arXiv:2611.32099v1 [cs.CV] 24 Nov 2026

PreferenceScene：基于反馈的个性化室内场景合成

SeongraeNoh RyunHa Lee HyeongYeop Kang†

高丽大学



图 1. PreferenceScene 为提出请求的人设计角色设定、PreferenceScene 所生成的三维场景一个房间。测试集中的四位模拟用户，每位用户都附有一段摘要，内容关于 e 经过三轮交互后，以及摘自指令的一段引文。插图：完整场景。

摘要

文本转场景方法可将简短指令转化为配有家具的 *3D*房间，但人们想要的特定房间很少能通过文字完整表达：许多偏好只有在人们看到具体房间，并能指出喜欢和不喜欢之处时才会显现。因此，我们研究存在隐藏用户偏好时的个性化室内场景合成：在多轮交互中，系统展示一个三维场景，而用户只能看到该场景的渲染图像，并以简短的自由形式反馈和满意度评分作答。因此，系统必须根据每次只针对场景部分内容的反馈，逐步使整个房间贴合用户偏好。系统还必须能够改变房间的每个部分，因为偏好不仅涉及单个物体，也涉及表面、照明、纺织品、家具及其布局共同营造的外观和氛围。我们提出了 *PreferenceScene*，其中大语言模型会明确记录用户对房间每项设计属性的偏好估计，并在每轮交互后推断反馈背后的偏好，修订整体估计，包括反馈未提及的部分。场景实例化模块会将每个估计值构建为房间中的一个独立部分，使这些部分都能遵循用户的偏好。我们在一个基准测试的测试集上进行评估，该测试集包含十种房间类型的 *320* 个实例，并使用模拟用户和基于渲染的评估器。与针对相同交互方式调整的五种已发表文本到场景基线相比，*PreferenceScene* 能更好地匹配隐藏偏好，且匹配度随着每轮交互不断提高。即使直接向这些基线提供偏好信息，*PreferenceScene* 仍然胜过它们；视觉语言模型也认为它生成的场景最真实、功能性最佳。

†通讯作者。

1. 引言

室内场景合成能够生成配有家具的 3D房间，而基于大语言模型（LLM）构建的系统如今可以根据简短的文本指令生成一个合理的房间 [12, 21,25, 70, 86]。然而，仅凭一条指令，很难得到一个适合特定个人的房间。一个人想要一间舒适的卧室，脑海中或许会浮现温暖的灯光、亚麻床品和浅色橡木地板，但只写下“舒适”二字。系统会补全其余内容，而生成的房间可能符合字面描述，却不符合这个人的心意。文生图模型的用户也会遇到同样的落差：文本描述只能传达目标图像的一部分，用户只能通过反复试错来得到想要的结果 [27, 33, 49]。与其期待一条指令就能描述完整个房间，为特定个人进行设计的系统不如与用户经过多轮交互来共同完成设计。

交互的作用不只是纠正错误。人们通常无法预先列出自己的偏好，其中许多偏好只有在评估具体选项时才会变得清晰 [43, 68],这也是推荐系统和设计工具让用户对示例做出反应，而不是一开始就说明需求的原因 [14, 67, 77]。当看到一个房间时，人们会注意到灯光太冷，或地板应该更深一些；这些偏好并不会通过简短的指令表达出来。我们将这类偏好称为隐藏偏好：系统并未获知它们，它们只会在用户对系统展示的场景做出反应时显现出来。这些反应也很稀疏。人们只会评论少数吸引他们注意的事物，而要求他们对房间的每个细节都做出判断，很快就会让他们感到疲惫；这正是依赖人工判断的交互式系统的已知局限 [10, 71]。因此，系统必须根据只提及部分部件的反馈，让整个房间逐步趋近于用户的偏好。

这些偏好涉及整个房间。房间的外观与氛围源自其表面、照明、纺织品、家具和布置的组合 [15],，而选择家具风格或配色方案的方法会针对整个场景优化兼容性，而非针对每个物体进行优化 [13, 48]。用户可以在这些部分之间自由表达反馈，从墙面颜色到床上用品，再到灯具的灯光。因此，个性化系统必须能够更改每个部分。我们比较的五种已发表的文本到场景基线方法 [12, 21, 25, 70, 86] 会选择并摆放物体，至多从固定库中挑选墙面和地面材料，因此无法更改其他部分。迭代式和编辑式系统只遵循其指令中明确说明的内容 [11, 31, 78, 87, 91], ；偏好引导仅根据用户对项目或图像的反馈来推断隐藏目标 [26, 28, 42, 53, 66], ，而不涉及3D房间的设计。

我们将这一设定表述为隐藏用户偏好下的个性化室内场景合成（第3节），用户只能看到每个展示场景的渲染图像，并以简短的自由形式反馈作答。我们的基准测试（第5节）为每位模拟用户设定一份隐藏的设计档案，指令会部分提及该档案，也可能完全不提及，这与使用模拟用户开展的偏好引出研究相同 [23, 27, 42], ，评估器则会检查最终场景体现了哪些隐藏值。

我们的方法 PreferenceScene 基于一项设计原则：系统将对用户偏好的估计保留为显式计划，为每项设计属性指定一个取值，并将每个取值实例化为场景中的一个独立部分。房间的每个部分都是关于用户偏好的一项假设，用户反馈可以确认或推翻它；由于每份计划都涵盖所有属性，系统会在每轮交互后重新考虑整个房间，包括反馈未提及的部分。大语言模型会在第一轮交互前为每项属性构建候选值空间，并在每轮交互后根据反馈重写计划。

在与这些系统采用相同交互方式的情况下，PreferenceScene 在两种基准测试形式中都取得了最高的偏好匹配度；即使对于反馈从未提及的属性也是如此，而且其匹配度随每轮交互不断提高。图1展示了 PreferenceScene 为四位用户生成的最终房间。仅凭反馈，它匹配到的隐藏偏好就多于所有获得文本形式偏好的已发表基线方法；视觉语言模型也认为它生成的场景最真实、最实用。在一项用户研究中，对于偏好和设计两方面，参与者都更喜欢它生成的场景，而非任何基线方法生成的场景（第6.4节）。

总而言之，我们的贡献如下：

- 我们提出了 PreferenceScene，它将用户的偏好建模为针对每个设计属性候选值空间的显式计划，并通过稀疏的自由形式反馈，使整个房间逐步趋向这些偏好（第4节）。
- 我们通过场景实例化模块，让该计划能够控制整个房间；该模块用于设置表面、照明、纺织品，以及家具的材料和颜色（第4.3节）。
- 我们在隐藏的用户偏好条件下，对个性化室内场景合成进行了形式化，并构建了一个涵盖十种房间类型、采用两种形式的基准测试，其中的模拟用户持有隐藏的设计档案，并且只能查看渲染图像（第3 节和5节）。
- 我们将 PreferenceScene 与五种经过调整以适应相同交互方式的已发表基线方法进行比较，并通过消融实验、第二评估者和用户研究对其进行测试（第6节）。

2. 相关工作

文本驱动的室内场景合成使用大语言模型预测布局 [21], ，为优化器提出约束 [70, 86]，或推理空间关系 [25], ；其他方法则将这些步骤组合成多步骤流程 [12, 60]，或通过自有评判器进行迭代 [58, 84, 87]。在这些方法中，输入文本定义目标场景，而编辑方法会按照指令修改场景 [11, 54, 91]。在使用偏好的方法中，I-Design [12]根据用户在指令中写下的偏好进行条件生成，而 OptiScene [88]



图2。一个实例（图3和4的用户；指令和角色设定已缩短）以及根据 G^*（第5节）构建的参考场景。隐藏值，例如腮红粉墙面和大理石棋盘格地板，在指令或角色设定中均未出现。

和 ReSpace [11]在偏好对上为所有用户训练同一个模型。交互式检索和偏好 elicitation 从反馈、比较或选择中推断用户心中的目标 [23, 26, 27, 38, 66], ， 而对话状态跟踪则在各轮次中持续估计用户的目标 [81]。我们的系统从用户对完整三维场景渲染图像的自由形式反馈中推断其未明确表达的偏好，并将这些偏好作为关于房间设计属性的计划加以维护。与这一研究方向相同，我们使用持有隐藏目标的模拟用户进行评估 [23, 27]，我们的评估器会针对渲染图像中的每个属性回答一个多项选择题，类似于 TIFA [30]。由于大语言模型评估器偏爱自身生成的内容 [56], ， 每个角色使用不同的模型家族。补充材料第I节扩展了本节内容。

3. 问题定义

在此任务中，一个系统与用户进行多轮交互，并必须生成一个展示用户隐藏偏好的最终三维场景；用户不会明确说明这些偏好，而只会通过对渲染图像的反馈来透露它们。

实例。 一个实例（图 2）由用户的隐含偏好和系统看到的输入组成：房间类型、用户提出房间需求的指令 q 以及描述用户的角色设定 $\mathcal{P}$ 。每种房间类型都有一套设计属性模式，例如地板，如地板，每个属性 a 都有一个由互斥取值组成的有限集合，例如仿木纹板材或水磨石。属性列表是去掉取值后的模式：其中包含 M 属性名称，每个名称都附有一句不提及任何取值的描述。这些信息告诉系统哪些设计选项对用户重要，但不说明用户想要哪些取值。 隐含偏好 G^* 为每个属性 $G^*[a] \in \mathcal{V}_a$ 指定一个取值 a，其形式与第 4节中的方案相同。指令和角色设定可以明确提及其中一些取值，并暗示另一些取值。

交互。 系统接收房间类型、 q、 $\mathcal{P}$ 和属性列表。它从未接收取值集合 $\mathcal{V}_a$ 或 G^*，也不会在交互过程中更新模型参数。随后，系统与用户进行 K 轮交互。针对一个实例进行的一次此类运行称为一次交互。在第 k轮中，系统展示一个三维场景。用户查看其渲染图像，并返回自然语言反馈 η_k 以及评分$z_k \in \{1, \ldots, 10\}$，说明用户对该房间的满意程度。在第 K轮之后，系统再输出一个场景，即最终场景 X，该场景不会收到任何反馈。

评估。我们通过感知属性匹配率（pAMR）衡量最终场景 X对 G^* 的呈现效果，即评估器从 X的渲染图像中恢复出的隐藏值所占的比例：

$$\mathrm{pAMR}(X, G^*) = \frac{1}{M} \sum_{a} \mathbf{1}\big[r(X, a) = G^*[a]\big], \qquad (1)$$

其中， $r(X, a) \in \mathcal{V}_a$ 是视觉语言模型（VLM）根据a在 X的渲染图像上针对属性提出的多项选择题所给出的答案，选项为 $\mathcal{V}_a$ 的取值。该指标对两种选项顺序取平均值（第6.1节）。pAMR仅根据渲染图像计算，因此适用于任何输出场景的系统；我们以百分比报告该指标，并对各实例取平均值。

4. 方法

PreferenceScene以显式形式保留对用户偏好的当前估计，因此每一轮反馈都可以改变这一估计，且每次变化都会体现在ˆ房间中（图3）。该估计是一个计划：第 S_k 轮的计划 k 为房间 M 设计属性 a 中的每一项ˆ指定一个取值 $S_k[a]$，例如用“浅色橡木板”这样的简短文本描述地板。计划由三个组件共同处理。第一轮开始前，大语言模型会为这些属性生成一组候选值，并据此选择一个初始计划（第4.1节）。每轮结束后，大语言模型会根据用户反馈修订计划（第4.2节）。场景实例化会将每个计划构建为一个三维场景，其中每个取值决定房间中相应的部分（第4.3节）。我们用 c 表示第3节中的系统输入。没有任何组件会读取隐藏偏好 G^* 或取值集合 $\mathcal{V}_a$。

4.1. 偏好空间与初始计划

计划会从每个属性的一组选项中选择取值。如果大语言模型每轮都重写所有取值，那么即使房间保持不变，取值的表述也可能改变，



图3。测试交互第一轮中 PreferenceScene 的概览（图1中的 隐式形式的第二位用户）。(1) 大语言模型编写设计方案，为每个设计属性指定一个取值。(2) 场景实例化将该方案构建为三维场景。(3) 用户通过反馈和评分进行回答。(4) 大语言模型推断用户的偏好并重写方案。“新”表示候选集之外的取值。在 K 轮之后，大语言模型再次调用，为最终场景进行家具布置。

这会使两个轮次的计划难以比较。然而，固定的选项列表无法适用于所有用户。因此，我们让大语言模型在第一轮开始前为每位用户编写一次选项，并在整个交互过程中保持不变。

给定房间类型、 q 和 $\mathcal{P}$后，一次大语言模型调用会列出房间中可见的设计方面，例如墙面饰面或座椅，并为每个方面列出十二个候选值。其中六个候选项是大语言模型认为最可能适合此人的选项，另外六个是其他常见选项，因此该空间也包含输入未提示的选项。每个候选项都有一个简短标签、一句描述其外观的描述，如果名称中包含颜色，还会提供一个十六进制代码。由于大语言模型使用自己的措辞命名这些方面，一个方面并不总是对应属性列表中的一个条目。因此，我们使用 Jev [76], 决策模型将这些方面分配给属性；该模型接收一个上下文和一组多项选择题，并为每道问题的每个选项返回一个概率。上下文是房间类型，每个方面对应一个问题，其中引用该方面的名称、描述及其前三个候选标签。选项是属性列表中的 M 条目，每个条目均由其名称和描述给出，另加“无”。如果某个属性的概率至少为 0.5，则将该方面分配给其概率最高的属性；此外，补充材料第B.1节给出的其他规则确保每个属性至少接收一个方面。分配给属性a 的各方面的候选项构成其候选集 C_a，所有候选集构成该交互的偏好空间（图4）。

计划中的值要么是来自 C_a 的候选项，要么是一个新值；当没有候选项符合反馈要求时，大语言模型会以与候选项相同的形式写入该新值。第二次大语言模型调用会读取 c 和所有候选集，简要分析该用户可能的偏好，然后ˆ选择初始计划 S_1。

4.2. 计划更新

一轮交互的反馈只会提及房间的部分部件，并且描述的是用户所看到的内容，而不是计划所规定的内容。例如“房间感觉太暗了”这样的评论，可能涉及墙面颜色、地板或灯具，也可能涉及场景中显示效果与计划不同的某个部分。因此，计划更新会将交互的完整历史记录提供给大语言模型调用，包括每个展示场景所包含的内容。

在第 k 轮之后，一次大语言模型调用会写出下一份计划，

$$\hat{S}_{k+1} = \mathrm{LLM}\big(c, \{(a, C_a)\}_a, (\hat{S}_{k'}, D_{k'}, \eta_{k'}, z_{k'})_{k'=1}^{k}\big), \tag{2}$$

其中 $\eta_{k'}$ 和 $z_{k'}$ 分别是第 k'轮的反馈和评分，$D_{k'}$ 而第轮的场景描述 k'则是：对于每个属性，用文本说明为其取值构建了什么（第4.3节）。

大语言模型首先根据所有反馈和评分推断用户的偏好，然后编写一份完整的计划。由于计划必须再次为每个属性指定一个值，大语言模型会针对房间的每个部分决定是保留现有值、切换到另一个候选项，还是填写一个新值，包括反馈中未提及的部分。因此，对某个部分的反应也可能改变其他部分：例如，觉得房间冷清的一句评论，可能会同时促使灯光变得更暖、木材色调变得更暖。ˆ 在第 K 轮之后，同一次调用会写出最终计划 S_{K+1}，我们据此构建最终场景 K。



图4. 第 1 轮反馈后计划如何变化（图3的交互）。反馈中提及的墙壁和地板会获得新取值；未被提及的窗户则会移动到符合推断出的暖度的候选项。

并由此构建最终场景 X。

4.3. 场景实例化

场景实例化必须让每个取值都能改变房间中对应的部分。取值是自由文本，因此我们首先将其转换为规格，即对该取值所决定部分的结构化描述：其中包含哪些物体及其数量、它们的位置，以及它们具有何种饰面、颜色或灯光。Jev 会回答构成该规格的各个问题。每个属性都属于某个组件类别，例如地板、座椅或照明；每个类别都有一组固定的多项选择题。每个问题都会引用该属性、其描述和取值文本，我们会为每个问题选取概率最高的选项。颜色取自该取值的十六进制代码。规格示例及更多说明见补充材料第C.1节和C.9节。

然后，我们根据规格，基于物体检索和 Holodeck 的基于约束的布局来构建场景 [86]。物体会从 Objaverse [19]资产中按照尺寸，以及其视觉和文本内容与取值文本的相似度进行检索；地板和墙的材料则从 ambientCG [20]中检索。我们的布局在约束来源方面有所不同：规格中的摆放位置答案，例如一组家具所倚靠的墙面及其朝向，会取代大语言模型编写的约束。搜索会按照概率顺序尝试这些组合，并保留第一个满足硬约束的布局，例如不存在碰撞的布局。由于从固定资产池中检索无法设置一件家具的材料或颜色，因此我们会在检索到的资产的包围盒内，以规格所设定的形状、材料和颜色，通过程序化方式重建家具。补充材料第 C.1至C.4 节给出了相关类别和问题、检索、程序化库以及放置规则。

只有最终场景 X 才会进行家具布置，因为计划决定了房间的外观，却不涵盖布置好的房间里包含的每一件物品。另一次大语言模型调用会在空闲空间中增添更多家具、家具台面上的物体以及墙面装饰物件，但不会移动计划中的任何物体。为避免新增内容与计划相矛盾，同一条回复会先列出每项属性的取值会排除哪些物件类型，并删除这些类型的物件。补充材料第B.3节和C.9节提供了这些清单示例及进一步说明。

最后，模板会根据已构建的内容撰写场景描述，其中涵盖 D_k 每项属性；也就是说，描述依据的是规格、摆放位置、检索到的资产的说明文字及其材料和颜色，而不是属性值本身的措辞，因此描述反映的是用户实际可见的内容。补充材料提供了场景描述的详细信息和示例（第C.8节和C.9节）、所有大语言模型调用的提示词（第B节）以及算法清单（第B.4节）。

5. 基准测试

诸如 3D-FRONT [22] 这样的室内场景数据集包含房间，却没有用户，因此我们构建了一个基准测试，其中每个实例都具有已知的隐藏偏好，以及一项说明其中受控部分的输入。该基准测试涵盖十种房间类型，每种类型都有一个包含 $M = 10$ 属性的模式，每个属性都是房间的一个可见组成部分，例如墙面涂料、地板或床上用品（共 100 项属性，包含 483 个取值）。每种房间类型都有八个偏好配置文件：每个偏好配置文件代表一种一致的设计，由大语言模型根据模式中的十项取值编写，并作为其对应实例的隐藏偏好 G^*。对于每个偏好配置文件，GPT-5.4-mini [55]会撰写两种人物设定及其对应的两种形式的指令，因此每种形式有 160 个实例，总计 320 个实例。在显式形式中，人物设定和指令会说明十项隐藏值中的四到六项，因此系统可以从输入中读取大约一半的偏好配置文件内容；在隐式形式中，它们不说明任何隐藏值，只暗示人物的偏好，因此必须通过交互推断偏好配置文件。我们按偏好配置文件划分实例，得到一个开发子集和一个测试子集：开发子集每种形式包含 40 个实例，用于开发方法；测试子集每种形式包含 120 个实例，本文第 6 节的所有结果均在该子集上测量。

模拟用户 Qwen3.8-27B [61], 以约 120 字的自由形式品味叙述表达其偏好，涵盖所有隐藏值。每轮中，它都会收到所展示场景的渲染图像、角色设定、叙述以及此前的反馈和评分，并写下三至五句反馈以及 1 到 10 分的满意度评分；它看不到系统的计划或模式。与我们的场景实例化相分离的基于规则的构建会直接抽取每个模式取值。对于视觉语言模型无法从其渲染图像中读回的取值，我们进行了修订，并且相同的构建会从其渲染

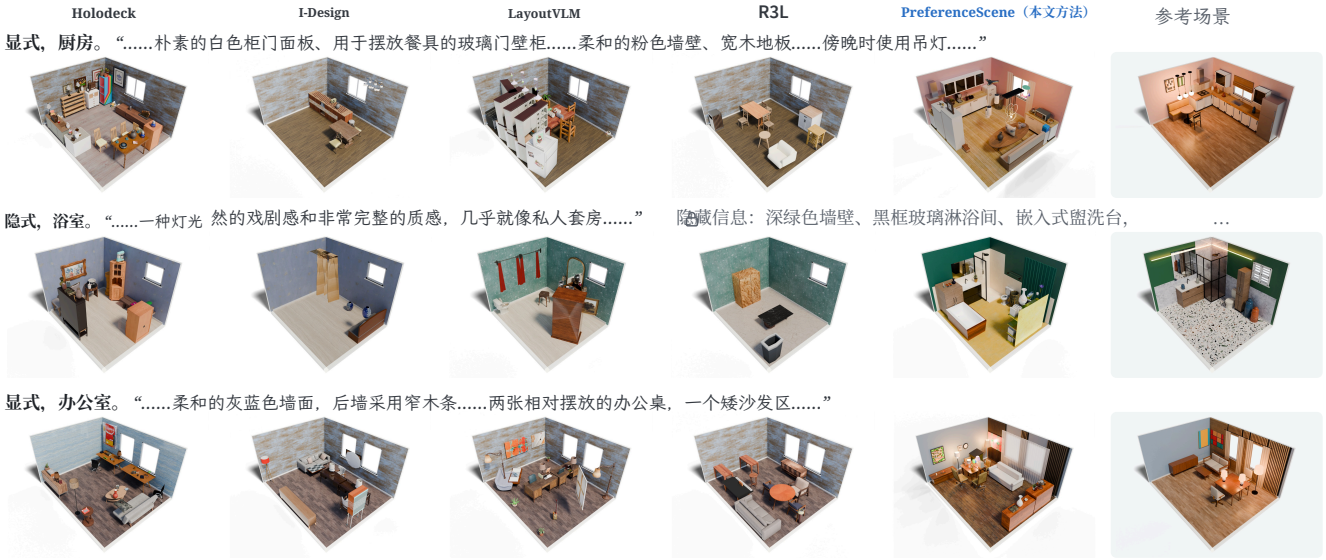


图 5。在三个测试实例上经过 $K = 3$ 轮次后的最终场景。每一行给出了形式、房间类型和指令摘录，而隐式 行还给出了指令中未明确说明的三个隐藏值。

图像中构建参考场景，为每个档案各生成一个。第 6.1 节的评估器从参考场景中恢复出 89.3% 的隐藏值，而随机猜测的准确率为 21.4%。补充材料第 A节和第 J 节给出了模式、统计数据、一个实例示例和提示词。

6. 实验

6.1. 设置

运行与模型。每次交互包含 $K = 3$ 轮次。每个系统对每个实例运行四次，模拟用户和系统的大语言模型分别使用 0–3 号随机种子，并报告这些随机种子下的均值和标准差。GLM-4.6V-Flash [29] 是评估器，而 Gemma-4-31B [75] 是每个系统所使用的大语言模型。

基线方法。 我们与五种已发表的文本到场景基线进行比较：Holodeck [86], LayoutGPT [21], I-Design [12], LayoutVLM [70] 和 R3L [25]。它们都不接受反馈，因此我们让每种方法都在与我们的方法相同的交互中运行。在第 1 轮，系统根据输入 c生成一个场景。在之后的每一轮中，它根据 c以及交互历史记录生成一个新场景，即它所展示的每个场景的描述、用户对此的反馈以及评分。我们还基于 Holodeck 构建了一个编辑器，*Holodeck-Edit*，该编辑器在每一轮中编辑其上一场景的地板、墙和对象规格，并保留未改变对象的资产。在 *Oracle* 条件下，与 PRELUDE [23], 一样，系统以文本形式接收隐藏偏好 G^* 并生成一个场景，以展示它在不受限时能够实现的效果。必须进行推断。在没有反馈的情况下，Holodeck-Edit 会展示 Holodeck 的场景，因此其 Oracle 值与 Holodeck 相同（* 见表 1）。所有系统使用相同的房间外壳（每种房间类型的墙壁、地板、门和窗户）、相同的 6,793 个 Objaverse [19] 资产池以及相同的渲染器（Blender Cycles [9]）和摄像机。补充材料第D.2 节和E.4节介绍了已发表的基线方法的适配、失败生成的处理方式以及进一步的变体。LayoutGPT 仅针对卧室和客厅提供上下文示例，因此在每种形式下，它都运行于这两种房间类型的 24 个测试集实例上。

指标。 我们的主要指标是 pAMR（公式（1）），评估器会根据最终场景的两张渲染图像——一张从上方俯视的剖视图和一张平视图——回答每个属性对应的问题。由于视觉语言模型的回答会受到选项顺序的影响，评估器会按给定顺序和相反顺序各回答一次每个问题，pAMR 取这两次回答的平均值。对于场景质量，我们遵循 R3L [25]: 真实度、 功能性 和指令遵循度 分别是评估器根据其评估提示词对真实性、功能性和指令遵循情况给出的 1–10 分。*Coll.* 是根据 SceneEval 的网格相交测试得出的发生碰撞的物体百分比 [72]。对于成本，我们报告每次交互的大语言模型调用次数，以及每个场景的生成时间（时间），即系统自身各步骤的运行时间除以一次交互中的 $K + 1$ 场景数。我们在单独的 计时测试中测量时间：每个系统进行八次交互，在相同硬件上逐次运行。在每张表中，每项指标的最佳值以粗体标出，次佳值加下划线。

表 1. 最终场景的偏好匹配度：pAMR（%，↑），四个随机种子上的均值± 标准差（Oracle：一个随机种子；来自显式输入和隐藏值的一个场景）。粗体：最佳；下划线：次佳。 *与 Holodeck 使用相同场景。 †仅限卧室和客厅。

方法	隐式	显式	Oracle
Holodeck [86]	41.4±0.6	44.0±0.2	45.3
LayoutGPT† [21]	25.2±1.3	27.9±1.0	25.8
I-Design [12]	41.0±0.5	44.3±1.1	43.8
LayoutVLM [70]	43.1±1.4	45.1±0.3	47.8
R3L [25]	40.8±0.6	43.9±1.0	45.5
Holodeck-Edit	42.6±1.4	45.6±0.7	45.3*
PreferenceScene（本文方法）	**49.6±1.7**	**56.6±0.3**	**59.4**

表2。两种形式的最终场景质量（粗体：最佳；下划线：次佳），四个随机种子上的均值 ± 标准差。真实度、功能性、指令遵循度：1–10分的评估者评分。Coll.：碰撞物体的百分比。时间：计时测试中每个场景所需的秒数。 †仅限卧室和客厅。

方法	真实度。↑	功能性。↑	指令遵循度。↑	Coll. ↓	时间 ↓
Holodeck	7.9±0.0	7.2±0.0	7.9±0.1	5.7±0.5	102
LayoutGPT†	7.0±0.1	6.6±0.1	6.5±0.1	31.2±2.1	**12**
I-Design	7.6±0.0	6.9±0.0	7.5±0.1	2.8±0.4	340
LayoutVLM	7.6±0.0	6.9±0.0	7.6±0.0	41.9±0.8	361
R3L	7.6±0.0	7.0±0.0	7.4±0.1	**1.4±0.4**	332
Holodeck-Edit	7.8±0.1	7.2±0.1	7.8±0.0	4.5±0.3	79
PreferenceScene（本文方法）	**8.2±0.0**	**7.5±0.0**	**8.4±0.0**	2.8±0.5	43

6.2. 与基线方法的比较

偏好匹配。 PreferenceScene 在显式 形式下达到 56.6 分，在隐式 形式下达到 49.6 分，在两种形式下均为最高匹配度（表 1）。表现最佳的已发表基线方法 Lay-outVLM 得分分别为 45.1 和 43.1，Holodeck-Edit 得分分别为 45.6 和 42.6。PreferenceScene 在显式 形式下至少高于每个基线方法 11 分，在隐式 形式下至少高出 6 分。

定性结果。 在 显式 厨房（图5）中，PreferenceScene 展示了指令要求的粉色墙壁、宽木地板和白色柜门面板，与参考场景一致。在 隐式 浴室中，隐藏值包括深绿色墙壁、黑框玻璃淋浴间和嵌入式盥洗台，而输入中并未说明这些内容。反馈后，PreferenceScene 展示了这三项，而已发表的基线方法展示的是蓝灰色或蓝绿色墙壁。在办公室中，只有 PreferenceScene 展示了指令要求的木格栅墙和琥珀色灯光。补充材料第G 节和H 节展示了更多结果、失败案例以及该方法的局限性。

对整个房间的控制。 由于已发表的基线方法至多只能从固定库中选择墙面和地面材料，因此无法根据反馈调整房间的其他部件。即使在为其提供隐藏偏好的预言机条件下，其场景中最多也只能呈现 47.8% 的隐藏值（表 1）。PreferenceScene 在预言机条件下展示了 59.4%，仅凭反馈，其结果在两种形式下都超过了每个已发表基线方法的预言机取值。补充材料第 E.11 节和 E.12节报告了基线方法使用我们的场景实例化时的匹配度，以及所有系统都能更改的属性上的匹配度。

反馈的作用。 PreferenceScene 的匹配度随着每一轮而提升（见补充图 S3），显式 形式下从 51.9 提升至 56.6，隐式 形式下从 37.5 提升至 49.6 的 隐式 形式。在隐式 形式中，反馈是主要证据，从第 1 轮到最终场景的提升幅度 +12.1 个百分点，高于所有基线方法（最多 9.7 个百分点）。反馈也较为稀疏：大语言模型标注器（Ministral-3-14B）发现，每轮反馈会提及十种属性中的四到五种，约三分之一的隐藏属性在交互中从未被提及。对于这些从未被提及的属性，PreferenceScene 对隐藏值的匹配度分别为 54.0% 和 43.2%，而基线方法最高分别为 44.4% 和 37.6%；它还会使房间中未被提及的部件更符合用户偏好。补充材料第 E.14 节测试了噪声反馈。

场景质量与成本。 在表2中，PreferenceScene 的真实感、功能性和指令得分最高（分别为 8.2、7.5 和 8.4，而其他方法最高分别为 7.9、7.2 和 7.9）。其物体中有 2.8% 发生碰撞，且没有物体越界。与 PreferenceScene 的一次交互需要进行六次大语言模型调用，并向 Jev 发出约 25 个请求。其每个场景的生成时间在七个系统中排名第二低。

鲁棒性。 我们使用来自另一个模型系列的第二评估者 Ministral-3-14B [47],以及一种允许评估者回答“不可见”的变体，对所有最终场景重新评分。在这两种情况下，PreferenceScene 的得分都高于每个基线。使用第二评估者时，其显式得分为 52.5（显式），隐式得分为 46.5（隐式），而其他方法最高分别为 44.9 和 43.0。使用“不可见”选项时，其得分为 47.8 和 40.9，而其他方法最高分别为 35.8 和 33.4。补充材料第 E.9 节和 E.13节还报告了将 Gemma- 4-31B 替换为托管大语言模型，以及将 Jev 替换为开放模型时的方法表现。

6.3. 消融实验

表 3 在餐厅和办公室测试实例上（每种形式 24 个）一次移除一个组件。 无迭代对初始计划的场景进行评分，因此与完整方法之间的差距完全体现了反馈的效果：在 显式 形式下约为 3 分，在 隐式

表3. 餐厅和办公室测试实例上的消融实验：pAMR（%，↑），四个随机种子上的均值 ± 标准差。粗体：最佳；下划线：次佳。

变体	显式	隐式
PreferenceScene（完整）	**67.1±1.3**	**56.5±2.9**
无迭代（初始计划）	64.0±1.4	40.6±2.2
不包含较早轮次	66.4±2.9	50.1±2.9
不含新取值	64.0±2.8	48.4±2.4

形式下。 不使用较早轮次时，计划更新（式（2））只读取历史记录中的最新轮次。匹配度随之降至 66.4 和 50.1：大语言模型会利用所有轮次的反馈，而不只是最新一轮。 不使用新值时，候选集之外的取值 C_a 会被拒绝，属性则保留原有取值。匹配度随之降至 64.0 和 48.4，不过在完整方法的最终计划中，新值仅约占六分之一（16.2% 和 16.6%）：当用户要求候选项均无法提供的内容时，计划必须能够跳出候选集。补充材料第 E.15 节报告了更多消融实验，包括不使用显式计划、不在历史记录中使用场景描述，以及不使用场景实例化规则的变体。

6.4. 用户研究

20 位参与者各完成了 36 次试验，每种基线方法进行六次。在每次试验中，他们阅读指令、人物角色和品味叙述，并查看 PreferenceScene 与某种基线方法针对同一实例生成的最终场景；场景不标注系统名称，且以随机顺序呈现，如 Scene-Smith [58]所示。他们选择更符合人物偏好的场景，以及总体设计更好的场景，不考虑这些偏好。在 720 次试验中，79.7% 的情况下他们选择了 PreferenceScene 作为更符合偏好的场景，80.8% 的情况下选择它作为设计更好的场景（图 6a）。与各个基线方法相比，这两项比例分别为 74.2–91.7% 和 68.3–91.7%；在对 12 项测试进行 Benjamini-Hochberg校正的双侧二项检验中，每项比例均与 50% 存在显著差异（$p < 0.0001$）。参与者还使用五点评分量表，对 24 个单独场景的真实感、功能性和指令遵循情况进行评分，评分时依据评分标准、区组设计以及 R3L [25]中的 1–10 分映射方式。参与者的平均评分显示，在七个系统中，PreferenceScene 在真实感、功能性和指令遵循方面均排名第一（图 6b）。与 R3L [25],一样，我们将表 2 中的评估器与参与者进行比较：场景的平均评分与其评估器评分之间的 Kendall's τ_b 在真实感、功能性和指令遵循方面分别为 0.17、0.33 和 0.36；每个场景的评分者随机两半之间则分别为 0.53、0.42 和 0.48。补充

补充章节F 给出了各系统的比率、威尔逊区间，以及与评估者评分并列的评分。补充章节F.1 报告了与真实用户的交互。

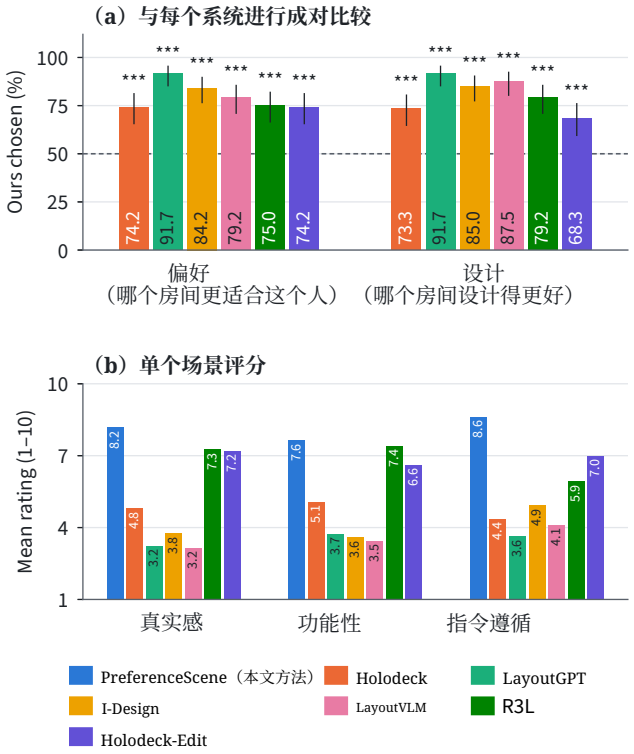


图 6。用户研究（20 名参与者）。（a）在每个系统的对比中选择 PreferenceScene 的答案占比，带有 95% Wilson 置信区间；虚线：50%；* $p < 0.05$、** $p < 0.01$、*** $p < 0.001$（经校正的二项检验）。（b）平均评分（1–10）。

7. 结论

我们研究了隐藏用户偏好下的个性化室内场景合成：系统通过一个人对所展示房间的反应，学习此人对房间的需求。PreferenceScene 将其对用户偏好的估计表示为一份显式计划，涵盖每项设计属性的一系列候选值；它根据稀疏的自由形式反馈修订整份计划，并将每个值实例化为房间中的独立部分。在我们的基准测试中，与五种已发表的文本到场景基线方法相比，PreferenceScene 在获得相同反馈的情况下能更好地匹配隐藏偏好；即使这些系统直接获得偏好，PreferenceScene 的表现仍更好。用户研究的参与者认为，无论在偏好还是设计方面，PreferenceScene 生成的场景都优于所有基线方法。

局限性与未来工作。 我们的研究仅限于一个房间和一位用户。完整住宅、由多人组成的家庭住户以及不断变化的偏好仍有待未来研究。

大语言模型使用声明

Claude（Anthropic）通过 Claude Code 协助作者完成了实验和本文写作。

作者对其内容承担全部责任。

参考文献

[1] Ryuya Akase 和 Yoshihiro Okada。基于交互式进化计算的自动三维家具布局。载于第七届复杂、智能与软件密集型系统国际会议（CISIS），第726–731页。IEEE，2013年。37[2] Ryuya Akase 和 Yoshihiro Okada。结合联合分析的交互式进化计算网络多用户三维房间布局系统。载于第七届视觉信息通信与交互国际研讨会（VINCI）论文集，第178–187页，2014年。37[3]Stephane Aroca-Ouellette、Natalie Mackraz、Barry-John Theobald 和 Katherine Metcalf。PREDICT：通过评估从候选轨迹中推断出的分解偏好进行偏好推理。arXiv预印本 arXiv:2410.06273，2024年。35[4] Raul Astudillo、Zhiyuan Jerry Lin、Eytan Bakshy 和 Peter Frazier。qEUBO：一种用于偏好贝叶斯优化的决策论获取函数。载于第二十六届人工智能与统计国际会议（AISTATS）论文集，第1093–1114页，2023年。35[5] David Eric Austin、Anton Korikov、Armin Toroghi 和 Scott Sanner。采用基于大语言模型的获取函数进行自然语言偏好引出的贝叶斯优化。载于第十八届ACM推荐系统会议（RecSys）论文集，第74–83页。ACM，2024年。28[6] Krisztian Balog 和 ChengXiang Zhai。生成式人工智能时代的用户模拟：用户建模、合成数据生成和系统评估。arXiv预印本 arXiv:2501.04410，2025年。37[7] Erdem Bıyık 和 Dorsa Sadigh。基于批量主动偏好的奖励函数学习。载于第2 528，2018 519 35届机器人学习会议（CoRL）论文集，第–页。[8] Erdem Bıyık、Aditi Talati 和 Dorsa Sadigh。APReL：主动偏好奖励学习算法库。载于第十七届ACM/IEEE人机交互国际会议（HRI）论文集，第613–617页，2022年。35[9] Blender在线社区。Blender——三维建模与渲染软件包。https://www.blender.org，2025年。4.5版。6，21[10] Craig Boutilier。偏好引出问题的部分可观测马尔可夫决策过程表述。AAAI/IAAI，1：239–246，2002年。2[11] Martin JJ Bucher 和 Iro Armeni。ReSpace：文本驱动的自回归三维室内场景合成与编辑。arXiv预印本 arXiv:2506.02459，2025年。2，3，34[12] 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。载于欧洲计算机视觉会议论文集，第217–234页。Springer，2024年。2，6，7，21，33，34[13] Kang Chen、Kun Xu、Yizhou Yu、Tian-Yi Wang 和 Shi-Min Hu。Magic decorator：自动材料推荐，用于室内数字场景。《ACM图形学汇刊》（TOG），34(6):1–11，2015年。2[14] Li Chen 和 Pearl Pu。《基于批评的推荐系统：综述与新兴趋势》。《用户建模与用户适应交互》，22(1):125–150，2012年。2, 35[15] Francis DK Ching 和 Corky Binggeli。《室内设计图解》。John Wiley & Sons，2012年。2, 13[16] Jaemin Cho、Yushi Hu、Jason Baldridge、Roopal Garg、Peter Anderson、Ranjay Krishna、Mohit Bansal、Jordi Pont-Tuset 和 Su Wang。《戴维森场景图：提升文本到图像生成细粒度评估的可靠性》。载于《国际学习表征会议》（ICLR），2024年。37[17] Paul F. Christiano、Jan Leike、Tom Brown、Miljan Martic、Shane Legg 和 Dario Amodei。《基于人类偏好的深度强化学习》。载于《神经信息处理系统进展》（NeurIPS），2017年。35[18] Matt Deitke、Eli VanderBilt、Alvaro Herrasti、Luca Weihs、Kiana Ehsani、Jordi Salvador、Winson Han、Eric Kolve、Aniruddha Kembhavi 和 Roozbeh Mottaghi。《ProcTHOR：利用程序化生成实现大规模具身人工智能》。载于《神经信息处理系统进展》（NeurIPS），第5982–5994页，2022年。18[19] Matt Deitke、Dustin Schwenk、Jordi Salvador、Luca Weihs、Oscar Michel、Eli VanderBilt、Ludwig Schmidt、Kiana Ehsani、Aniruddha Kembhavi 和 Ali Farhadi。《Objaverse：带标注的三维物体宇宙》。载于《IEEE/CVF计算机视觉与模式识别会议》，第13142–13153页，2023年。5, 6, 18, 21[20] Lennart Demes。ambientCG：公有领域基于物理的渲染材质。https://ambientcg.com，2025年。5, 18, 21[21] Weixi Feng、Wanrong Zhu、Tsu-jui Fu、Varun Jampani、Arjun Akula、Xuehai He、Sugato Basu、Xin Eric Wang 和 William Yang Wang。《LayoutGPT：利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》，36 18250，2023年 18225 2, 6, 7, 21, 33 : – .[22] Huan Fu、Bowen Cai、Lin Gao、Ling-Xiao Zhang、Jiaming Wang、Cao Li、Qixun Zeng、Chengyue Sun、Rongfei Jia、Binqiang Zhao 等。《3D-FRONT：带布局和语义信息的三维配家具房间》。载于2021年IEEE/CVF国际计算机视觉会议（ICCV），第10913–10922页。IEEE，2021年。5, 22, 33[23] Ge Gao、Alexey Taymanov、Eduardo Salinas、Paul Mineiro 和 Dipendra Misra。《通过从用户编辑中学习潜在偏好来对齐大语言模型智能体》。《神经信息处理系统进展》，37:136873–136896，2024年。2, 3, 6, 35, 37[24] Javier Gonz´alez、Zhenwen Dai、Andreas Damianou 和 Neil D. Lawrence。《偏好贝叶斯优化》。载于第34届国际机器学习会议（ICML），第1282–1291页，2017年。35[25] Zhifeng Gu、Yuqi Wang 和 Bing Wang。《R^3L：基于相对空间关系推理三维布局》。arXiv预印本 arXiv:2605.06758，2026年。2, 6, 7, 8, 21, 22, 31, 33, 37[26] Xiaoxiao Guo、Hui Wu、Yu Cheng、Steven Rennie、Gerald Tesauro 和 Rogerio Feris。《基于对话的交互式图像

检索。《神经信息处理系统进展》，31，2018年。2, 3, 35[27] Meera Hahn、Wenjun Zeng、Nithish Kannen、Rich Galt、Kartikeya Badola、Been Kim 和 Zi Wang。《不确定条件下用于多轮文本生成图像的主动式智能体》。arXiv预印本 arXiv:2412.06771，2024年。2, 3, 36, 37[28] Kunal Handa、Yarin Gal、Ellie Pavlick、Noah Goodman、Jacob Andreas、Alex Tamkin 和 Belinda Z Li。《利用语言模型进行贝叶斯偏好引出》。arXiv预印本 arXiv:2403.05534，2024年。2, 35[29] Wenyi Hong、Wenmeng Yu、Xiaotao Gu、Guo Wang、Guobing Gan、Haomiao Tang、Jiale Cheng、Ji Qi、Junhui Ji、Lihang Pan 等。《GLM-4.5 V和 GLM-4.1 V-Thinking：利用可扩展强化学习实现通用多模态推理》。arXiv预印本 arXiv:2507.01006，2025年。6[30] Yushi Hu、Benlin Liu、Jungo Kasai、Yizhong Wang、Mari Ostendorf、Ranjay Krishna 和 Noah A Smith。《TIFA：通过问答实现准确且可解释的文本生成图像忠实度评估》。载于2023年IEEE/CVF国际计算机视觉会议（ICCV），第20349–20360页。IEEE，2023年。3, 37[31] Ziniu Hu、Ahmet Iscen、Aashi Jain、Thomas Kipf、Yisong Yue、David A Ross、Cordelia Schmid 和 Alireza Fathi。《SceneCraft：一种将三维场景合成为Blender代码的大语言模型智能体》。载于ICML，第19252–19282页，2024年。2, 33[32] J. D. Hunter。《Matplotlib：二维图形环境》。《Computing in Science & Engineering》，9(3):90–95，2007年。18[33] Ben Hutchinson、Jason Baldridge 和 Vinodkumar Prabhakaran。《场景描述到图示任务中的欠规格化》。载于亚太计算语言学协会第2届会议暨自然语言处理国际联合会议第12届会议（第1卷：长篇论文集），第1172–1184页，2022年。2[34] Gabriel Ilharco、Mitchell Wortsman、Ross Wightman、Cade Gordon、Nicholas Carlini、Rohan Taori、Achal Dave、Vaishaal Shankar、Hongseok Namkoong、John Miller、Hannaneh Hajishirzi、Ali Farhadi 和 Ludwig Schmidt。《Open-CLIP》。软件，2021年。18[35] Hong Jun Jeon、Smitha Milli 和 Anca Dragan。《奖励理性（隐式）选择：奖励学习的统一形式体系》。载于《神经信息处理系统进展》，第4415–4426页，2020年。36[36] Haiyan Jiang、Deyu Zhang、Dongdong Weng、Weitao Song 和 Henry Been-Lirn Duh。《HOG-Layout：通过视觉语言模型实现分层三维场景生成、优化与编辑》。载于IEEE/CVF计算机视觉与模式识别会议（CVPR），2026年。33[37]Adriana Kovashka、Devi Parikh 和 Kristen Grauman。《WhittleSearch：利用相对属性反馈进行图像搜索》。载于IEEE计算机视觉与模式识别会议（CVPR），第2973 2980页，2012年 35 – .[38] Yuki Koyama、Issei Sato 和 Masataka Goto。《用于交互式视觉设计优化的序列画廊》。《ACM图形学汇刊》（TOG），39(4):88–1，2020年。3, 35[39] Woosuk Kwon、Zhuohan Li、Siyuan Zhuang、Ying Sheng、Lianmin Zheng、Cody Hao Yu、Joseph E. Gonzalez、Hao Zhang 和 Ion Stoica。利用 Paged Attention 高效管理大语言模型服务的内存。载于 SOSP，2023年。15[40] Hyeonmin Lee、Zheng Wei、Kyungmin Kwon、Jumin Seo、Jiwon Park 和 Hayoung Oh。SPHERE：通过大语言模型增强的空间偏好学习和人在回路强化学习实现自适应虚拟现实室内场景生成。IEEE 可视化与计算机图形学汇刊，2026年。35[41] Matan Levy、Rami Ben-Ari、Nir Darshan 和 Dani Lischinski。聊天成就完美：基于聊天的图像检索。载于神经信息处理系统进展（NeurIPS），2023年。35[42] Belinda Li、Alex Tamkin、Noah Goodman 和 Jacob Andreas。利用语言模型引出人类偏好。载于国际学习表征会议，第80984–81013页，2025年。2, 36[43] Sarah Lichtenstein 和 Paul Slovic（编）。偏好的建构。剑桥大学出版社，2006年。2[44] Chenguo Lin 和 Yadong Mu。InstructScene：基于指令并利用语义图先验的三维室内场景合成。载于国际学习表征会议（ICLR），2024年。22, 33[45] Lu Ling、Chen-Hsuan Lin、Tsung-Yi Lin、Yifan Ding、Yu Zeng、Yichen Sheng、Yunhao Ge、Ming-Yu Liu、Aniket Bera 和 Max Li。Scenethesis：用于三维场景生成的语言与视觉智能体框架。载于国际学习表征会议（ICLR），2026年。33[46] Gabrielle Littlefair、Niladri Shekhar Dutt 和 Niloy J. Mitra。FlairGPT：重新利用大语言模型进行室内设计。Computer Graphics Forum，44(2)：e70036，2025年。33[47] Alexander H Liu、Kartik Khandelwal、Sandeep Subramanian、Victor Jouault、Abhinav Rastogi、Adrien Sad´e、Alan Jeffares、Albert Jiang、Alexandre Cahill、Alexandre Gavaudan 等。Ministral 3 08584，2026 2601 7。arXiv 预印本 arXiv:。。[48] Tianqiang Liu、Aaron Hertzmann、Wilmot Li 和 Thomas Funkhouser。三维家具模型的风格兼容性。ACM图形学汇刊（TOG），34(4)：1–9，2015年。2[49] Vivian Liu 和 Lydia B Chilton。文生图生成模型提示工程设计指南。载于2022年 CHI 计算系统人因会议论文集，第1–23页，2022年。2[50] Xiujin Liu、Tianyu Yang、Yilun Zhao 和 Xiangliang Zhang。DesignAgent3D：通过设计师式多模态推理实现交互式三维场景编辑。arXiv 预印本 arXiv:2608.21438，2026年。34[51] Aman Madaan、Niket Tandon、Prakhar Gupta、Skyler Hallinan、Luyu Gao、Sarah Wiegreffe、Uri Alon、Nouha Dziri、Shrimai Prabhumoye、Yiming Yang、Shashank Gupta、Bodhisattwa Prasad Majumder、Katherine Hermann、Sean Welleck、Amir Yazdanbakhsh 和 Peter Clark。Self-Refine：利用自我反馈进行迭代优化。载于神经信息处理系统进展（NeurIPS），2023年。37[52] Paul Merrell、Eric Schkufza、Zeyang Li、Maneesh Agrawala 和 Vladlen Koltun。使用室内设计指南进行交互式家具布局。ACM图形学汇刊，30 (4)：87，2011年。

内设计指南。ACM图形学汇刊，30 (4)：87，2011年。33[53] Ofir Nabati、Guy Tennenholtz、ChihWei Hsu、Moonkyung Ryu、Deepak Ramachandran、Yinlam Chow、Xiang Li 和 Craig Boutilier。偏好自适应的序列式文生图生成。arXiv 预印本 arXiv:2412.10419，2024年。2, 35[54] Seongrae Noh、SeungWon Seo、Gyeong-Moon Park 和 HyeongYeop Kang。编辑即行动：面向开放词汇三维室内场景编辑的目标回归规划。arXiv 预印本 arXiv:2603.17583，2026年。2, 34[55] OpenAI。GPT-5.4 mini。https : / / developers . openai.com/api/docs/models/gpt-5.4-mini，2026年。访问日期：2026-11-17。5[56] Arjun Panickssery、Samuel R Bowman 和 Shi Feng。大语言模型评估器能识别并偏好自身生成的内容。神经信息处理系统进展，37：68772–68802，2024年。3, 37[57] Despoina Paschalidou、Amlan Kar、Maria Shugrina、Karsten Kreis、Andreas Geiger 和 Sanja Fidler。ATISS：用于室内场景合成的自回归变换器。载于神经信息处理系统进展（NeurIPS），第12013–12026页，2021年。22, 33[58] Nicholas Pfaff、Thomas Cohn、Sergey Zakharov、Rick Cory 和 Russ Tedrake。SceneSmith：用于生成可仿真室内场景的智能体方法。载于第四十三届国际机器学习会议，2026年。2, 8, 22, 31, 33[59] Poly Haven。Poly Haven：公有领域HDRI、纹理和模型。https://polyhaven.com，2025年。19[60] Hou In Derek Pun、Hou In Ivan Tam、Austin T Wang、Xiaoliang Huo、Angel X Chang 和 Manolis Savva。HSM：用于多尺度室内场景生成的层次化场景模式。载于2026年国际三维视觉会议（3DV），第1356–1367页。IEEE，2026年。2, 33[61] Qwen团队。Qwen3.8-27B。https://huggingface.co/Qwen/Qwen3.8-27B，2026年。访问日期：2026-11-17。5[62] Rafael Rafailov、Archit Sharma、Eric Mitchell、Christopher D. Manning、Stefano Ermon 和 Chelsea Finn。直接偏好优化：你的语言模型其实暗中是一个奖励模型。载于神经信息处理系统进展（NeurIPS），2023年。35[63] Rajalaxmi Rajagopalan、Debottam Dutta、Yu-Lin Wei 和 Romit Roy Choudhury。通过人在回路贝叶斯优化实现个性化图像生成。arXiv 预印本 arXiv:2602.02388，2026年。35[64] Xingjian Ran、Yixuan Li、Linning Xu、Mulin Yu 和 Bo Dai。通过空间推理直接生成三维室内场景的数值布局。载于神经信息处理系统进展（NeurIPS），2025年。33[65] Nils Reimers 和 Iryna Gurevych。Sentence-BERT：使用孪生 BERT 网络生成句子嵌入。载于2019年自然语言处理中的经验方法会议和第九届国际自然语言处理联合会议（EMNLP-IJCNLP）论文集，第3980–3990页。计算语言学协会，2019年。17[66] Dorsa Sadigh、Anca Dragan、Shankar Sastry 和 Sanjit Se- shia。基于主动偏好的奖励函数学习。载于《Robotics: Science and Systems》会议录，马萨诸塞州剑桥，2017年。2、3、35[67] Donald A Sch¨on。将设计视为与设计情境中的材料进行反思性对话。Knowledge-based systems，5(1):3–14，1992年。2[68] Paul Slovic。偏好的建构。American psy- chologist，50(5):364，1995年。2[69] Theodore R. Sumers、Mark K. Ho、Robert D. Hawkins、Karthik Narasimhan 和 Thomas L. Griffiths。通过语言反馈学习奖励。载于AAAI人工智能会议论文集，第6002–6010页，2021年。36[70] 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。载于2025年IEEE/CVF计算机视觉与模式识别会议（CVPR），第29469–29478页。IEEE，2025年。2、6、7、21、22、33、37[71] Hideyuki Takagi。交互式进化计算：融合进化计算优化能力与人工评估能力。IEEE会议录，89(9):1275–1296，2001年。2、36[72] Hou In Ivan Tam、Hou In Derek Pun、Austin T Wang、An- gel 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。载于2026年IEEE/CVF冬季计算机视觉应用会议（WACV），第7355–7365页。IEEE，2026年。6、22、37[73] Fuwen Tan、Paola Cascante-Bonilla、Xiaoxiao Guo、Hui Wu、Song Feng 和 Vicente Ordonez。Drill-down：使用自然语言查询交互式检索复杂场景。载于神经信息处理系统进展（NeurIPS），2019年。35[74] Jiapeng Tang、Yinyu Nie、Lev Markhasin、Angela Dai、Justus Thies 和 Matthias Nießner。DiffuScene：用于生成式室内场景合成的去噪扩散模型。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第20507–20518页，2024年。22、33[75] Gemma Team、Sherif El Abd、Vaibhav Aggarwal、Robin Al- gayres、Alek Andreev、Olivier Bachem、Ian Ballantyne、Cormac Brick、Victor C˘arbune、Michelle Casbon 等。Gemma 4技术报告。arXiv预印本 arXiv:2607.02770，2026年。6[76] TypeSafe。Jev，System One模型（模型标识符 jev- 1.13.0）：文档。https://docs.typesafe.ai/models，2026年。访问日期：2026-10-27。4、16、21、27[77] Paolo Viappiani、Boi Faltings 和 Pearl Pu。使用示例批评与建议进行基于偏好的搜索。Journal of artificial intelligence Research，27 503，2006 465 2 : – 。[78] Can Wang、Hongliang Zhong、Menglei Chai、Mingming He、Dongdong Chen 和 Jing Liao。Chat2Layout：借助多模态大语言模型进行交互式三维家具布局。IEEE可视化与计算机图形学汇刊，2025年。2、34

[79] Xingyao Wang、Zihan Wang、Jiateng Liu、Yangyi Chen、Li- fan Yuan、Hao Peng 和 Heng Ji。MINT：评估大语言模型在结合工具与语言反馈的多轮交互中的表现。载于国际学习表征会议（ICLR），2024年。37[80] Yian Wang、Xiaowen Qiu、Jiageng Liu、Zhehuan Chen、Jit- ing Cai、Yufei Wang、Tsun-Hsuan Wang、Zhou Xian 和 Chuang Gan。Architect：通过分层二维图像修复生成生动且可交互的三维场景。载于神经信息处理系统进展（NeurIPS），2024年。33[81] Jason D Williams、Antoine Raux 和 Matthew Henderson。对话状态跟踪挑战系列：综述。Dialogue & Discourse，7:4–33，2016年。3、35[82] Ga Wu、Kai Luo、Scott Sanner 和 Harold Soh。面向推荐系统的深度语言批评。载于第13届ACM推荐系统会议（RecSys）论文集，第137–145页，2019年。35[83] Hui Wu、Yupeng Gao、Xiaoxiao Guo、Ziad Al-Halah、Steven Rennie、Kristen Grauman 和 Rogerio Feris。Fashion IQ：一个旨在通过自然语言反馈检索图像的新数据集。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第11307–11317页，2021年。35[84] Hongchi Xia、Xuan Li、Zhaoshuo Li、Qianli Ma、Jiashu Xu、Ming-Yu Liu、Yin Cui、Tsung-Yi Lin、Wei-Chiu Ma、Shen- long 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年。2、33[85] 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（CVPR）论文集，第16262–16272页，2024年。33[86] Yue Yang、Fan-Yun Sun、Luca Weihs、Eli VanderBilt、Alvaro Herrasti、Winson Han、Jiajun Wu、Nick Haber、Ranjay Krishna、Lingjie Liu 等。Holodeck：语言引导的具身人工智能三维环境生成。载于2024年IEEE/CVF计算机视觉与模式识别会议（CVPR），第16277–16287页。IEEE，2024年。2、5、6、7、18、19、21、33[87] Yandan Yang、Baoxiong Jia、Shujie Zhang 和 Siyuan Huang。SceneWeaver：采用可扩展自反思智能体的一体化三维场景合成。神经信息处理系统进展，38:140319–140351，2026年。2、22、33、37[88] Yixuan Yang、Zhen Luo、Tongsheng Ding、Junru Lu、Mingqi Gao、Jinyu Yang、Victor Sanchez 和 Feng Zheng。OptiScene：通过规模化人类对齐数据合成与多阶段偏好优化，实现大语言模型驱动的室内场景布局生成。神经信息处理系统进展，38:42499–42529，2026年。2、23、34、37[89] Lap-Fai Yu、Sai-Kit Yeung、Chi-Keung Tang、Demetri Terzopoulos、Tony F. Chan 和 Stanley J. Osher。让家具布置成家：家具布局的自动优化。ACM图形学汇刊，30(4):86，2011年。33[90] He Zhang、Ying Sun、Weiyu Guo、Yafei Liu、Haonan Lu、Xi- aodong Lin 和 Hui Xiong。通过由粗到细的多模态强化实现交互式室内设计推荐学习。在第*31*届 *ACM* 多媒体国际会议（*MM*）论文集，第6472–6480页，2023年。35，37[91] Kaizhi Zheng、Xiaotong Chen、Xuehai He、Jing Gu、Lin-jie Li、Zhengyuan Yang、Kevin Lin、Jianfeng Wang、Lijuan Wang 和 Xin Wang。Editroom：用于可组合三维房间布局编辑的大语言模型参数化图扩散。载于国际学习表征会议，第86791–86808页，2025年。2，34[92]Lianmin Zheng、Wei-Lin Chiang、Ying Sheng、Siyuan Zhuang、Zhanghao Wu、Yonghao Zhuang、Zi Lin、Zhuohan Li、Dacheng Li、Eric P. Xing、Hao Zhang、Joseph E. Gonzalez 和 Ion Stoica。使用 MT-Bench 和 Chatbot Arena 评估大语言模型裁判。载于神经信息处理系统进展（*NeurIPS*），数据集与基准赛道，第46595–46623页，2023年。37[93] Yifei Zhou、Song Jiang、Yuandong Tian、Jason Weston、Sergey Levine、Sainbayar Sukhbaatar 和 Xian Li。SWEET-RL：在协作推理任务上训练多轮大语言模型智能体。*arXiv* 预印本 *arXiv:2503.15478*，2025年。35

Prefere nceScene：根据反馈进行个性化室内场景合成 k

补充材料

目录

A. 基准详情 13B. 方法细节 16C. 场景实例化详情 17D. 实验细节 21E. 补充结果 23F. 用户研究 30G. 定性结果和失败案例 32H. 局限性 33I. 相关工作（完整版） 33提示词与示例输出 38J. 完整模式 45

隐藏画像：画廊（⋆：以显式 形式陈述）：悬挂方式为轨道悬挂，艺术品类型为黑白照片，墙面颜色为画廊白⋆，特色墙为木质护墙板，地板类型为大规格瓷砖⋆，展示装置为独立隔断，座椅配置为简易长凳⋆，家具布局为分成两组⋆，照明方案为暗处的画作照明灯⋆，装饰主题为落地植物。**角色设定：** 年龄 28 岁，城市规划师。“他每天都在绘图、审阅方案，并步行或骑自行车往返于各个会议之间。”显式 **指令：** “我希望墙壁是亮白色，地面铺设大块、素净的浅色瓷砖，家具分成两组，中间留出一条清晰的通道。请用各自的小灯照亮作品，而不要使用日光；长凳请保持朴素低调。”隐式 **指令：** “请让画廊保持整洁、富有当代风格，并留出足够的空间，让人感觉思路清晰。我希望整体低调、雅致，不要装饰过度。”**品味叙述**（节选；仅提供给模拟用户）：“.....吸顶灯应当关闭，每件作品都由自己的小灯照亮，还应摆放几盆大型盆栽植物，让空间显得宁静而富有生气。”

图 S1。开发子集中的一个实例及其两种形式。系统接收角色设定和两条指令中的一条。模拟用户接收角色设定、品味叙述和渲染图像。

A. 基准详情

本节详细介绍主论文第 5 节中的基准测试：一个示例实例（第 A.1 节）、属性模式（第 A.2 节）、偏好配置文件和输入（第 A.3 节）、模拟用户（第 A.4 节）、可见性检查（第A.5 节）、参考场景（第 A.6 节）以及评估器（第 A.7 节）。参考文献之后的代码清单逐字列出了提示词及其模板占位符。

A.1. 示例实例

图 S1展示了 两种形式下的一个实例，图 S2展示了 四个测试偏好配置文件及其参考场景，表 S1给出了基准测试的统计数据。

A.2. 属性模式

属性和值。 十种房间类型分别是浴室、卧室、儿童房、餐厅、画廊、厨房、图书馆、客厅、办公室和候诊室。每种房间类型都有一个包含 10 项属性的模式。属性是房间的一个可见组成部分，其取值彼此互斥：一个场景只会呈现其中一种。每个取值都有名称，以及一句描述观者所见内容的话。列表 7展示了一项属性的取值。每项属性都有两段文本。其定义 指出了该属性所指的组成部分

并被基准测试构建所使用的提示词以及评估器的问题引用。其在属性列表 中的描述是一个不指明任何取值的句子，例如。“镜子类型决定洗脸盆上方的镜子”。这是系统接收的关于该属性的唯一文本（主论文第 3 节）。第 J 节列出了每种房间类型的属性及其组、属性列表中的描述和取值名称。

统计数据。 这 10 个模式包含 100 个属性和 483 个取值，每种房间类型有 40 至 55 个属性。每个属性有 3 至 6 个取值，平均为 4.8 个；其中分别有 5、31、40 和 24 个属性具有三个、四个、五个和六个取值。这些属性遵循室内设计教科书 [15] 中设计元素的组：30 个属性描述表面，10 个描述照明，26 个描述家具，11 个描述纺织品，13 个描述配饰，10 个描述空间布局。

A.3. 隐藏档案和输入

档案。 每种房间类型有 8 个档案，共 80 个。GPT- 5.4-mini 根据模式（清单 1）在一次调用中编写某种房间类型的所有档案。对于每个档案，模型会先选择一种设计风格和一种目标氛围，然后为每个属性选择一个同时符合两者的取值。某种房间类型的任意两个档案在 10 个属性中至少有 5 个不同，平均有 8.4 个不同。这 80 个档案使用了 483 个取值中的 460 个。

人物设定。 每个人物设定包含 8 个字段：姓名、年龄、职业、生活方式、性格特征、居住状况、期望和不喜欢的内容。



图 S2. 基准测试中四种房间类型的测试集档案示例。每行展示一个档案：其参考场景、十个隐藏值，以及一个包含显式 和隐式 指令的角色设定。带点的粗体取值在显式 指令中有所说明。

对于每种房间类型，GPT-5.4-mini 会调用一次，根据偏好配置文件（列表2）编写人物设定；随后再调用一次，将每个人物设定分配给最适合它的偏好配置文件，确保每个配置文件都分配到两个人物设定。两次调用还会列出“混合型”配置文件，即结合该房间类型下两个配置文件取值的配置文件；这些配置文件及其人物设定不属于基准测试。因此，每种房间类型有 16 个实例，总计 160 个。

两种形式的输入。 两种形式共享档案、品味叙述和角色设定的前六个字段；它们在期望、不喜欢的内容和指令上有所不同（表。S1）。在显式 形式中，每个实例会列出 4 到 6 项属性；分别有 50、60 和 50 个实例列出四项、五项和六项属性。明示属性的分布均衡：对于每种房间类型，每项属性都会在 16 个实例中的 8 个实例里明示。GPT-5.4-mini 根据角色设定、档案以及需要明示的属性来撰写期望、不喜欢的内容和指令（列表3）。指令至少明示两项取值，其余取值则在期望或不喜欢的内容中明示。在隐式 形式中，期望和不喜欢的内容保留角色设定生成时的原文，GPT-5.4-mini 则仅根据角色设定撰写指令（列表 4）。

两种形式共有		
属性	$M = 10$在 10 种房间类型中各有；总计 100 个	
每个属性的取值	3–6（4.8）；总计 483 个	
隐藏档案	80（每种房间类型 8 个）	
每种形式的实例	160（40 个开发集 / 120 个测试集）	
品味叙述词	94–137（115.9）	
	显式	隐式
明示属性	4–6（5.0）	0
指令词	37–76（57.2）	25–46（33.3）
角色词	126–380 (183.7)	100–240 (161.4)

表 S1. 基准测试的统计数据。每个单元格给出范围，并在括号中注明平均值。角色设定的词数不包括姓名和年龄。

未明示属性数量。

品味叙述。模拟用户被设计为以散文而非模式取值的形式了解自己的偏好。因此，对于每个角色设定，GPT-5.4-mini 都会根据角色设定以及带有描述的 10 个隐藏值，以该角色的口吻撰写品味叙述（列表5）。提示词要求篇幅为 90 至 140 词，实际叙述为 94 至 137 词，平均 115.9 词。提示词要求叙述涵盖每个隐藏值，并展现整体风格。只有模拟用户会读取这段叙述。

子集。 每种房间类型中，有两个偏好配置文件构成开发子集，该子集包含 20 个偏好配置文件，每种形式有 40 个实例。其余 60 个偏好配置文件构成测试子集，每种形式有 120 个实例，每种房间类型有 12 个。

A.4. 模拟用户

模型与输入。 模拟用户使用 Qwen3.8-27B（检查点 `Qwen/Qwen3.8-27B-FP8`），由 vLLM [39]提供服务。在每个轮次中，它都会接收一条系统消息和一条用户消息（列表 6）。系统消息包含任务描述、角色设定和偏好叙述。用户消息包含所示场景的两张渲染图像，随后是反馈请求。这些图像分别是从上方观察的剖视图和一张平视图，每张图像均为 1024×768 像素。从第二轮开始，用户消息还会列出此前每个轮次的反馈和评分，并询问模拟用户要求更改的房间部件是否有所改善。模拟用户从未接收系统的方案或模式。

输出。 答案是一个包含两个字段的 JSON 对象，其格式在解码过程中受到强制约束。第一个字段是反馈，内容为 10 到 1200 个字符的文本，提示词要求写三到五句话。第二个字段是评分，取值为 1 到 10 的整数，提示词将评分标准划分为五个等级，每级相差两分。系统只接收反馈和评分。

采样。 模型以温度 0.7、top-p 0.8、top-k 20、min-p 0 和 1.5 的存在惩罚进行采样，在非思考模式下最多写入 480 个词元。每个答案的采样随机种子由实例、轮次和交互的随机种子推导得出。

A.5. 可见性检查

可见性检查用于衡量能否从渲染图像中读取模式的取值。对于每种房间类型，我们固定三个基础方案，每个方案为每个属性分配一个取值。对于每个基础方案和每个取值，基于规则的基准测试构建会生成具有该取值的场景，而其他属性则采用基础方案的取值。因此，每个取值都会在三个场景中进行渲染，检查总计包含 1179 个场景。Qwen3.8- 27B 会接收某个场景的两张渲染图像，以及第 A.7 节中关于变化属性的多项选择题，选项分别按给定顺序和相反顺序排列，因此每个取值有六次作答机会。对于模型在作答中命中率低于 70% 的取值，我们进行了修订、合并或替换，并在修订后的模式上重复了检查。对于基准测试的模式，95.6% 的回答指出了渲染出的取值；483 个取值中有 454 个至少在六次回答中的五次被正确指出。

A.6. 参考场景

对于每个档案，基于规则的基准构建方法会直接根据 10 个隐藏值生成一个参考场景，总计 80 个场景。该构建方法与我们方法的场景实例化过程不同，区别在于如何将取值转换为规格：针对每种房间类型，手写表格会将每个模式取值映射到固定参数，且不会调用大语言模型或 Jev。我们的方法从不读取模式取值，而是从 Jev 和检索中获取规格（第 C 节）。两种方法都使用相同的代码构建房间、检查碰撞并搜索布局，从相同的资源池中抽取物体，并采用相同的照明、白平衡和摄像机进行渲染。在两种形式中，同一档案的各个实例共享其参考场景。主论文第 5 节报告了两种形式下 240 个测试实例的参考场景 pAMR。

A.7. 评估器与 pAMR

评估器。 评估器为 GLM-4.6V-Flash（检查点 `zai-org/GLM-4.6V-Flash`），由 vLLM 提供服务 [39]。对于一个场景，每个属性都会回答一个多项选择题（清单 7）。每个问题是一条消息，其中包含该场景的两张渲染图像，后接问题文本。文本会注明属性及其定义，并将该属性的取值列为带字母编号的选项，每个选项都附有描述。答案限定为一个选项字母，并以温度 0 的非思考模式进行解码。

pAMR。 每个问题都会分别按给定顺序和相反顺序排列选项，因此每个场景有 20 个答案。pAMR（主论文中的公式（1））是这些答案中指出隐藏值的比例。对于有 n 个取值的属性，均匀随机猜测指出隐藏值的概率为 $1/n$。100 个属性的 $1/n$ 均值为 21.4%。由于每种房间类型的测试实例数量相同，因此该均值就是随机猜测在测试子集上的 pAMR。

B. 方法细节

本节介绍提示词、设计方面到属性的分配、大语言模型调用的设置，以及 PreferenceScene 的算法列表（主论文第4节）。

B.1. 候选值

大语言模型 Gemma-4-31B（`gemma-4-31b-it-qat`）在第一轮之前调用一次，生成候选项（列表 8）。其提示词包含房间类型、指令 q 和角色设定的八个字段$\mathcal{P}$，每个系统都会收到这些信息，但不会收到属性列表。采样温度为 1.0，top-p 0.95 和 top-k 64，禁用思考，输出词元数最多为 14,000。此次调用的随机种子由实例、提示词和交互随机种子推导得出。如果回复不是有效的 JSON，则不会重复调用；系统会通过匹配所请求格式中的字段名称，从文本中读取各方面及其候选项。

设计方面与属性的分配。向 Jev [76] (jev-1.13.0) 发出的一次请求，会针对每个设计方面 g_j提出一个单选问题。选项是属性列表中的 M 条目，每项均以属性名称及其描述呈现，另有“无”这一选项。问题会引用该方面的名称、大语言模型为其撰写的一句话描述，以及其前三个候选标签。令 $\lambda_j(a)$ 表示 Jev 针对关于 g_j的问题返回选项 a 的概率。如果 a 不是“无”且 $\lambda_j(a) \geq 0.5$，则将方面 g_j 分配给属性 $a = \arg\max_b \lambda_j(b)$。随后，如果 $\lambda_{j^*}(a) \geq 0.3$，每个属性 a也会接收方面 $j^* = \arg\max_{j \notin A(a)} \lambda_j(a)$，其中 $A(a)$ 是已分配给 a 的方面集合。这样的方面也可以属于另一个属性。应用这两条规则后仍没有方面的属性，会接收具有最高 $\lambda_j(a)$的方面，前提是该概率大于 0。候选集 C_a 串联了 $A(a)$ 中各方面的候选项，每个方面最多 12 个，并按概率最高的方面优先排列。在规划提示词（第 B.2 节）中，候选项少于两个的属性会不列出候选项；场景实例化会使用其类别的默认规范（第 C.7 节），除非大语言模型为其写入新值。

B.2. 初始计划和计划更新

初始计划和每次计划更新的调用都会向大语言模型发送相同的系统消息，以及格式相同的用户消息；二者在历史记录、任务和回答格式中的一个短语（列表 9 和10）方面有所不同。用户消息包含输入 c （即房间类型、指令、角色设定，以及每个属性名称及其描述）和每个属性的候选项；候选项按方面分组，每项包含其标签、描述和十六进制代码。消息说明：候选项来自多个方面的属性取一个值，而没有候选项的属性保持其默认值，除非大语言模型写入新值。从第 2 轮开始，历史记录会针对每个较早的轮次，列出每个属性所显示的值及其场景描述（第C.8节）、反馈和评分。较早轮次产生的新值会在历史记录中以其标签和描述呈现。历史记录会省略对较早回复的分析。第 $k \leq K$ 轮次的任务会要求给出该轮次的计划，而第 K 轮次之后的任务会要求给出最终计划，最终计划不会收到反馈。每次调用使用候选调用的采样设置，最多生成 3,072 个输出词元，并使用一个由实例、输入、交互的随机种子和轮次推导出的随机种子。

读取回复。 如果回复中的 JSON 字段“plan”对每个属性都给出 null、非空文本，或带有非空标签或描述的对象，则该回复有效。文本或对象标签若与某个候选标签相同，则选择该候选项，除非对象提供了不同的描述。null 以及“default”或“none”之类的文本会保留默认规格。任何其他文本都是一个新值，其标签和描述均为该文本；任何其他对象都是一个新值，包含其标签、描述和十六进制代码。若文本与先前轮次中新值的标签相同，或对象的标签、描述和十六进制代码均相同，则重新选择该先前的值，并保留其规格。

无效回复。 如果回复在达到词元上限时中止、无法解析或无效，则使用新的派生随机种子重复该请求；每个计划最多请求三次。三次回复均无效后，第一轮会为每个有候选项的属性显示第一个候选项，并为其他每个属性显示默认值；后续轮次则重复之前的计划。

B.3. 布置最终场景

计划决定房间的外观，而场景实例化会放置取值所指定的物件，因此，仅根据计划构建的房间中物体很少。最后一轮结束后，还会进行一次大语言模型调用，为最终场景提出更多物件（Listing11）。其提示词会提供计划的取值、房间大小以及房间中已有的物件。它还会列出可承载物体的家具台面，

并提供这些台面的尺寸、高度和目标物体数量，以及允许使用的物件类型，即检索所接受的类型名称。提示词要求在房间中添加约十件地面物件，在每件新增且带有台面的物件上放置两到五个物体，至少添加六件墙体部件，并确保新增物件至少涵盖八种类型。提示词还要求，在供人用餐的桌子上摆放餐具，且桌面装饰品最多一个。对于装饰、艺术品或桌面类别的某项属性，回复会先说明其取值是要求空置表面、指定具体物件，还是两者皆非。如果该取值要求空置表面，则不会在台面上添加物体或墙体部件。

与计划的一致性。 在提出物件之前，回复会针对每个属性指出其取值所决定的物品类别，例如座椅或储物家具，并列出该取值未提及的该类别允许类型。所有提出的、属于所列类型的物件都会被移除，其台面上提出的物体也一并移除。进行此项检查时，提示词只包含计划的自由文本取值和允许的类型名称，绝不包含基准测试的取值集合 $\mathcal{V}_a$。完成这些检查后，只有存在书桌时才会添加书桌座椅，只有存在电视时才会添加电视柜。

摆放位置。 最终计划采用与各轮方案相同的程序进行实例化，而家具布置不会移动由此产生的任何物体。新增物件的检索方式与所有其他物体相同（第 C.2节），并使用摆放位置中的碰撞检测将其放置在剩余的空闲空间中（第C.4节）。地面物件沿着墙壁的空闲区域以及地面空位摆放，物体放在台面的空闲区域，墙面装饰则沿着墙壁的空闲区域摆放。没有空闲位置的物件不会被添加。

调用与回复。 家具布置调用使用温度 0、用于回复的 JSON 模式、由计划派生的随机种子，以及最多 4,000 个输出词元。如果请求失败四次，则交付不含新增内容的最终场景。在各轮次中展示的场景（用户的反馈基于这些场景）不会进行家具布置。对于每轮次的 pAMR（图 S3和S7以及表S12和S13），以及主论文表 3中不含迭代的变体，每一轮次的场景都按照该轮次结束后交付的状态进行家具布置。

B.4. 算法

算法S1 列出了 PreferenceScene 的一次交互步骤。

算法 S1 PreferenceScene 的一次交互。

要求： 房间类型、指令 q、角色设定 $\mathcal{P}$、属性列表、数量 K轮次

大语言模型调用：根据房间类型、q 和 $\mathcal{P}$

Jev 将这些方面分配给属性，为每个属性提供候选集 C_a 每个属性 a

大语言模型调用：初始计划 $\hat{S}_1$，来自 c和 $\{C_a\}$，每个属-致谢（候选项、新值或默认值）

用于 $k = 1, \ldots, K$ **do**

  实例化 $\hat{S}_k$ 为场景 X_k（Jev 回答规格中—— 首次显示的每个值对应的问题），并渲染它

  接收反馈 η_k 和评分 z_k

  编写场景描述 D_k 的 X_k

  大语言模型调用：$\hat{S}_{k+1}$ 来自 c，$\{C_a\}$ 和（ $\hat{S}_{k'}, D_{k'}, \eta_{k'}, z_{k'})^{k' \le k}$ 由主论文中的公式（2）

结束循环

实例化最终计划 $\hat{S}_{K+1}$ 作为最终场景 X

大语言模型调用：要添加到 X 中的物件，以及每个属性对应的类型排除（第 B.3节）；删除属于这些类型的物件，并放置其余物件且不移动 X中的任何对象

C. 场景实例化详情

场景实例化会根据计划设置房间的每个部分：墙壁和地板的材料与颜色；墙面饰面，例如面板、图案和图像；灯光及其对应的一天中的时段、来源、功率和色温；窗户覆盖物、地毯和织物；家具及其数量、尺寸、材料和颜色；配饰；以及家具布局。

C.1. 组件类别与问题

每个属性都属于 20 个组件类别之一。在十种房间类型的 100 个属性中，各类别所含的属性数量分别为：涂料、地板、墙面饰面和灯光各 10 个，装饰 9 个，窗户、储物和座椅各 7 个，布置和风格各 5 个，织物 4 个，固定装置和桌子各 3 个，床、饰面和艺术品各 2 个，活动、桌面、展示和柜台各 1 个。每个类别定义了规格中的字段，这些字段可由该类别问题的答案设定。每个关于候选项的问题都以“A design for a(n) {房间}描述其 {属性} ({属性描述})}，标题为‘{方面}’的内容如下：‘{标签}: {描述}’。”对于新值，标题就是属性名称。大多数问题以“If the {主题}未提及，请选择最符合它的选项。”结尾。只有当取值的文字指出或暗示了黄昏或夜晚时，时间问题才允许选择黄昏或夜晚；仅有灯光并不暗示时间。表 S2 列出了各类别的问题，表S3 列出了选项集及其来源。每轮首次展示的取值对应的所有问题，都会以每次最多 40 个问题的请求发送给 Jev。每个答案都是概率最高的选项；如果概率并列，则选择排在最前的选项。布局答案的概率会保留，用于布局搜索（第C.4节）。

数量问题。 数量问题的表述为：“ {subject} name 对于 {attribute}有多少{type plural} ？ ”问题还会列出按 Sentence-BERT [65] similarity计算得出的五个最近类别池，并说明每个物体必须

类别	问题	数量
涂料、织物、饰面，风格、柜台	无（仅检索）	0
墙面饰面	覆盖层、图案	2
地板	覆盖层、图案、地毯	3
窗户	形式、不透明度、长度、高度	4
灯光	时间、来源、色温，功率、灯具、日光，数量为 4	12
	灯具类型、暗槽灯、画灯	
布置	每项的 7 个摆放问题	9–12
	组、混合型，以及在适用情况下，程序参数	
物体类别	每种池塘类型、尺寸的数量，相同或混合型物件、支撑表面，以及在适用情况下，7	93–105
	每组的摆放问题	
	座椅和程序参数	

表 S2。按组件类别划分的每个值对应的问题。类别池是资产池中的 90 个类型名称。

问题	选项	来源
每种类型的数量	无、一个、两个、三个或四个，五个或更多、一些	池类型名称；本研究的分箱
size	小、中、大（×0.8，1.0、1.25 倍的中位数池尺寸）	本研究
锚点、分组，沿着、对称性，朝向、间距	后方、左墙或右墙、最大的墙 窗户，或中间；边缘或中间；0.2/0.5/0.8；是/否；四种朝向；三种间距	Holodeck 约束 [86]；本研究的分箱
组，每组，mix	1–3 组；1、2、3–4、5–9、10+；相同或混合型	数量和方差 Holodeck [86]；本研究的分箱
色温	2700–6500 K，六个取值	ANSI C78.377 名义值
时间、来源	天、黄昏、夜晚；日光，电力	ProcTHOR [18]
功率、灯具、日光	三个类别；是/否；三个类别	本研究
窗饰	无、窗帘、百叶窗、遮阳帘，百叶窗；薄纱、灯光，全遮光；窗台、地板；折叠、卷起	本研究
墙面覆盖物、图案	涂料、材料、浮雕，面板、图案、图像；九图案类型	本研究
程序化参数	九个房间专用集合中的 14 个房间类型	.....的参数程序化部分

表 S3. 规格问题的选项集及其来源。

在这些类型中的某一类型能更好地命名它时，不得重复计数。此外还说明，仅表示另一个物体的形状、材料、用途或部分的词，不命名任何物体。如果某个物体类别的某一取值的所有数量答案都是“无”，第二个请求会询问哪种类别池是所述属性的主体部分，并将该类型计为一个部分。

C.2. 检索

物体。 我们使用 Holodeck [86] 在基线方法所用的 6,793 个 Objaverse [19] 资产中的检索结果。查询为“ {类型}、 {文本}的三维模型”，其中，文本是值的标签和描述。资产的评分为 100 $\max_v \cos(\text{CLIP}_{\text{text}}, \text{CLIP}_{\text{image}, v})$ +SBERT(文本) · SBERT(说明文字)−−~~10~~$|d - d'|$。对资产的三个视图 v 取最大值； $|d - d'|$ 是资产尺寸与目标尺寸排序后各维度之间以米为单位的平均绝对差。目标尺寸为该类型的中位数池尺寸乘以尺寸选项。编码器为在 LAION- 2B 上训练的 OpenCLIP [34]ViT-L/14 和 Sentence-BERT all-mpnet-base-v2，与 Holodeck 中相同。与 Holodeck 不同，检索范围限定为规格中数量答案所指定的池类型。低于 Holodeck 的 CLIP 评分阈值的资产不会被剔除，而是排在最后。资产通过确定性的 top-k 方式选出，而非从前十名中采样；若评分相同，则采用确定性规则打破平局。Holodeck 的房间尺寸和地面放置筛选器不予使用。实例化家具部件时，其查询还会包含风格和家具饰面属性的文本；对于软垫类，还会包含织物属性的文本，并排除场景中其他位置已使用的资产。

表面与颜色。 对于材质查询，每种材料的评分由 $\cos(\text{CLIP}_{\text{text}}, \text{CLIP}_{\text{swatch}})$ 与查询和材料名称之间的字符串相似度相加得出，与 Holodeck 中的方法相同。地板和墙面饰面类别会搜索全部 236 种 ambientCG [20]材料，其他类别则只搜索适用于其表面类型的材料。若大语言模型写出了颜色的十六进制代码，则采用该代码；否则，像 Holodeck 中那样，根据 CLIP 文本相似度，从 Matplotlib [32]的 148 种命名颜色中选取最接近的一种。检索到的材料会在 HSV（线性 RGB）空间中向目标颜色调整色调，其中色相偏移限制在色相环的 ± 0.15 范围内，饱和度和明度的缩放系数则分别限制在 $[0.3, 1.6]$ 和$[0.5, 1.5]$范围内。厨房台面的材料会根据台面值的文本重新选择，仅依据 CLIP术语在全部 236 种材料中进行选择；金属材质则渲染为金属。由于 CIELAB L^*低于 25 的彩色墙面颜色在渲染图像中不会呈现可见色相，因此会将其替换为文本所指的另一种颜色，前提是该颜色的色相与墙面颜色相差不超过 $60°$，且其亮度在 L^*中高出 10 以上。

图像覆盖层。墙面饰面规范若采用“图像”覆盖层，则会从资产池的 111 幅画作中选取一幅，其画布纹理与该值文本的 CLIP 相似度最高。画布会按墙面的宽高比进行中心裁剪。不会生成图像。

C.3. 程序化家具库

在六个组——座椅、椅子、桌子、储物、搁架和床——中，属于 27 种类别池的家具，会由我们的程序化库在检索到的资产的包围盒内，按其位置和偏航角重新构建（主论文第 4.3 节）。Jev 会回答有关该库参数的问题。每个部件（如框架、台面、正面板或软包）的材料，均根据属性文本，从 ambientCG 检索；检索仅依据 CLIP术语，并限定在该部件的表面类型和材质类别内。规格中的十六进制代码用于设定部件的颜色。对于未包含颜色名称的文本，若 CLIP 颜色搜索推断出一种颜色，则不会将其应用于织物部件。对于木质和金属部件，如果色调无法通过色调调整达到，则会舍弃该颜色，因为截断的色相偏移会呈现第三种颜色，而材料和属性值中都没有这种颜色。皮革软包会保留其材料的粗糙度贴图；若部件的材料具有金属质感，则会渲染为金属。若程序化部件的网格与其包围盒沿某一轴的差异超过 5 毫米，则会重新适配；若重新适配失败，则保留检索到的资产。程序化部件还提供了资产池中缺少的物体类型：带水槽和炉灶的厨房台面组、中岛柜、吊柜和抽油烟机；浴室的淋浴底盘、隔屏和长凳；卧室和儿童房的嵌入式衣柜组；画廊的基座和独立面板；图书馆的隔屏和嵌入式搁架；以及餐厅和厨房的嵌入式长凳。窗饰、墙面板条、凸起面板、平面图案表面、地毯、暗槽灯带和画作照明灯则是各房间类型共用的程序化部件。

C.4. 摆放位置

与 Holodeck 的差异。在 Holodeck 中， [86], 大语言模型会为成对物体编写约束，然后通过在网格上进行深度优先搜索，每次摆放一个物体，耗时 30 秒，并保留得分最高的布局。我们的软约束是规格的布局答案，用于描述一组家具。搜索会按这些答案的概率顺序枚举其组合，依据固定规则布置每种组合，并保留第一个可行布局。墙体部件、家具台面上的物体和天花板灯光均按照固定规则摆放，而不是采用 Holodeck 的大语言模型编写的高度约束，以及在支撑物上的随机布局。

每种房间类型都会定义必需物件：当没有值指定其类型时，这些物件便会被添加；还会定义默认座椅：当没有座位值指定座椅时，这些座椅便会被添加。它也会定义家具组的中心家具类型，以及用于启动搜索并作为回退方案的中性布局答案。在客厅中，这些分别是茶几和边桌、沙发和两把扶手椅、茶几，以及房间中央的一个对称组。地面家具遵循布置属性的规范（如有），否则遵循座位属性的规范，再否则采用中性答案。搜索根据 Jev 的答案概率，对锚点、分组、组数、朝向、对称性、间距和沿墙位置的选项进行排序。它会在最多 400 次试验中尝试每个参数最可能的三个选项，按软评分 $\sum\log(p+10^{-3})$ 排序，并最后改变答案置信度较高的参数（$p > 0.5$）。硬约束确保每件物品都位于房间内，并排除碰撞。为保持门、窗户和窗帘周围畅通，并确保平视图视野开阔，搜索还会排除门的净空区（深 1.0 米，且比门宽 0.2 米）、窗户的净空区（窗台上方深 0.35 米）、窗帘所在的带状区域，以及距离平视摄像机 1.2 米以内的地面家具。靠墙摆放的物件会放置在墙面段上，即与门至少相距 0.1 米、与窗户相距 0.22 米、与墙端相距 0.05 米的墙面空段。如果所有试验都失败，且有些试验仅因地面物件找不到墙面段而失败，则说明墙面太短，无法容纳请求的物件。此时，搜索会移除失败试验中未能摆放的物件数量最少的那次试验里的未摆放物件，然后重新运行，之后才回退到中性答案。对于高度至少为 1.2 米的物件，会优先尝试剖视图中未被遮挡的墙壁。随后，桌子以及活动、展示和收纳属性中的物件会根据各自的布局答案摆放；小物体则会排列在天花板排布中、放在墙上或支撑物上，或置于地面空位。布置规范要求每组至少包含“每组物件数 × 组数”件物品，墙上最多 18 件，地面最多 12 件。为填满墙面，密集墙面悬挂若包含十幅或更多图片，就会要求摆放 24 幅不同尺寸的图片，并将它们沿墙面高度带排布。当地面布置要求在房间中央设置 $k \geq 2$ 组时，每组都会配有自己的矮桌，各组分散在房间宽度的 0.8 范围内，并在墙边放置一张工作桌，例如接待台。当地面布局包含 $k \geq 2$ 组且未指定组大小时，会按指定座椅数量的 k 倍提出请求，并将彼此相对的座椅放在各组矮桌的两条长边上。若对称的座椅、展示或收纳规范只指定一件物品，则会要求一对配对物件。由于单排厨房需要一段较长的空闲墙面，如果其所在墙面的空闲墙面段短于墙长的一半，且另一段至少长 0.5 米，则厨房布局会移至空闲墙面最长的可见墙面。

C.5. 照明和窗户

每个房间都由 Poly Haven [59]环境贴图“阅览室”照明。黄昏和夜晚时，环境光和日光会减弱，灯具功率加倍，所有窗饰都会关闭；如果照明文本提到了灯具或隐藏灯光，但没有提到吸顶灯，则吸顶灯会调暗至 40W。

灯光，但没有吸顶灯。黄昏和夜晚时灯具会点亮；如果照明文本说明灯具已点亮、提到固定装置或隐藏灯光，或包含发光词（如烛光）或灯具词（如吊灯），则无论何时灯具都会点亮。若照明文本描述的是夜间室内环境，也就是说，灯光照明的房间，那么其主要光源必须是电力，日光等级不能高，文本不能提及日光，且提及的固定装置不能只有吸顶灯；此外，文本还必须包含夜间提示：时间为黄昏或夜晚、包含灯具词或发光词，或提及隐藏灯光。此类文本采用夜间照明：没有日光、关闭窗饰、灯具以双倍功率和文本指定的色温照明，且吸顶灯功率至少为 200 W。如果照明文本说明灯具已点亮、包含发光词或描述夜间室内环境，但没有列出固定装置或隐藏灯光，则会添加文本中提到的每种灯具类型各两盏；若未提及任何类型，则添加两盏落地灯和两盏台灯。由于相机无法看到灯具的点光源，资产池中的每盏点亮的灯具下方、靠近其光源处都会放置一个半径为 5 cm 的发光球体，其颜色取自照明文本中的十六进制代码或最后出现的颜色词。渲染图像未进行相机白平衡（第C.10节），因此暖色吸顶灯会使表面呈橙色；色温为$T < T_0 = 5000$K 的吸顶灯会以T'的温度渲染，$1/T' - 1/T_0 = 0.4\,(1/T - 1/T_0)$。布质窗帘会采用检索到的织物。

C.6. 实例化规则

不含实例化规则的变体（第 E.15节）会移除以下规则，这些规则在对应的小节中有完整说明。

- 照明（第 C.5节）：由指定灯具点亮的灯、隐藏灯光以及发光或灯具词；晚间照明；新增的灯具；发光球体；以及将低于 5000 K 的吸顶灯色温调整至 5000 K。
- 材料（第 C.2 节和 C.3节）：仅依据 CLIP 术语为每个程序化部件选择材料，并根据文本再次选择厨房台面材质；排除织物部件的推断颜色，以及木质和金属部件的不可达色调；皮革的粗糙度贴图；以及将金属材质渲染为金属。
- 物体数量与摆放（第 C.4节）：密集墙面悬挂中的图片数量，以及桌子、座椅和匹配物件的数量；将座椅组分散摆放，并在墙边放置一张工作桌；以及移动单墙式厨房布局。
- 颜色（第 C.2节）：深色彩墙面颜色的替换。

C.7. 默认规格

计划保持开放的属性，或没有任何答案设置的字段，会采用一个仅取决于其类别的固定默认值。这些默认值为油漆颜色 #e9e5dd、ambientCG 地板材料 WoodFloor051、粉刷墙壁、织物颜色 #b9b2a6、3000K 中等功率的灯光（灯具已点亮并有日光）、#ece8e0 的浅色落地窗帘，以及房间类型的中性布局答案（第 C.4节）。

C.8. 场景描述

场景实例化后，模板会描述其规格中的每个属性、搜索所选择的摆放位置、检索到的资产的说明文字、ambientCG 材料的名称和标签，以及颜色名称。对于灯光属性，描述会说明色温、功率、一天中的时段、固定装置、点亮的灯，以及是否有日光进入。描述不会直接复制取值中的文字。

C.9. 规格、场景描述和家具清单示例

规格。 清单12 展示了图 3所示主论文交互中第 1 轮计划的三个取值的规格（餐厅、隐式 形式、种子 1）。对于一个取值，规格说明场景实例化所构建的内容。字段 类别 是属性的组件类别（第 C.1节），而标签 和 外观 是该取值的文本。字段 颜色 保存该取值的十六进制代码，而 表面 保存为表面检索到的 ambientCG 材料，以及该表面是单一材料还是图案。字段 物体列出该取值的物件类型及其数量、所处表面和以厘米为单位的尺寸。字段摆放位置 保存摆放位置答案：组所依靠的墙，或房间中央（锚点）、沿着该墙的位置、分组、组数（k）、对称性、朝向和间距（第 C.4节）。字段 灯光 保存以开尔文为单位的色温、功率、灯具是否点亮、日光、一天中的时段和灯具。除 标签、外观 和 颜色 之外，所有字段均来自 Jev 的回答或检索结果。

场景描述。 列表13 展示了相同三个属性的场景描述 D_1，大语言模型会在第 1 轮之后的更新中读取这些描述（主论文中的式（2））。模板根据已构建的内容编写这些描述：对于地板，使用带有标签和颜色的 ambientCG 材料；对于照明，使用灯光设置；对于座椅，使用检索到的物件及其部件材料、尺寸和摆放位置。地面铺装值要求使用带白色纹理的黑色大理石，而其描述则报告了检索到的材料 Marble 006，以及 black、dark 和 grey 等标签。因此，大语言模型读取的是用户看到的内容，而不是该值所要求的内容。

Role	模型
基准测试文本	GPT-5.4-mini
模拟用户	Qwen3.8-27B（开放权重）
评估器	GLM-4.6V-Flash（开放权重）
每个系统的大语言模型	Gemma-4-31B（开放权重）
多项选择决策	Jev（jev-1.13.0） [76]
第二评估者	Ministral-3-14B
托管大语言模型	GPT-6.1 Terra

表 S4. 模型及其作用。多项选择决策在 PreferenceScene 中作出。最后两行分别用于第 E.8 节和第 E.9 节。Ministral-3-14B 还负责标注反馈中提及的属性（第 E.3 节）。

家具清单。列表14 展示了大语言模型调用回复的部分内容；该调用为主论文图1 中第二位用户的最终场景布置家具，该场景是一个采用显式形式、随机种子为 3 的餐厅；主论文图3 展示了同一用户的隐式形式。对于每个属性，回复都会先说明其值所控制的内容，并列出该值排除的物件类型（第B.3节）。储物值“黑色漆面玻璃展示柜”会排除其他储物家具，因此，同一回复中建议的餐边柜会被移除，其台面上建议摆放的四件物品也一并移除。照明值不会排除任何类型。

C.10. 渲染

Blender Cycles [9] 从上方渲染剖视图，并以平视图方式渲染，分辨率为 1024×768 像素，每像素最多采样 128 次。每个系统的场景均使用相同的相机设置进行渲染：相机光线所见的背景颜色相同，不调整曝光，也不设置相机白平衡。

D. 实验细节

D.1. 模型和运行

表 S4 列出了各角色所使用的模型。在主要评估中，每个角色使用不同的模型家族，而 PreferenceScene 从不调用用于模拟用户、评估场景或生成基准测试文本的模型。每个系统都使用随机种子 0–3，运行每种形式的 120 个测试实例。随机种子决定模拟用户和系统大语言模型每次调用所用的采样种子。所有交互均包含 $K = 3$轮次：系统展示三个场景，接收针对每个场景的反馈和评分，然后输出其最终场景，该场景不会收到反馈。所有系统的场景均由同一渲染器使用相同的两台摄像机渲染（第C.10节）。

D.2. 基线方法的适配

我们从其发布的代码运行 Holodeck [86], LayoutGPT [21], I-Design [12],LayoutVLM [70] 和 R3L [25]，并按如下方式对其进行调整。

常规更改。 Gemma-4-31B 替代每个基线方法中的大语言模型，其采样温度采用基线方法发布配置中的设置：Holodeck 和 R3L 为 1.0，LayoutGPT、I-Design 和 LayoutVLM 为 0.7，但 I-Design 的一个智能体使用 0。每个基线方法都使用其实例房间类型对应的房间外壳，该外壳由地板多边形、墙高、门和窗户组成，并从包含 6,793 个 Objaverse [19] 资产的池中选取物体。我们将每个基线方法的输出转换为我们的场景格式，并使用我们的渲染器进行渲染。由于所有基线方法都不设置照明，渲染器为其场景使用固定的照明设置：3500 K 下功率为 200W 的吸顶灯、6000K 下功率为 250 W 的窗户灯光，以及一张环境贴图。

各系统根据文本设定的内容。 Holodeck 设定物体、物体的摆放位置以及墙面和地面材料，并从固定库中选择这些材料。LayoutGPT、I-Design、LayoutVLM 和 R3L 设定物体及其摆放位置。这五种基线方法均不会决定房间的照明、窗户覆盖材料或纺织品。灯具、窗帘和纺织品仅作为它们摆放的资产池中的资产出现在场景中。

输入文本。 在第 1 轮，系统接收房间类型、指令、角色设定和属性列表，并将其作为“该房间的设计属性（属性：其决定的内容）：”这一行添加，随后每个属性各占一行，并附带其描述。在之后的每一轮以及最终场景中，文本还会针对每个较早轮次包含一个区块，其中包括所展示场景的描述、反馈和评分。该描述由系统的输出生成，列出墙面和地面材料，以及物体名称和数量。在预言机条件下，第 1 轮的文本后面会针对每个属性各列一行，其中包含其隐藏值的名称和描述。

Holodeck。 Holodeck 根据文本描述选择墙面和地面材料。我们通过 CLIP 与 Holodeck 的材质文本之间的相似度，将每种选定材料映射到 236 种自有 ambientCG [20]材料中的一种，并将映射后的材料向 Holodeck 材料的颜色进行着色，着色范围不超出映射材料的着色边界，且色相偏移最多为 ± 0.25色相环的范围；超出此限制的色相不会发生偏移。

LayoutGPT。 LayoutGPT 仅有卧室和客厅的上下文示例，因此在这两种房间类型的每种形式中，对 24 个测试样本运行。LayoutGPT 输出不含材料的物体框。每个框都会选用资产池中与其尺寸相符的对应类别资产，而墙壁和地板则使用固定材料。

I-Design。 I-Design 不选择墙壁或地板材料，而是沿用 Holodeck 场景中与输入和随机种子相同的墙壁和地板。在 I-Design 构建场景图之前，我们会移除大语言模型回复中将墙、天花板或房间中央指定为物体的条目。我们将其回溯求解器的运行时间限制为 300 秒，并将每次边缘提取的运行时间限制为 60 秒。超过任一时限的生成均计为失败。

LayoutVLM。 LayoutVLM 使用相同输入和随机种子对应的 Holodeck 场景中的墙壁和地板。约束程序无法求解的生成将计为失败。

R3L。 R3L将房间外壳作为其房间，放置地板物体，并使用相同输入和随机种子对应的 Holodeck 场景中的墙壁和地板。

Holodeck-Edit。 在第 1 轮中，Holodeck-Edit 显示 Holodeck 的场景。第 1 轮之后，每轮进行一次大语言模型调用，接收已发表的基线方法的输入文本、地板设计、墙面设计以及上一个显示场景的物体列表，并接收保留反馈未要求更改的每个部分的请求。未更改条目的家具及其上的小物体保留原有资产，Holodeck 的求解器会再次放置所有物体。

失败的生成。 系统在一次生成中重复执行的请求，例如 PreferenceScene 的重复大语言模型请求（第 B.2 和 B.3 节），属于系统的操作流程。我们不会重新运行失败的生成。系统会保留之前的场景。如果其第一次生成失败，则会像 LayoutVLM [70], 一样随机放置选定的资产；如果没有选定资产，则显示空置的房间。同样的规则也适用于 PreferenceScene：当某一轮失败时，它会保留之前的计划和场景。因基础设施故障（例如连接中断）而被中断的生成不计为失败，而是使用相同的随机种子重新启动。表S6 给出了每个系统的成功率。

D.3. 我们未运行的系统

我们未运行 SceneSmith [58] 和 SceneWeaver [87], , 因为它们的论文报告称，每个场景分别需要约 3.4 小时和 64 分钟，且一次交互需要 $K+1=4$ 次生成。基于 3D-FRONT [22], 训练的方法，例如 ATISS [57], 、DiffuScene [74] 和 InstructScene [44], , 最多仅覆盖我们十种房间类型中的四种。

D.4. 指标

偏好匹配。 pAMR（主论文公式（1））由评估器按照第 A.7 节所述计算。

场景质量。 评估器使用 R3L 未作改动的评估提示词，对 R3L 提供给评估器的俯视图和侧视图进行评分，分别评估 真实度、功能性 和 指令遵循度。 提示词中的指令字段包含第 1 轮的文本，但不含属性列表：房间类型、指令和角色设定。提示词不包含系统名称、计划或反馈。与 R3L 一样，每个场景评分五次，并对评分取平均值。这五次查询的采样温度为 1.0，top- [25] .0。越界率 p 1 是 R3L 的越界率，即占地轮廓超出房间多边形超过 $0.0003m^2$ 的物体所占的百分比。碰撞率 是在 *SceneEval* 的网格相交测试中与其他物体发生碰撞的物体所占的百分比 [72]: 如果两个物体的网格相交，且将其中一个物体从接触点移开 5 毫米后仍然相交，则这两个物体发生碰撞。地毯等地面覆盖物不参与测试，因为 SceneSmith [58]会排除地毯。场景实例化将光源放置在点亮的灯具内部（第 C.5 节）；这些光源属于灯具本身，不计为 碰撞率中的物体。

成本。 一次交互的大语言模型调用次数和词元数按其所有 $K+1$ 次生成累计计算。对于 PreferenceScene，这包括最终场景的家具布置调用（第 B.3 节）。每个场景的生成时间是系统自身步骤的墙钟时间，包括模型调用和场景构建，但不包括渲染，再除以一次交互中的 $K+1$ 个场景。对于 PreferenceScene，这些步骤还包括最终场景的家具布置。每轮生成时间是从第 k 轮的反馈到下一场景的渲染图像之间的墙钟时间，并对 $k=1$、.....、K 取平均值。因此，它包括渲染，但不包括第 1 轮场景的生成。与先前渲染过的场景相同的场景不会再次渲染，其图像从缓存中读取。两种时间都不包括模拟用户。计时测试在一台配备两颗英特尔至强金牌 6526Y 处理器（64 个线程）的主机上，逐次处理一次交互，执行每个系统的所有步骤。大语言模型运行于一台使用 4 位权重、vLLM 0.32.0 和一块 NVIDIA H100 GPU 的 Gemma-4-31B 服务器上；渲染则使用 NVIDIA RTX 4090 和 H100 GPU。

统计数据。 每项指标首先按每个随机种子的测试交互取平均值。表格给出四个随机种子的均值和标准差，或标题中所列随机种子的均值和标准差。Oracle 行和使用托管大语言模型的行仅使用随机种子 0。
PreferenceScene 与其他行之间的差异，是在两行共有的实例和随机种子配对上计算的。

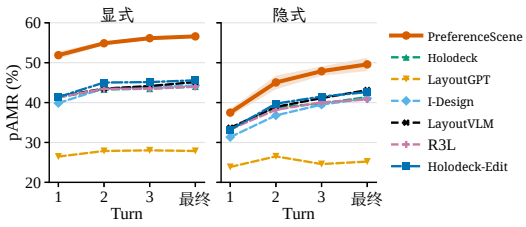


图 S3。PreferenceScene 中每轮场景及最终场景的 pAMR，其中每轮场景均已布置（第 B.3 节）、采用反馈的已发表的基线方法以及 Holodeck- Edit。曲线：四个随机种子上的均值；带状区域：PreferenceScene 的标准差。LayoutGPT 仅涵盖卧室和客厅。

E. 补充结果

E.1. 每轮匹配度

图 S3展示了所有系统每轮场景及最终场景的 pAMR，主论文则给出了 PreferenceScene 的取值。在隐式形式中，反馈是主要证据；PreferenceScene 从第 1 轮场景到最终场景的提升（+12.1 个百分点）超过了相同交互中所有基线的提升：两者提升幅度之差介于 +2.4 到 +4.6 个百分点之间，且每个区间都大于 0。每项差异均基于两个系统共有的实例、随机种子和属性三元组计算，先按每个随机种子的三元组取平均值，再对随机种子 0–3 取平均值；并基于实例计算 95% 分位数自助法区间（10,000 次重抽样）。

E.2. 各房间类型的结果

在测试实例（每种房间类型和形式各 12 个）中，PreferenceScene 在 显式形式的全部十种房间类型以及隐式 形式的十种房间类型中的九种上，最终场景的 pAMR（四个随机种子的平均值）最高。

E.3. 反馈中提及的属性

对于主论文“反馈的作用”段落中的数量，大语言模型标注器 Ministral-3-14B（`Ministral-3-14B-Instruct-2512`，温度为 0）读取表 1中每一轮交互的反馈。它看不到场景。对于给定其选项的房间类型的十个属性中的每一个，它都会回答反馈的任何部分是否涉及该属性，包括赞扬、抱怨、愿望或问题，并引用支持该答案的最短反馈片段。当标注器回答“是”时，该属性即在某一轮中被提及；如果交互中的任何一轮都未提及该属性，则该属性永远不会被视为在该交互中被提及。

E.4. 基线方法的其他变体

表S5 报告了已发表的基线方法的另外三种变体：有反馈历史但没有属性列表，有属性列表但没有反馈，以及两者都没有。没有反馈的变体根据其第 1 轮的输入生成一个场景。无论是否有属性列表，反馈历史都会提高每种已发表基线方法的 pAMR。在某种形式的十行反馈数据中，同一交互上某一行与不含反馈的相同变体之间的差异，在 +10.2 至 +50.2 分之间（显式 形式），在 +10.4 至 +90.7 分之间（隐式 形式）。属性列表对 pAMR 的影响小于 2 分。在两种形式中，PreferenceScene 都高于每种已发表基线方法的每一种变体。在 Oracle 条件下，系统除了第 1 轮的输入外，还会收到以文本形式表示的隐藏偏好，并生成一个场景。该显式 条件的取值与主论文表 1的 Oracle 列重复。每个 Δ 都使用两行共有的交互：Oracle 行使用随机种子 0，LayoutGPT 行使用卧室和客厅。

E.5. 生成成功率

遵循 OptiScene [88], 表 S6 报告了使用 Gemma-4-31B 时各系统成功生成的比例。I-Design 的大多数生成失败都超出了第D.2节规定的时间限制，而 LayoutVLM 的每次生成失败都是因为约束程序无解。

E.6. 物体数量中的光源

场景实例化会在资产池中一盏灯的光源下方放置一个半径为 5 cm 的发光球体（第 C.5 节）。该球体是这盏灯的光源，而非独立物体，因此主论文的表 2 将其计为这盏灯的一部分，列于 *Coll.*（第D.4节）。

E.7. 成本

表 S7 给出了每次交互的平均成本及其在各次交互中的范围。除六次大语言模型调用外，PreferenceScene 的每次交互平均还会向 Jev 发送 25 次请求，包含 21.3 万个输入词元（4 万至 45.5 万），费用为 0.0090 美元。每个场景的生成时间取自主论文中的计时测试；该测试按顺序逐一执行每个系统的 8 次交互，且所有系统均使用相同的设置和服务器。每轮生成时间在测试交互中测量；对于 PreferenceScene，不包括家具布置调用，该调用在最后一轮之后执行一次。

E.8. 评估器变化

我们通过两种方式重新评估所有交互的最终场景（主论文第6.2节）。第二评估者 Ministral-3-14B 在相同的渲染图像上回答相同的问题。“不可见”变体增加了“不可见”这一选项

方法	显式 pAMR	显式 Δ	隐式 pAMR	隐式 Δ
属性列表和反馈（如主论文所示）				
Holodeck 44.0±0.2	+12.6		41.4±0.6	+8.2
LayoutGPT	27.9±1.0	+26.4	25.2±1.3	+23.0
I-Design	44.3±1.1	+12.3	41.0±0.5	+8.5
LayoutVLM	45.1±0.3	+11.5	43.1±1.4	+6.4
R3L	43.9±1.0	+12.6	40.8±0.6	+8.8
反馈，不含属性列表				
Holodeck	43.0±0.1	+13.6	40.6±1.0	+9.0
LayoutGPT	27.1±0.6	+27.1	26.6±1.6	+21.6
I-Design	**45.6±0.3**	+11.0	41.7±1.0	+7.8
LayoutVLM	45.3±0.2	+11.3	**43.1±0.8**	+6.4
R3L	43.9±0.8	+12.7	41.2±0.3	+8.4
属性列表，不含反馈				
Holodeck 41.3±1.0		+15.3	33.2±0.5	+16.4
LayoutGPT	26.5±0.9	+27.8	23.9±2.3	+24.4
I-Design	39.9±1.3	+16.7	31.3±0.4	+18.2
LayoutVLM	**41.5±0.9**	+15.1	**33.7±0.9**	+15.9
R3L	41.3±0.7	+15.3	33.3±0.6	+16.3
没有属性列表，也没有反馈				
Holodeck	40.8±0.8	+15.8	34.2±1.1	+15.4
LayoutGPT	25.8±0.8	+28.4	24.7±1.2	+23.5
I-Design	40.4±1.3	+16.2	32.9±0.4	+16.7
LayoutVLM	**43.4±0.4**	+13.2	**35.1±0.4**	+14.5
R3L	41.2±0.5	+15.4	33.5±0.5	+16.1
Oracle：给定隐藏偏好（一个随机种子）				
Holodeck	45.3	+11.1	44.0	+7.0
LayoutGPT	25.8	+29.2	25.4	+24.8
I-Design	43.8	+12.7	40.3	+10.7
LayoutVLM	47.8	+8.7	47.5	+3.5
R3L	45.5	+11.0	45.2	+5.8
PreferenceScene	**59.4**	–	**60.7**	–
PreferenceScene（本文方法）	**56.6±0.3**	–	**49.6±1.7**	–

表 S5. 已发表的基线方法在有无属性列表和反馈以及预言机条件下的 pAMR（%，↑）；四个随机种子（预言机行使用一个随机种子）的均值 ± 标准差。 Δ：PreferenceScene 的反馈条件结果减去相同交互下对应行的结果，单位为百分点。粗体表示最佳；下划线表示次佳；每个区块内均标出这两者，最后一行与第一个区块一起排名。

系统	显式	隐式
Holodeck	98.1	98.5
LayoutGPT	**100.0**	**100.0**
I-Design	66.6	66.3
LayoutVLM	74.5	75.8
R3L	91.9	90.6
Holodeck-Edit	99.3	99.4
PreferenceScene（本文方法）	**100.0**	**100.0**

表 S6。在使用四个随机种子的测试实例中，每次交互的四次生成结果中，成功生成的比例（%，↑）。加粗：最佳；下划线：次佳。

对评估者提出的每个问题都回答“图像中不可见”，且此回答计为不匹配。桌子S8 显示

PreferenceScene 在两种变化条件下的两种形式中，pAMR 都是最高的。

E.9. 使用托管大语言模型替代 Gemma-4-31B

表 S9 比较了 PreferenceScene 分别使用 GPT-6.1 Terra 和 Gemma-4-31B 作为其大语言模型时的结果。每种形式包含 30 个随机抽取的测试实例，每个实例在随机种子 0 下进行一次交互（第 D.1 节）。使用 GPT-6.1 Terra 的 PreferenceScene 不执行家具布置步骤（第 B.3 节），因此两行结果均对未经过该步骤的最终场景进行评分。使用 GPT-6.1 Terra 时，显式 形式的 pAMR 为 59.3%，而 隐式 形式为 52.0%；使用 Gemma-4-31B 时，这两个数值分别为 60.5% 和 48.3%。

E.10. 消融实验详情

主论文的表 3 给出了消融变体最终场景的 pAMR。在第 1 轮中，我们根据相同的交互，以及交互中展示的未布置家具的场景，计算某个变体与完整方法之间的差异。该差异以百分点表示，为 +0.3（显式形式）和 −0.4（没有较早轮次历史记录的隐式形式），以及在没有新取值时的 −0.9 和 −0.2。因此，在第 1 轮中，这些变体与完整方法的差异小于 1 个百分点，最终场景中的差异源于计划更新。

E.11. 采用我们的场景实例化的基线方法

我们的场景实例化会根据各自的取值设置每面墙的颜色、光源和窗户覆盖材料，而基线方法则为所有墙壁设置同一种材料，在固定的照明设置下渲染每个场景，并且最多将窗户覆盖材料作为现成素材放置（第 D.2 节）。我们将其应用于每个基线方法的最终场景，使用该基线方法自行决定的这些部件的取值，并将这些场景与该基线方法的原始场景以及 PreferenceScene 的场景进行比较。

设置。 我们使用表格中最终场景。 1 这些场景来自主论文：每个基线方法在所有测试实例上生成的场景（每种形式 120 个），使用随机种子 0–3。我们将 新增属性定义为墙壁、照明和窗饰的属性，以及基线方法无法放置其物体的属性。对于每个最终场景，我们为每个新增属性构建一个取值。对于每个基线方法，这些属性包括每种房间类型中的墙面颜色和照明、具有这些属性的房间类型中的特色墙和窗饰、浴室中的墙砖、厨房中的防溅板和吊柜，以及候诊室中的声学处理。对于 R3L，还包括浴室中的镜子、餐厅中的餐桌布置、画廊中的艺术品类型和悬挂方式、厨房中的装饰主题，以及候诊室中的墙面装饰；这些属性均无法由 R3L 更改。我们不添加

方法	LLM 调用 ↓	大语言模型输入 词元（10^3）↓	大语言模型输出 词元（10^3）↓	生成时间 每个场景（秒）↓	生成时间 每轮（秒）↓
Holodeck	19.5 (14.0–20.0)	28.9 (14.8–35.0)	11.5 (3.6–19.1)	102 (54–134)	361 (133–1653)
LayoutGPT	**4.0** **(4.0–4.0)**	18.6 (15.2–22.0)	**1.3** **(0.3–2.0)**	**12** **(9.6–18)**	**155** **(59–578)**
I-Design	85.8 (24.0–339.0)	123.8 (36.7–474.0)	26.1 (5.1–84.7)	340 (131–639)	823 (170–4611)
LayoutVLM	184.9 (103.0–332.0)	454.0 (250.2–783.5)	7.9 (3.0–17.6)	361 (112–1588)	753 (257–3729)
R3L	59.7 (24.0–92.0)	120.8 (33.1–287.4)	28.4 (3.3–74.8)	332 (185–510)	695 (163–1618)
Holodeck-Edit	13.9 (11.0–15.0)	**17.9** **(8.5–22.8)**	8.5 (1.4–14.1)	79 (43–98)	358 (86–1829)
PreferenceScene（本文方法）	6.0 (6.0–8.0)	35.1 (24.2–51.7)	9.2 (7.5–14.5)	43 (34–65)	269 (5.7–562)

表 S7。测试集两种形式的测试实例上每次交互的成本（$K = 3$，四个场景），以所有交互的平均值（最小值–最大值）表示。每个场景的生成时间在主论文的计时测试中测量（每个系统进行八次交互），每轮生成时间则在测试交互中测量。词元数量以 Gemma-4-31B 的词元为准。粗体：最佳；下划线：次佳。

	显式			隐式		
方法	评估器	第二评估器	“Not 可见”	评估器	第二评估器	“Not 可见”
Holodeck	44.0±0.2	42.9±1.1	35.6±0.9	41.4±0.6	40.8±1.1	33.4±0.6
LayoutGPT	27.9±1.0	26.4±1.9	22.2±1.0	25.2±1.3	25.5±1.2	19.9±1.3
I-Design	44.3±1.1	43.8±1.7	34.0±1.3	41.0±0.5	41.0±0.6	30.6±0.4
LayoutVLM	45.1±0.3	44.0±0.3	33.0±0.4	43.1±1.4	43.0±1.3	32.2±1.0
R3L	43.9±1.0	42.4±1.1	33.6±1.0	40.8±0.6	39.9±1.0	31.2±0.3
Holodeck-Edit	45.6±0.7	44.9±0.5	35.8±1.0	42.6±1.4	41.9±1.2	32.3±1.4
PreferenceScene（本文方法）	**56.6±0.3**	**52.5±1.0**	**47.8±0.6**	**49.6±1.7**	**46.5±0.8**	**40.9±1.1**

表 S8. 主论文中的评估器（GLM-4.6V-Flash）、第二评估者（Ministral-3-14B）以及提供“不可见”选项的评估器对最终场景的 pAMR（%，↑）；四个随机种子上的均值 ± 标准差。粗体：最佳；下划线：次佳。

PreferenceScene 的大语言模型	显式	隐式
GPT-6.1 Terra	59.3	**52.0**
Gemma-4-31B	**60.5**	48.3

表 S9. PreferenceScene 使用两个大语言模型时，在每种形式的相同 30 次测试交互（随机种子 0）上的 pAMR（%，↑）；评分基于未经过家具布置步骤的最终场景。粗体：最佳。

基线方法所放置的地板或物体的属性，例如家具风格、储物和装饰。

每个值都是基线在其最终生成文本中自行确定的值，并由固定关键词规则提取，无需模型调用。这些文本包括墙面设计、基线所选物体的描述，以及（如果其大语言模型写出的话）设计策略；对于 I-Design、LayoutVLM 和 R3L，墙面设计取自构建其墙壁的 Holodeck 生成结果（第 D.2节）。这些文本未确定的新增属性采用其所属类别的默认规格（第 C.7节）。墙壁、特色墙或窗饰的颜色，是 Matplotlib 的 148 种命名颜色中，文本所指值点名的第一种颜色，并带有其明度词。如果文本点名如果这些颜色中没有任何一种颜色，或其他颜色或色调词修饰该名称或与其连写，则 Gemma-4-31B（所有系统所使用的大语言模型）仅阅读此文本，并从这些颜色中选择一种作答，或回答“无”。如果答案为“无”，则该取值采用默认颜色；对于特色墙，则采用通过 CLIP 相似度与文本最接近的颜色（第 C.2 节）。在此构建过程中，没有任何模型读取反馈。

PreferenceScene 的场景实例化会将每个取值转换为规格，并使用构建器以及 PreferenceScene 针对该属性所采用的高度和放置方式来构建它（主论文第 4.3 节和第 C 节）。照明值会替换基线场景中的固定照明（第 D.2 节），添加该取值所描述的灯具，并点亮基线场景中的灯具（第 C.5 节）。餐桌布置会放置在基线场景的餐桌上；当场景中没有餐桌或餐桌已摆满时，则放置在其他家具上；防溅板和吊柜会沿着基线场景的台面布置，或者在没有台面的厨房中，沿着最长的墙面空段布置。除这些变化外，基线场景中的任何物体都不会被移动、替换、重新着色或移除。与 PreferenceScene 的场景一样，每个新增物体都会检查其是否与基线场景中的物体发生碰撞；如果发生碰撞，则不会放置该物体；对于发生碰撞的落地窗帘，则会再尝试一次

	显式					隐式				
基线	原始	Own	Ours	Own − 原始	Ours − 自有	原始	Own	Ours	Own − 原始	Ours − 自有
所有属性										
Holodeck	44.0±0.2	50.6±0.7	58.1±0.5	+6.6±0.7 [+5.2, +7.9]	+7.5±0.9 [+5.5, +9.4]	41.4±0.6	46.0±0.7	49.7±1.9	+4.6±0.7 [+3.5, +5.8]	+3.7±2.3 [+1.7, +5.8]
I-Design	44.3±1.1	50.9±1.6	58.1±0.5	+6.6±0.7 [+5.6, +7.7]	+7.1±1.1 [+4.8, +9.6]	41.0±0.5	45.4±0.6	49.7±1.9	+4.3±0.2 [+3.4, +5.3]	+4.4±2.0 [+1.8, +6.9]
LayoutVLM	45.1±0.3	51.2±1.1	58.1±0.5	+6.0±1.0 [+4.8, +7.3]	+6.9±1.4 [+4.8, +9.0]	43.1±1.4	47.1±1.1	49.7±1.9	+3.9±0.4 [+3.0, +4.8]	+2.7±2.3 [+0.4, +4.9]
R3L	43.9±1.0	51.2±0.6	58.1±0.5	+7.3±0.9 [+6.1, +8.4]	+6.9±0.4 [+4.6, +9.1]	40.8±0.6	45.1±0.9	49.7±1.9	+4.3±1.0 [+3.2, +5.3]	+4.7±2.5 [+2.4, +6.9]
Holodeck-Edit	45.6±0.7	52.5±1.1	58.1±0.5	+6.9±0.5 [+5.7, +8.2]	+5.6±1.3 [+3.6, +7.6]	42.6±1.4	46.7±1.7	49.7±1.9	+4.1±0.5 [+3.1, +5.2]	+3.0±1.7 [+0.9, +5.2]
新增属性										
Holodeck	29岁.9±0.2	46.9±1.1	58.4±1.5	+17.0±1.0 [+13.9, +20.2]	+11.5±1.3 [+8.1, +14.9]	30.2±1.3	42.8±0.6	49.0±2.0	+12.6±1.6 [+10.1, +15.1]	+6.2±2.3 [+2.8, +9.6]
I-Design	28.6±1.0	45.0±2.1	58.4±1.5	+16.4±1.1 [+13.7, +19.1]	+13.3±2.7 [+9.8, +16.9]	29岁.8±0.9	40.6±0.5	49.0±2.0	+10.8±0.7 [+8.6, +12.9]	+8.5±1.6 [+5.1, +11.9]
LayoutVLM	31.1±0.8	47.3±0.8	58.4±1.5	+16.2±1.6 [+13.2, +19.0]	+11.1±2.1 [+7.9, +14.2]	32.2±2.6	42.7±1.7	49.0±2.0	+10.5±1.5 [+8.3, +12.7]	+6.3±3.6 [+3.0, +9.6]
R3L	28.8±1.3	44.2±0.6	58.2±1.2	+15.4±1.0 [+13.0, +17.7]	+14.0±1.2 [+10.8, +17.3]	29岁.5±0.8	39.4±1.9	48.7±2.2	+9.9±2.1 [+7.8, +12.1]	+9.4±3.3 [+6.1, +12.6]
Holodeck-Edit	32.8±0.6	50.8±1.4	58.4±1.5	+18.0±0.9 [+15.2, +20.9]	+7.6±0.9 [+4.6, +10.5]	32.8±1.0	44.5±0.9	49.0±2.0	+11.7±0.9 [+9.5, +14.1]	+4.6±1.7 [+1.0, +8.1]

表 S10。pAMR（%， ↑）分别表示每个基线场景的最终场景（原始）、采用我们对新增属性进行场景实例化后的同一场景（自有）以及 PreferenceScene 在家具布置步骤之前的最终场景（我们的方法），并给出差异的百分点数；统计对象为主论文表 1 中的所有测试实例（每种形式 120 个），随机种子为 0–3。自有采用基线自行确定的取值，否则采用默认值。均值 ± 四个随机种子上的标准差；每个差异下方给出跨实例的 95% 自助法区间。新增属性：墙面颜色和照明，以及在具有这些属性的房间类型中包含的特色墙、窗饰、墙砖、防溅板、吊柜和声学处理。对于 R3L，新增属性还包括镜子（浴室）、餐桌布置（餐厅）、艺术品类型和悬挂方式（画廊）、装饰主题（厨房）以及墙面装饰（候诊室）；因此，在此区块中，R3L 的我们的方法取值不同于其他基线方法。

窗台长度越长。

比较。 我们使用基线方法的自有取值渲染场景，并使用评估器（第 A.7 节）对其进行评分，方式与其他最终场景相同。在相同的实例和随机种子上，我们比较基线方法的最终场景（原始）、使用其自有取值的相同场景（自有），以及 PreferenceScene 交互后的最终场景（我们的方法）。基线方法没有家具布置步骤（第 B.3 节），因此我们的方法在该步骤之前即为最终场景。我们针对实例、随机种子和属性构成的三元组计算每个差值，并在每个随机种子的三元组上取平均值。我们报告四个随机种子上的均值 ± 标准差，以及在实例上计算的 95% 分位数自助法区间（10,000 次重抽样）。

结果。 表格S10 给出了所有属性和新增属性上的 pAMR。对于所有属性，采用基线自身取值减去原始值后得到的场景，其差值介于+6.0 到 +7.3 个点之间，形式为 显式 ；形式为 +3.9 到 +4.6 个点之间，形式为隐式 ；且每个区间均大于 0。对于新增属性，该差值介于 +15.4 到 +18.0 个点之间，形式为 显式 ；介于 +9.9 到 +12.6 个点之间，形式为 隐式 ，且每个区间均大于 0。对于未新增、但仍可能受到替换后的照明影响的属性，该差值介于 −0.2 到 +0.9 个点之间，形式为 显式 ，介于 −0.6 到 +0.4 个点之间，形式为 隐式 ，且没有区间小于 0。

对于新增属性，PreferenceScene 与使用基线方法自有取值的场景之间的差值，在<strong?Explicit 形式下介于 +7.6 和 +14.0 分之间，在 隐式 形式下介于 +4.6 和 +9.4 分之间，且每个区间均高于 0。在所有属性上，该差值在显式 形式下介于+5.6 和 +7.5 分之间，在 隐式 形式下介于 +2.7 和 +4.7 分之间，且每个区间均高于 0。在相同的场景实例化条件下，PreferenceScene 在两种形式中匹配到的隐藏值都多于所有基线方法。

E.12. 匹配每个系统都能更改的属性

已发表的基线方法会选择并摆放物体，还会选择墙壁和地板的材料（主论文第 6.2 节）。我们将 PreferenceScene 与这些方法进行比较，比较的属性仅限于每个系统都以这些方式设置房间相应部分的属性。

系统	显式		隐式	
	pAMR	PreferenceScene −系统	pAMR	PreferenceScene − 系统
Holodeck	48.5±0.2	+10.1±0.4 [+7.6, +12.6]	45.9±0.7	+5.8±2.3 [+3.2, +8.4]
I-Design	49.2±1.1	+9.3±1.6 [+6.3, +12.4]	45.7±0.6	+6.0±2.0 [+2.9, +9.2]
LayoutVLM	48.5±0.2	+10.0±0.7 [+7.2, +12.9]	46.5±1.4	+5.2±2.0 [+2.5, +8.0]
R3L	49.5±1.1	+9.0±1.5 [+6.2, +11.9]	45.5±0.6	+6.2±1.8 [+3.4, +9.1]
Holodeck-Edit	50.4±1.1	+8.2±1.3 [+5.6, +10.8]	46.8±1.6	+4.9±0.7 [+2.3, +7.4]
PreferenceScene（本文方法）	**58.5±0.5**	–	**51.7±1.8**	–

表 S11. pAMR（%，↑）在除 LayoutGPT 外所有系统均可更改的 76 个属性上的表现，以及 PreferenceScene 与各基线之间的差值（百分点）。场景：主论文表1 中交互的最终场景（每种形式 120 个测试实例，随机种子 0–3）。均值 ± 四个随机种子上的标准差；每个差值下方为基于实例计算的 95% 分位数自助法区间。

属性。对于每种房间类型、属性和系统，我们会查看经过我们适配的系统代码（第D.2节）及其最终场景。当除 LayoutGPT 外的六个系统都根据输入，通过摆放物体或设置墙壁、地板的材料来设置该属性所描述的房间部分，且该部分在该系统针对该房间类型生成的最终场景之间有所不同时，该属性便纳入比较范围。100 个属性中有 76 个符合这一条件，涵盖全部十种房间类型。其余 24 个属性包括十种房间类型的照明，其取值描述了点亮的光源，而每个基线场景都采用一种固定的照明设置进行渲染（第 D.2节）；有窗饰的七种房间类型的窗饰，而 Holodeck、R3L 和 Holodeck-Edit 不会摆放窗饰；厨房的吊柜，因为共享资源池中没有可安装在墙上的资产；以及六种墙面装饰和装饰品属性，而 R3L 只摆放地板上的物体，因此不会摆放这些物件：镜子、餐桌布置、艺术品类型、悬挂方式、厨房的装饰主题和墙面装饰。LayoutGPT 仅用于卧室和客厅，并且只摆放物体框，不包含材料；它生成的房间墙壁和地板材料固定（第 D.2节），因此不纳入此次比较。

比较。我们使用主论文表1 中交互的最终场景，并在实例、随机种子和属性三元组上计算 PreferenceScene 与基线的差值，再对每个随机种子的三元组取平均。我们报告四个随机种子的均值 ±和标准差，并给出基于实例的 95% 百分位数自助法区间（10,000 次重抽样）。

结果。 表格 S11 给出了这些属性上的 pAMR。PreferenceScene 在两种形式下的 pAMR 均为最高。它与基线方法的差异在 +8.2 到 +10.1 个百分点之间，适用于显式 形式；在 +4.9 到 +6.2 个百分点之间在隐式 形式下，且每个区间均大于 0。

E.13. 用开放模型替代 Jev

PreferenceScene 会向 Jev 发送两类多项选择问题 [76], ——Jev 是按请求付费的托管决策模型（第E.7节）：一类问题在第 1 轮之前将设计方面分配给属性（第B.1节），另一类是在每一轮进行场景实例化的具体说明问题（第C.1节）。为了测试 PreferenceScene 是否需要该托管模型，我们使用开放模型回答这两类问题，并比较其与 PreferenceScene 在相同实例和随机种子下使用 Jev 时的匹配度（表S12）。

设置。 开放模型是 Gemma-4-31B，即 PreferenceScene 的开放权重大语言模型（表 S4），因此大语言模型与决策模型是同一个模型。向 Jev 发出的请求包含一个上下文和若干道选择题。开放模型通过单独调用的方式接收 Jev 请求中的每一道问题；调用内容包括请求的上下文、带有 A、B、C、.....标签的选项，并要求模型给出最佳选项的字母。对于每个选项，我们取回复的首个词元为该选项字母的概率，并将其除以所有显示选项的这些概率之和。每道问题都会按照给定顺序和相反顺序分别呈现选项，随后对两种分布取平均值。对于选项超过 26 个、超出字母数量的问题，会将其拆分为若干道关于各组选项的问题，再加上一道关于每组最可能选项的问题；然后通过 Luce选择模型的最大似然拟合，将这些问题的分布合并为涵盖所有选项的一个分布。因此，开放模型会像 Jev 一样，为每道问题的每个选项返回一个概率。问题、选项，以及将这些概率转换为分配或规格的每条规则（包括分配的阈值（第 B.1节））均保持不变。PreferenceScene 不使用针对 Jev 概率拟合的校准，因此不会为开放模型重新拟合任何参数。

	显式			隐式		
场景	Jev	开放模型	Δ	Jev	开放模型	Δ
第 *E.9* 节的测试实例（每种形式 *30* 个）						
第 1 轮	51.8±1.7	52.8±2.1	+0.9±0.7 [−1.4, +3.0]	35.9±2.3	37.0±2.0	+1.1±1.1 [−1.1, +3.2]
最终场景	55.8±1.9	57.2±1.1	+1.4±2.4 [−1.4, +4.2]	47.9±1.0	46.3±0.3	−1.6±1.1 [−5.1, +2.1]
餐厅和办公室测试实例（每种形式 *24* 个）						
第 1 轮	63.8±1.6	64.7±0.9	+0.9±0.7 [−2.6, +4.2]	40.3±2.7	40.2±0.7	−0.1±2.3 [−2.3, +2.2]
最终场景	67.4±1.5	65.8±1.9	−1.5±1.9 [−4.7, +1.9]	55.6±2.7	55.8±0.9	+0.3±2.4 [−3.9, +4.4]

表 S12。PreferenceScene 使用 Jev，以及使用 Gemma-4-31B 替代 Jev（开放模型）时的结果；两者均在随机种子 0–2 下的测试实例的两个子集上运行，并在家具布置步骤完成后对所有场景进行评分：pAMR（%，↑）以及开放模型与 Jev 之间的差异（Δ，百分点）；均值 ± 三个随机种子上的标准差，每个差异下方给出跨实例的 95% 自助法区间。

因此，不会为开放模型重新拟合任何参数。

比较。 我们使用开放模型运行 PreferenceScene，使用 $K = 3$以及随机种子 0–2，在两组测试实例上进行测试：第E.9节中每种形式的 30 个实例（每种形式 90 次交互），以及消融实验中的每种形式的 24 个餐厅和办公室实例（每种形式 72 次交互）。对照组是在相同实例和随机种子上使用 Jev 的 PreferenceScene，即主论文表 1 中的交互。两种配置均从候选项调用的同一条回复开始，即在第 1 轮之前写入候选值的大语言模型调用（第B.1节），因此具有相同的候选值。两种配置从分配开始可能有所不同。如图 S3所示，每个场景均由评估器（第 A.7节）在家具布置步骤（第 B.3节）完成后进行渲染和评分。每个差异均基于实例、随机种子和属性三元组计算。先对每个随机种子的三元组取平均，再报告为均值 ± 三个随机种子上的标准差，并给出跨实例的 95% 百分位自助法区间（10,000 次重抽样）。

结果。表 S12 给出了匹配度。在第 E.9 节测试实例的最终场景中，开放模型与 Jev 之间的差异在显式形式下为 +1.4 分（95% 区间为 [−1.4, +4.2]），在隐式形式下为 −1.6 分（[−5.1, +2.1]）。在餐厅和办公室测试实例的最终场景中，差异在显式形式下为 −1.5 分（[−4.7, +1.9]），在隐式形式下为 +0.3 分（[−3.9, +4.4]）；两个区间均包含 0。在第 1 轮，即分配后的第一个场景，以及最终场景中，表 S12 中两个实例集的所有区间均包含 0，第 2 轮和第 3 轮的区间也包含 0。用开放模型替代 Jev 后，未检测到匹配度差异：最终场景的四个区间均包含 0，且排除了损失超过 5.1 分的可能性。开放模型回答一次交互中的多项选择题时，

平均调用 Gemma-4-31B 1,681 次（574–2,972 次），因为它会在单独的调用中接收每个问题，而使用 Jev 的 PreferenceScene 在测试实例中每次交互会向 Jev 发送约 25 个请求（第 E.7 节）。在表S12 的交互中，使用开放模型进行一次交互无需付费请求，平均向 Gemma-4-31B 发送 564.6k 个输入词元；而使用 Jev 进行一次交互则向 Gemma-4-31B 发送 36.1k 个输入词元，并向 Jev 发送 226.2k 个。

E.14. 噪声反馈

模拟用户根据其偏好叙述撰写反馈（第 A.4 节），但真实用户也可能表达自己并不具备的偏好，例如要求采用自己并不喜欢的墙面颜色。因此，我们测试 PreferenceScene 对包含此类错误陈述的反馈会作出何种回应。

设置。 在每一轮中，PreferenceScene 收到的反馈会以 p 的概率被独立重写。系统会使用固定随机种子，从房间的十个属性中随机抽取一个属性，并随机抽取该属性的一个非隐藏值。Gemma-4-31B（所有系统所使用的大语言模型）会将反馈重写为温度为 0 的请求，使其要求使用抽取值，不再陈述该属性的隐藏值，同时保留反馈的其余内容。如果反馈未提及该属性，重写内容会添加对抽取值的请求。模拟用户保持不变：它读取相同的品味叙述，其早先轮次的历史记录（第 A.4 节）保存的是它写下的反馈，而不是重写后的反馈。我们使用 $p = 0.25$ 和 $p = 0.5$，这是基于大语言模型的偏好 elicitation 研究中的非零噪声水平 [5], , 在该研究中，模拟用户的一定比例回答会被随机回答替代。例如，在餐厅的一轮噪声轮次中，抽取的属性是餐桌布置，隐藏值为 *Bare Ta-ble*，抽取值为 花卉中心装饰（第 A.2 节）。句子“另外，请移除桌子中心的装饰性物品；我确实需要让桌面保持

	显式				隐式			
	所有属性		错误陈述的属性		所有属性		错误陈述的属性	
场景	pAMR ↑	Δ	pAMR ↑	Δ	pAMR ↑	Δ	pAMR ↑	Δ
带有 $p = 0.25$ 的噪声反馈								
第2轮	64.6±2.6	−0.3±3.3 [−2.2, +1.5]	22.9±25.3	−59.0±21.3 [−83.3, −33.3]	48.0±2.0	0.0±1.0 [−3.5, +3.1]	18.1±6.4	−32.6±33.4 [−53.9, −7.6]
第3轮	63.7±2.7	−2.2±3.4 [−4.2, −0.1]	25.0±16.9	−49.9±27.2 [−68.1, −29.2]	51.9±1.3	−2.6±4.2 [−5.8, +0.6]	27.8±12.7	−28.2±28.6 [−42.8, −15.3]
最终场景	65.3±3.1	−2.0±2.9 [−4.5, +0.5]	48.5±11.7	−19.1±3.3 [−27.8, −9.7]	53.1±0.7	−2.4±3.1 [−6.3, +1.2]	40.7±11.7	−19.9±14.3 [−33.4, −3.4]
带有 $p = 0.5$ 的噪声反馈								
第2轮	62.3±2.8	−2.6±3.0 [−4.7, −0.6]	23.8±3.4	−51.8±13.8 [−66.6, −33.8]	48.1±1.4	+0.1±2.1 [−2.5, +2.8]	18.8±8.6	−29.5±18.9 [−45.4, −14.4]
第3轮	62.3±2.6	−3.6±3.6 [−6.7, −0.6]	36.0±8.3	−34.8±10.5 [−45.2, −24.5]	50.8±2.8	−3.6±4.9 [−6.5, −0.7]	28.1±4.3	−22.1±15.1 [−31.3, −12.9]
最终场景	63.1±1.5	−4.2±0.3 [−7.3, −1.3]	49.4±14.6	−21.9±7.8 [−30.7, −12.8]	53.1±2.9	−2.5±3.6 [−5.6, +0.7]	38.9±2.6	−19.3±6.7 [−29.4, −8.3]
第 *1* 轮中错误陈述的属性，在后续场景中								
$p = 0.25$，第3轮	–	–	41.0±30.4 (13)	−43.1±21.4 [−70.8, −11.1]	–	–	35.4±13.0 (17)	−17.4±24.0 [−35.6, +6.5]
$p = 0.25$, 最终场景	–	–	66.0±9.4 (13)	−12.5±12.5 [−26.2, +2.1]	–	–	42.4±8.4 (17)	−16.0±9.4 [−36.5, +11.1]
$p = 0.5$, 第3轮	–	–	56.4±1.7 (39)	−24.4±6.9 [−40.6, −10.0]	–	–	40.3±16.7 (34)	−12.7±22.8 [−22.0, −4.4]
$p = 0.5$，最终场景	–	–	55.6±11.4 (39)	−20.2±13.4 [−37.0, −3.5]	–	–	45.1±28.5 (34)	−10.9±25.7 [−23.9, +2.4]

表 S13. 当每轮反馈以概率 p 被改写为要求某一属性的错误值时，PreferenceScene 在餐厅和办公室测试实例上的 pAMR（%），以及与主论文表 1 中干净交互的配对差值（Δ，百分点）；均值 ± 三个随机种子上的标准差，下面给出实例层面的 95% 自助法区间。错误陈述的属性：在场景生成前的噪声轮次中抽取的属性；该集合从第 2 轮起逐渐扩大，直至最终场景。底部区块：仅包含在噪声第 1 轮中抽取的属性，pAMR 下方标出交互数量；在第 2 轮时，该集合对应“第2轮”各行。

完全空置…… ”被替换为“另外，请在桌子中间放置一个装满鲜花的大花瓶。”，其前的两个句子保持逐字不变。

我们在餐厅和办公室测试实例上（每种形式各 24 个），使用随机种子 0–2，在两个层级上运行 PreferenceScene，并采用 $K = 3$，总计 288 次交互。在一种形式、一个噪声水平下的 216 轮交互中，随机抽样使其中 56 轮（显式形式）和 44 轮（隐式形式）在 $p = 0.25$时受到噪声影响，而在 $p = 0.5$时分别有 112 轮和 108 轮受到噪声影响。对照组由主论文表 1 中相同测试实例和随机种子上的 PreferenceScene 交互组成，其反馈未被改写；我们称之为干净交互。与图 S3一样，每个场景在家具布置步骤（第 A.7 节）之后，由评估器渲染并评分（第 B.3 节）。

对改写后反馈的检查。在每个噪声轮次中，Gemma-4-31B 都会阅读 PreferenceScene 收到的反馈，并判断该反馈要求的是隐藏值、抽取值，还是两者都不是。它在全部 320 个噪声轮次中都识别出抽取值，在任何轮次中都没有识别出隐藏值。在这些轮次的原始反馈中，它在 132 个轮次中识别出隐藏值，在另外 188 个轮次中则认为两者都不是；在这 188 个轮次中，改写加入了错误陈述。在第二个问题中，它会并排阅读原始反馈和改写后的反馈，并列出改写删除、新增或更改的所有请求、赞扬或批评（不涉及抽取属性的部分）。在 320 个噪声轮次中的 303 个（94.7%）中，它没有列出任何内容。

比较。 每个差值均基于噪声交互与干净交互共有的实例、随机种子和属性三元组计算。先对每个随机种子的三元组求平均，再将各随机种子的均值报告为 ± 个随机种子的均值 ± 标准差，并给出基于实例的 95% 分位数自助法区间（10,000 次重抽样）。我们报告所有属性以及错误陈述的属性上的匹配度，即场景生成前噪声轮次中抽取的属性：对于第 2 轮场景，是第 1 轮抽取的属性；对于第 3 轮场景，是第 1、2 轮抽取的属性；对于最终场景，则是第 1 至第 3 轮抽取的属性。对于错误陈述的属性，干净交互也基于相同的属性评分。由于这些属性集在不同场景间有所不同，我们还追踪一个固定属性集，即噪声第 1 轮中抽取的属性，并考察它在第 2、3 轮场景及最终场景中的表现（表S13底部部分）。对于这个属性集，我们还使用相同的三元组和相同的估计量，计算第 2 轮场景到最终场景的差值变化。

结果。 表 S13 给出了匹配度。第 1 轮场景生成于任何反馈之前，与干净交互生成的场景相比，在两个层级和两种形式下，所有属性的评分均相差 −0.1 分，且所有区间均包含 0。对于所有属性，噪声反馈下最终场景的 pAMR 比干净交互下的最终场景低 2.0 至 4.2 分（显式形式为 67.4%，隐式形式为 55.6%）。只有在显式 形式下，该差值的区间低于 0，且为显式 形式，在 $p = 0.5$错误陈述的属性占最终场景三元组的 6.0% 至 14.4%。

对于错误陈述的属性，损失在噪声轮次之后立即变大。在第2轮的场景中，Prefer- enceScene 根据结合第 1 轮噪声反 馈更新后的计划构建场景，差异介于 −59.0 和 −29.5 个百分点之间。在最终场景中，错误陈述的属性是第 1 至第 3 轮抽取的属性，差异介于 −21.9 和 −19.1 个百分点之间，并且这两行的每个区间都低于 0。表 S13 的底部区块对应于第 1 轮抽取的属性，每种等级与形式的组合包含 13 至 39 次交互。在这组数据中，第2轮的场景中的差异介于 −59.0 和 −29.5 个百分点之间，第3轮的场景中的差异介于 −43.1 和 −12.7 个百分点之间，最终场景中的差异介于 −20.2 和 −10.9 个百分点之间；在最终场景中，只有 $p = 0.5$ 处的显式形式区间低于 0。从第2轮的场景到最终场景，这些三元组上的差异变化了 +16.7 至 +46.5 个百分点；在四种等级与形式组合中的三种组合里，该变化的区间高于 0，而在 $p = 0.25$ 处的隐式形式中包含 0（+16.7 个百分点，[−9.6, +46.7]）。因此，对于第 1 轮错误陈述的属性，后续轮次抵消了第2轮场景中的部分损失。模拟用户会将自有反馈保留在其历史记录中，并根据其品味叙述为每个场景撰写反馈。因此，它可以对抽取值提出异议，并再次询问其隐藏值。由此，该设置模拟了这样一种用户：在某一轮中错误陈述偏好，并能在后续轮次中对结果提出异议。

总而言之，错误陈述会大幅降低下一场景中对应属性的匹配度。对于第 1 轮中被错误陈述的属性，后续轮次会弥补部分损失；在四种水平与形式的组合中，有三种组合的变化区间高于 0。所有属性的匹配度变化则小得多。

E.15. 更多消融实验

我们移除了 PreferenceScene 的其余组件：明确的计划、计划更新历史记录中的场景描述以及实例化规则。如主论文表 3所示，每个变体都在餐厅和办公室测试实例上运行（每种形式 24 个），使用 $K = 3$ 以及随机种子 0–3，并且与完整方法一样，在布置步骤完成后的最终场景上进行评分（第 B.3节）。完整方法这一行使用的是主论文表 1 中 PreferenceScene 的交互在相同的实例和随机种子上运行，如表 3所示。

- 不使用显式计划时，大语言模型会在完整方法生成计划的地方，以自由文本形式撰写房间描述。另一次大语言模型调用会将该描述转换为每个属性的一个取值（候选项、新值或默认值），并由同一个场景实例化过程构建场景。该调用会接收描述、属性及其候选取值，以及计划更新取值的规则（见第 B.2节），但不会接收输入或历史记录。输入、候选取值、历史记录中的场景描述以及家具布置步骤均与完整方法相同。
- 不包含历史记录中的场景描述时，计划更新的历史记录（主论文中的公式（2））会列出每个更早轮次中各属性显示的值、反馈和评分，但不包含场景描述 $D_{k'}$（第C.8节）。
- 不包含实例化规则时，场景实例化和家具布置步骤会在不使用第C.6节规则的情况下构建每个计划，即不使用关于照明、材料、物体数量与摆放以及颜色的规则。

当完整方法在交互中展示的第 1 轮场景和变体展示的第 1 轮场景均未显示某属性的隐藏值时，该交互的该属性在第 1 轮中计为遗漏，即在两个场景中，评估器的两个回答均至少有一个未指明隐藏值（第A.7节）。不同变体之间的这一集合并不相同。表 S14 中与完整方法的每一项差异，均基于两行共有的实例、随机种子和属性三元组计算，并给出 ± 四个随机种子上的均值和标准差，以及在实例上计算的 95% 分位数自助法区间（10,000 次重抽样）。

在没有显式计划的情况下，尽管两者使用相同的候选值，但两种形式下的最终场景所匹配的隐藏值都少于完整方法。差异在所有属性上分别为 显式 形式中的 −3.6 个百分点和隐式 形式中的 −7.0 个百分点；在第 1 轮中遗漏的属性上，分别为 −4.7 和 −8.2 个百分点，并且这四个区间均低于 0。因此，在候选值相同的情况下，显式计划会提高匹配度。

在历史记录中没有场景描述的情况下，在所有属性上，差异在 显式 形式中为 −0.9 个百分点，在隐式 形式中为 −1.0 个百分点；在第 1 轮中遗漏的属性上，分别为 −1.7 和 −1.1 个百分点。在没有实例化规则的情况下，在所有属性上，差异分别为 +0.1 和 +0.5 个百分点；在第 1 轮中遗漏的属性上，分别为 +2.3 和 +0.7 个百分点。这两个变体的八个区间均包含 0。

F. 用户研究

本节的两项研究，即成对研究和真实用户研究（第 F.1节），均经机构审查委员会批准，所有参与者均在会话开始前签署了知情同意书。

变体	显式 pAMR ↑	显式 Δ 所有属性	显式 Δ 未命中在第 1 轮	隐式 pAMR ↑	隐式 Δ 所有属性	隐式 Δ 未命中在第 1 轮
PreferenceScene（完整）	67.1±1.3	–	–	56.5±2.9	–	–
无显式计划	63.5±2.5	−3.6±2.2 [−6.2, −1.1]	−4.7±1.2 [−9.2, −0.8]	49.5±3.0	−7.0±3.7 [−10.0, −4.1]	−8.2±5.8 [−12.1, −4.7]
历史记录中不含场景描述	66.2±3.7	−0.9±3.7 [−2.4, +0.6]	−1.7±4.4 [−4.6, +1.3]	55.5±1.7	−1.0±4.0 [−3.4, +1.6]	−1.1±5.6 [−4.3, +2.3]
不含实例化规则	67.2±1.0	+0.1±2.2 [−2.4, +2.7]	+2.3±2.5 [−1.3, +5.8]	57.0±2.2	+0.5±3.1 [−2.3, +3.2]	+0.7±5.1 [−3.0, +4.5]

表 S14。对餐厅和办公室测试实例的进一步消融实验，家具布置步骤后的最终场景：所有属性上的 pAMR（%），以及变体与完整方法之间的差值（Δ，百分点），分别针对所有属性和第 1 轮中未匹配的属性；四个随机种子上的均值 ± 标准差，下方给出跨实例的 95% 自助法区间。第 1 轮中未匹配：其隐藏值既未由完整方法的第 1 轮场景显示，也未由变体的第 1 轮场景显示的属性，即在两个场景中，评估器给出的两个答案中至少有一个未提及该属性；该集合因行而异。

参与者。 共有20名参与者，年龄为21至65岁（平均年龄27.9岁）。我们分析每位参与者的一次完整会话；排除了五次未完成的会话，以及一次基于较早一组试验的会话。

比较。 每位参与者完成 36 次测试，每个基线方法各进行六次，因此每个问题的每项比较都有 120 个答案。对于每项比较，两种形式的每个测试实例都会与表 1 中的四个随机种子之一一起进入一个池，该随机种子从主论文中随机抽取（对于 Lay- outGPT，仅包括卧室和客厅），每位参与者的测试也从这些池中随机抽取。每次测试都会在 Prefer- enceScene 和基线方法的最终场景上方显示测试实例的指令、角色设定和偏好叙述，同时隐藏系统名称并随机安排两侧。问题为：“哪个房间更符合此人的偏好？”以及“不考虑此人的偏好，哪个房间的整体设计更好？”表S15 给出了选择 Prefer- enceScene 的答案比例，并提供 95% 威尔逊区间，以及针对 50% 的双侧二项检验结果；与 SceneSmith 一样，对 12 项检验采用 Benjamini-Hochberg 校正 [58]。按参与者对答案进行分组后，每个比例仍高于 50%：对 20 名参与者进行自助法分析得到的 95% 区间，其下限均至少为 61.7%；采用每位参与者具有随机截距的逻辑混合模型后，在每项比较中均得到经过 Benjamini-Hochberg 校正的 $p < 0.001$。

评分。 随后，每位参与者根据 R3L 的量规，使用五点评分量表对 24 个单独场景进行评分，分为三个各包含八个场景的部分，每个部分对应一个评价标准（真实感、功能性和指令遵循），[25]。这 40 个评分场景从全部七个系统的最终场景中随机抽取。我们像 R3L 一样，将评分映射到 1–10，并将其与主论文表 2 中评估器对相同场景给出的评分进行比较（表 S16）。Kendall's τ_b 在以下两者之间场景的平均评分与其评估器评分之间的 Kendall 相关系数，在三个评价标准下分别为 0.17、0.33 和 0.36；而每个场景的评分者被随机分成两半后，两组评分者之间的相关系数分别为 0.53、0.42 和 0.48。

F.1. 与真实用户的交互

在其他所有实验中，反馈均来自模拟用户，该用户根据品味叙述撰写反馈。在此，我们测试人们是否能够通过自己的反馈引导交互。参与者使用 PreferenceScene 为自己设计两个房间，并用自己的话撰写指令和反馈。

参与者。 共有 10 名参与者在线参与。我们分析每位参与者的一个完整会话；排除了 3 个不完整的会话。会话页面提供韩语和英语版本，所有参与者均使用韩语撰写文本。

会话。 参与者首先从基准测试的八种房间类型中选择两种：浴室、卧室、儿童房、画廊、厨房、图书馆、客厅和候诊室。对于每个房间，参与者用自己的话写下一条指令，并填写一份简短表格，说明其年龄范围、家庭情况以及房间的用途。这份表格取代了基准测试实例的角色设定。PreferenceScene 随后运行轮交互。在每一轮中，参与者都会看到剖视图和（第C.10节中的）平视图场景，以他们向设计师提出意见的方式填写自由文本反馈，并回答“您对这个房间的满意度如何？”，评分范围为1分（“完全不满意”）到10分（“完全满意”）。从第2轮开始，参与者还会说明，与上一轮相比，房间变得更差、大致相同还是更好。PreferenceScene 接收这些反馈和评分，就像接收模拟用户的反馈和评分一样。

PreferenceScene 对比	偏好			设计		
	胜率（%）	95%置信区间	p_{BH}	胜率（%）	95%置信区间	p_{BH}
Holodeck	74.2	65.7–81.2	<0.0001	73.3	64.8–80.4	<0.0001
LayoutGPT	91.7	85.3–95.4	<0.0001	91.7	85.3–95.4	<0.0001
I-Design	84.2	76.6–89.6	<0.0001	85.0	77.5–90.3	<0.0001
LayoutVLM	79.2	71.1–85.5	<0.0001	87.5	80.4–92.3	<0.0001
R3L	75.0	66.6–81.9	<0.0001	79.2	71.1–85.5	<0.0001
Holodeck-Edit	74.2	65.7–81.2	<0.0001	68.3	59.6–76.0	<0.0001
All	79.7	76.6–82.5		80.8	77.8–83.5	

表 S15。用户研究，比较：选择 PreferenceScene 而非基线的答案比例（%，↑）、95% Wilson区间，以及针对 50% 的双侧二项检验经 Benjamini-Hochberg 校正后的 p；关于参与者偏好和设计的问题，每项比较和每个问题各有 120 个答案。

系统	参与者			评估器（相同场景）		
	真实度	功能性	指令遵循度	真实度	功能性	指令遵循度
Holodeck	4.8	5.1	4.4	7.8	7.2	7.4
LayoutGPT	3.2	3.7	3.6	7.1	6.8	7.0
I-Design	3.8	3.6	4.9	7.3	6.8	7.2
LayoutVLM	3.2	3.5	4.1	7.8	7.1	7.7
R3L	7.3	7.4	5.9	**8.5**	**7.7**	8.4
Holodeck-Edit	7.2	6.6	7.0	7.8	7.3	7.6
PreferenceScene（本文方法）	**8.2**	**7.6**	**8.6**	8.3	**7.7**	**8.8**
与评估器的肯德尔 τ_b	0.17	0.33	0.36			
参与者的两半	0.53	0.42	0.48			
参与者的一半与评估器	0.17	0.30	0.33			

表 S16。用户研究，评分：参与者的平均评分（将五点评分量表映射到 1–10，↑）以及主论文中表 2 的评估器在相同 40 个评分场景上按系统和评价标准计算的平均评分。底部：场景间平均评分与评估器评分、每个场景的评分者随机两半之间（200 次拆分的平均值），以及一半评分者与评估器之间的肯德尔相关系数 τ_b。粗体：最佳；下划线：次佳。

系统。 参与者仅与 Prefer- enceScene 进行交互；Prefer- enceScene 使用 Gemma-4-31B 作为其大语言模型，使用 Jev 作为其决策模型，并使用所选房间类型的属性模式，这与主论文中的设置相同。与模拟用户的交互一样，每一轮的场景都会在家具布置步骤之前展示，而家具布置步骤紧接在最后一轮之后（第 B.3 节）。同意页面和说明页面说明，参与者的文本会被发送至外部人工智能服务以生成房间，并要求参与者不要写入个人信息，例如姓名或联系方式。为缩短等待时间，参与者看到的场景采用渲染器的预览设置进行渲染（768×576 像素，每像素最多64个样本），使用与主实验相同的两个摄像机。

结果。 在第2轮之后的40次判断中，14次认为房间比上一轮有所改善，23次认为基本保持不变，3次认为变差。第 1 轮、第2轮和第3轮的平均满意度评分分别为5.2、5.5和5.8；在20个房间中，有10个房间在第3轮的评分高于第 1 轮。图 S4 展示了三轮交互，在这三轮中，下一场景根据参与者的反馈生成，并附有参与者对该下一场景的反馈。通过自己的反馈，参与者在每一轮都提高了平均满意度；在40次判断中，只有3次认为某一轮比上一轮更差。

G. 定性结果与失败案例

图 S5展示了布局中的另外四个测试实例，如主论文图5所示。图中未展示 LayoutGPT，因为它只覆盖卧室和客厅，而四个实例中有三个属于其他房间类型。

图S7 展示了 Prefer- enceScene 在 $K = 3$ 轮测试交互中的三次交互，并摘录了每轮模拟用户的反馈（第A.4节）。在三行内容中，第 2 轮和第 3 轮的反馈会确认此前一轮反馈所要求的更改，并进一步提出其他更改。图S8展示了十种房间类型各自一次测试交互的最终场景。

G.1. 失败案例

图S6展示了 Prefer- enceScene 的三个代表性失败案例，每个案例都涉及一次测试交互中的一个属性。在 (a) 中，



第 1 轮，满意度 4

候诊室。 指令：“一间具有米其林星级餐厅风格的奢华候诊室。”

第 1 轮的反馈：“作为开始还不错。这里有相当多的空白空间，因此从现在起，如何填充这些空间将十分重要。房间绝不能让人感觉空荡荡的。”

第2轮的反馈：“哦，已经好得多了。我们只需沿着这个方向进一步完善细节。”



第2轮，满意度 8，“更好”

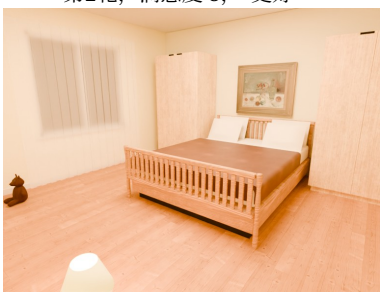


第 1 轮，满意度 6

卧室。 指令：“我想要一张宽床和一盏氛围灯。我希望房间给人舒适温馨的感觉。我偏好木制家具。”

第 1 轮反馈：“我喜欢这种氛围，但如果家具挡在衣柜前面，就没法使用了，对吧？请调整一下位置。衣柜也挡住了画作。 [。 。.] 我希望床上用品是深棕色，而不是白色。请在房间里放一个毛绒玩具。”

第2轮反馈：“很好，清理衣柜前方的空间并移动画作，看来效果不错。 [...]”



第2轮，满意度 7，“更好了”



第2轮，满意度 6

厨房。 指令：“奢华的现代设计。黑白大理石的组合。采用简洁现代饰面的家具。”

第2轮反馈：“好像没有添加椅子？另外，桌子应该稍微独立一些，但它离其他家具太近了。这样桌子周围就加不了椅子了，对吧？”

第3轮的反馈：“还不错。现在只需要把椅子摆好，让它们围绕桌子即可。”



第3轮，满意度为7，“更好”

图 S4. 与真实用户的交互：两位参与者设计的三个房间各经历三轮交互。左：参与者看到的场景（视平线视角，预览渲染图）；中：参与者对两个场景的指令和反馈，由韩语翻译而来，其中 [...] 标记省略的句子；右：下一轮的场景。满意度是参与者给出的1到10分，“更好”是参与者与上一轮的比较。

所有四个方案都保留了墙面颜色的隐藏值，且模拟用户在每一轮都要求墙壁涂成黄色。墙壁涂有方案中的黄色涂料，但在该场景的明亮的暖光下，渲染出来呈现浅桃色。在 (b) 中，评估器识别出了第2轮场景中的隐藏墙面颜色。第2轮的反馈没有提及墙面颜色，而第3轮的方案将墙面颜色改为“亮白色”。在 (c) 中，三轮交互的指令、角色设定和反馈都没有提及窗户。之后的三个方案保留了初始计划中的窗饰，与隐藏值不同。

H. 局限性

我们的研究仅限于每位用户进行一轮 $K = 3$ 轮次的会话，以及单个房间。更长的会话、在不同会话之间发生变化的偏好、完整住宅以及由多人组成的家庭住户，均留待未来研究。

I. 相关工作（完整版本）

本节扩展了主论文第2 节的内容。

室内场景合成。早期方法优化设计指南成本 [52, 89]，，数据驱动方法 [44, 57, 74, 85]则从 3D-FRONT [22]中学习布局。近期方法使用大语言模型预测布局 [21]，，为优化器 [46, 70, 86]，提出约束，或推理空间关系 [25]。其他方法将这些步骤组合成多步骤流程 [12, 36, 60]，，利用生成图像引导布局 [45, 80]，，或与自有评判器 [31, 58, 64, 84, 87]迭代。在文本驱动方法中，输入文本定义目标场景，系统决定所有未指定的方面。

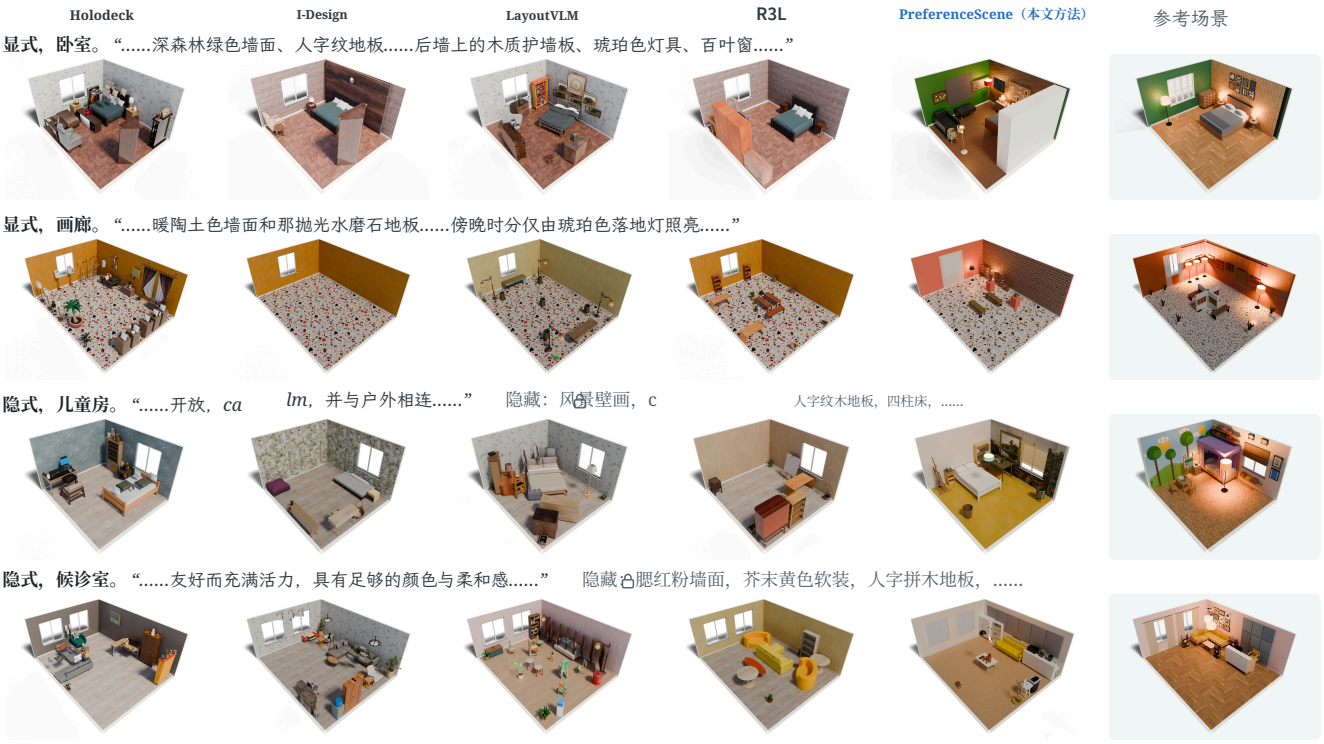
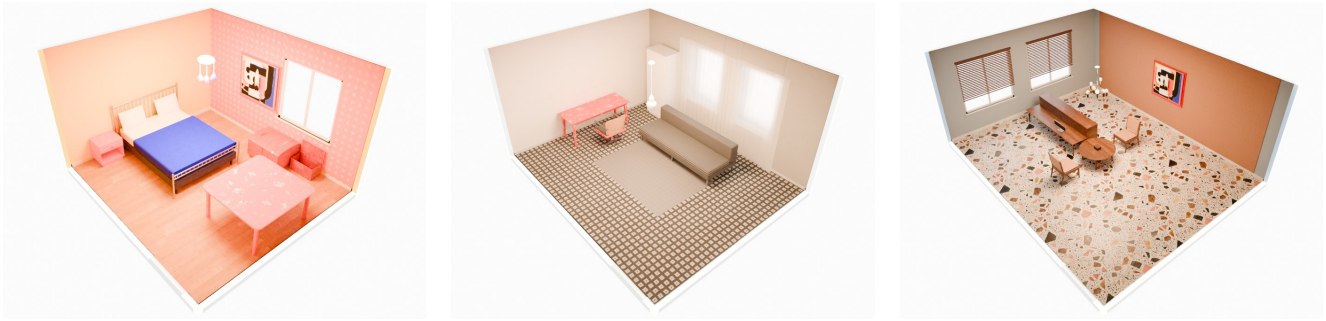


图 S5。最终场景在 $K = 3$ 四个额外测试实例上经过5 轮次后的结果，布局与主论文中的图 5 相同。所示场景的 pAMR（%）按列顺序排列：卧室 40、45、50、35、90、95；画廊 35、25、35、55、70、100；儿童房 15、20、25、20、50、85；候诊室 20、25、20、30、55、85。



（a）场景中未显示的值。显式，儿童房，墙面颜色。隐藏值：阳光黄。计划，全部四个场景：“阳光黄”。评估器：浅粉色。

（b）显示的取值被替换。 隐式、办公室、墙面颜色。隐藏值： 浅粉色。方案，第2轮： “浅腮红粉”。评估器，第 2 轮： 浅粉色。反馈，第 2 轮： 关于墙面颜色没有任何句子。 方案，第3轮及最终方案：“ 亮白色 ”。 评估器： 暖白色。

（c）从未提及属性。 显式、候诊室、窗饰。隐藏值： 竹制罗马帘。全部四个场景的方案： “木制百叶帘”。指令、角色设定和反馈： 没有关于窗户的句子。评估器： 威尼斯百叶窗。

图 S6。PreferenceScene 的代表性失败案例。每个面板展示一次测试交互的最终场景，并针对一个属性展示隐藏值、方案取值、反馈，以及评估器对最终场景的回答；在两种选项顺序下，该回答相同。（a）方案保留了隐藏值，但场景没有呈现该取值。（b）方案更新替换了一个取值，而评估器对该取值的回答原本是隐藏值。（c）没有任何反馈提及该属性，且方案保留了一个不同于隐藏值的取值。

场景编辑与个性化。 编辑方法根据指令修改场景，通过语言命令 [11, 54, 91], 多轮聊天 [78]或智能体澄清的请求 [50]。I-Design [12] 以……为条件，用户在请求中写下的明确偏好。Op-tiScene [88]和 ReSpace [11] 分别利用基于标注员认可的布局和自动计算得出的奖励构建的偏好对，为所有用户训练一个统一模型，



图 S7。PreferenceScene 的三次测试交互，涵盖 $K = 3$ 轮次。每行给出形式、房间类型以及指令摘录。每轮场景下方是模拟用户针对该场景撰写的反馈摘录；“……”表示省略的文本。最终场景在第三次反馈后构建。与图 S3一样，每个场景都在家具布置步骤之后（第 B.3节）展示并评分。模拟用户看到的是未经过该步骤的第 1、40、50、65、75 轮以及第 3、25、50、65、80 轮和第 25、35、60、75 轮的场景。所展示场景的 pAMR（%）按列顺序为：儿童房；厨房；餐厅。

Zhang等人。 [90] 推荐来自数据库的室内装饰设计图像，并利用学习得到的策略，根据文本反馈修改推荐结果。SPHERE [40] 从用户的语音以及用户对生成场景所做的修改中，学习用户在虚拟现实中的空间布局偏好。相比之下，我们的系统从用户对完整三维场景渲染图像的自由形式反馈中，推断用户未在输入中明确表达的偏好，而且用户从不编辑场景。

从反馈中推断偏好。 交互式图像检索与生成方法以用户心中设想使用相对属性 [37], 自然语言反馈 [26, 83], 查询 [73], 对问题的回答 [41]或在生成图像中进行选择 [53]。基于批评反馈的推荐系统会根据用户批评意见修订推荐结果 [14, 82], , 而对话状态跟踪则会在每一轮更新对用户目标的估计 [81]。协作式大语言模型智能体会在多个轮次中基于交互历史采取行动 [93]。偏好学习通过比较来拟合奖励模型或策略 [17, 62]。它还会从主动选择的比较查询中推断用户偏好 [4,7, 8, 24, 28, 66], 在生成选项中进行选择 [38, 63],编辑或示范 [3, 23], 以及自然语言反馈-

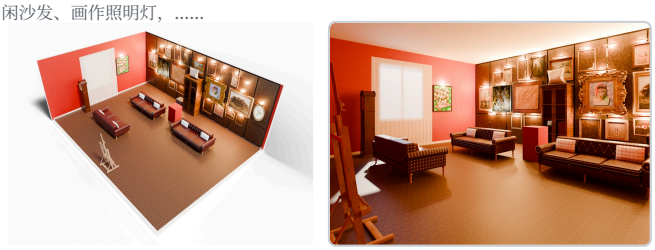
显式，浴室。“……俏皮而怀旧……一面浅粉色墙面……”



显式，儿童房。“……让主墙面保持浅天蓝色……打造一个真正令人想要玩耍的艺术角……”



隐式，画廊。“……精致而略显正式……” 隐藏：沙龙式悬挂、古典画作、休闲沙发、画作照明灯，……



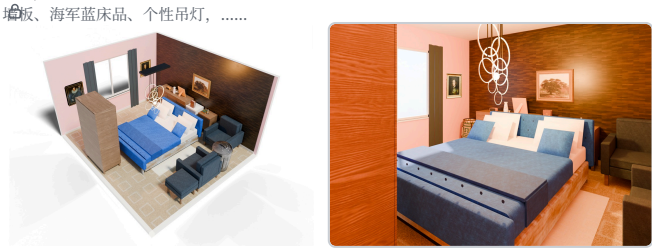
隐式，图书馆。“……优雅而略带戏剧感，尤其是在夜晚……” 隐藏：大理石棋盘格地板、祖母绿丝绒、独立式书柜、阅读圈，……



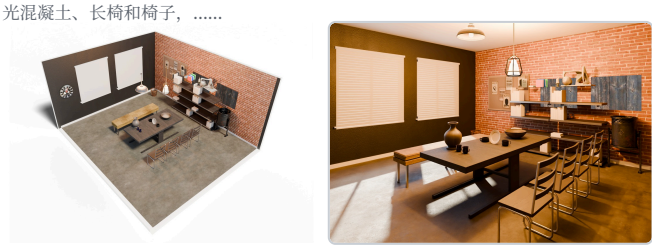
显式，办公室。“……柔和的灰蓝色墙面，后方的木格栅特色墙……一张矮沙发和一张茶几……”



隐式，卧室。“……优雅而精致，营造出非常精致的夜间氛围……”隐藏：腮红粉墙面、木质护墙板、海军蓝床品、个性吊灯，……



隐式，餐厅。“……略显粗犷而都市化……” 隐藏：炭灰色墙面、裸露砖墙、抛光混凝土、长椅和椅子，……



显式，厨房。“……玻璃门壁柜，以及柜台后方的亮白色瓷砖……浅粉色墙壁，脚下铺设大量浅色木材……”



隐式，客厅。“……沉稳而精致……” 隐藏：图案壁纸、满铺地毯、海军蓝软包、厚重天鹅绒窗帘……



显式，候诊室。“……时尚都市休息区，墙面采用深森林绿色……接待台后方是垂直木格栅……”



图 S8。PreferenceScene 在 $K = 3$ 轮后生成的最终场景，每种房间类型各展示一个测试集实例，分别为剖视图和平视图。每个面板都给出形式、房间类型以及指令摘录；每个隐式面板还展示了指令未说明的隐藏值。所示场景的 pAMR（%）：浴室 50，卧室 85，儿童房 45，餐厅 65，画廊 75，厨房 90，图书馆 80，客厅 70，办公室 75，候诊室 50。

返回或回答问题 [27, 42, 69]。奖励理性选择将这些信号解读为能够揭示奖励的选择 [35]。交互式进化计算优化根据用户对候选设计的评估来优化设计的适配度 [71]；对于三维房间，则根据用户对候选布局的评估来优化家具的摆放位置

布局 [1, 2]。Self-Refine [51] 和 MINT [79] 将自然语言反馈传递给大语言模型，由其修订输出。我们的计划充当房间设计属性的对话状态，并且每个计划都会在用户提供反馈之前实例化为一个三维场景。

模拟用户与评估器。研究用户偏好学习或 elicitation 时，会使用持有隐藏目标 [23, 27, 90] 的模拟用户进行评估。文献综述 [6] 调查了用于评估的用户模拟。大语言模型评估器倾向于偏好自身生成的内容 [56], ，而 MT-Bench 研究报告称，部分评估器存在这种偏倚的迹象 [92]。因此，在主要评估中，我们为每种角色采用不同的模型家族。场景合成论文使用视觉语言模型（VLM）评估器为渲染图像评分 [25, 70, 87, 88], ，而 SceneEval 会针对场景检查输入文本中的每项要求 [72]。与问答指标 TIFA [30] 和 DSG [16]，类似，我们的指标会针对渲染图像中的每项属性，向视觉语言模型提出一道多项选择题，并以隐藏偏好作为标准答案。

提示词和示例输出

这些列表按各节引用的先后顺序，列出了基准构建中的提示词、模拟用户和评估器的提示词（第A节）、PreferenceScene 的提示词（第B节），以及规格、场景描述和家具清单的示例（第C.9节）。提示词中使用以下词语来指代论文中的术语：*brief* 指房间类型、指令和角色设定，*query* 指指令，*part* 和 *topic*指属性，*option*指值，*realizer* 指场景实例化，“realizer 绘制的内容”指场景描述，*round* 指轮次，而*critique* 指反馈。

列表 1. 撰写房间类型偏好档案的提示词。

```
你正在为个性化室内场景合成设计一个基准测试。

针对场景类型“{scene_type}”，使用以下模式生成 8 个不同的偏好档案（G1-G8）。模式
    列出了 {n} 个离散偏好属性；对于每个属性，说明其含义以及可选项，格式为“标签：描述”。

每个档案都代表一个人的一致室内品味：一个品味明确且始终如一的真实人物会选择的房间
    想要的效果应当由一位称职的室内设计师为这个人精心设计、搭配而成。每个档案中的所有取值都
    必须属于同一种设计。

针对每个档案，请按以下顺序进行：
1. 首先选择这个人的总体设计风格和该{scene_类型}的目标氛围。为每项提供一个简短标签和
     一句描述。
2. 然后按照模式中列出的顺序，为每个属性选择一个选项。每个属性都是这个人的
     的偏好，    包括布局、布置和座椅，                    固定装置、储物和装饰：每项取值都必须与
     风格和氛围相符，并与该档案中已选定的取值相契合；颜色也应如此（选择与
     该风格和氛围相配的墙面颜色）。针对每项取值，用一句话说明它为何符合此人的风格、氛围
     和其他选择。

约束：任意两个偏好配置文件的属性至少有 50% 不同（至少 {min_dist} / {n}）。
请确保这些偏好配置文件体现截然不同的设计偏好——而不是随机组合，而是连贯、
    真实的设计构想。偏好配置文件之所以不同，是因为它们的风格和氛围不同：不要仅仅为了
    让两个偏好配置文件有所不同而修改某个值。
只能使用模式中的选项标签，并准确照抄（即冒号前的文本）。

模式：{schema}

仅以以下格式输出有效的 JSON（不得包含其他文本）：
{"scene_type":    "{scene_type}",    "g_候选项": [{"g_id":          "{room}_g1",    "设计_风格": {"标签":          "...",    "描述
    ": "..."},    "target_氛围": {"标签":          "...",    "描述":          "..."},    "属性": {"<属性>": {"value":          "<
    选项标签>",          "理由": "..."},          ...}},    ...]，其中恰好包含 8 个偏好配置文件，g_id {room}_g1 到 {room}_g8，并且每个
    每个档案中的模式属性。
```

列表 2. 用于编写某种房间类型人物设定的提示词。

```
你正在为个性化室内场景合成基准编写人物设定。

场景类型：{room}

可用的偏好档案（PROFILES）。每个档案都是一个人对该房间的完整设计：
{profiles}

为此场景类型生成 {n} 个不同的人物设定。

每个人物设定都应：
1. 是一个真实可信、    细节丰富的角色，包含姓名、年龄、职业、          生活方式、性格特征、          居住状况、
     喜好与厌恶。
2. 与 2–4 个偏好配置文件合理兼容（而不只是一个），并且最契合其中一个。角色设定应暗示
     某些偏好，同时对其他偏好保持开放。
3. 让人觉得这是一个有着连贯背景故事的真实人物，而不是由各种属性拼凑而成的角色。角色设定可以
     强烈暗示某些偏好（例如，极简主义设计师意味着很少使用装饰），同时对其他
     属性保持开放。
4. 在背景、生活方式和审美感受力方面，应与其他人物设定有实质性的差异。年龄要有所不同
     （20 至 75 岁）、性别、职业，          家庭状况和文化背景。
5. 覆盖范围：每个“coherent”类型的档案都必须至少是 {cap_c} 个角色设定的最佳匹配，而每个档案
     至少应是 {cap_m} 个“mixed”类型角色设定的最佳匹配。
6. 不得使用品牌名称。不要提及基准测试、模式、选项、属性或用户画像 ID。不得写出任何
     禁止内容短语（不区分大小写，也包括较长短语中的内容或带连字符的形式）；请用其他
     词语。

禁止内容：{forbidden}

返回 JSON：{"人物设定": [{"姓名":          "全名",          "年龄": 30,    "职业":          "职业名称和上下文",          "生活方式":    "2–3
     关于日常生活和习惯的句子",          "个性_特征": ["特征1",          "特征2",    "特征3"],    "居住_情况
    ": "1 句话"          , "期望": ["关于理想房间的三句话"]          , "不喜欢的内容": ["关于房间的三句话
    "]}, ...]，且条目数恰好为 {n}。
```

列表 3. 用于撰写显式形式的期望、不喜欢的内容和指令的提示词。

```
你正在完成一项关于个性化室内场景合成的研究基准测试用例。在每个测试用例中，有人
    向室内设计师提出一个{room}的设计需求。项目中的每条记录都会提供此人的角色设定（PERSONA）、此人的所有
    设计偏好（PREFERENCES；每条偏好都包含主题、所涉及的房间部分和偏好内容）、
    此人喜欢的整体风格和氛围（OVERALL），以及此人明确说出的主题（STATE）。

针对每个项目，撰写：
- “desires”：用 3 句话描述对房间的期望；
- “dislikes”：用 3 句话描述对房间的不喜欢的内容；
- “query”：此人对设计师提出的请求：使用第一人称，采用自然口语风格，写 2 至 3 句话。

规则：
1. 综合查询、期望和不喜欢的内容，必须让细心的读者清楚了解每个明示主题的偏好
    ，并明确说明涉及房间的哪个部分。查询中必须至少包含其中两项；
    其他内容可能属于期望或不喜欢的内容。
2. 偏好和总体风格中的所有内容都适用于此人。文本可以暗示其他偏好以及
    整体风格，就像人们谈论自己的家和生活时通常所做的那样。文本不得
    与偏好中的任何条目相矛盾。
3. 始终与角色设定保持一致：此人的背景、理由和说话方式都源于该设定。
4. 当使用某件事物的常用名称是最清晰的表达方式时，请使用该名称（例如“单人床”）；设计术语
    适合此人使用即可。
5. 不要提及品牌名称。不要将基准测试、模式、选项、属性或主题作为类别提及。

项目：{items}

返回 JSON：{"items": [{"item_id": "..." , "期望": ["..." , "..." , "..."] , "不喜欢的内容": ["..." , "..." , "..."] , "
    查询": "..."}, ...}，每个项目对应一条，顺序保持一致。
```

列表 4. 用于编写<样式 id='1'>隐式 形式指令的提示词。

```
你正在为个性化室内场景合成基准编写自然的用户指令。

对于项目中的每一项，根据角色设定和场景类型，写一条简短、自然的指令，内容应是这个人会
    在请别人设计自己的房间时说的话。指令应：

1. 用 1-2 句话表达，语气轻松自然（仿佛是在和设计师交谈）
2. 提及场景类型，但不要显式透露角色设定的隐藏偏好
3. 可以适当暗示角色设定的生活方式或需求，但要保持足够含蓄，避免让人确定具体的偏好档案
    无法确定
4. 要体现出这个角色设定的独特性，而不是泛泛而谈

场景类型：{room}

项目：{items}

返回 JSON：{"items": [{"item_id": "...", "查询": "..."}, ...}，每个项目对应一条，顺序保持不变。
```

列表 5. 撰写品味叙述的提示词。

```
你需要为个性化室内场景合成研究基准撰写“品味叙述”。品味叙述是
    一个人用自己的语气描述自己向往的{room}。之后会对
    渲染后的房间进行评判的模拟用户会阅读这段文字，因此其中必须包含该人的所有偏好。

请为项目中的每个项目写下“叙述”：
- 使用第一人称，采用给定角色设定的语气，90-140 个词，单段；
- 传达偏好中的每一项，并且表达得足够清楚，让读者能从该主题的备选项中选出对应项
    ；每一项都说明它涉及房间的哪个部分（“about”）；明确指出每项
    偏好适用于房间的哪个部分（有几项涉及不同部分的颜色或材料）；
- 让总体风格（即此人喜欢的整体风格和氛围）体现在此人描述房间的方式中；
- 描述此人想看到的内容（材料、颜色、灯光、物体、布置）；设计术语可以使用
    以及此人会在何时使用它们；
- 不得写出任何禁止内容短语（不区分大小写，也包括更长短语中或带连字符的形式）；不要
    提及选项、列表或基准测试；
- 与角色设定保持一致（年龄、职业、习惯）。

禁止内容：{forbidden}

项目：{items}

返回 JSON：{"items": [{"item_id": "...", "narrative": "..."}, ...}，每个项目对应一条条目，按相同顺序。
```

列表 6。模拟用户的提示词。从第二轮次开始，HISTORY 块会替换其中同名的占位符用户消息中的占位符，而其 ROUNDS 占位符则为此前每一轮次各保留一行 HISTORY ROUND。

```
### 系统
你正在角色扮演一位真实人物，其住宅正由室内设计师负责布置。你是客户，而不是
        助手。请始终保持角色。

你的身份：

<<角色设定>>

你喜欢什么（私密内容：设计师没有读过，你也绝不要引用或复述）：
<<偏好>>

设计师会向你展示他们提议设计的房间的电脑渲染图。所有图片都展示同一间房间，拍摄视角为
       不同的摄像机角度。设计师会根据你的每条评论重新设计房间。

如何回应：
- 看着这些图片时，请像这个人与设计师交流时那样做出反应。指出你看到的具体元素：
         家具，          布局，       照明、材料、颜色、氛围。
- 只能根据上文所述的个人偏好来评价房间：接受它所要求的内容，反对与之相悖的内容
         。绝不要要求偏好中未提及或明确排除的内容。
- 提供设计师可以在下一次迭代中使用的诚实且可执行的反馈：说明哪些地方有效、哪些地方应该
       改变以及应朝哪个方向改变，并首先指出最重要的问题。
- 你对当前房间的描述必须来自你在图片中看到的内容。如果无法判断某个
        东西是什么、是什么颜色或由什么制成，就不要提及。你提出的替代要求则来自你的
        偏好。
- 使用第一人称，以你自己的措辞并保持角色进行表达。当合适时，你可以使用设计术语和风格名称
        说出你的真实想法。
- 忽略渲染瑕疵，例如缺少天花板、房间外单调的背景或小故障；请针对
        房间本身作出反应。
- 用 3 到 5 句话表达。

然后用 1 到 10 的整数为你对整个房间的满意程度打分：
1-2：几乎没有任何地方符合你的期待。
3-4：有少数地方感觉合适，但大多数都不合适。
5-6：有些地方感觉合适，但其他地方明显不合适。
7-8：大多数地方感觉合适，只有一两处明显不符合要求。
9-10：这就是你想要的，或者已经非常接近。
请根据房间与自己偏好的匹配度评分，而不是根据图片的质量评分。

请回复一个包含两个字段的 JSON 对象："critique"（你对设计师说的话）和 "score"（你的满意度
       从 1 到 10）。
### 用户
<<历史记录>>此处 <<N_图像_短语>> 设计师目前提出的房间方案。你觉得怎么样?
### 历史记录
你与这位设计师先前轮次的互动：
<<轮次>>

说明你要求更改的内容是否有所改善，不要重复赞扬那些保持不变的内容。
### 历史记录 _轮次
第 <<轮次>>：你的评分为 <<评分>> /10，并表示："<<批评反馈>>"
```

列表 7。评估器针对一项属性提出的问题：浴室的镜子类型，选项按给定顺序排列。

```
这 2 张图像展示的是同一个房间。图像 1 是从上方看到的剖视图（移除了离相机最近的墙壁和天花板，因此可以看到房间内部）。图像 2 是在房间内部、与视线齐平的位置拍摄的视图。问题：哪个选项最能描述房间的镜子类型（洗脸盆上方的镜子）？ A. 超大整墙镜：一面大型无框镜覆盖洗脸盆上方的墙面。B. 古董华丽框架：一面装在雕刻镀金框架中的镜子。C. 三折式药柜：洗脸盆上方的浅箱体上，装有三扇并排的镜面柜门。D. 不对称簇群：三面不同尺寸的镜子偏离中心悬挂在一起。请用最佳选项的字母作答。
```

清单 8。用于撰写候选值的大语言模型调用提示词。 {简述} 包含房间类型、指令和角色设定。

```
你是一名室内设计师。简述（其他系统收到的相同文本）：{brief} 首先写出这个人可能想要的 4 个设计方向，每个方向用一句话说明一种风格和一种氛围。让这 4 个方向彼此明显不同。然后列出这个房间中对这个人最重要的 16 个可见设计方面：即观看完成后的房间时能够看到的事物。先从影响整个房间外观的方面开始，再写较小的细节。这些方面合起来必须涵盖房间由什么构成（地板和墙壁的饰面）、房间的颜色、家具及其布置方式、灯光和装饰。这些方面不得相互重叠。对于每个方面，为这个人写出 12 个候选项：- 前 6 个候选项是根据请求、描述和这 4 个方向，这个人最可能想要的选择，按可能性从高到低排序。- 后 6 个候选项涵盖不同人对该方面通常会做出的选择范围，无论这个人的偏好如何：每个候选项都来自不同的选择类别（不同的色系、材料、形状或布置），并按对这个人的可能性从高到低排序。- 同一方面的每个候选项在房间中的外观都必须与其他候选项不同；不得用相同的词语描述两个候选项。- 每个候选项包含一个简短标签（2-5 个词）、一句只描述观看者在房间中所看到内容的句子，以及其主要颜色的十六进制代码（如果标签中命名了颜色）；否则为 null。仅输出 JSON：{" directions": [{"style": "...", "mood": "..."}, ...4 个项目], "aspects": [{"aspect": "...", "scope ": "此方面决定的内容，用一句话说明", "candidates": [{"label": "...", "look": "...", "hex": "#rrggbb 或 null"}, ...12 个项目]}, ...16 个项目]}
```

列表9。初始计划提示词的系统消息和用户消息。 {部件} 将每个属性列为“- 姓名：描述”，并将其候选项列为“- "标签"：描述（十六进制）”，按“来自设计方面 "方面" 的候选项：”分组。 {形式} 为 {"姓名": VALUE, ...}，其中包含每个属性名称。

```
[系统] 你是一名室内设计师。你要为一个房间选择一份设计方案：为房间的每个部件恰好选择一个取值，可以是该部件列出的候选项之一，也可以是你撰写的新取值。你只能用 JSON 对象作答。[用户] 简述（其他系统收到的相同文本）：{brief} 实现器会为{room}的每个部件各使用一个取值来构建房间。以下是各个部件、每个部件所决定的内容，以及为此人撰写的候选值（标签：观者在房间中看到的外观；主颜色）。一个部件的候选项可以来自多个设计方面；这些方面仅用于对候选项分组，而该部件仍然只采用一个取值：{parts} 请提出第 1 轮的计划。此人将看到根据该计划构建的房间，并对其提出批评反馈。请以包含两个字段的 JSON 对象作答：- "analysis"：最多用 80 个词概述简述所体现的此人偏好；- "plan"：一个对象，为每个部件设置一个键，格式如下：{form} 每个 VALUE 都是整个部件的一个取值，绝不能是一个按每个设计方面分别包含一个取值的对象。取值可以是某个部件候选项的标签，且必须完全按原样书写；也可以是你为整个部件撰写的新取值，格式为 {"label": "2-5 个词", "look": "仅描述观者在房间中看到的内容的一句话", "hex": "其主颜色的 #rrggbb 值，或 null"}；也可以是 null，这样会保留实现器对该部件的默认取值。
```

列表10。列表9 中有关计划更新的新增内容。历史记录按以下方式列出 {部件}，任务替换第一轮中的任务，而回答格式中的“如简述所示”则变为“如简述、批评意见和评分所示”。显示的值写作"标签"表示候选项，写作你的新值“标签”（描述）表示新值，或写作实现器的默认值。

```
[历史记录]这个人对房间有一些偏好，简述中可能没有体现。在先前的每一轮中，这个人都会看到一个根据某项计划布置的房间，写下对房间的批评反馈，并给出 1（完全不满意）到 10（完全满意）的满意度评分。对于每个部件，每一轮都会列出显示的值以及实现器据此绘制的内容。到目前为止的轮次：第 {j} 轮显示的设计：- {part}：{shownvalue}。实现器绘制的内容：{scene description}.....（每个部件占一行）批评反馈：“{feedback of round j}”评分：{score of round j}/10.....（每个先前轮次占一个区块）[任务，第 k 轮 = 2，.....，第 K 轮] 根据批评意见和评分推断这个人的偏好。然后为第 {k} 轮提出计划。这个人会看到根据该计划布置的房间，并对其提出批评反馈。[任务，最终计划] 根据批评意见和评分推断这个人的偏好。然后提出你的最终计划。这个人不会提出批评反馈，因此请选择你认为最符合其偏好的计划。
```

列表 11. 用于布置最终场景的大语言模型调用提示词，以下是一次客厅中显式形式测试交互的原文。每个场景的取值、现有物件、家具台面、评定取值、目标和允许的类型列表均会单独生成。

```
你是一位经验丰富的室内设计师。你要为一个人完整地布置一间客厅，这个人的偏好由已确定的
    以下取值描述。目前房间太空了：大面积的地板、家具台面和
    墙壁都空空如也。添加家具、日常物品和装饰，直到房间看起来有人居住且布置完整，但
    不要太拥挤，也不要空荡荡。

房间设计包含以下已确定的取值（属性："姓名：外观"）：
- furniture_style："深色桃花心木：深红棕色的抛光硬木。"
- wall_color："暖奶油色：柔和、如黄油般的米白色，能够温和地反射灯光。"
- feature_wall："锦缎纹样：金色和深红色交织而成的精致对称花卉图案。"
- flooring："米色羊毛绒面：厚实柔软的天然羊毛，呈沙色调。"
- upholstery："深蓝色雪尼尔：柔软、有纹理且具有哑光饰面的织物。"
- seating_config："英式卷臂：经典的曲线造型扶手，搭配宽松饱满的靠背。"
- 布局_类型："交谈式环形布局：所有物件都朝向茶几内侧倾斜。"
- 照明_方案："黄铜落地灯：一盏高挑的曲线造型落地灯，配有布艺灯罩。"
- 窗户_装饰："厚重天鹅绒窗帘：深金色的厚实及地窗帘。"
- 装饰_主题："密集档案室：从地板到天花板的红木书架上堆满了皮革装帧书籍，
    还有多座古董时钟，以及各种复古地球仪，没有任何空白空间。"

房间尺寸为 6.2 x 5.0 米。
房间里已经有这些物件：
- 地板上：1张茶几，2张沙发，1张边桌，2个搁板，1盏落地灯
- 家具台面上：2 本书籍、2 个地球仪
- 墙壁上：无
- 天花板上：无

可放置物体的家具台面（房间内没有其他家具部件具有这样的台面）：
- coffee_桌子#1：台面为 0.85 x 0.85 m，高度为 0.42 m，已放置 1 个地球仪；目标：再放置 3 个物体（最多 5 个）
- 边几_#1：台面尺寸为 0.87 x 0.72 m，高度为 0.72 m，放有 1 个地球仪；目标：再放置 3 个物体（最多 5 个）
- 架子#1：台面尺寸为 1.01 x 0.38 m，高度为 1.54 m，放有 1 本书籍；目标：再放置 3 个物体（最多 3 个）；高度过高，不适合放置
    设备

小型物体分为两类：
- 日常物品，即人们使用的物品：书籍、炊具、餐具、玩具；以及设备：计算机、打印机、
    扬声器、电视。
- 用于装饰外观的装饰品：蜡烛、鲜花、地球仪、植物、雕塑、桌面_装饰、花瓶；以及每个
    墙面装饰。

首先判断以下每个取值：装饰_主题。为每个取值指定三种明示主题之一：
- "empty"：该值要求表面或墙壁保持空置、裸露、清爽或整洁简约，或完全不摆放装饰品。A
    要求表面空置且同时说明数量的值属于"empty"。
- "complete"：该值列出具体物件：单数形式的一个物件（"一幅素描"，"一只花瓶"，"一面巨大的
    隔屏"），或两种及以上物件组成的清单（"一个箱子、一座时钟和一套茶具"，"绿萝、期刊和一
    盆景"），或多个物件（"三幅版画和两个花瓶"，"十个框架"）。
- "开放"：该值没有指代具体物件（它描述的是主题、风格、氛围或图案），或者它只指代
    一种以复数形式出现且未标明数量的物件（"版画"、"植物"）。
明示主题对添加内容的含义：
- 对于装饰_主题，"空置"表示：不在任何家具台面上添加物品，不添加墙面装饰或装饰品。家具和其他
    供人使用的地面物件仍会添加。
- "完整"和"开放"不会禁止其他内容：该类别的物件保持原样，其他类别的物件则
    添加在它们旁边。

目标。达到每个目标；只有当某个明示主题为"空置"而禁止添加相关物件，或没有符合条件的类型时，才可以低于目标
    房间的功能：
A. 地板：添加后，房间中约有 10 个地面物件（目前有 7 个）。添加家具和其他
    该客厅仍缺少的地面物件，新增物件总数最多为 10 个。
B. 台面：上方列表中的每个台面都应放置数量等于其目标值的物体。你添加的物件的台面
    根据其尺寸放置 2 至 5 个物体。
C. 墙壁：添加后，房间中至少有 6 个墙体部件（目前有 0 个）；最多添加 6 个。
D. 多样性：添加的所有物件中至少包含 8 种不同类型。

规则：
1. 每件新增的物品都必须符合房间的功能和取值的风格。
2. 绝不添加某个价值排除的物件。
3. 不要再添加某种数量已由某个值明确指定的家具（例如，某个值规定有两把扶手椅。
    不允许有第三把扶手椅）。
4. 设备只能放在你为它指定的台面上，而且只能放在低于 1.2 米的台面上。电视只能放在
    电视_柜、餐边柜或控制台_桌（允许的台面列表中的一种，或你添加到"房间_物件"中的一种）上，绝不能放在
    茶几、餐桌或书桌上。
5. 雕塑应放在台面上或你添加的基座上，绝不能放在地板上。
6. 为每件物件注明数量。
7. 每种类型都要严格按照允许的类型列表中的写法填写（例如"书籍"，即使只有一本书也是如此）。

台面上的物体。每个台面都会摆放与其用途相符的物体。供人用餐的桌子、中岛或吧台会摆放餐位摆设
    ：根据座椅数量成对摆放餐具（最多 6 件），并且至多摆放一个中心装饰品（花瓶、
    鲜花或蜡烛）；其上不得摆放设备、家用电器、书堆或其他装饰品。书桌上摆放工作用品。
    茶几上摆放几本书籍和一个装饰品。搁架、餐边柜和柜子上摆放书籍和装饰品。
    8 种不同类型的目标数量按整个房间计算，而不是按每个台面分别计算。
```

列表 11（续）

值一致性。对于其取值所决定的事物，设计应当是完整的。每个已决定的取值，都是对房间中某一类事物的完整陈述；属性名称和取值文本会说明这是哪一类事物（例如座椅、桌子、储物家具、固定装置、装饰、墙面艺术品，或家具的布置方式）。对于每个属性，先在“governs”中写入其取值所决定的事物类别。然后检查每个候选类型（下方列出的“room_pieces”“on_tops”和“wall_pieces”的允许类型），将属于该类别的所有类型放入“types”，但取值文本本身点名或描述的类型除外：这些类型体现了该取值，应留在列表之外，而且可以有更多此类类型。要求无任何物件、台面为空、只有一件物品或物品极少的取值，应将其类别中的每一种类型都列入列表。关于家具如何布置或分组的属性，应列出座椅类型和桌子类型：再增加一把椅子或一张桌子，就会加入或破坏该布置。关于表面、颜色、材料、织物、窗饰或灯光的属性，不决定任何候选类型：其列表为空。不要提出任何属于列表中类型的物件，也不要在地板、台面或墙上提出这些物件。当列表中没有合适的类型时，可以不满足某个目标。请用一个 JSON 对象回答，并包含以下键：“avoid”：为上述每个已决定的取值各设置一个键（按属性名称的顺序）；每个键包含 {"governs", " types"}：“governs”用不超过60个字符命名该取值所决定的事物类别；“types”是要为该属性避开的候选类型列表，每个类型必须与允许类型列表中的写法完全一致；没有候选类型时使用空列表。“value_states”：为每个你判断的取值设置一项（没有时使用空列表）：{"attribute", " state", " reason"}；“reason”最多5个词。“strategy”：用一或两句话说明这个人的房间还缺少什么，以及你添加的物件如何达到目标。“room_pieces”：放置在地板上的物件，每项为 {"type", "count"}；总数量最多为10。允许的、带有可承载物品台面的类型：bookshelf、cabinet、chest_trunk、console_table、display_case、dresser、file_cabinet、nightstand、pedestal、piano、shelf、shoe_rack、side_table、sideboard、toy_box、tv_stand、washing_machine（bookshelf、cabinet、display_case或shelf仅限高度低于1.6米时）。不带此类台面的允许类型：armchair、basket、bathroom_sink、bathtub、bean_bag、bench、chair、coat_rack、easel、fan、fridge、grandfather_clock、high_chair、kids_chair、kitchen_cabinet、kitchen_island、ladder、laundry_hamper、magazine_rack、office_chair、ottoman、radiator、range_hood、rocking_chair、rocking_horse、room_divider、sofa、stool、stove、stroller、toilet、towel_rack、trash_can、vanity、wardrobe、water_dispenser。“on_tops”：为每个放置物品的家具台面设置一项：{"top", "objects"}。“top”是上述台面列表中的标签，或者是你添加的带台面房间物件的台面，格式为“<type>#<n>”；“ <n> ”沿用列表中该类型的编号（若列表中没有该类型，添加的第一个物件为“<type>#1”）。“objects”是由 {"type", "text", "count"} 组成的列表；“text”以取值文本的风格，用不超过8个词描述用于检索的物品。允许的类型包括上方列出的日常物品、设备和装饰品。“wall_pieces”：要添加的画作或其他墙面装饰，每项为 {"type", "text", "count"}；总数量最多为6。允许的类型：painting、board、wall_clock。空列表表示不添加任何物件。

列表 12。主论文图 3 中交互所涉及的第 1 轮计划的三个取值的规格。省略了布局答案的检索查询、资产标识符、空字段和概率。

```
{
  "flooring_type": {
    "属性":              "flooring_type",
    "类别":            "地板",
    "标签": "黑色马奎纳大理石",
    "外观": "高度抛光的黑色
        石材，带有醒目的白色纹理脉络。",
    "颜色": {
      "地板":       "#000000"
    },
    "表面": {
      "地板": {
        "acg":     "acg:Marble006",
        "覆盖层":    "材料"
      }
    }
  },
  "照明_方案": {
    "属性": "照明_方案",
    "类别":            "灯光",
    "标签":       "水晶卫星吊灯
        ”
    "外观": "由金色
        手臂和闪耀水晶。",
    "灯光": {
      "相关色温": 3000,
      "功率":       "高",
      "灯具": false,
      "日光":            "低",
      "时间":     "天",
      "来源":       "电力"         ,
      "固定装置":      [
        {
          "类型":     "天花板_灯",
          "数量": 1      ,
          "表面":          "天花板"
        }
      ]
    }
  },
  "座椅_布置": {
    "属性": "座椅_布置",
    "类别":            "座椅",
    "标签":       "圆形光环",
    "外观": "椅子等距排列
        在一张轮次
        桌子。
    "物体":          [
      {
        "类型":     "椅子",
        "数量": 2,
        "表面":           "地板",
        "尺寸_厘米": [46.9,     52.1,     92.0]
      },
      {
        "类型":     "餐厅_桌子",
        "数量": 1,
        "表面":           "地板"    ,
        "尺寸_厘米": [156.2,     100.5,     75.6]
      }
    ],
    "摆放位置": {
      "锚点":         "中心"      ,
      "沿着": 0.5,
      "分组":           "中间",
      "k": 1,
      "对称性": true,
      "朝向":         "焦点",
      "间距":           "正常"
    }
  }
}
```

列表 13. 场景描述 D_1 主论文图 3 中交互涉及的三个属性。座位安排的描述在摆放位置之后、布局答案的概率之前被截断。

```
地板_类型：地板：Marble 006（黑色、深色、灰色、灰、marble、大理石、石材；黑色）。地板为黑色（#000000）。照明_方案：3000 K灯光（暖色）；白天显示；固定装置：1盏吸顶灯；灯具关闭；窗户透入的日光开启。座椅_布置：2把椅子：采用一体成型外壳制作，配备全软垫座椅、细金属管腿、开放式扶手、雕刻曲线、绗缝、装饰线脚，整体为完全圆润的曲线形状；框架和椅腿为Metal 007（金属表面凹凸不平、金色、金属、划痕、银色、钢；棕色）；椅背正面为Metal007（金属表面凹凸不平、金色、金属、划痕、银色、钢；棕色）；软包为织物026（地毯、布、织物、地板、红色、瓷砖；棕色）；尺寸约为50 x 53 x 100厘米，独立放置在房间中，沿0.5，位于中间成组，对称布置，朝向焦点，正常间距。[.....]
```

列表14. 主论文图 1中第二位用户最终场景的大语言模型调用回复摘录（显式 形式，随机种子 3），其中列出了三个属性的取值、这些取值排除的物件类型，以及被舍弃的拟议物件。其余七个属性排除的类型列表均为空。


```
{
 "取值": {
  "座位安排_": "对称方形",
  "储物_方案": "黑色漆面玻璃展示柜",
  "照明_方案": "装饰艺术风格球形灯"
 },
 "避免": {
  "座椅_布置": {
   "控制": "座椅布局",
   "类型": ["扶手椅", "长凳", "豆 袋", "高 椅子", "儿童 椅子", "办公室 椅子", "脚凳", "摇摇椅
     ", "凳子"]
  },
  "储物_解决方案": {
   "控制": "储物家具",
   "类型": ["书架", "胸部 箱子", "斗柜", "文件 柜子", "床头柜", "搁板", "鞋 架", "餐边柜",
     "玩具 箱", "电视 架", "衣柜"]
  },
  "照明_方案": {
   "控制": "天花板灯具",
   "类型": []
  }
 },
 "已移除": [
  {
   "部分": "房间_物件",
   "类型": "餐边柜",
   "数量": 1,
   "by": "储物_解决方案"
  },
  {
   "部分": "在 台面上",
   "类型": "扬声器",
   "数量": 1,
   "由": "储物 解决方案",
   "台面": "餐边柜#1"
  },
  {
   "部分": "在 台面上",
   "类型": "桌面_装饰",
   "数量": 1,
   "by": "储物 解决方案",
   "台面": "餐边柜#1"
  },
  {
   "部分": "在 台面上",
   "类型": "蜡烛",
   "数量": 2,
   "by": "储物_solution",
   "台面": "餐边柜#1"
  }
 ]
}
```

J. 完整模式

本节的桌子列出了每种房间类型的模式（第A.2节）。每一行对应一个属性：其姓名、所属组、属性列表中的描述，以及其取值的名称，各项之间以分号分隔。未列出属性的定义和取值的描述；列表7 展示了一个属性的这两项内容。

表 S17。浴室房间类型的模式：10 个属性，50 个取值。

属性	组	属性列表中的描述	取值
墙砖	表面	此属性决定了墙壁和淋浴间墙壁的下部。	白色地铁砖；大型米色瓷砖；白色大理石；图案水泥砖；光滑混凝土
墙面颜色	表面	此属性决定上部墙面位于瓷砖。	暖白色；灰蓝色；赤陶色；深绿色；炭灰色；浅粉色
地板类型	表面	地板类型决定了浴室的地面表面。	大型浅色瓷砖；黑白棋盘格；小六边形马赛克；仿木纹板材；抛光混凝土；水磨石
洗手台布局	家具	洗手台布局决定了洗脸盆柜的布置。	悬浮式双人洗手台；落地腿式洗手台；壁挂式洗手台；嵌入式洗手台柜；立柱式洗脸盆
淋浴间配置	家具	淋浴配置决定淋浴间或浴缸及其淋浴房。	无框玻璃步入式淋浴房；黑框玻璃淋浴房；带浴帘的浴缸；带长凳
储物方案	家具	此属性用于确定房间其他位置的储物设施房间中的其他地方。	高型封闭式柜；开放式毛巾架；墙壁龛；马桶上方柜；无额外收纳空间
镜子类型	配饰	镜子类型决定了洗脸盆上方的镜子。	超大整墙镜；古董华丽框架；三折式药柜；不对称簇群
照明方案	照明	照明方案决定房间的光照。	明亮日光；冷色明亮顶灯；镜前壁灯；裸灯泡吊灯；隐藏式灯带光效
窗饰	纺织品	窗饰决定窗户。	磨砂玻璃；彩色玻璃；竹制罗马帘遮阳帘；威尼斯百叶窗；木制百叶窗
装饰主题	配饰	装饰主题决定了墙上的装饰物墙壁 ，柜台 ，和地板 ，以及浴室地垫。	裸露表面；盆栽植物；装框版画；编织篮子；陶瓷器皿

表 S18。卧室房间类型的模式：10 个属性，47 个取值。

属性	组	属性列表中的描述	取值
墙面颜色	表面	墙面颜色决定墙壁的饰面，不包括拥有独立饰面的任何特色墙。	暖白色；灰蓝色；赤陶色；深绿色；炭灰色；浅粉色
特色墙	表面	特色墙决定墙面的饰面与门相对。	纯色涂料；图案壁纸；木质护墙板；裸露砖墙；清水混凝土
地板类型	表面	地板类型决定了房间。	硬木地板；厚绒地毯；抛光混凝土；人字拼木地板
床架材料	家具	床架和床头板。	实木；软包织物；金属框架
床上用品	纺织品	床上用品决定床的外观。	清爽白；鼠尾草绿；石板灰；海军蓝；芥末黄；印花被子
储物方案	家具	储物方案决定了.....的布置用于存放衣物和物体的家具。	嵌入式衣柜墙；独立式五斗柜；开放式书架；开放式挂衣架；床底抽屉
家具布局	空间布置	家具布局决定了房间内床和其他家具的摆放位置。	居中对称；角落床；窗边横向摆放；窗户；抬高平台
照明方案	照明	照明方案决定了房间。	明亮日光；温暖灯光；床头阅读灯；个性吊灯；隐藏式凹槽 Glow
窗饰	纺织品	窗饰决定窗户的风格。	薄纱窗帘；亚麻侧帘；威尼斯百叶窗；竹制罗马帘；木制百叶窗
装饰主题	配饰	装饰主题决定了摆放在墙壁、桌子和地板上的装饰物。	裸露表面；醒目艺术品；画廊墙；室内丛林

表 S19。房间类型儿童房的模式：10 个属性，44 个取值。

属性	组	属性列表中的描述	取值
特色墙	表面	特色墙属性决定了进入时左侧长墙的饰面。	纯色涂料；波点；几何图形；风景壁画；花卉壁纸
墙面颜色	表面	墙面颜色决定了墙壁。	暖白色；浅粉色；鼠尾草绿；阳光黄；天蓝色
地板类型	表面	该属性决定了房间。	铺地毯的硬木地板；满铺地毯；软木地板；泡沫游戏垫
床型	家具	床型决定了床的形状和结构。	单人床；双层床；屋架床；四柱床；低矮地板床
家具饰面	家具	家具饰面决定了床架、储物家具和活动家具的外观。	天然浅色木材；白色涂装；鲜艳的原色；藤条与藤编
储物方案	家具	它决定储物区域。	墙架；封闭式橱柜；格子收纳架；玩具胸部
活动区域	空间布置	活动区域决定了房间的活动角。	艺术与手工角；阅读角；音乐角；学习书桌区
照明方案	照明	照明方案决定了房间。	明亮日光；温暖灯光；串灯；俏皮吊灯
窗饰	纺织品	窗饰决定窗户。	薄纱窗帘；印花窗帘；遮光卷帘百叶窗；木制百叶窗；竹制罗马帘
装饰主题	配饰	装饰主题决定墙壁、床、搁架和地板。	裸露表面；装框版画；毛绒玩具；学习道具

表 S20。餐厅房间类型的模式：10 个属性，48 个取值。

属性	组	属性列表中的描述	取值
墙面颜色	表面	墙面颜色决定墙壁的饰面，不包括具有自有饰面的特色墙。	暖白色；灰蓝色；赤陶色；深绿色；炭灰色；浅粉色
特色墙	表面	特色墙属性决定了门对面的墙。	纯色涂料；图案壁纸；木质护墙板；裸露砖墙；清水混凝土
地板类型	表面	地板类型决定了房间。	宽板硬木地板；人字拼木地板；大理石棋盘格；赤陶瓷砖；抛光混凝土
照明方案	照明	照明方案决定了房间。	明亮日光；个性吊灯；吊灯一排吊灯；壁灯；温暖灯光
家具风格	家具	家具风格决定了桌子的设计，椅子、长凳、内置座椅底座和储物家具。	斯堪的纳维亚风格；工业风格；世纪中叶现代风格；传统风格；装饰艺术风格；当代风格
桌子形状	家具	桌子形状决定餐桌。	长方形长桌；椭圆桌；圆桌；方桌
座位安排	空间布置	座位安排决定了桌子的摆放位置以及桌子及其座椅在房间内的摆放位置。	四周摆放椅子；长椅和椅子；转角卡座；窗边卡座
储物方案	家具	储物方案决定了储物家具在房间中的位置。	矮餐边柜；玻璃门柜；开放式餐盘架搁架；壁挂式酒架
窗饰	纺织品	窗饰决定窗户区域。	薄纱窗帘；亚麻侧帘；厚重天鹅绒窗帘；威尼斯百叶窗；木制百叶窗
餐桌布置	配饰	餐桌布置决定项目在餐桌。	空餐桌；白色桌布；完整餐具摆设；花卉中心装饰

表 S21。房间类型图库的模式：10 个属性，40 个取值。

属性	组	属性列表中的描述	取值
悬挂方式	空间布置	悬挂方式决定了.....的布置墙上的艺术作品。	稀疏单排；沙龙式悬挂；对称网格；轨道悬挂
艺术品类型	配饰	艺术类型属性决定了艺术品及其装框。	抽象画；黑白照片；古典画作；纺织品挂饰
墙面颜色	表面	墙面颜色决定墙壁的饰面，不包括具有自有饰面的任何特色墙。	画廊白；赤陶色；深红色；海军蓝；炭灰色；浅粉色
特色墙	表面	特色墙决定墙的饰面入口对面。	纯色涂料；裸露砖墙；木质护墙板；清水具体的

属性	组	属性列表中的描述	取值
地板类型	表面	地板类型决定了房间的地面表面 房间。	抛光混凝土；人字拼木地板；大 规格瓷砖；水磨石
展示装置	家具	展示装置决定地面展台和 展柜。	白色基座；玻璃展柜；独立隔断
座椅配置	家具	座椅配置属性决定座椅 房间面积。	无座位；简易长凳；休闲沙发
家具布置	空间 布置	家具布局决定展示装置的摆放位置 以及座椅在地板上的摆放位置。	沿着墙壁；一个中央组合；分成两组 组
照明方案	照明	照明方案决定了 房间。	明亮日光；温暖灯光；画作照明灯 暗处的画作照明灯；冷色明亮顶灯
装饰主题	配饰	装饰主题决定了装饰物 但墙面艺术品和展示装置除外。	无装饰；落地植物；落地雕塑；时钟、 箱子和地球仪；视频屏幕

表 S22。房间类型厨房的模式：10 个属性，47 个取值。

属性	组	属性列表中的描述	取值
橱柜颜色	家具	此属性决定了.....的外观 橱柜门和抽屉面板。	白色；浅橡木色；鼠尾草绿；海军蓝；哑光 黑色
吊柜	家具	吊柜决定其上方的墙面空间 台面。	封闭式壁柜；玻璃门柜；开放 搁架；无上部储物柜
台面材料	家具	该属性决定了 台面和中岛台面。	人造石英石；屠夫台木材；抛光 混凝土；天然大理石；经典花岗岩；不锈钢 钢
挡墙类型	表面	它决定台面之间的墙面 以及吊柜下方的墙面。	整板延伸；裸露砖墙；白色地铁砖 ；六边形马赛克；图案水泥砖
墙面颜色	表面	墙面颜色决定了墙面的外观 未被柜子或防溅板覆盖的表面。	暖白色；灰蓝色；赤陶色；深绿色； 炭灰色；浅粉色
地板类型	表面	该属性决定厨房的地面表面。	宽木板地板；黑白棋盘格； 赤陶瓷砖；抛光混凝土；大型浅色瓷砖
空间布局	空间 布置	空间布局决定了 柜台在房间内的摆放位置。	岛台居中；L形转角；直线 单排墙面
照明方案	照明	照明方案决定了 房间。	明亮日光；冷色明亮顶灯；吊灯 一排吊灯；橱柜底部灯光；壁灯
用餐座位	家具	用餐座位决定了 厨房中的用餐座位区。	吧台凳；小餐桌；嵌入式 卡座；无座位
装饰主题	配饰	装饰主题决定装饰性和日常 台面、窗台和 墙壁。	空置台面；陈列炊具；盆栽香草； 艺术品与陶瓷

表 S23。房间类型库模式：10 个属性，50 个取值。

属性	组	属性列表中的描述	取值
家具风格	家具	家具风格决定了 书柜、桌子、书桌和椅子框架的设计。	斯堪的纳维亚风格；工业风格；世纪中叶现代风格； 传统风格；装饰艺术风格；当代风格
墙面颜色	表面	墙面颜色决定墙壁的饰面， 不包括采用自有饰面的任何特色墙。	暖白色；灰蓝色；赤陶色；深绿色； 炭灰色；浅粉色
特色墙	表面	特色墙决定了后 墙（与门相对）。	纯色涂料；图案壁纸；裸露砖墙； 木质护墙板；织物面板
地板类型	表面	该属性决定房间的地面表面 。	抛光硬木；大型图案地毯；抛光 混凝土；大理石棋盘格；软木瓷砖
软包	纺织品	软包决定扶手椅的覆盖材料， 沙发和靠垫的覆盖材料。	干邑色皮革；祖母绿丝绒；燕麦色亚麻； 石板灰；芥末黄；海军蓝
搁架配置	家具	置物架配置决定了书柜的摆放位置 在房间内的位置。	落地到顶嵌入式书架；独立式书柜； 悬浮式壁架；中央书籍堆架
座位安排	空间 布置	座位安排决定了摆放位置 和座椅的朝向。	阅读圈；窗边座位；休闲座椅组； 阅读角
工作空间整合	家具	工作空间整合决定了 书桌和工作桌的摆放位置。	写字台；长型共用桌；学习隔间；否 Desk

属性	组	属性列表中的描述	取值
照明方案	照明	照明方案决定了房间。	明亮日光；温暖灯光；绿色银行家台灯；点亮的书架；隐藏式灯槽光
装饰主题	配饰	装饰主题决定了搁架、桌子、墙壁和地板上的装饰物，但不包括书籍。	仅书籍；室内植物；时钟、箱子和地球仪；抽象画；古典半身像

表 S24。客厅房间类型的模式：10 个属性，55 个取值。

属性	组	属性列表中的描述	取值
家具风格	家具	家具风格决定了座椅的设计，桌子、搁架、餐边柜和电视柜。	斯堪的纳维亚风格；工业风格；世纪中叶现代风格；传统风格；藤条与藤编；当代风格
墙面颜色	表面	墙面颜色决定了.....的外观墙壁。	暖白色；灰蓝色；赤陶色；深绿色；炭灰色；浅粉色
特色墙	表面	特色墙决定墙的饰面位于房间入口的对面。	纯色涂料；裸露砖墙；木条；图案墙纸；清水混凝土
地板	表面	地板决定了房间的地面表面。	宽板硬木地板；人字拼木地板；抛光混凝土；大规格瓷砖；满铺地毯地毯；大型图案地毯
软包	纺织品	软包决定座椅的覆盖材料家具。	干邑色皮革；祖母绿丝绒；燕麦色亚麻；石板灰；芥末黄；海军蓝
座位配置	家具	座位配置属性决定了其中的座椅部件房间。	大型U形组合沙发；沙发和装饰椅；经典三件套；低矮地垫坐垫；成对双人沙发；长凳和脚凳组合
布局类型	空间布置	布局类型决定座椅的摆放位置和朝向。	交谈环形布局；面向电视；沿墙排列；休息区 + 餐厅
照明方案	照明	照明方案决定了房间。	明亮日光；温暖灯光；大型吊灯；冷色明亮顶灯；隐藏式灯槽光
窗饰	纺织品	窗饰决定窗户区域。	薄纱窗帘；亚麻侧帘；厚重天鹅绒窗帘；威尼斯百叶窗；竹制罗马帘；木制百叶窗
装饰主题	配饰	装饰主题决定了墙壁、搁架、桌子和地板上的装饰物。	裸露表面；焦点装饰品；画廊墙；室内丛林；收藏家搁架

表 S25。办公室房间类型的模式：10 个属性，49 个取值。

属性	组	属性列表中的描述	取值
墙面颜色	表面	墙面颜色决定墙壁的饰面，不包括具有自有饰面的任何特色墙。	暖白色；灰蓝色；赤陶色；深绿色；炭灰色；浅粉色
特色墙	表面	特色墙决定后墙（门对面的墙）的饰面。	纯色涂料；裸露砖墙；木条；织物面板；清水混凝土
地板类型	表面	该属性决定房间的地板。	宽板硬木地板；人字拼木地板；地毯瓷砖；抛光混凝土；软木瓷砖
照明方案	照明	照明方案决定了房间。	明亮日光；冷色明亮顶灯；温暖灯光光；绿色银行家台灯；隐藏式灯槽光
家具风格	家具	家具风格决定了书桌的设计，椅子和座椅框架、桌子以及储物家具。	斯堪的纳维亚风格；工业风格；世纪中叶现代风格；传统风格；装饰艺术风格；当代风格
书桌类型	家具	书桌类型决定办公桌及其配套座椅。	行政办公桌；L形办公桌；双人对坐办公桌；站立式办公桌
储物方案	家具	储物方案决定了.....的摆放位置房间内的储物家具。	文件柜；顶天立地书架；立方格架；矮边柜
座位区	空间布局	座位区决定房间中的哪一部分被指定用于座椅。	会议桌；休闲沙发；阅读扶手椅；否额外座位
窗饰	纺织品	窗饰决定窗户区域。	薄纱窗帘；亚麻侧帘；威尼斯百叶窗；遮光卷帘；木制百叶窗
装饰主题	配饰	装饰主题决定了上面的装饰物墙壁、搁架、书桌和地板。	裸露表面；室内植物；抽象画；白板和软木公告板；装框照片

表 S26。候诊室房间类型的模式：10 个属性，53 个取值。

属性	组	属性列表中的描述	取值
家具风格	家具	此属性决定了椅子的设计	斯堪的纳维亚风格；工业风格；世纪中叶现代风格；
		椅架、桌子和接待台。	传统风格；藤条与藤编；当代风格
墙面颜色	表面	墙面颜色决定墙壁的饰面，	暖白色；灰蓝色；赤陶色；深绿色；
		不包括接待台后方的区域。	炭灰色；浅粉色
软包	纺织品	软包决定座椅的覆盖材料。	干邑色皮革；祖母绿丝绒；燕麦色亚麻；
			石板灰；芥末黄；海军蓝
座椅布局	空间	座椅布局决定了座椅的摆放位置和	围桌成圈；沿墙排列成行；小型
	布置	房间座椅的朝向。	聚簇；长椅排列成行；L形沙发角
地板类型	表面	该属性决定房间的地板。	木纹乙烯基地板；方块地毯；抛光
			混凝土；水磨石；大块陶瓷砖；
			人字拼木地板
照明方案	照明	照明方案决定了	明亮日光；冷色明亮顶灯；大型
		房间。	吊灯；壁灯；隐藏式灯槽光
墙面装饰	配饰	它决定后墙的空余部分。	一幅大型画作；风景照片壁画；
			几何色块；生机绿墙；
			画廊式组合
配饰主题	配饰	配饰主题决定了摆放物体的位置	杂志架；醒目雕塑；边桌；
		——座椅之间以及侧墙上。	数字信息屏；陶器与书籍架
声学处理	表面	它决定接待台后方的墙面区域。	织物面板；木条；素墙；软木墙
窗饰	纺织品	窗饰决定窗户区域。	薄纱窗帘；亚麻侧帘；威尼斯百叶窗；
			竹制罗马帘；木制百叶窗