arXiv:2602.01298v1 [cs.CV] 1 Feb 2026

基于多模态大语言模型推理的交互一致的物体移除

Ching-Kai Huang & Wen-Chieh Lin&
Yan-CenLee国立阳明交通大学



图1：我们提出交互一致的物体移除（ICOR）任务，其目标不仅是移除目标对象本身，还包括其相关的交互元素，从而生成语义连贯且视觉真实的输出。这些交互效果可大致分为四种类型：依赖光照的效果、物理连接的物体、目标产生的元素以及情境关联物体。

摘要

基于图像的物体移除通常只会擦除指定的目标，却会留下交互痕迹，导致结果在语义上不一致。我们将这一问题形式化为交互一致的物体移除（ICOR）：它不仅要求移除目标物体，还要求移除相关的交互元素，例如依赖光照的效果、物理连接的物体、由目标产生的元素以及上下文关联的物体。为解决这一任务，我们提出基于多模态大语言模型的推理增强对象移除（REORM），这是一种推理增强型物体移除框架，利用多模态大语言模型推断哪些元素必须一并移除。REORM 采用模块化设计，整合了由多模态大语言模型驱动的分析、掩码引导移除和自我纠正机制，并提供本地部署变体，以支持在资源有限的情况下进行精准编辑。为支持评估，我们推出 ICOREval，这是一个包含具有丰富交互依赖关系的指令驱动移除任务的基准测试。在 ICOREval 上，REORM 的表现优于最先进的图像编辑系统，证明了其生成交互一致结果的有效性。

1 引言

物体移除是图像编辑中的一项基础任务，支持广泛的现实应用。其目标是从图像中移除目标物体，并对缺失区域进行图像修复，以生成视觉上逼真的结果。尽管近期物体移除（Jiang 等人（2025年）；Ekin 等人（2024年）；Sun 等人（2025年））和指令引导图像编辑（Fu 等人（2024年）；Huang 等人（2024年））领域的进展利用强大的扩散模型和多模态大语言模型提升了可控性与视觉质量，但这些方法大多只关注移除目标物体本身，而不推理它与周围场景之间的交互。一些近期方法开始处理阴影和反射等特定视觉效果，但其范围仍然有限，无法涵盖更广泛的对象–场景交互（Winter 等人（2024年）；Wei 等人（2025年）；Zhao 等人（2025年））。因此，相关元素或上下文依赖的物体往往会被忽略，导致编辑后的输出出现语义不一致。

为解决这一局限，我们提出交互一致的物体移除（ICOR）的概念，通过明确考虑物体与场景之间的交互关系，扩展传统的物体移除方法。ICOR 要求全面理解整幅图像以及物体之间的交互关系。收到移除特定物体的指令后，处理过程不只是简单地消除目标对象，而是推断该对象不存在时场景可能呈现的样貌。这包括识别并移除所有在目标对象移除后会变得不一致或不合逻辑的次要元素，从而生成连贯、逼真的结果。

我们将交互效果分为四种类型：依赖光照的效果、物理连接的物体、目标产生的元素，以及情境关联物体。图 1 展示了每种类型的示例。(i) 中，移除人物时，也必须移除其影子。(ii) 中，移除狗时，也需要移除它佩戴的牵引绳。(iii) 中，如果移除人物，那么其体重造成的湿沙色调变化以及留下的脚印就会变得不合逻辑。(iv) 中，移除充电站后，旁边的标识在语义上便不再相关，因为它指向一个已不存在的物体。为保持场景连贯性，也应移除该标志。

本文提出了基于多模态大语言模型的推理增强对象移除（REORM），这是一种旨在解决 ICOR 任务的框架。REORM 的核心思想是利用多模态大语言模型强大的常识推理能力，分析移除目标对象所带来的影响。给定移除某个物体的指令，多模态大语言模型会推断场景中还有哪些元素会因目标对象的缺失而在语义上变得不一致，或在物理上变得不合理。随后，这些关联元素会被识别并归为移除候选对象。

为实现这一构想，我们将 REORM 设计为一个模块化框架。在主要阶段——多模态大语言模型驱动的分析与移除中，多模态大语言模型会解析指令、识别目标对象，并通过推理找出为保持一致性而必须一并移除的关联元素。随后，利用开放词汇分割定位这些元素，并通过掩码引导的物体移除模型将其移除。为支持本地部署，我们采用提示链策略，将推理过程分配给一个紧凑型多模态大语言模型和一个大语言模型。最后，为提升输出质量，我们引入多模态大语言模型控制的自我校正阶段，通过模拟预期结果并将其与编辑后的图像进行比较，解决剩余的不一致之处。REORM 将强大的常识推理能力、模块化灵活性与实用的可部署性相结合，无需额外训练即可在多种场景中实现交互一致的物体移除。

由于 ICOR 是一项新提出的任务，我们进一步构建了一个新的基准测试 ICOREval，其中包含成对图像、指令和真实标注结果。在该基准测试上的实验评估表明，与现有方法相比，REORM 在处理交互一致性案例方面表现更优。总之，我们的主要贡献如下：

- 我们提出了交互一致的物体移除（ICOR），这是一项新任务，通过明确处理物体交互和上下文依赖关系，拓展了传统的物体移除任务。
- 我们提出了基于多模态大语言模型的推理增强对象移除（REORM）框架，该框架利用大语言模型进行常识推理，以解决 ICOR 任务。
- 我们提出了 ICOREval，这是首个面向 ICOR 的基准测试。实验结果表明，REORM 在图像质量和鲁棒性方面表现出色。

2 相关工作

基于指令的图像编辑旨在根据自然语言指令修改图像。Nguyen 等人（2024）对早期基于指令的编辑方法进行了全面综述，这些方法涵盖了从基于生成对抗网络到基于扩散的方法。较新的方法利用多模态大语言模型（多模态大语言模型）增强语义理解能力和可控性。Fu 等人（2024）使用多模态大语言模型改写并落地用户指令，再据此引导编辑过程；Huang 等人（2024）则将多模态大语言模型与双向交互模块相结合，以推理复杂指令并更有效地定位编辑区域。尽管引入了多模态大语言模型，这些方法并未充分利用其常识推理能力来建模交互关系。因此，它们无法修改关联元素，限制了生成交互一致编辑的能力。

物体移除作为一项图像编辑任务已得到广泛研究，大多数方法依赖于在用户提供的掩码内进行修补。基于扩散模型的编辑器推动了大掩码补全的发展（Banitalebi-Dehkordi 和 Zhang（2021）、Rombach 等人（2022）），而近期无需调优的方法或插件式移除方法则利用模型内部的视觉特征或注意力机制来引导物体移除（Ekin 等人（2024）、Sun 等人（2025）、Jiang 等人（2025））。这些方法大多主要在所提供掩码的边界内进行操作。
近期研究开始纳入物体及其影响移除，旨在不仅移除目标物体，还消除其依赖光照的效果，例如阴影和反射。Winter 等人（2024）采用监督式反事实方法，拍摄移除单个物体前后的同一场景，并对扩散模型进行微调。Wei 等人（2025）通过明确的引导来增强物体与背景的分离，以抑制残留伪影。Zhao 等人（2025）提出物体及其影响注意力，以联合擦除物体及其引发的效果。然而，这些方法的范围仍局限于依赖光照的效果，且缺乏对更高层次交互关系的显式建模，例如物理连接和上下文关联的物体。这一局限促使我们提出交互一致的物体移除方法，以处理掩码和效果边界之外更广泛的交互关系。

图层分解 为物体移除提供了另一种途径：通过显式地将场景分离为可编辑图层，可以在保留背景的同时，隔离并移除目标对象及其相关效果。Lu 等人（2021）估计将主体及其时变影响（例如阴影、反射、烟雾）整合在一起的 RGBA 图层，从而实现视频中的物体移除或重新合成，并保留这些效果。Lee 等人（2025）通过更强的生成先验和多图层输出扩展了这一思路，以实现更可控的视频编辑。这些基于视频分解的方法依赖时间线索。对于静态图像，Yang 等人（2024）采用一种基于扩散的模型，该模型使用模拟分层数据和重新合成损失进行训练，以分离出干净的背景和能够保留阴影与反射的透明前景。然而，尽管这些方法在保留依赖光照的效果和烟雾方面表现有效，但它们并未建模更高层次的交互关系，例如物理连接或上下文依赖关系。

3 任务概述：交互一致的物体移除

我们提出交互一致的物体移除（ICOR），该方法在传统物体移除的基础上，要求模型考虑场景元素之间的交互与依赖关系。例如，移除一只正在被遛的狗时，也应移除牵引绳、影子和脚印；否则，结果将不一致。因此，ICOR 对视觉系统提出了挑战，要求其全面建模物体与场景的关系，并生成既连贯又逼真的编辑结果。
任务定义。 给定一张输入图像 $I \in \mathbb{R}^{H\times W\times 3}$和一条指令 T，ICOR 的目标是ˆ生成一张编辑后的图像 I，其中目标物体及其移除所导致的所有不一致元素均被消除，同时保持整体场景合理性，ICOR : $(I, T) \mapsto I$。
我们将目标物体 定义为指令中明确提及的实体，将关联元素 定义为与目标物体发生交互的所有元素。

范围。 ICOR 中考虑的关联元素分为四类：

(i) 依赖光照的效果——目标对象与场景光照相互作用所产生的视觉伪影，例如阴影或反射。(ii) 物理连接的物体——由目标持有、使用或以其他方式获得物理支撑的物品，例如手中提着的包。(iii) 目标产生的元素——由目标直接留下的有形痕迹，例如沙子上的脚印。(iv) 情境关联物体——若没有目标便失去用途或语境的元素；例如，没有灭火器的灭火器标志。

鉴于这类交互可能层层嵌套或串联，因此需要进行递归推理，以识别所有应移除的元素，从而确保结果连贯。

4 方法：基于多模态大语言模型的推理增强型物体移除

我们提出了基于多模态大语言模型的推理增强对象移除（REORM）框架，如图2所示。该框架的核心是MLLM驱动的分析与移除（第4.1节）：多模态大语言模型通过推理识别目标对象及其关联元素，并指导移除过程。我们进一步提出多模态大语言模型控制的自我校正（第4.2节），通过额外一轮推理与验证来优化编辑后的图像。最后，我们实现了用于本地部署的多模态大语言模型与大语言模型协作（第4.3节），旨在应对较小型多模态大语言模型推理能力下降的问题。

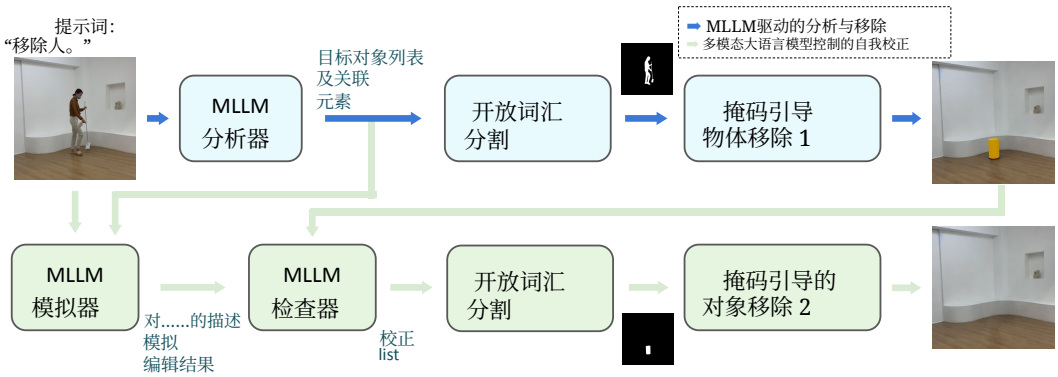

图2：所提出的REORM框架概览。该框架利用多模态大语言模型的常识推理能力，实现交互一致的物体移除。

4.1 MLLM驱动的分析与移除

鉴于交互一致的物体移除任务十分复杂，我们采用任务分解策略，将问题划分为三个依次进行的子任务：指令理解与推理、物体定位与分割，以及基于图像修复的移除。

多模态大语言模型驱动的分析与移除过程首先由多模态大语言模型解读指令，以识别输入图像中的目标物体及其关联元素。我们使用 OpenAI GPT-4o 作为多模态大语言模型分析器。

为提高推理过程的质量和完整性，我们不会直接提示多模态大语言模型返回关联元素列表。相反，我们采用思维链 提示策略（Wei 等人（2023）），鼓励多模态大语言模型明确推理移除目标物体所带来的影响。多模态大语言模型首先描述交互依赖关系，并评估是否应移除每个元素。之后才收集所提及的元素，并汇总成最终的移除列表。这种结构化推理提高了可靠性，也更有助于捕捉细微或间接的关联。完整提示词请参见附录B。

给定多模态大语言模型分析器识别出的目标物体及相关元素列表，我们随后使用开放词汇分割模型 Grounded SAM（Ren 等人（2024年））为

每个已识别的元素生成二值掩码。物体移除的一项关键考量是尽可能保留原始背景。通过使用精确的掩码明确指定待移除的区域，我们确保后续编辑操作集中于目标区域。这种局部控制有助于尽量减少对无关区域的意外更改。

最后一步使用掩码引导的对象移除模型 ObjectClear（Zhao 等人（2025年）），该模型根据生成的掩码从图像中移除所有对应元素。由于掩码既包含目标物体，也包含关联元素，并且移除过程由基于扩散模型、具备上下文感知能力的修复技术引导，因此生成的图像往往能够保持全局视觉与语义一致性。扩散模型具有更强的生成能力，即使在复杂场景中也能合成逼真的内容。

该模块化框架由多模态大语言模型分析器、开放词汇分割模型和基于掩码的物体移除模型组成，能够在各种真实场景中实现强大的泛化能力。每个组件都可以灵活替换为更先进的方案。该框架旨在完全无需额外训练即可运行，同时能够有效适应多样的指令和物体关系。多模态大语言模型驱动的设计具有良好的可解释性，因为可以轻松检查各阶段的中间输出，从而识别潜在的故障点。

4.2 多模态大语言模型控制的自我纠正

ON

我们在掩码引导的物体移除中使用的扩散模型可能会引入幻觉内容或留下残留伪影（Aithal 等人（2024），Cao 等人（2025））。受 Wu 等人（2024）提出的概念启发，我们引入了一种由多模态大语言模型控制的自我校正机制，通过多模态大语言模型进行额外轮次的推理，来验证并优化物体移除结果。

多模态大语言模型分析器生成包含目标物体及相关元素的移除目标列表后，基于多模态大语言模型的模拟器会将输入图像和移除列表作为输入，并生成对预期移除后场景的文本描述。该描述用作语义参考。编辑后的图像生成后，多模态大语言模型审查器会将其与模拟描述进行比较，以识别任何残留或不一致的元素，并将这些元素汇总到纠正列表中。我们的多模态大语言模型模拟器和多模态大语言模型审查器均使用 GPT-4o 实现。

随后，使用开放词汇分割模型对每个校正项进行掩码处理。为进一步优化输出，采用 Attentive Eraser（Sun 等人（2025））进行第二轮掩码引导的对象移除。与第 4.1 节中强调基于语义的物体移除的移除模型不同，这种图像修复方法通过将注意力重新引导至周围上下文来填充掩码区域，从而确保移除新识别出的元素。

我们提出的多模态大语言模型控制的自我校正利用多模态大语言模型的推理能力，对输出执行单轮验证，检测并校正任何残留的不一致之处或伪影。当与较弱的图像修复模型结合使用时，该机制尤其有益，因为它能够通过提升一致性和伪影去除效果来弥补这些模型的局限性。

4.3 MLLM–LLM协作实现本地部署

在实际部署中，由于隐私、成本、可复现性或长期访问基于云的 API 的需求，物体移除系统通常需要在本地设备上运行。然而，本地设备有限的硬件资源限制了大多数用户运行大规模模型的能力，尤其是需要大量内存和计算资源来进行高级推理的大语言模型。为此，我们开发了该框架的本地版本，以较小的开源替代方案取代大型专有模型，并使其能够在单个图形处理器上高效运行。

这一设定中的关键挑战之一在于，较小型多模态大语言模型的推理能力有所不足。在我们的实验中，紧凑型多模态大语言模型难以可靠地充当多模态大语言模型分析器（第4.1节）。为缓解这一局限，我们采用提示链策略（Wu 等人（2022）），将推理过程分解为多个子步骤。我们不再要求模型一次性完成所有推理，而是让每个步骤专注于一项更简单的任务，并将某一步的输出作为下一步的输入。这降低了各阶段的推理复杂度，并在模型容量受限时提升了可靠性。

此外，我们采用多模态大语言模型与大语言模型相结合的混合系统，发挥不同模型类型各自的优势。由于多模态大语言模型将大量参数用于学习视觉表征，其在常识推理任务上的表现往往不及规模相近的纯文本大语言模型。为解决这一局限，我们将图像相关问题交由多模态大语言模型处理；另一方面，纯文本推理则由紧凑型纯文本大语言模型负责。这种协作架构使我们既能保留 ICOR 所需的推理语义深度，又能兼容有限的计算资源。第5.3节验证了这一设计的有效性。具体而言，我们分别部署了量化至4位的多模态大语言模型 llava-hf/llava-v1.6-vicuna-13b-hf（Liu等（2024）），以及量化至8位的大语言模型 meta-llama/Llama-3.1-8B-Instruct（Grattafiori等（2024））。

如图3所示，我们将分析阶段分解为一系列顺序步骤。首先，大语言模型处理指令，以识别应当移除的主要物体。然后，多模态大语言模型检查输入图像并枚举所有关联元素。基于该集合，大语言模型进行推理，想象目标物体不存在时的场景，并确定哪些元素会变得不一致，因而也应被移除。最后，大语言模型将目标物体与关联元素整合为经过精炼且结构化的移除列表，随后将其传递给分割模型以进行掩码生成。为简便与高效起见，本地变体中省略了自我纠正模块。整个框架可在单块24 GB GPU上高效运行。

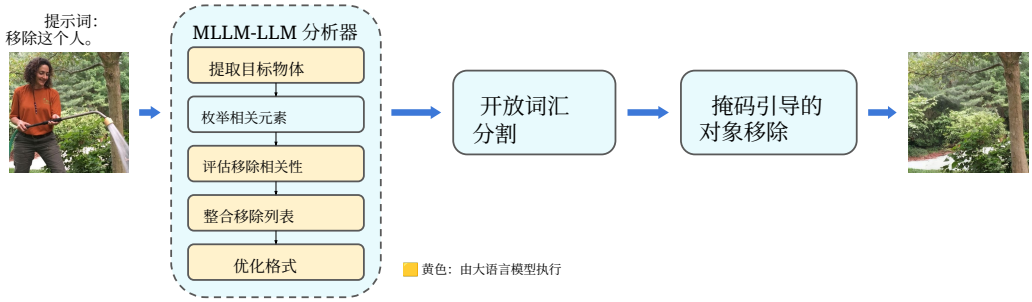

图3：面向本地部署的REORM概览，其采用提示链策略，并协调一个紧凑的多模态大语言模型与一个大语言模型。

5 实验

5.1 实验设置

评估数据集。现有的物体移除和图像编辑数据集提供了有价值的基准测试，但并非旨在评估模型在交互一致性条件下的表现。为填补这一空白，我们推出了**ICOREval**，这是一个新的基准数据集，用于评估目标对象与周围环境存在情境或物理交互时的物体移除效果。

ICOREval 包含 110 个示例，每个示例都包括一张输入图像，其中包含参与交互的物体；一条指定要移除物体的自然语言指令；以及一张真实输出图像，其中物体已被移除，同时保留了场景的一致性。指令由人工撰写，旨在反映真实的用户意图，且不会描述交互细节，因此模型需要在移除过程中推断必要的上下文调整。

该数据集由三个来源构建：（1）从公共数据集中筛选的交互案例，例如 RORD（Sagong等（2022年）），该数据集将真实场景记录为固定视角视频，并在不同拍摄片段中添加或移除物体；以及 RemovalBench（Wei 等人（2025年）），该基准测试使用固定摄像机对同一场景拍摄两次（一次有物体，一次没有物体）；（2）使用 Tewel等（2025年）和 Google Nano Banana 的物体插入模型生成的合成图像：先生成干净的背景图像，再以符合上下文的方式插入物体进行扩充；（3）复制粘贴合成，即手动将目标物体及其关联元素粘贴到真实世界背景图像上，以模拟交互场景。

通过结合真实数据与合成数据，ICOREval 提供了一个规模精简但颇具挑战性的基准测试，不仅评估移除操作的视觉质量，还评估模型理解指令并推理对象–场景交互的能力。

评估指标。 我们采用四种广泛使用的指标来评估物体移除结果：用于衡量感知相似度的 DINO 分数和 LPIPS，以及用于衡量像素级保真度的 PSNR 和 SSIM。这些指标全面评估语义一致性和重建质量。

基线方法。 我们将本方法与若干最先进的语言引导图像编辑基线方法进行比较。比较对象包括 MGIE（Fu 等人（2024））和 SmartEdit（Huang 等人（2024）），它们利用多模态大语言模型理解用户指令并引导编辑过程。这两种方法均为开源，可在本地运行。此外，我们还纳入了两种专有模型：OpenAI GPT Image 1[1] 和 Google Nano Banana[2]，，它们通过商业 API 支持基于指令的图像编辑。这些闭源模型代表了当前的最先进水平，可作为强有力的比较基线。

5.2 与语言引导的图像编辑方法比较

定量结果。 表 1 列出了所有受评估方法在 ICOREval 上的定量比较结果。我们提出的方法的两个变体 REORM（GPT-4o）和 REORM（LLaVA+Llama）在所有评估指标上均优于所有基线方法。值得注意的是，REORM 的表现超过了专有商业系统 GPT Image 1 和 Nano Banana，展现了其在实现交互一致的物体移除方面的强大推理能力。

我们进一步评估了所有对比方法的推理运行时间。如表 1 所示，除 GPT Image 1 外，与其他方法相比，我们的方法推理时间更长。这主要归因于迭代推理过程和自我纠正机制。尽管这些设计增加了运行时间，但它们显著提升了最终结果的交互一致性和视觉保真度，使 REORM 的表现优于现有方法。[3]

方法	DINO ↑	LPIPS ↓	PSNR ↑	SSIM ↑	运行时间（秒/图像）
MGIE*	0.709	0.261	19.217	0.634	5.20
SmartEdit*	0.760	0.275	20.064	0.615	10.09
GPT Image 1†	0.789	0.347	18.352	0.535	21.12（应用程序接口）
Nano Banana†	0.873	0.124	25.225	0.771	8.45（应用程序接口）
我们的方法（LLaVA +Llama）	<u>0.897</u>	<u>0.121</u>	<u>25.674</u>	<u>0.820</u>	12.54
我们的方法（GPT-4o）	**0.937**	**0.104**	**27.063**	**0.825**	13.15（应用程序接口）+4.33

表 1：**ICOREval 上的定量比较**。我们的 REORM 在四项评估指标（DINO、LPIPS、PSNR 和 SSIM）上均取得最佳性能，但需要更长的运行时间。 *可在本地运行的开源方法。†闭源方法。

定性结果。图 4 展示了 ICOREval 中代表性示例的定性比较，说明了我们的方法在实现交互一致的物体移除方面的有效性。（1）在第一个示例中，REORM 正确推断出该人正在骑自行车，并同时移除了骑乘者和自行车，避免留下自行车而形成不符合物理规律的场景。（2）在第二个示例中，我们的方法将人与其投射阴影一同移除。（3）在第三个示例中，我们模型的两个变体都将人与其手持的（物理连接的）洒水壶一同移除。（4）在第四个示例中，我们的方法将猫与其正在玩耍的玩具一同移除，包括连接在玩具上的细绳。

为进一步评估 REORM 在 ICOR 中的能力，我们从网络上收集了一组真实场景案例，涵盖范围更广的交互场景。图 5 展示了对比结果。(1) 在第一个示例中，两个 REORM 变体都成功移除了目标人物及其投射阴影（依赖光照）和脚印（目标产生）。(2) 在第二个示例中，在

[1]https://platform.openai.com/docs/guides/image-generation?image-generation-model=1 gpt-image-1

[2]https://ai.google.dev/gemini-api/docs/image-generation [3]通过应用程序接口依赖远程推理的方法会相应标注。GPT Image 1 和 Nano Banana 的运行时间报告为三次运行的平均值。所有数值均以每张图像所需的秒数为单位进行测量。

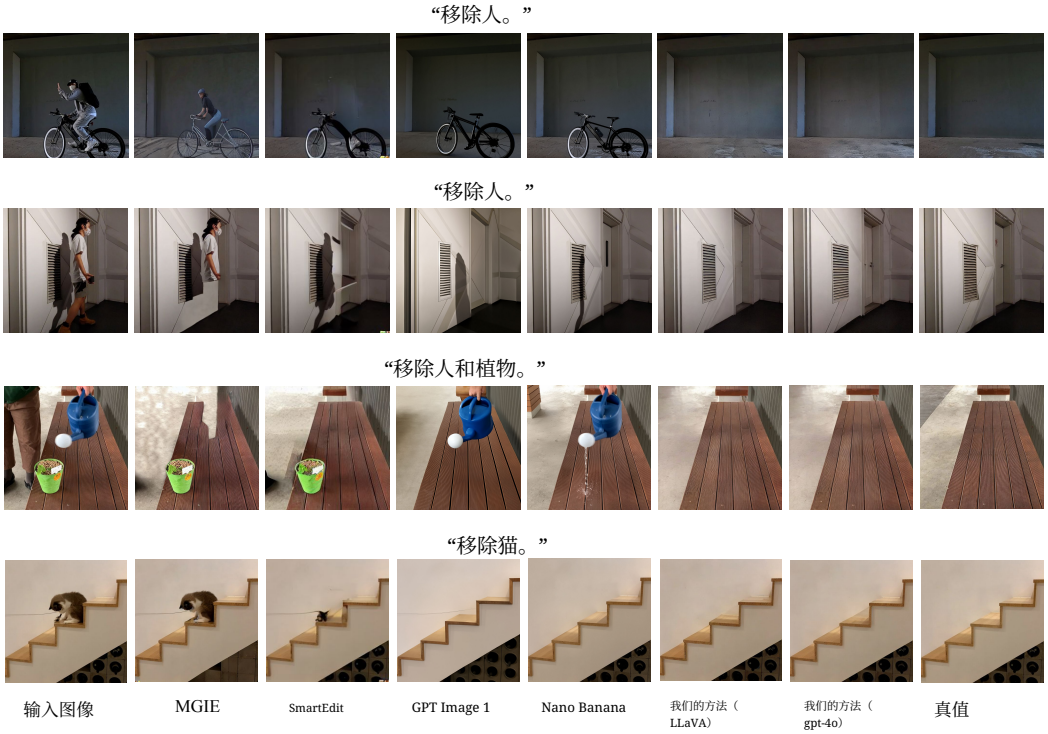

图 4：ICOREval 上的定性比较。我们的方法实现了 ICOR，成功移除了骑乘者及其自行车、带有投射阴影的人、手持洒水壶的人，以及正在玩玩具和细绳的猫。

移除手之后，我们的方法正确推断出，与手物理连接的洒水壶会呈现悬浮状态，并且还移除了正在倾倒的水流（目标产生）。(3) 第三个示例展示了两个变体都移除了灭火器及其相关标识（语境关联）。(4) 在第四个示例中，只有 REORM（GPT-4o）成功移除了灭火器及其标志，这表明更强大的多模态大语言模型能够提升该任务的泛化能力。另请参阅附录A中的其他结果。

5.3 本地部署设计验证

为验证第4.3节提出的本地部署设计的有效性，我们开展了一项研究，考察两种关键技术：提示链和MLLM–LLM协作。这些技术旨在缓解紧凑型多模态大语言模型部署在资源受限设备上时的推理局限。表2比较了三种实验设置。在实验（a）中，LLaVA单独处理与GPT-4o相同的提示词，不进行任何形式的提示分解。在实验（b）中，采用提示链将推理过程分解为多个步骤，但仍仅依靠LLaVA执行所有子任务。值得注意的是，我们观察到，与（a）相比，（b）的性能有所下降。这一下降主要归因于误差累积：紧凑型多模态大语言模型缺乏足够的推理能力，无法准确处理分解后的步骤，导致各阶段的错误沿链式流程逐步传播。

最后，实验（c）结合了提示链和 MLLM–LLM协作。通过将逻辑推理阶段交由推理能力更强的大语言模型处理，我们降低了出错的概率，有效解决了误差累积问题。结果表明，这一完整设计在所有指标上都取得了显著改进，证实这两种技术对于在受限条件下实现稳健推理都是必要的。

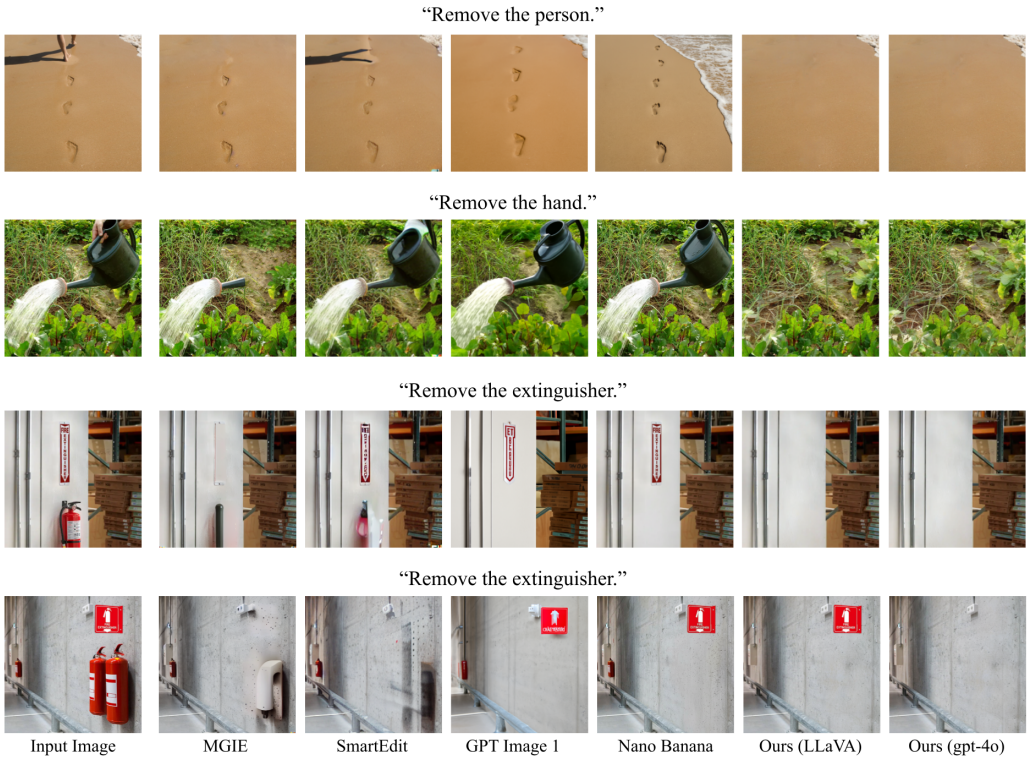

图 5：无真值图像上的定性比较。REORM 不仅移除了目标物体，还移除了阴影、脚印、洒水壶和水流、标识等关联元素；在具有挑战性的案例中，GPT-4o 版本表现出更强的泛化能力。

Exp.	MLLM	提示链	LLM	DINO ↑	LPIPS ↓	PSNR ↑	SSIM ↑
(a)	✓			0.921	0.125	25.266	0.810
(b)	✓	✓		0.913	0.128	25.230	0.813
(c)	✓	✓	✓	**0.943**	**0.109**	**26.033**	**0.830**

表2：**本地部署设计的验证**，在 ICOREval 上进行。本实验基于 ICOREval 的初始版本（72 个示例）开展。

实验（a）仅使用 LLaVA；（b）仅使用 LLaVA 应用提示链；（c）则将提示链与 MLLM–LLM协作相结合。（c）中的完整设计在所有指标上都表现出更好的性能，表明这两种技术对于在受限条件下实现稳健推理都是必要的。

5.4 失败案例分析

如图 6 所示，我们的方法偶尔会误删本应保留的背景元素。例如，背景中的黑色物体（左）和沙发上的装饰品（右）都被错误地擦除了。这一现象源于自我纠正机制：该机制会将初始结果与预期的移除后场景描述进行比较，以触发第二轮移除。然而，生成的描述有时未能捕捉背景物体的所有细节。当某个背景元素未出现在场景描述中时，系统会将其误认为是需要移除的残留物，从而产生假阳性，继而导致过度编辑。例如，在右侧的案例中，预期的移除后场景描述没有提到圆锥形装饰品，导致多模态大语言模型审查器错误地将其视为需要移除的物体。

值得注意的是，自我纠正模块纠正的错误数量超过了它引入的新瑕疵。尽管如此，如何减少此类假阳性并进一步降低过度编辑，仍是未来研究的一个重要方向。



图6：**失败案例**。 我们的方法偶尔会错误地移除背景元素。最右侧一列展示了模拟编辑结果的描述（多模态大语言模型审查器的输入），而对话气泡中包含相应的输出。

5.5 讨论：面向意图感知与隐私保护的物体移除

尽管交互一致的物体移除（ICOR）旨在通过不仅移除目标对象，还移除其相关交互效果，来生成语义一致且视觉真实的结果，但这一目标未必总是与用户的编辑意图一致。在某些场景中，用户可能希望保留某些上下文元素——例如，在制作特效、生成有意不真实的图像，或仅对目标对象进行最小编辑时。

在此类情况下，强制执行交互感知移除可能导致过度编辑，或产生与用户偏好相冲突的意外修改。这凸显了在执行编辑指令时考虑用户意图的重要性。理想情况下，物体移除或图像编辑系统应能适应因任务、图像内容或个人编辑习惯而异的多样化意图。然而，在没有额外输入的情况下隐式推断用户意图仍然是一个具有挑战性的问题。另一个日益受到关注的问题是隐私，它在实际部署中正引起越来越多的关注。我们的本地部署设计也代表了应对这一挑战的初步尝试。这些问题共同指向未来的研究方向：开发更智能的图像编辑系统，在资源受限条件下平衡用户自主权、效率（通过意图预测和个性化）与隐私保护。

6 结论

我们提出了交互一致的物体移除，这一方法通过同时移除相关交互元素，拓展了物体移除的范畴，从而避免这些元素损害场景合理性。为解决这一任务，我们提出了基于多模态大语言模型的推理增强对象移除（REORM）框架，利用多模态大语言模型的推理能力识别需要移除的相关元素。我们构建了 ICOREval 来评估这一任务，并证明 REORM 的表现优于现有图像编辑模型。我们认为，交互一致的物体移除这一任务的提出，为物体移除和图像编辑开辟了新的方向。

参考文献

Shashank K Aithal 等人。通过模式插值理解扩散模型中的幻觉。发表于*NeurIPS*，2024年。

AminBanitalebi-Dehkordi 和 Yong Zhang。Repaint：通过生成多个训练样本实例，提升下游视觉任务的泛化能力，2021年。

Yitong Cao 等人。《时间评分分析：理解并纠正扩散伪影》。载于*CVPR*，2025年。

Yi˘git Ekin、Ahmet Burak Yildirim、Erdem Eren Caglar、Aykut Erdem、Erkut Erdem 和 Aysegul Dundar。《CLIPAway：协调聚焦嵌入，通过扩散移除物体》。载于神经信息处理系统进展，第37卷，第17572–17601页，2024年。Tsu-Jui Fu、Wenze Hu、Xianzhi Du、William Yang Wang、Yinfei Yang 和 Zhe Gan。《通过多模态大语言模型引导基于指令的图像编辑》。载于国际学习表征会议（*ICLR*），2024年。Aaron Grattafiori、Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle、Aiesha Letman 等人。《Llama 3 模型家族》，2024年。Yuzhou Huang、Liangbin Xie、Xintao Wang、Ziyang Yuan、Xiaodong Cun、Yixiao Ge、Jiantao Zhou、Chao Dong、Rui Huang、Ruimao Zhang 等人。《SmartEdit：探索多模态大语言模型驱动的复杂基于指令的图像编辑》。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第8362–8371页，2024年。

Longtao Jiang、Zhendong Wang、Jianmin Bao、Wengang Zhou、Dongdong Chen、Lei Shi、Dong Chen 和 Houqiang Li。《SmartEraser：利用掩码区域引导从图像中移除任意内容》，2025年。

Yao-Chih Lee、Erika Lu、Sarah Rumbley、Michal Geyer、Jia-Bin Huang、Tali Dekel 和 Forrester Cole。《生成式Omnimatte：学习将视频分解为多个图层》。发表于计算机视觉与模式识别会议论文集，2025年。

Haotian Liu、Chunyuan Li、Yuheng Li、Bo Li、Yuanhan Zhang、Sheng Shen 和 Yong Jae Lee。《LLaVA-NeXT：改进推理、OCR 和世界知识》，2024年1月。网址：https://llava-vl.github.io/blog/2024-01-30-llava-next/。

Erika Lu、Forrester Cole、Tali Dekel、Andrew Zisserman、William T Freeman 和 Michael Rubinstein。《Omnimatte：关联视频中的物体及其影响》。发表于*CVPR*，2021年。

Thanh Tam Nguyen、Zhao Ren、Trinh Pham、Thanh Trung Huynh、Phi Le Nguyen、Hongzhi Yin 和 Quoc Viet Hung Nguyen。大语言模型时代图像与多媒体的指令引导编辑控制：综述，2024年。

Tianhe Ren、Shilong Liu、Ailing Zeng、Jing Lin、Kunchang Li、He Cao、Jiayu Chen、Xinyu Huang、Yukang Chen、Feng Yan、Zhaoyang Zeng、Hao Zhang、Feng Li、Jie Yang、Hongyang Li、Qing Jiang 和 Lei Zhang。Grounded SAM：面向多样化视觉任务的开放世界模型集成，2024年。网址：https://arxiv.org/abs/2401.14159。

Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bj¨orn Ommer。利用潜扩散模型进行高分辨率图像合成。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第10684–10695页，2022年6月。

Min-Cheol Sagong、Yoon-Jae Yeo、Seung-Won Jung 和 Sung-Jea Ko。RORD：真实世界物体移除数据集。载于第*33*届英国机器视觉会议 *2022*年，2022年。

Wenhao Sun、BenleiCui、Xue-MeiDong 和 Jingqun Tang。Attentive Eraser：通过自注意力重定向引导释放扩散模型的物体移除潜力。载于*AAAI*人工智能会议论文集，第39卷，第20734–20742页，2025年。

Yoad Tewel、Rinon Gal、Dvir Samuel、Yuval Atzmon、Lior Wolf 和 Gal Chechik。Add-it：使用预训练扩散模型在图像中进行无需训练的物体插入。载于第十三届国际学习表征会议，2025年。

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Brian Ichter、Fei Xia、EdChi、Quoc Le 和 Denny Zhou。思维链提示能够激发大语言模型中的推理，2023年。

Runpu Wei、Zijin Yin、Shuo Zhang、Lanxiang Zhou、Xueyi Wang、Chao Ban、Tianwei Cao、Hao Sun、Zhongjiang He、Kongming Liang 和 Zhanyu Ma。《OmniEraser：利用配对视频帧数据移除图像中的物体及其影响》。*arXiv*预印本 *arXiv:2501.07397*，2025年。

Daniel Winter、Matan Cohen、Shlomi Fruchter、Yael Pritch、Alex Rav-Acha 和 Yedid Hoshen。《ObjectDrop：为照片级真实感物体移除与插入引导反事实数据生成》，2024年。Tongshuang Wu、Ellen Jiang、Aaron Donsbach、Jeff Gray、Alejandra Molina、Michael Terry 和 Carrie J Cai。《PromptChainer：通过可视化编程串联大语言模型提示词》，2022年。网址：https://arxiv.org/abs/2203.06566。Tsung-Han Wu、Long Lian、Joseph E. Gonzalez、Boyi Li 和 Trevor Darrell。《自我纠正的、由大语言模型控制的扩散模型》。载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），第6327–6336页，2024年6月。Jinrui Yang、Qing Liu、Yijun Li、Soo Ye Kim、Daniil Pakhomov、Mengwei Ren、Jianming Zhang、Zhe Lin、Cihang Xie 和 Yuyin Zhou。《结合视觉效果的生成式图像图层分解》。*arXiv*预印本 *arXiv:2411.17864*，2024年。Kai Zhang、Lingbo Mo、Wenhu Chen、Huan Sun 和 Yu Su。《Magicbrush：用于指令引导图像编辑的人工标注数据集》。载于神经信息处理系统进展，2023年。Jixin Zhao、Shangchen Zhou、Zhouxia Wang、Peiqing Yang 和 Chen Change Loy。《ObjectClear：通过物体及其影响注意力实现完整物体移除》。载于*arXiv*预印本 *arXiv:2505.22636*，2025年。

A 附录：补充结果

图 7 展示了 ICOREval 基准上的更多定性结果。(1) 在第一个示例中，我们的方法既移除了人物，也移除了玻璃上的倒影；而 Nano Banana 未能移除倒影，GPT Image 1 则造成明显的背景变化，影响了地板和墙面的外观。(2) 在第二个示例中，GPT Image 1 在扶手区域产生了伪影，而 Nano Banana 未能移除人物。(3) 在第三个示例中，我们的方法（ LLaVA ）未能完全移除人物的手臂，但如我们的方法（GPT-4o）的结果所示，自我纠正机制弥补了这一不足，生成了更完整且一致的结果。(4) 在第四个示例中，GPT Image 1 不必要地改变了人物的外貌（移除了帽子），而 Nano Banana 和我们的方法均保留了人物身份，并成功移除了狗和牵引绳。(5) 在第五个示例中，GPT Image 1 添加了箱子并将其移到了地板上。虽然视觉上并非不合理，但这凸显了用户意图预测的重要性，正如第 5.5 节所讨论的那样。地毯和沙发靠垫上也能观察到细微的背景变化。(6) 最后，在最后一个示例中，GPT Image 1 轻微改变了前景附近的瓷砖；与 Nano Banana、我们的方法和真值相比，只有仔细检查才能发现这一变化。

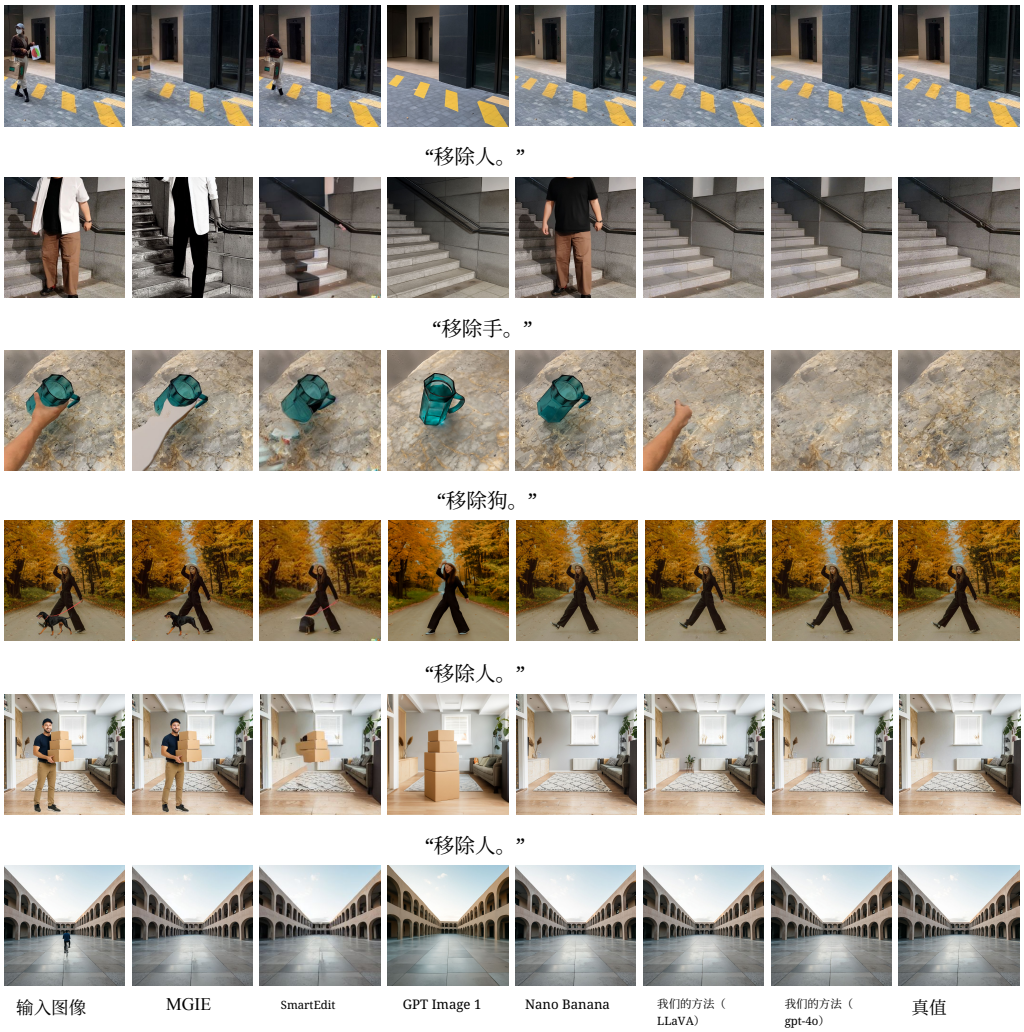

图 7：不同方法在 ICOREval 上的定性比较。

图8 展示了没有真实标注的图像上的更多定性比较。(1) 在第一个示例中，只有我们的方法成功移除了人物及其脚印。(2) 在第二个示例中，只有我们的方法同时移除了垃圾桶及其相关标识；其他方法要么留下了标志，要么只移除了部分垃圾桶。(3) 在第三个示例中，GPT Image 1 的结果留下了花园软管，使其看起来像是不自然地悬浮着。(4) 在第四个示例中，其他方法均未能成功移除猫玩具。此外，Nano Banana 改变了玩具的外观。(5) 在第五个示例中，GPT Image 1 留下了一股不合理的水流，而 Nano Banana 未能移除洒水壶，使其不自然地悬浮在空中。(6) 在第六个示例中，我们的方法（LLaVA）未能移除狗的影子，但我们的方法（GPT-4o）生成了连贯且完整的结果，展现了更强推理能力带来的优势。

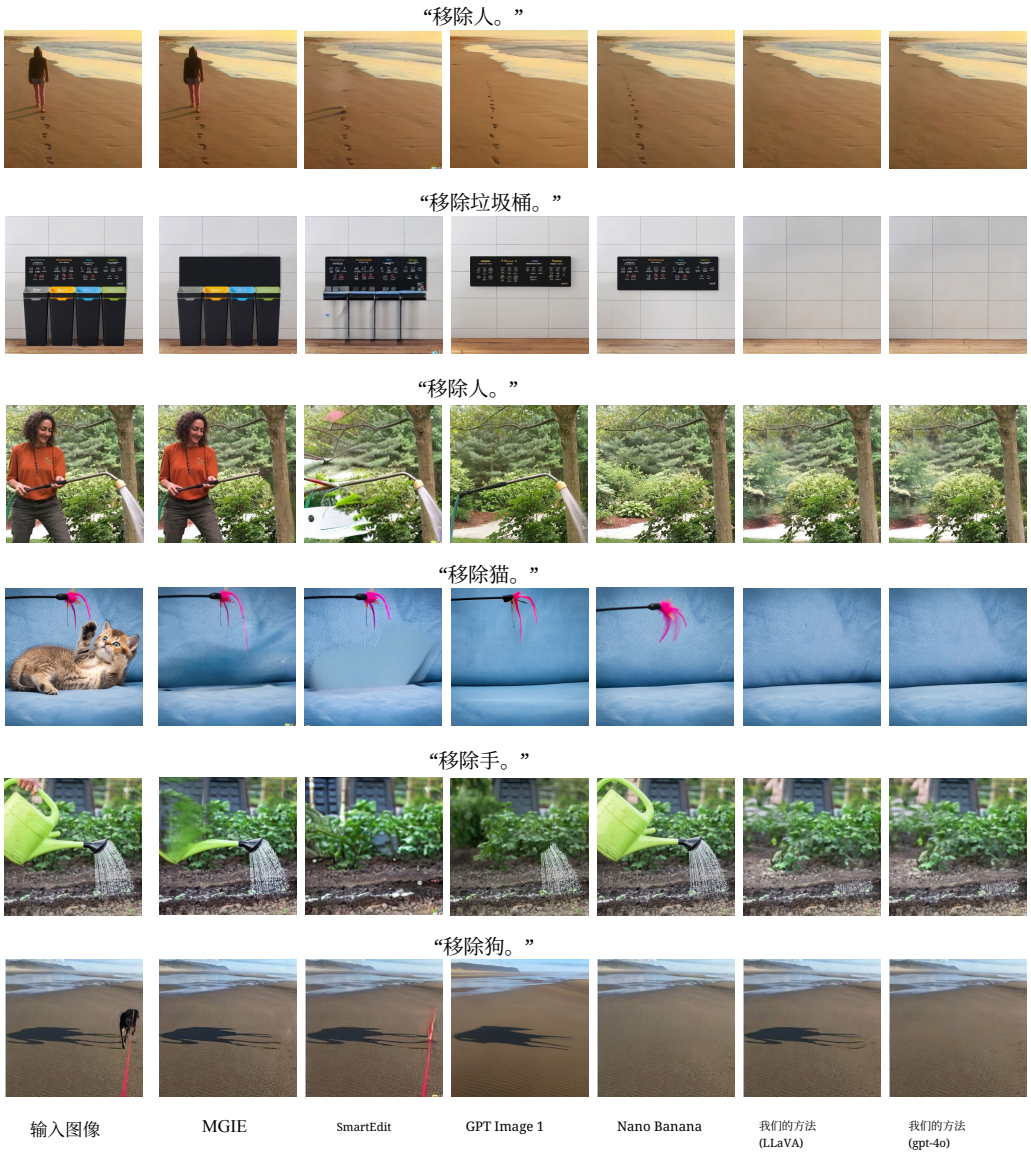

图8：无真值图像上不同方法的定性比较。

B 附录：我们框架中使用的多模态大语言模型提示词

B.1 多模态大语言模型分析器

```
# 身份
能够帮助用户识别应从图像中移除的目标物体及其关联
元素的助手，确保最终结果在语境上
连贯。

# 你的任务
给定一张图像和一个提示词，你需要分析并识别图像中应移除的目标物体以及所有相关物体和效果。你需要提供一份应在
图像中移除的物体列表。


# 处理步骤。
1
2．  阅读用户提示词，并识别带有相应属性的目标物体。
    仔细检查图像，以识别任何其他物体或效果，其存在
若主要目标对象被移除，将变得在物理上不合理或在语境上不合逻辑
已被移除。          这包括          ，但不限于：
  * 依赖光照的效果：                    阴影、反射。
    * 物理连接的物体：目标正拿着、穿戴着或由其提供物理支撑的物品。

  * 目标      –生成的元素：                目标产生的效果或物体，例如来自
洒水器的水。
  * 情境关联物体：                    失去其用途或上下文的项目，除非
目标，例如物体的标志、房子的邮箱或狗的牵引绳。
3．  请说明你的理解，然后按照示例所示的格式作答。
4．  避免使用模糊或抽象的名词短语后接关系从句，例如“the
object that...”          或“any item that...”。          只使用具体的物体名称。                例如，
写      “狗的玩具”而不是“它正在与之交互的物体”。
5．  避免描述细微的背景变化或纹理。

# 示例
<产品评论 id="example-1">
移除这个人。
</产品评论>-
<助手回复 ID="example-1">
推理：          “目标对象是这个人。                    如果移除这个人，他的影子和
他的滑板车在情境中会显得不协调，因为滑板车看起来像是
没有骑乘者却直立着。”
目标物体：          ["人",          "这个人的影子",          "滑板车"]
</助手回复>    -

<产品评论 id="example-2">
移除白狗。
</产品评论>-
<助手回复 id="example-2">
推理：          “目标对象是一只白狗。                    移除这只狗会使它的
水中的倒影和它正在玩的玩具在情境上显得不协调，
因为没有这只狗，玩具不太可能会在那里。”
目标物体：          ["白狗",          “白狗的倒影”,          “狗玩具”]
</助手回复>    -
                                    <                    >
<产品评论 id="example-3">  移除这个人。<                    /产品评论助手回复 id="example-3"> 推理：“目标对象
是一个人。移除这个人后，他拿着的包和杯子会看起来像是漂浮在半空中，这在物理上是不合理的。”目标物体：["人", "这
些包", "这个杯子"]</助手回复>

                    -
```

表3：我们为多模态大语言模型分析器设计的完整提示词。

B.2 多模态大语言模型模拟器

身份 你是一名视觉语言助手，负责描述移除用户要求移除的物体后，图像会呈现什么样子。 # 你的任务 给定一张图像和用户的移除请求，描述生成的图像。 # 输出规则 1. 尽可能详细地描述生成图像中的所有可见元素。包括背景、前景以及图像中其他所有可见部分的元素。 2. 避免描述抽象的情绪或心理状态。 3. 不要提及原始图像、移除过程，也不要使用“在给定图像中”之类的短语。

表 4：我们为多模态大语言模型模拟器设计的完整提示词。

B.3 多模态大语言模型检查器

身份 你是一位视觉语言助手，负责检查编辑后的图像是否符合描述。 # 你的任务 给定一张图像和一段描述，检查图像中是否包含描述中未提及的意外物体或元素。如果有，请列出这些元素；否则返回空列表。 # 输出规则 1. 先提供推理，然后列出应该从 t 中移除的物体 he 图像。 。 2 仅包含应该移除的物体 . 3. 忽略物体数量上的差异。 4. 仅包含物体的具体名称，不要包含其位置。 说出名称 直接（例如， “一个蓝色花瓶”），而不是以模糊的方式（例如， “一个看起来像 花瓶的物体”）。如果你确定，请包含颜色或形状。 5. 如果你不确定某个物体的确切身份，就不要列出它。 6. 避免使用含糊或范围较大的位置词，例如 ``area’’, “空间”，或 “地区”。 只关注具体的有形物体。 # 示例 <产品评论 id="example-1"> 编辑后的图像呈现了一间厨房，里面有炉灶、水槽和冰箱。 炉灶 干净而闪亮，水槽是空的，冰箱则关着。 没有 图像中有人或宠物。 </产品评论>- <助手回复 ID="example-1"> 推理: “图像展示了一间温馨的厨房。 里面有一台炉灶、一把椅子和一台 冰箱。 描述中没有提到椅子，因此应该将其移除。 图像中有个类似手的东西，但描述中并未提及。” 要移除的物体: [“椅子”, “手”, “手臂”] </助手回复> -

表 5：我们为多模态大语言模型审查器设计的完整提示词。

附录 C：数据集分布与多样性

为全面了解数据集的构成与多样性，我们在表 C 中给出了统计细分。对于左侧所示的关联元素类别，我们按照第 3 节中的定义，将元素分为四种类型。需要注意的是，这些类别并非相互排斥；单张图像通常会同时包含多个关联元素，这增加了评估的复杂性与真实性。对于右侧的数据构建来源，图像通过三种不同的方法收集而来，具体方法已在第 5.1 节中介绍。

关联元素类别		**数据构建来源**	
Type	数量	Type	数量
依赖光照	53	公共数据集	46
物理连接	35	合成图像	31
目标生成	16	复制粘贴	33
上下文关联	28	–	–

表6：数据集分布统计。左侧详述交互类别，右侧展示数据收集来源。请注意，单张图像可能同时包含多个关联元素。

为评估 ICOREval 的视觉多样性和语义覆盖度，我们将其特征分布与三种基线方法进行比较：MagicBrush Zhang 等人（2023）、RemovalBench Wei 等人（2025）和 OBER-Test Zhao 等人（2025）。我们采用预训练的 CLIP（ViT-B/32）模型，将图像映射到高维语义空间。对于每张图像，我们提取全局嵌入并进行L_2 归一化，以确保分析聚焦于语义方向性。为进行公平、无偏的密度比较，我们从较大的数据集中随机抽取一个子集，使其样本数量与较小数据集相同（$N = \min(N_{ours},\ N_{baseline})$）。随后，使用带有 PCA 初始化的 t-SNE 将这些高维嵌入投影到二维流形，以保留全局结构。
最终的可视化结果如图9所示，将 ICOREval（蓝色）置于现有基准测试的语义空间中。开放领域的 MagicBrush 展现出最广泛的分布，而 ICOREval 也达到了具有竞争力的多样性水平。更重要的是，ICOREval 优于对象移除基准；如中间图和右图所示，OBER-Test 和 RemovalBench（灰色）往往聚集在有限的子区域中，而 ICOREval 则保持了更广泛的分布。这表明，与其他基线方法相比，我们的数据集具有出色的语义覆盖度。

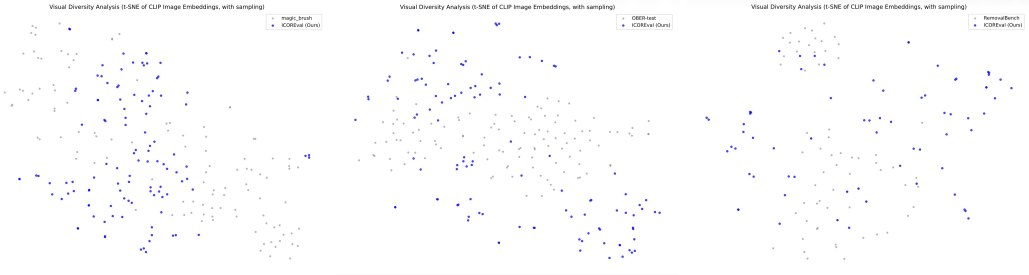

图9：使用 t-SNE 对 CLIP 嵌入进行视觉多样性分析。我们将 ICOREval（蓝色）与 MagicBrush、OBER-Test 和 RemovalBench（灰色）进行比较。可视化结果显示，ICOREval 保持了广泛的语义分布，其多样性高于 OBER-Test 和 RemovalBench，但不及开放领域的 MagicBrush。

为了进一步量化 ICOREval 超越二维可视化的语义复杂性，我们使用主成分分析（PCA）对 512 维 CLIP

嵌入进行内在维度分析。我们分析累积解释方差比，以确定表示数据集信息所需的主成分数量。图 10 显示，前 5 个主成分仅占总方差的 32.44%，这表明数据分布并未由少数几个主要特征主导。此外，分别需要 52 个和 67 个主成分才能捕获总方差的 90% 和 95%。对大量维度的这一需求表明，ICOREval 具有丰富的语义结构和多样的视觉属性，而不是局限于低维流形。

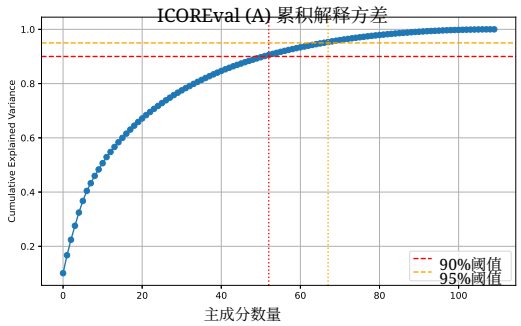

图 10：通过对 CLIP 嵌入进行主成分分析，分析 ICOREval 的内在维度。曲线展示了累积解释方差比。垂直虚线表示分别需要 52 个和 67 个主成分才能捕获总方差的 90%（红色）和 95%（橙色）。