下条件，则覆盖物体 n： $[h-\tau, h+\tau]\cap I_n \neq \emptyset$，其中 τ=0.10m 是预处理中使用的栅格化容差。算法1以 0.01 米的间隔采样候选高度，从固定的地板和天花板锚点开始，然后反复添加能够覆盖最多当前未覆盖物体的候选高度，同时确保其与已选高度之间至少相隔 δ=0.10m。

算法1： 贪心切片高度选择

输入： 物体高度区间 $\mathcal{I} = \{I_n = [z_n^-, z_n^+]\}_{n=1}^{N}$；候选高度 $\mathcal{H}$；切片数量S；容差 τ；最小间距 δ；固定锚点 $\mathcal{A}$**输出：** 选定切片高度 $\mathcal{Z}$1 $\mathcal{Z} \leftarrow \mathcal{A}$；2 $\mathcal{U} \leftarrow \{1,\ldots,N\}$；// 尚未覆盖的物体3 从 $\mathcal{U}$ 中移除任何被 $\mathcal{Z}$中的高度覆盖的物体；4 **当** $|\mathcal{Z}| < S$ **时**

5 $\mathcal{H}_{\text{ok}} \leftarrow \{h \in \mathcal{H} :\ |h - z| \geq \delta$ 对所有 $z \in \mathcal{Z}\}$；6 $g(h) \leftarrow |\{n \in \mathcal{U} : [h-\tau, h+\tau] \cap I_n \neq \emptyset\}|$ 对每个 $h \in \mathcal{H}_{\text{ok}}$；7 $h^\star \leftarrow \text{argmax}_{h\in\mathcal{H}_{\text{ok}}}\, g(h)$；

8 $\mathcal{Z} \leftarrow \mathcal{Z} \cup \{h^\star\}$；9 从 $\mathcal{U}$ 中移除任何被 $h^\star$覆盖的物体；

10 **返回**排序后的 $\mathcal{Z}$；

实例着色分配。实例颜色仅用于区分语义类别相同且彼此相邻的家具实例；它们不是全局实例 ID。对于每个场景和语义类别 k，我们构建一个邻接图 $G_k = (V_k,\ E_k)$，其顶点为类别 k中的对象实例。当两个不同的类别$(u,\ v) \in E_k$ 实例在至少一个高度切片中占据水平相邻的网格单元时，就添加一条边。算法2 会分别为每个 G_k 着色，按实例规模从大到小处理，并为每个实例分配一个未被其已着色邻居使用的最先可用颜色。

算法2：贪心实例着色分配 输入：语义切片 Y；预处理实例 ID 切片 P；家具类别 $\mathcal{K}$；颜色调色板$\Gamma = \{0, \ldots, C-1\}$输出：实例着色分配 γ 及冲突记录 $\mathcal{F}$ $\gamma \leftarrow \emptyset$、 $\mathcal{F} \leftarrow \emptyset$； 对每个语义类别 $k \in \mathcal{K}$ 执行 $V_k \leftarrow$收集所有以类别 k出现的实例； $E_k \leftarrow \emptyset$； 对每个高度切片和水平邻接对 (p, q) 执行 如果 $Y_p = Y_q = k, P_p \neq P_q$ 和 P_p, P_q 是有效的家具实例 ID，则 将 $(\min(P_p, P_q), \max(P_p, P_q))$ 添加到 E_k； 按占用单元格数量降序排列 V_k； 对每个实例 $u \in V_k$ 执行 $N_u \leftarrow \{v : (u, v) \in E_k$ 或 $(v, u) \in E_k\}$； $A \leftarrow \Gamma \setminus \{\gamma(v) : v \in N_u$ 且 v 已着色$\}$； 如果 $A \neq \emptyset$则 $\gamma(u) \leftarrow A$ 中的第一个颜色； 否则 $\gamma(u) \leftarrow \Gamma$中的第一个颜色； // 回退 $\mathcal{F} \leftarrow \mathcal{F} \cup \{(u, v) \in E_k : \gamma(u) = \gamma(v)\}$； **返回** γ， $\mathcal{F}$；

以下单切片结果说明，与其为每个对象分别设置标签，不如使用小型调色板来区分相邻实例。

定理 1（单个切片中连通区域的四色充分性）。 考虑一个高度切片，其中每个对象实例占据一组通过共享边相连的网格单元。对于每个语义类别 k，当两个实例所占单元共享一条网格边时，就将它们相连。则存在一种至多使用四种颜色的分配方式，使每对相邻实例的颜色均不相同。

证明。 在单个切片中，将共享边的网格单元相连所形成的图是平面图。移除属于其他语义类别的单元，并将每个实例中相连的单元收缩为一个顶点。边收缩和删除操作均保持平面性，因此得到的实例邻接图是平面图。由四色定理可知，存在所需的着色方案（Appel 和 Haken，1977；Robertson 等人，1997）。

我们利用这一着色原理，通过紧凑调色板编码实例信息。在预处理阶段，贪心着色会为每个实例分配一种在所有切片中保持一致的颜色。跨切片的组合邻接图不一定是平面图，而且贪心着色无法保证找到四色分配。我们根据经验选择三种颜色：99.08% 的相邻同类别不同实例词元对具有不同颜色（附录C），且与四种颜色相比，三种颜色可带来更低的 COL 和更高的物体覆盖率（表 3）。空标签表示不适用属性，并非第四种实例颜色。

F 条件化、推理与场景组装详情

建筑掩码条件。 为使去噪记法保持简洁，正文将式（4）中的条件化元组写作 $\kappa = (r, \mathbf{a})$。其中， r 表示房间类型，$\mathbf{a}$ 表示输入建筑掩码。在正向掩码和反向去噪过程中，带有 $a_i = 1$ 的单元格均被固定为其建筑标签，而带有 $a_i = 0$ 的单元格则由模型生成。这样，MDIRT 就能在保留所提供房间几何结构的同时，为家具采样语义、实例颜色和朝向场。

采样。 三个词元字段在同一个反向扩散过程中生成。建筑词元在整个过程中保持固定；所有未固定的单元格起初都将其语义、实例颜色和朝向字段设为掩码。每个反向步骤如下进行：

1. 根据反向调度预测并更新被掩蔽的语义标签，同时保留已揭示的标签。

2. 使用更新后的语义状态和共享的去噪器特征，重新计算实例颜色和朝向预测。3. 如果语义仍被掩蔽，则两个属性都保持掩蔽状态。对于已揭示的家具语义，根据调度揭示符合条件的属性字段；对于建筑/背景语义，则赋予空属性。4. 在保留未解析掩码的同时，确保字段值彼此兼容，并恢复固定的建筑词元。

例如，当语义标签变为“椅子”后，实例颜色或朝向可能会在该步骤中显现，也可能继续保持掩蔽，直至后续步骤。这类部分掩蔽的词元是有效的中间状态。掩码表示某个值未知；空值表示某项属性不适用。附录 B 报告了 256 步去噪过程中的去噪轨迹。最终网格会保留给定的建筑标记，然后再进行分组和资产放置。

标记到对象的组装。 我们将高度切片中语义标签和实例颜色标签相同且相连的 RoomTokens 分组为对象实例。每个实例的词元决定其类别、水平中心和网格分辨率下的水平尺寸。被占用的高度切片提供粗略的垂直范围，而主体/前向标签则指示朝向。

MDIRT 中的 MLP头根据去噪特征估计连续高度和垂直偏移。对于每个实例，我们构建一个二维掩码，其中包含该实例出现在任意高度切片中的所有网格单元。我们对该掩码内的去噪特征取平均，并将所得向量输入这些头。当主体/前向标签无法确定方向时，偏航角头会提供朝向。这些预测与直接从词元获取的属性共同确定每个对象的类别、位置、尺寸和朝向，以便进行资产检索和放置。

字段移除消融实验的场景组装。借助 RoomTokens，我们的实例颜色使我们能够生成具有空间范围和方向信息的对象实例。我们的完整模型按语义和实例颜色标签对词元进行分组，并利用主体/正面线索预测朝向。为了消融移除实例颜色和朝向等字段的影响，我们遵循 SemLayoutDiff Sun 等人(2026）处理实例识别和朝向的方式。移除实例颜色字段时，我们使用连通分量分析，将相连且属于同一语义类别的词元组合成一个对象。在这种情况下，相邻的同类别词元会合并为同一个实例。移除朝向字段时，我们使用辅助实例级MLP头，根据池化后的实例特征预测偏航角。请注意，偏航角并非固定为零。所有这些变体的资产检索和位置摆放均遵循相同流程。

资产检索。 最终场景通过从 3D-FUTURE 检索资产来实例化Fu 等人(2021b)。对于每条对象记录，我们首先按类别筛选候选资产；若没有完全匹配的标签，则退而使用兼容的粗粒度标签。我们将预测的对象尺寸 (w, d, h)转换为 3D-FUTURE 半尺寸 $(w/2, h/2, d/2)$，并选择在三维尺寸空间中平方误差最小的候选项。随后，沿各轴缩放所选资产以匹配估计范围，按预测偏航角旋转，将其平移至恢复出的对象位置，并与重建的地板、墙壁、门和窗户一同导出。基于类别和尺寸的检索与位置摆放流程遵循 SemLayoutDiff。我们直接评估生成的场景，不进行后续裁剪、优化或人工设计的修正。报告的零越界表示这些最终场景的边界合规性。

旋转和缩放的影响。 即使位置固定，改变对象的偏航角或尺寸也可能使其包围框超出墙壁（越界），造成碰撞（COL），或阻塞通道（NAV）。在表 10 中，MDiRT 位置摆放实现了零越界。固定偏航角时，30.90% 的场景中存在越界对象，COL 从 12.75% 升至 35.82%，NAV 则从 99.30% 降至 90.92%。在图 6(a) 中，旋转后的家具与咖啡桌发生碰撞，并阻断了部分可行走地面。取消缩放后，场景越界和 COL 分别升至 10.50% 和 19.85%，NAV 则降至 97.02%。在图 6(b) 中，两个电视柜相互重叠，尽管此例中的地板仍保持连通。

表 10：旋转和缩放对相同生成场景的影响。所有行均使用相同的检索到的资产，并放置在相同的位置。固定偏航角会将每个物体的偏航角设为零。不进行缩放表示每个资产均采用检索到时的尺寸。如果物体放置后的包围框延伸到房间之外，则该物体超出边界。KID 和 CKL 分别按 10^3 和 10^2 缩放。

资产放置	FID↓	KID↓	CKL↓	成对关系↓	对象数量	越界$_S$↓	越界$_O$↓	COL↓	NAV↑	物体覆盖率↑
MDIRT 布置	**49.84**	**10.41**		**6.43**	**12.13**	8.36	**0.00**	**0.00**	**12.75**	**99.30 82.04**
固定偏航角	65.56	17.45		**6.43**	18.09	8.36	30.90	20.64	35.82	90.92 79.42
不进行缩放	58.20	14.43		**6.43**	12.34	8.36	10.50	5.38	19.85	97.02 80.17

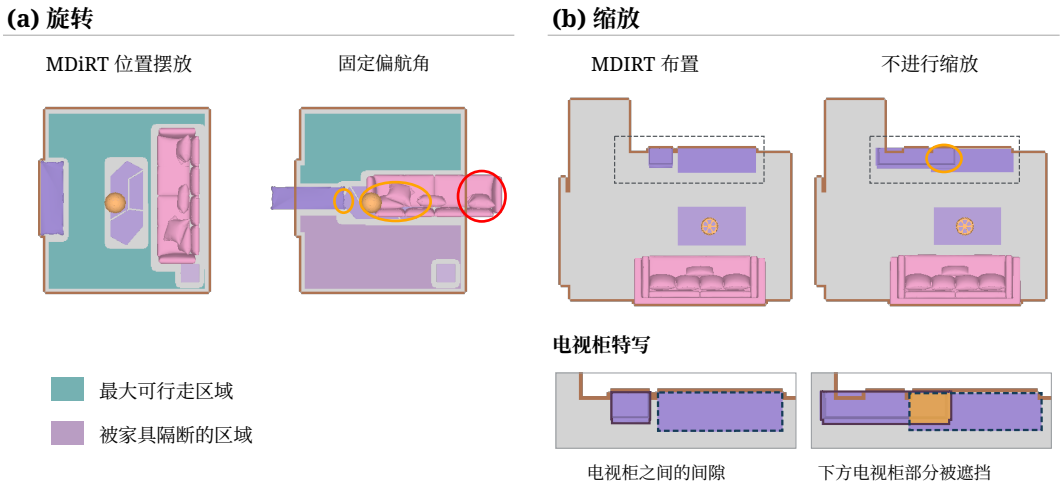

图6：旋转和缩放对资产放置的影响。两个原始 MDIRT 场景均具有无碰撞家具摆放，且未测得边界违规。橙色圆圈标记物体间相交，红色标记物体与墙壁相交，如图 4 所示。(a) 在固定偏航角下，沙发和电视柜与咖啡桌发生碰撞，并将可行走地面分割开来，从而降低 NAV。地面叠加图以青绿色显示最大的连通可行走区域，以紫色显示被家具隔断的区域。(b) 采用 MDIRT 布置时，电视柜之间存在间隙。不进行缩放时，较高的电视柜会与较低的电视柜重叠。虚线轮廓描绘出部分被遮挡的电视柜，橙色阴影标记重叠区域。该碰撞会提高 COL，而可行走地面仍保持连通。

G 补充定性结果

在图 7中，我们使用等轴测投影展示了 DiffuScene（Tang 等人，2024）、MiDiffusion（Hu 等人，2026）、SemLayoutDiff Sun 等人（2026）和 MDIRT 的输出场景示例。

在图8中，我们提供了更多采用我们的方法生成的等轴测视图投影结果。

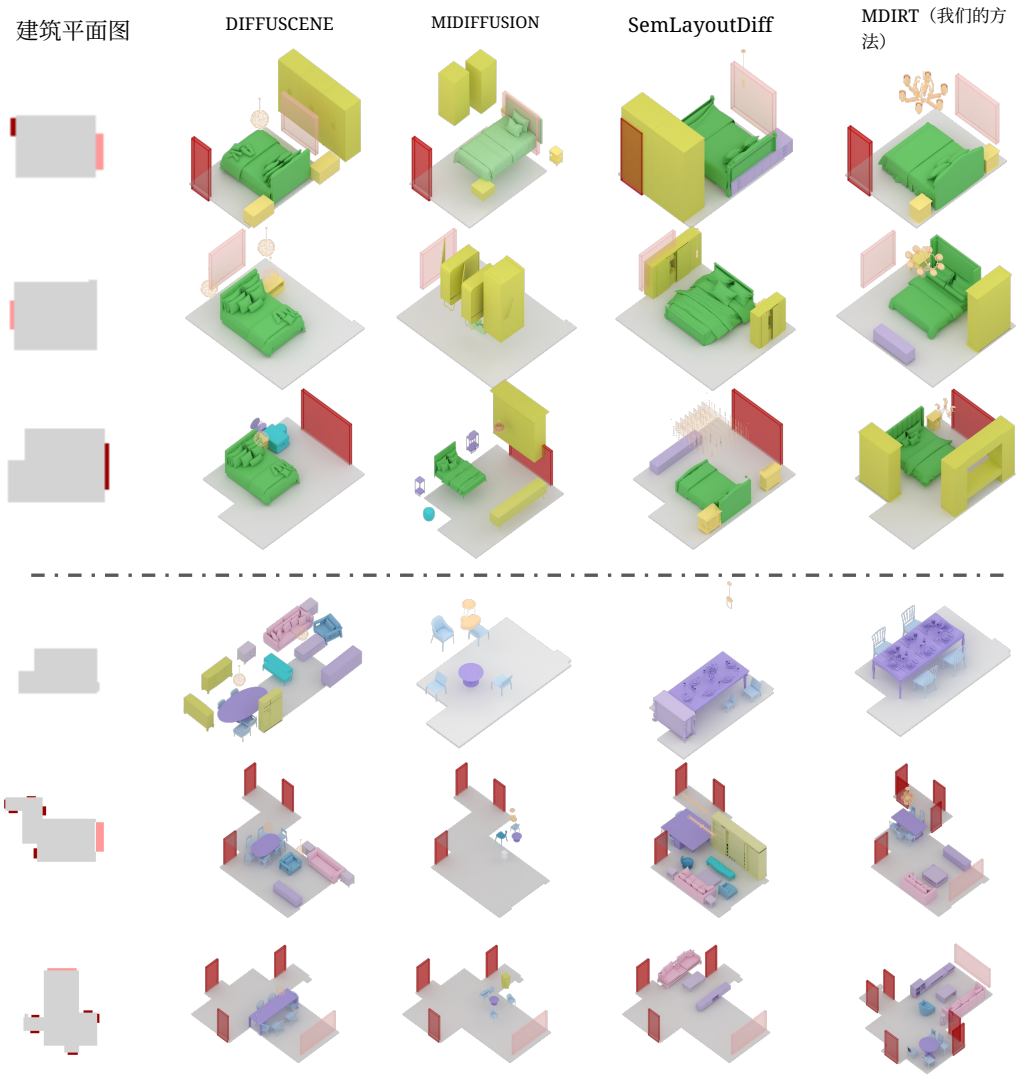

图 7：在等轴测视图中，基于建筑平面图掩码条件的 3D-FRONT 定性比较。

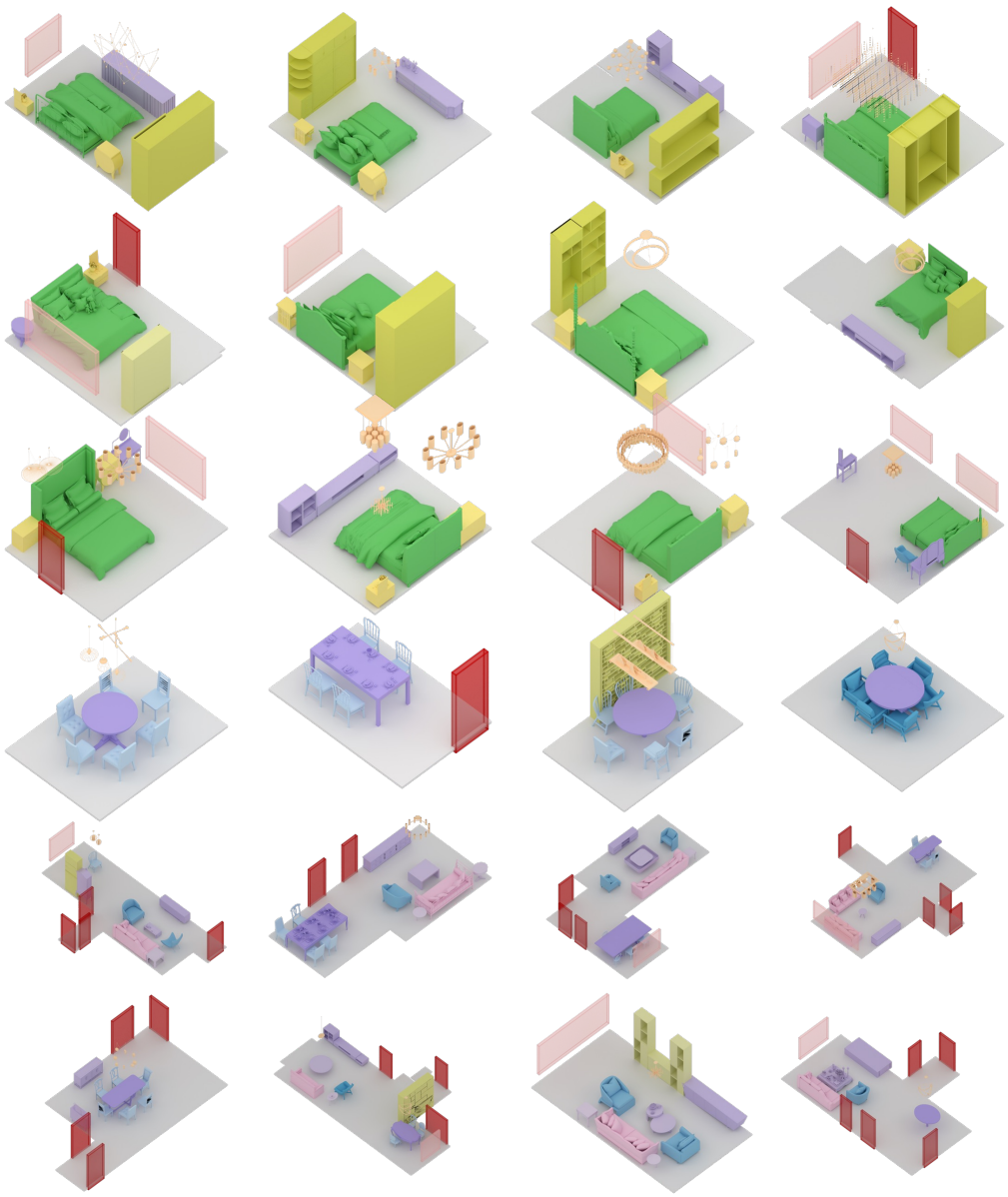

图 8：MDIRT 在等轴测视图下的更多拱形掩码条件生成结果。