arXiv:2603.01194v1 [cs.CV] 1 Mar 2026

RnG：用于从部分观测进行完整三维建模的统一 Transformer

Mochu Xiang [1, 2‡] Zhelun Shen[2*†] Xuesong Li[3,4] Jiahui Ren[1] Jing Zhang[3]
Chen Zhao[2] Shanshan Liu[2] Haocheng Feng[2] Jingdong Wang[2] Yuchao Dai[1†]

[1]西北工业大学，中国 [2]百度公司，中国 [3]澳大利亚国立大学，澳大利亚 [4]澳大利亚联邦科学与工业研究组织，澳大利亚

shenzhelun@pku.edu.cn daiyuchao@nwpu.edu.cn



图 1. **RnG 能做什么？** 给定物体的少量无位姿图像，VGGT 等三维重建基础模型可以恢复观测区域的结构，但无法对未观测部分建模。RnG 使用单个前馈Transformer，可在 A800 GPU 上不到一秒内估计物体的完整三维几何。RnG 以隐式方式重建三维，并在新的视点下渲染出具有外观和几何信息的结果。通过累积这些渲染出的点图，RnG 可以生成完整的三维物体，宛如一台虚拟三维扫描仪。

摘要

人类通过来自有限视点的二维观测感知三维世界。尽管近期的前馈式可泛化三维重建模型在从稀疏图像中恢复三维结构方面表现出色，但其表示通常局限于观测区域，无法对未见几何进行建模。这引出了一个关键且根本的挑战：我们能否根据部分二维观测推断出完整的三维结构？我们提出了 *RnG*（重建与生成），这是一种新颖的前馈 *Transformer*，通过预测隐式的完整三维表示，将这两项任务统一起来。*RnG* 的核心是我们提出的重建引导的因果注意力机制，该机制在注意力层面上分离重建与生成，并将键值缓存视为隐式三维表示。随后，任意姿态都可以高效查询此缓存，以渲染高保真的新视角 *RGBD* 输出。因此，*RnG* 不仅能准确重建可见几何，还能生成合理连贯的未见几何与外观。我们的方法在可泛化三维重建和新视角生成两项任务中均达到最先进性能，同时运行效率足以支持实时交互应用。项目页面：
https://npucvr.github.io/RnG

* 项目负责人。† 通讯作者。
‡ 本工作是在百度实习期间完成的。

1. 引言

稀疏视觉数据上的三维感知是现代计算机视觉的基石，推动了机器人 [12], 、增强现实 [31], 和大规模数字内容创作 [54]的发展。近年来，可泛化三维重建 [50, 52] 取得了显著进展，证明仅凭少量无位姿输入图像即可恢复场景几何和外观。然而，大多数重建流程仍存在一个根本局限：这些模型经过专门训练，只能恢复输入相机视角中可见区域的几何。因此，其几何输出并不完整，阻碍了物理仿真 [27] 或内容生成 [10]等要求完整物体表示的下游应用。

新视角合成（NVS）通过根据观测数据生成来自未见视点的图像，拓展了这些能力。虽然某些新视角合成流程能够泛化到未观测区域，但通常缺乏一致的三维结构。如表 1 所示，这些通用新视角合成方法能够渲染照片级真实新视角，但往往受限于对有位姿输入的依赖、缺乏几何重建或相机控制受限。最近，Matrix3D [30]探索了一种统一的重建与新视角生成方法。然而，其基于扩散模型的设计计算成本高昂，因此不

表 1. 代表性三维重建与新视角合成方法的比较。

Task	三维重建		新视角合成			统一模型	
方法	VGGT [50]	DUSt3R [14]	LVSM [18]	LGM [44]	RayZer [17]	Matrix3D [30]	RnG Ours
无位姿推断	✓	✓	✗	✗	✓	✓	✓
相机控制	N/A	N/A	✓	✓	✗	✓	✓
生成未见内容	✗	✗	✓	✗	✓	✓	✓
显式三维	✓	✓	✗	✓	✗	✓	✓
实时推理	✓	✓	✓	✓	✓	✗	✓

适用于实时应用。

尽管存在这些挑战，我们假设，前馈三维重建基础模型 [50, 52],虽然主要通过训练来恢复可见几何，但可能具备在其潜在空间中推断更完整三维表示的潜力。如果能有效利用这一潜力，并通过视点条件神经解码进行渲染，就有望揭示模型内部全面的三维理解。

我们提出了**RnG**，这是一种旨在激活并显式呈现这种潜在感知的前馈Transformer。通过将视点条件神经渲染视为对模型潜在特征的查询，RnG 在单一统一框架中衔接了重建与生成。具体而言，为了处理无位姿输入，同时保持显式相机控制，我们基于三维重建基础模型 [50]并继承其具有意义的潜在三维表示。我们的单一模型联合重建潜在三维内容，并生成新视角外观与几何。为此，我们采用重建引导的因果注意力机制，在注意力层面将这两个过程解耦，同时让两项任务共用同一组参数。这代表了一种将三维重建知识迁移至图像生成的新方向。尽管以往工作主要关注将生成先验适配于重建 [21, 30], ，我们证明，将重建先验反向迁移至生成是可行且有效的。

为推广到未观测的视点，我们采用隐式表述：将新视角用作对潜在三维空间的查询，网络的生成阶段会在这些视点上渲染隐式三维表示。此外，我们会在每个查询视点渲染一张点图，将隐式潜在三维表示转换为显式几何表示。由于 RnG 是一种确定性的前馈 Transformer，因此具备实时推理能力。因果注意力设计支持键值缓存机制，使模型能够先从输入图像中完成一次重建（用时 ∼0.2 秒），随后以极高效率生成新视角外观和几何（用时 ＜0.1 秒）。RnG 的运行速度比基于扩散模型的方法快 100× 倍 [30]。

实验结果表明，RnG 在多个三维重建和新视角生成基准测试中取得了最先进的性能，涵盖相机姿态估计、输入图像与新视角的深度预测，以及新视角合成。值得注意的是，它大幅超越了专为单项任务设计的专用模型。RnG 的主要亮点总结如下：

• **通过因果注意力实现统一架构**。 我们引入了一种重建引导的因果注意力机制。在注意力层面上，这种机制将两项任务解耦，使单个前馈Transformer能够连贯地执行隐式三维重建、新视角合成和显式几何生成。

• **将键值缓存用作隐式三维表示**。 得益于因果注意力设计，网络的键值缓存可充当灵活的隐式三维表示。这有助于高效生成新视角（渲染），并在内部记忆与重建几何之间建立透明的联系。

• **重建驱动生成**。 RnG 表明，相较于生成先验，利用重建先验能够提升新视角合成性能，同时显著降低计算成本。

2. 相关工作

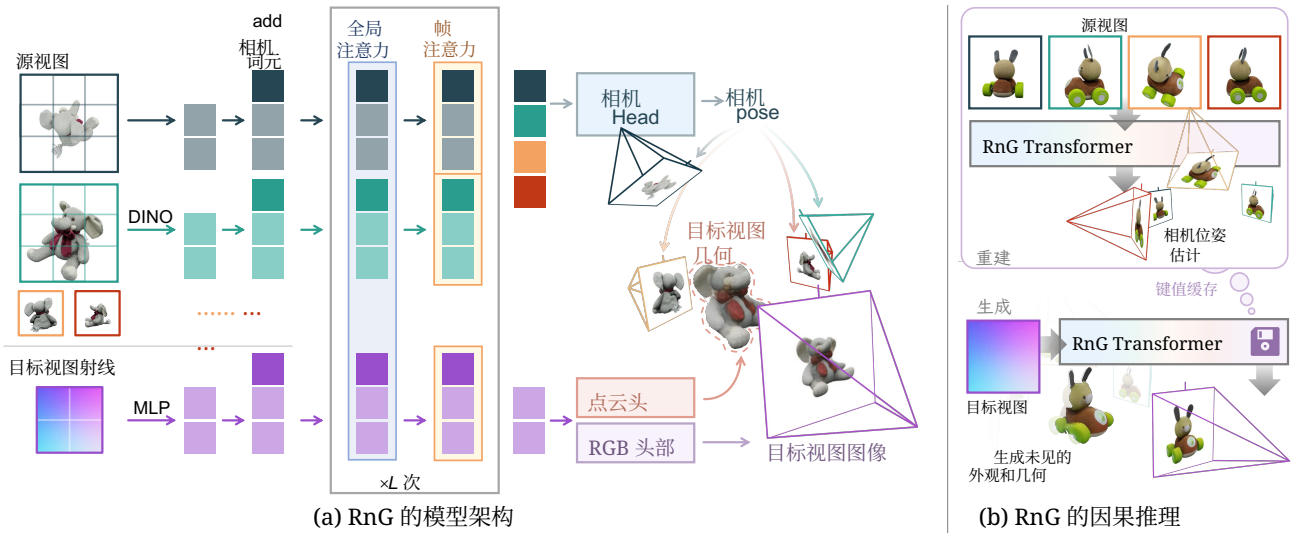
2.1. 神经渲染

神经渲染通过模拟相机捕捉真实世界的方式来表示场景 [47]。一些可微光栅化方法 [25] 或光线追踪方法 [15]直接根据三维网格描述三维到二维成像过程。为了从图像集合中恢复场景，辐射场能够生成逼真的图像。这些表示编码在神经网络权重 [32],、体素网格 [11]或哈希网格 [34]中。3D 高斯泼溅使用类似点云的三维高斯基元显式地对场景建模。为了获得新视图图像，所有这些方法都需要一个物理可解释的渲染过程。近来，也有研究尝试直接使用神经网络生成新视角图像，无论输入是多视图图像 [18, 40]还是三维网格 [63]。

2.2. 新视角合成

新视角合成（NVS）旨在根据给定观测生成新视点处的图像。该任务通常在两种不同的设定下进行处理。当可获得覆盖充分的密集观测时，利用隐式 [32, 34]或显式 [23, 25]表示的方法可以过拟合到特定场景，从而通过神经渲染或光栅化实现高保真新视角。

在更具挑战性的稀疏观测设置中，需要采用具备泛化能力的方法。基于扩散模型的方法在生成多样的新颖



(a) RnG 的模型架构　　(b) RnG 的因果推理

图 2. **RnG 网络架构**。(a) 首先使用 DINO 视觉 Transformer 对源视图图像进行词元化；表示目标视点的 Pl¨ucker 射线图则经过一个线性层。为每个视图添加相机令牌后，所有词元将交替参与全局注意力块和帧级注意力块。最后，利用输入视图中的相机令牌估计相机位姿，同时由点云头和 RGB 头部处理目标视图的射线令牌，从而估计几何和外观。(b) 在推理过程中，模型可以缓存源视图中的 K/V 令牌，以更快的速度合成新视图几何和几何信息。

视图方面展现出潜力，但往往难以维持严格的三维一致性 [28]。后续研究致力于通过隐式 [24] 或显式 [57] 3三维建模来提升多视图一致性。这一研究方向还包括面向动态场景的视频扩散模型 [1, 66]以及多模态方法 [30, 67]，后者会生成显式三维几何（例如带位姿 RGBD）以及外观。

作为高成本扩散模型的替代方案，前馈方法能够显著加快推理速度。一种流行策略是针对稀疏输入调整 3D 高斯泼溅 [4, 6, 7, 45, 55, 60]。为更好地合成遮挡区域，其他隐式方法 [18, 40] 使用神经网络直接估计新视图图像。这些方法通常采用专门的网络设计，例如多视图注意力机制 [33, 39] 或极线几何编码 [43, 62]。尽管其中许多方法需要已知相机位姿，但有一类独特的无位姿方法 [17, 41] 无需已知位姿即可运行，不过其估计出的潜在位姿较难解释。

2.3. 3D 视觉基础模型

Dust3R [52]和 VGGT[50] 展示了针对三维计算机视觉任务的有效数据驱动解决方案。受其实践启发，近期的尝试得益于其广泛的预训练数据，并在各种任务中展现出卓越性能。添加高斯参数预测头，可以将点图预测转化为照片级真实感场景重建。一些尝试探索了对象级别的 [56]和场景级别的 [42, 60] 适配。将模型输入扩展到更多视图 [46, 58] 或使模型适应流式输入 [26, 51]，可将模型应用扩展到更大尺度。大量训练数据同样有助于域适配过程，使模型在其他领域 [13, 37, 49] 或模型提示 [16]中取得更佳性能。尽管训练数据主要涵盖静态场景，但模型可以轻松适配以执行动态重建 [5, 19, 29, 53, 64]。

生成先验 [38] 已被证明在迁移学习中极为有效。将基于扩散的图像生成骨干网络重新用于密集预测任务 [21] 和多模态学习 [30,65, 67]，已取得显著进展。尽管采用生成先验能够有效辅助感知任务，但基于扩散的去噪范式会大幅增加推理时间，给实时交互应用带来沉重负担。

3. 方法

给定少量无位姿源图像 $\{\mathbf{I}_s\}$，RnG 会重建源视图的几何，并合成新视图的几何与外观。具体而言，RnG 首先执行隐式三维重建，并将其存储在网络的键值缓存中。同时，估计源图像的相机位姿 $\{\hat{\mathbf{g}}_s\}$，以供后续使用。随后，RnG 从该键值缓存中读取信息，从而高效地在任意视点 $\mathbf{g}_t$ 进行渲染，生成新颖外观$\mathbf{I}_t$ 和点图 $\hat{\mathbf{p}}_t$。通过累积从多个视点查询到的结构，或根据估计的源视图相机位姿 $\hat{\mathbf{g}}_s$直接映射纹理，即可生成完整三维结构。整个过程只需一秒。

为实现这一目标，我们首先概述 RnG 的整体架构，具体见第 3.1 节。RnG 使用单个网络执行源视图重建和新视图生成；我们设计了一种重建引导的因果注意力机制，在注意力层面将这两个过程分解开来（见第 3.2 节）。此外，该设计使网络的键值缓存能够充当隐式三维表示（第 3.3 节）。

3.1. 重建与生成统一 Transformer

我们的目标是根据有限数量的无位姿图像，在任意视点恢复三维结构和外观。这要求我们通过共享的隐式表示，将三维重建与新视角生成衔接起来。为此，我们提出了重建与生成统一 Transformer（RnG）。整体架构如图2所示。我们继承了 VGGT 的模型架构和权重 [50]。如图所示，图像级特征提取器 [35] 首先将源视图图像 $\{\mathbf{I}_s\}$ 转换为词元。随后，我们将目标视图编码为 Pl¨ucker 射线图，并通过线性层投影，得到目标视图词元。源视图和目标视图的词元分别与相机词元和寄存器词元拼接，形成的序列再经过交错的全局注意力层和帧注意力层处理。

经过 $L = 24$ 层交错的全局注意力和帧注意力后，源视图词元由相机头用于预测相机位姿 $\{\hat{\mathbf{g}}_s\}$。对于目标视图词元，我们连接两个 DPT 头 [36]，分别生成新视角外观和几何。具体而言，由目标视图的 Pl¨ucker 射线图（从$\mathbf{g}_t$编码得到）生成的词元会输入 RGB 头 $\mathcal{D}_{\text{RGB}}$，以生成新视图图像 $\hat{\mathbf{I}}_t$；而点云头 $\mathcal{D}_{\text{pmap}}$ 则输出像素对齐点图 $\hat{\mathbf{p}}_t$，表示相应区域的几何。

遵循 VGGT [50] 的做法并更好地保留其所学知识，我们为第一个源视图分配专用的相机词元和寄存器词元，并让其余源视图与目标视图共享相同的词元。我们训练模型始终将第一个视图回归到固定位置：

$$\hat{\mathbf{g}}_{s=1} = [I_{3\times3}\,|\,[0,0,-1]^\intercal]\,, \tag{1}$$

这隐式地为重建指定了一个世界坐标系。目标视点 $\mathbf{g}_t$ 和估计的点图 $\hat{\mathbf{p}}_t$ 也都在该世界坐标系中表示。

3.2. 重建引导的因果注意力

直观而言，重建应引导生成过程，
过程，但生成不应干扰重建。这将促进源视图感知的一致性，并利用潜在特征中编码的几何理解来增强新视图生成。

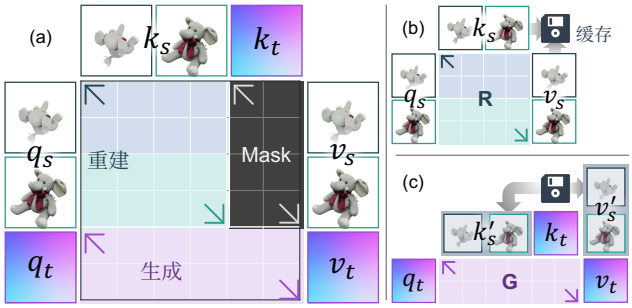


图 3. **重建引导的因果注意力。(a)**训练期间，我们在全局注意力块内部的注意力层面将重建与生成解耦。推理时，注意力过程分为两个步骤：**(b)** 源视图键值令牌被缓存为隐式三维表示；**(c)** 通过目标视图位姿查询 KV 缓存，以生成新视角。

受语言 [2, 48]和视觉 [59, 68] 模型中所采用的因果注意力启发，我们提出了重建引导的因果注意力，以实现这一目标。具体而言，我们引入二进制掩码 M 来控制注意力计算期间的信息流：

$$M_{i,j} = \begin{cases} 0 & \text{if } i \in \{s\} \text{ and } j \in \{t\} \\ 1 & \text{elsewhere} \end{cases}, \tag{2}$$

其中，$\{s\}$ 和 $\{t\}$ 分别是来自源视图和目标视图的词元索引。由此得到注意力输出（d_k 为查询词元和键词元的特征维度）：

$$\text{Out} = \text{softmax}\left(\frac{M \odot QK^\intercal}{\sqrt{d_k}}\right)V. \tag{3}$$

图3(a) 详细展示了所提出的机制。它禁止源视图中的查询词元关注目标视图中的键。换言之，源视图中的查询词元只能关注源视图中的键，而目标视图词元则可以关注源视图和目标视图中的词元。通过使用掩码注意力，我们的模型在给定任意目标视图时都能提供一致的重建结果。

值得注意的是，尽管源视图和目标视图中的词元分别服务于任务的不同方面，网络仍使用共享参数来处理它们：源视图中的词元负责感知并估计相机位姿（重建），而目标视图中的词元则负责合成新颖外观和几何（生成）。因此，我们需要对网络进行联合训练，使其完成这两项任务。重建引导的因果注意力有助于在注意力层面解耦这两项任务，同时保持解决方案的参数高效性。

至关重要的是，这种因果设计不仅是实现任务分离的机制；它还直接支持高效的两阶段推理过程，并使网络的键值缓存得以

重新诠释为一种完整的隐式三维表示，详见下一节。

3.3. 键值缓存作为隐式三维表示

重建引导的因果注意力还有一项重要特性：它使我们能够将每个注意力块中缓存的键/值词元重新诠释为一种隐式三维表示，即一个编码场景几何与外观、且独立于观察方向的潜在记忆空间。这使我们能够将推理过程解耦为两个不同阶段，如图3 (b) 和 (c) 所示：

阶段 1：重建与缓存。由于源视图词元的注意力计算独立于目标视图词元是否存在，因此模型可以在推理时仅使用源视图重建场景，无需任何目标视图。具体而言，我们在每个全局注意力块中缓存源视图的键和值词元，随后将其作为可复用的场景表示，用于新视图生成。

$$K'_s = \text{Cache}(K_s), \quad V'_s = \text{Cache}(V_s). \tag{4}$$

阶段 2：生成与查询。随后，为了合成新视图，我们只需直接传入目标视图的 Plücker 射线，无需计算源视图的全局注意力和帧注意力，而是直接从缓存中读取：

$$\text{Out}_t = \text{softmax}\left(\frac{Q_t \cdot [K'_s; K_t]^\intercal}{\sqrt{d_k}}\right)[V'_s; V_t], \tag{5}$$

其中 [;] 表示沿令牌维度进行拼接。这种两阶段方法显著降低了计算量，并加快了渲染过程。

经过 L层交错的全局注意力与帧注意力后，目标视图令牌 Out_t 由DPT头 [36]$\mathcal{D}_{\text{RGB}}$ 和 $\mathcal{D}_{\text{pmap}}$ 解码，以预测RGB图像和点图：

$$\hat{\mathbf{I}} = \mathcal{D}_{\text{RGB}}(\text{Out}_t), \quad \hat{\mathbf{P}} = \mathcal{D}_{\text{pmap}}(\text{Out}_t). \tag{6}$$

最后，我们可以通过累积来自多个视点查询的结构来恢复完整的三维结构。

3.4. 训练策略

我们使用多任务损失训练RnG：

$$\mathcal{L} = \mathcal{L}_{RGB} + \lambda_{pmap}\mathcal{L}_{pmap} + \lambda_c\mathcal{L}_{cam}. \tag{7}$$

为监督新视图图像，我们结合使用均方误差损失和感知损失 [20]，作用于真实值$\mathbf{I}_t$ 与渲染视图$\mathbf{I}_t$之间：

$$\mathcal{L}_{RGB} = \left|\mathbf{I}_t - \hat{\mathbf{I}}_t\right|_2 + \lambda_{\text{p}} \cdot \text{Perceptual}\left(\mathbf{I}_t, \hat{\mathbf{I}}_t\right). \tag{8}$$

点云头解码一张四通道图：点ˆ图$\mathbf{P}_t$ 表示可见区域的 xyz位置

以及一张单通道不确定性图 Σ_t，由遵循 [50, 52]的偶然不确定性损失 [22], 进行监督：

$$\begin{aligned}\mathcal{L}_{pmap} = &\left\|\Sigma_t \odot (\mathbf{P}_t - \hat{\mathbf{P}}_t)\right\| \\ &+ \left\|\Sigma_t \odot (\nabla\mathbf{P}_t - \nabla\hat{\mathbf{P}}_t)\right\| - \alpha \cdot \log \Sigma_t.\end{aligned} \tag{9}$$

使用 Huber 损失对输入帧的估计相机位姿 ˆ$\mathbf{g}_s$ 进行监督：

$$\mathcal{L}_{cam} = \sum_s |\mathbf{g}_s - \hat{\mathbf{g}}_s|_\epsilon. \tag{10}$$

此处，将一个批次中的真实相机姿态归一化，使其满足 $\mathbf{g}_{s=1} = [I_{3\times3}\,|[0,0,-1]^\intercal]$。

4. 实验

4.1. 实验细节

数据集。 我们使用三维物体数据集获取多视角 RGBD 渲染。模型训练使用 Objaverse[8] 数据集。我们使用 LVIS 子集和 LGM[44]，筛选得到的列表，共计 113.5K 个物体。评估时，我们使用 Google Scanned Objects (GSO) [9]数据集，其中包含 1030 个物体。我们对三维物体进行归一化，并将其置于世界原点。然后，在物体周围随机放置相机；所有相机均朝向世界原点，但距离各不相同。

对比方法。 我们将本方法与新视角合成和三维重建领域的先进方法进行比较，包括需要将精确相机位姿作为输入的方法 [18, 30, 44] 以及无位姿方法 [30]。由于 LVSM 的物体级模型未公开，我们使用官方实现，在自己的训练数据上训练 LVSM，以确保比较公平。

评估指标。 我们使用重建指标和新视角生成指标评估模型。重建指标在输入视图上进行评估，包括相机位姿和深度。生成指标则在新视点处评估：我们使用光度指标评估新视角的外观，并使用新视角深度图评估几何。此外，我们采用倒角距离评估整体重建三维几何的质量，而非逐视图评估几何。在所有实验中，每个模型均接收从随机位置采样的四个输入视图，并在十个目标视点上进行评估。

4.2. 实现细节

我们从多视图渲染中随机选取 5 张图像，组成一个训练批次（4 个输入视图和 1 个目标视图）。我们以 256×256 分辨率训练 RnG，图块大小为 8。模型在 8 个 A800 GPU 上训练 40K 步，总批大小为 96。我们使用 bfloat16 和梯度检查点来提高训练效率。多任务损失中各项的权重超参数设为

表 2。**定量比较**。我们在 GSO 数据集上评估了所有模型的重建与生成能力。“—”表示模型无法提供该结果。

		重建					生成					完整
		Pose			源视图深度		新视图深度		新视角合成			*3D*
		RA@5↑	RT@5↑	AUC@30↑	Rel↓	a1↑	Rel↓	a1↑	峰值信噪比↑	结构相似性指数↑	学习感知图像块相似度↓	CD↓
已知位姿	LVSM[18]	相机位姿已知			—	—	—	—	**27.522**	**0.902**	**0.0895**	—
	LGM[44]				13.704	85.646	16.804	82.205	16.677	0.751	0.2248	0.1130
	Matrix3D[30]				8.782	95.383	8.897	93.858	19.941	0.803	0.1684	0.0580
	pixelSplat[4]				33.594	33.241	39.819	20.723	12.904	0.718	0.4514	0.4898
无位姿	MUSt3R[3]	21.683	21.683	33.490	10.941	92.663	—	—	—	—	—	0.7757
	VGGT[50]	74.239	65.680	77.234	5.960	97.723	—	—	—	—	—	0.0260
	VGGT[50] + LVSM[18]	74.239	65.680	77.234	5.960	97.723	—	—	19.394	0.796	0.1980	0.0260
	Matrix3D[30]	43.770	65.922	66.389	9.430	92.261	9.964	90.282	18.736	0.786	0.1930	0.0670
	RnG（我们的方法）	**85.146**	**86.019**	**86.942**	**0.584**	**99.929**	**0.717**	**99.850**	26.276	0.891	0.0975	**0.0067**

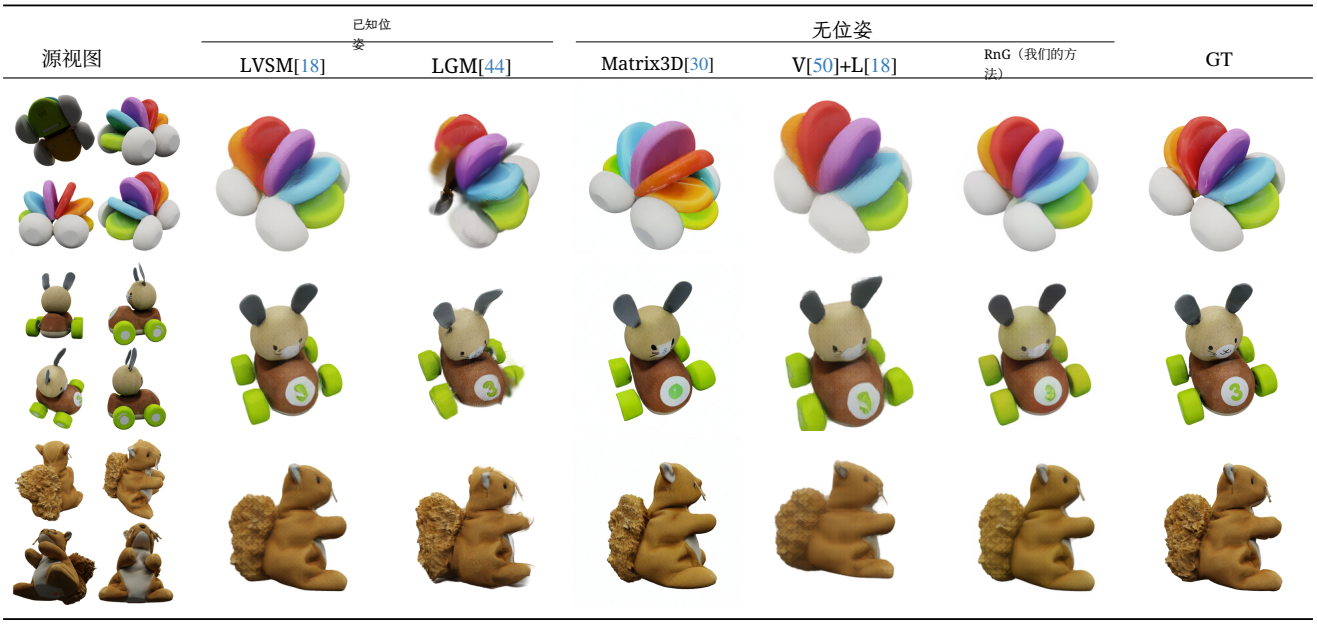


图4. **新视角合成的视觉比较**。尽管 RnG 不需要将准确的姿态作为输入，但其视觉质量可与 LVSM 等最先进的依赖位姿的方法相媲美。我们的模型能够生成未见区域的幻觉，同时保持较高的三维一致性。

$\lambda_{pmap} = 0.2$、 $\lambda_c = 1$ 和 $\lambda_p = 0.5$。不确定性重新加权项 $\alpha = 0.2$。更多训练细节请参阅补充材料。

4.3. 实验结果

4.3.1. 与现有方法比较

我们将 RnG 与最先进的重建和新视角合成方法进行比较。在表2中，我们同时基于源视图重建指标和新视角生成指标评估我们的方法。下面，我们针对每个部分提供详细分析。

输入视角重建。 我们首先评估各方法的重建能力，涵盖输入视角的相机姿态估计和深度预测精度。如表2所示，与统一模型 Matrix3D [30], RnG 在这两项任务中均表现出色。与最先进的三维重建方法 VGGT 相比， [50],RnG 在相机位姿和深度精度方面均有显著优势。

输入视图的重建点云在图5中进行了可视化。RnG 的多视角反投影点云在三维中高度一致。

新视角外观生成。 接下来，我们评估 RnG 的新视图外观生成能力。尽管 RnG 是一种无位姿方法，但其新视角合成性能可与表现最佳的需要位姿的模型 LVSM [18]，相媲美。由于 LVSM 无法直接处理无位姿渲染，我们使用 VGGT [50]获取姿态估计结果，然后将这些结果输入 LVSM [18]进行新视角合成，该方法记为“VGGT + LVSM”。

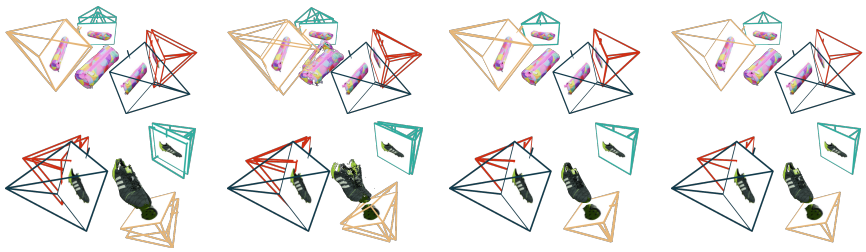


图5. **相机位姿与点云可视化**。 重建结果已归一化以匹配真实值的尺度，并对齐到首帧的位置（深蓝色）。RnG估计的相机位姿与真实值高度对齐。我们从源视图得到的反投影点云不存在分层伪影，呈现出准确的物体结构。

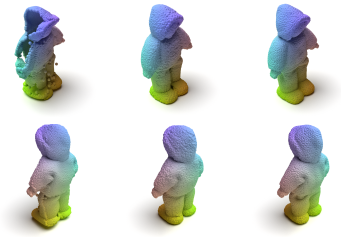


图6. **由4个视图得到的三维结构**。VGGT重建出的观测区域存在分层伪影。RnG能够通过多个视图查询逐步生成完整三维结构。

在所有无位姿流程中，RnG 的照片级真实感和视觉一致性均显著更优。

我们还在图4中展示了新视角合成结果。LGM [44] 根据源图像重建三维高斯，因此，当观测无法覆盖整个三维物体时，新视角也不完整；多视图堆叠高斯也存在分层伪影。Matrix3D [30] 生成的新视角图像“赏心悦目”，但往往无法保持物体三维结构的一致性，还可能臆造不存在的细节。使用 LVSM [18]和 VGGT [50], 估计的相机位姿时，生成的物体往往会发生扭曲。

新视图几何生成。最后，我们通过深度图评估 RnG 的新视图几何生成能力。VGGT [50]等重建方法只能从源视点恢复三维结构，而 Matrix3D [30] 和 RnG 则能够从任意新视图生成深度。值得注意的是，RnG 的深度误差比 Matrix3D 的无位姿版本甚至需要姿态版本低一个数量级。

此外，我们的方法能够生成高质量的多视图融合点云。如表2中报告的倒角距离所示，RnG 在整体三维几何重建方面达到了最先进的性能。这一结果进一步验证了 RnG 生成的点云在不同视点下具有高度一致性。图6 清楚地展示了生成未见几何方面的优势。VGGT 重建的三维结构 [50] 不完整（只能恢复观测区域），并且存在分层伪影。相比之下，RnG 能够为源视点和目标视点生成合理的几何结构，从而使整体三维重建更加完整且结构一致。

总而言之，RnG 在多项三维重建和新视角生成任务中均达到了最先进的性能——
.....指标，包括相机姿态估计、源视图和新视图深度预测、重建的三维几何以及 RGB 渲染。值得注意的是，它甚至能显著超越针对单项任务设计的专用模型。

4.3.2. 泛化到任意输入视图

尽管 RnG 使用固定数量的输入视图进行训练，但它可以直接泛化到任意数量的输入视图。如图 7所示，提供更密集的源视图时，目标视图深度和图像合成质量都会提升。值得注意的是，即使输入视图数量较少，RnG 仍能取得合理的结果。

我们还在图7中提供了可视化结果。随着源图像数量增加、可用观测增多，合成的目标图像呈现出更丰富的细节和更高的视觉保真度。此外，当物体具有对称结构时，即使只有一张源图像，RnG 也能生成不错的新视角。更多实验结果见补充材料。

4.4. 消融研究

我们进一步开展消融研究，以证明模型架构的有效性。结果见表3。

模型扩展。 在 表 3中，“Ours-15K”模型使用包含 44.5K 个物体的 LVIS 子集进行训练，而主模型“ours-40k”则使用完整数据集（113.5K 个物体）进行训练，并增加了训练迭代次数，从而取得更优的性能。这些结果凸显了 RnG 的可扩展性和

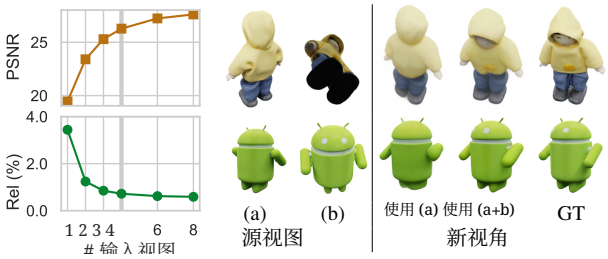


图7. **泛化到不同数量的输入视图**。虽然我们的模型并未经过处理不同数量输入视图的训练，但它对不同数量的源图像仍展现出很强的泛化能力。即使仅提供1张图像，RnG也能取得不错的效果。

表 3. **消融研究**。我们 通过与 LVSM 比较，研究 RnG 的训练效率以及模型架构的有效性 .

模型	重建 Pose RA@5↑	Pose RT@5↑	Pose AUC@30↑	源视图深度 相对↓	源视图深度 a1↑	新视图深度 Rel↓	新视图深度 a1↑	生成 新视角合成 峰值信噪比↑	结构相似性指数↑	学习感知图像块相似度↓
LVSM-100K	不适用							27.522	0.902	0.0895
LVSM-40K								24.619	0.864	0.1544
Ours-40K	85.146	86.019	86.942	0.584	99.929	0.717	99.850	26.276	0.891	0.0975
Ours-15K-scratch	8.252	17.184	33.654	2.331	99.475	2.331	99.364	20.784	0.819	0.2035
Ours-15K	81.650	82.233	82.631	0.742	99.884	0.882	99.788	24.856	0.873	0.1243
Ours-15K-w/o cam	不适用			0.739	99.890	0.878	99.789	24.850	0.873	0.1235
Ours-15K-FullAttn	82.718	81.456	82.686	0.727	99.906	0.892	99.792	24.861	0.874	0.1193

强大的扩展潜力。

训练效率。 我们训练了 LVSM 的两个变体。LVSM-100K 经过 100K 次迭代训练，体现了 LVSM 所能达到的最佳性能。LVSM-40K 经过 40K 步训练，与 RnG 相同，以提供公平的比较。如表3所示，我们的方法-40K 在新视角合成指标上优于 LVSM-40K，表明当训练迭代次数更多时，我们提出的方法有望超越最先进的位姿条件新视角合成方法。此外，尽管训练所用的数据集小得多，我们的方法-15K 的表现仍已超过 LVSM-40K，凸显了利用三维重建先验生成新视角外观和几何的有效性。

重建先验。 我们还训练了一个不使用预训练 VGGT [50] 权重的 RnG 变体，记为 Ours-15K-scratch。与使用预训练权重初始化的模型相比，该模型的性能显著下降。这表明，融入重建先验知识能够有效提升我们的生成任务表现。此外，我们还训练了一个不含相机头的变体（Ours-15K-w/o cam），结果表明，移除相机姿态监督并不会降低新视角合成性能。这意味着，在我们的多任务学习范式中，学习重建相机位姿与生成新视角并不相互冲突。通过估计得到的相机位姿这一额外成果，我们可以轻松地将源图像的纹理映射过去，从而获得更精细的细节。

重建引导的因果注意力。我们的核心设计选择——重建引导的因果注意力——旨在将重建与生成解耦。如第 3.3 节所述，这一设计实现了利用键值缓存的两阶段推理流程，大幅提升了效率（在 A800 GPU 上，推理时从 213 毫秒缩短至 85 毫秒；相比之下，Matrix3D [30] 依赖去噪扩散过程，生成单个新视角需要 27 秒）。详细统计数据见表 4。为验证这种高效架构并未以牺牲准确性为代价，我们通过将

表 4. **采用和不采用 KV- 缓存的模型在推断单个新视图外观和几何方面的效率比较**。

	峰值内存	FLOPS	time
Matrix3D [30]	11.31 G	2228 T	27 s
不使用键值缓存的 RnG	8.12 G	12.26 T	213 毫秒
使用键值缓存的 RnG	7.10 G	2.29 T	85 毫秒

我们的因果掩码替换为完整的双向注意力（Ours-15K-FullAttn）。如表3所示，这一改动带来的性能差异微乎其微。这证实了我们的因果设计在实现显著架构优势的同时，能够确保合理的信息流并支持键值缓存，且不影响性能。

5. 局限性

尽管 RnG 在重建源视图结构以及生成新视图几何和外观方面均展现出领先性能，但它也存在局限性。**细节不足**。 与能够生成引人注目且细节丰富图像的基于扩散的图像生成器不同（尽管生成结果可能有误），RnG 难以呈现精细纹理。这也是 LVSM [18] 面临的问题。引入图像生成预训练或自编码组件或许可以解决这一问题。

世界原点定义。 对于物体级重建任务，理想的世界原点应当是物体中心。然而，在数据准备过程中，所有相机都朝向世界中心，这可能会使网络误将源视图光轴的交点视为世界原点。这也限制了模型在真实世界中的应用，因为使用手持设备拍摄的多视图图像必须朝向同一个三维点，才能获得最佳结果。改进数据处理流程可以解决这一问题。

从多个视点累积三维信息。 尽管我们的模型能够根据视点查询生成高度一致的点图，但仍需要累积多视图观测，以获得完整三维表示。多视图几何整合也可能引入噪声和冲突。

6. 结论

本文提出了 RnG（重建与生成），这是一种新颖的前馈 Transformer，旨在解决可泛化三维重建中因观测不完整而产生的局限。其重建引导的因果注意力机制将 Transformer 的键值缓存用作一种隐式的完整三维表示。因此，RnG 能够根据任意查询位姿，高效预测并渲染高保真的新视角 RGBD 输出，从而将隐式的结构理解有效转化为显式的几何实现。因此，RnG 不仅能够准确重建可见几何，还能成功生成合理、连贯的未见几何与外观。我们的方法在重建和新视图生成基准测试中均达到了最先进的性能，同时效率足以支持实时交互式应用。未来，我们计划扩展 RnG 框架以应对更广泛的场景，包括动态和非刚性内容，使 RnG 成为统一三维重建与生成的通用基础。

致谢

本研究部分得到了国家自然科学基金委员会（62525115、U2570208）的资助。

参考文献

[1] Jianhong Bai、Menghan Xia、Xiao Fu、Xintao Wang、Lian- rui Mu、Jinwen Cao、Zuozhu Liu、Haoji Hu、Xiang Bai、Pengfei Wan 等。Recammaster：基于单个视频的相机控制生成式渲染。*arXiv* 预印本 *arXiv:2503.11647*，2025。3[2] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Sub- biah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakan- tan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是少样本学习者。载于*NeurIPS*，第 1877–1901 页，2020 年。4[3] Yohann Cabon、Lucas Stoffl、Leonid Antsfeld、Gabriela Csurka、Boris Chidlovskii、Jerome Revaud 和 Vincent Leroy。Must3r：用于立体三维重建的多视图网络。载于电气电子工程师学会/计算机视觉基金会计算机视觉与模式识别会议论文集，第 1050 1060 页，2025 年6– 。[4] David Charatan、Sizhe Lester Li、Andrea Tagliasacchi 和 Vincent Sitzmann。PixelSplat：基于图像对的三维高斯泼溅，实现可扩展的可泛化三维重建。载于*CVPR*，第 19457–19467 页，2024 年。3，6[5] Xingyu Chen、Yue Chen、Yuliang Xiu、Andreas Geiger 和 Anpei Chen。Easi3r：从 dust3 中估计解耦运动，*24391*，2025*25033*r，无需训练。*arXiv* 预印本 *arXiv:*。。[6] Yuedong Chen、Haofei Xu、Chuanxia Zheng、Bohan Zhuang、Marc Pollefeys、Andreas Geiger、Tat-Jen Cham 和 Jianfei Cai。Mvsplat：从稀疏多视图图像。在 ECCV 中，第 370–386 页。施普林格，2024。3[7] Yuedong Chen、Chuanxia Zheng、Haofei Xu、Bohan Zhuang、Andrea Vedaldi、Tat-Jen Cham 和 Jianfei Cai。Mvsplat360：从稀疏视图进行前馈式 360度场景合成。在 NeurIPS 中，第 107064–107086 页，2024。3[8] Matt Deitke、Dustin Schwenk、Jordi Salvador、Luca Weihs、Oscar Michel、Eli VanderBilt、Ludwig Schmidt、Kiana Ehsani、Aniruddha Kembhavi 和 Ali Farhadi。Objaverse：一个带标注的三维物体宇宙。在 CVPR 中，第 13142–13153 页，2023。5[9] Laura Downs、Anthony Francis、Nate Koenig、Brandon Kin- man、Ryan Hickman、Krista Reymann、Thomas B McHugh 和 Vincent Vanhoucke。Google Scanned Objects：一个高质量的 3D扫描家居用品数据集。在国际机器人与自动化会议中，第 2553–2560 页。电气电子工程师学会，2022。5, 12[10] Paul Engstler、Aleksandar Shtedritski、Iro Laina、Christian Rupprecht 和 Andrea Vedaldi。Syncity：无需训练的三维世界生成。arXiv 预印本 arXiv: 16420, 2025 2503 1。[11] Sara Fridovich-Keil、Alex Yu、Matthew Tancik、Qinhong Chen、Benjamin Recht 和 Angjoo Kanazawa。Plenoxels：无需神经网络的辐射场。在 CVPR 中，第 5501–5510 页，2022。2[12] Andreas Geiger、Philip Lenz 和 Raquel Urtasun。我们准备好自动驾驶了吗？KITTI视觉基准套件。在 CVPR 中，第 3354–3361 页。电气电子工程师学会，2012。1[13] Jiaxin Guo、Wenzhen Dong、Tianyu Huang、Hao Ding、Ziyi Wang、Haomin Kuang、Qi Dou 和 Yun-Hui Liu。Endo3r：从动态单目内窥镜视频进行统一在线重建。arXiv 预印本 arXiv:2504.03198，2025。3[14] Jisang Han、Honggyu An、Jaewoo Jung、Takuya Narihira、Junyoung Seo、Kazumi Fukuda、Chaehyun Kim、Sunghwan Hong、Yuki Mitsufuji 和 Seungryong Kim。Dˆ 2ust3r：利用动态场景的 4D点图增强三维重建。arXiv 预印本 arXiv:2504.06264，2025。2[15] Wenzel Jakob、S´ebastien Speierer、Nicolas Roussel 和 De- lio Vicini。Dr.jit：用于可微渲染的即时编译器。ACM TOG，41(4)，2022。2[16] Wonbong Jang、Philippe Weinzaepfel、Vincent Leroy、Lour- des Agapito 和 Jerome Revaud。Pow3r：借助相机与场景先验，赋能不受约束的三维重建。在 CVPR 中，第 1071–1081 页，2025。3[17] Hanwen Jiang、Hao Tan、Peng Wang、Haian Jin、Yue Zhao、Sai Bi、Kai Zhang、Fujun Luan、Kalyan Sunkavalli、Qixing Huang 等。Rayzer：一种自监督大型视图合成模型。arXiv 预印本 arXiv:2505.00702，2025。2, 3[18] Haian Jin、Hanwen Jiang、Hao Tan、Kai Zhang、Sai Bi、Tianyuan Zhang、Fujun Luan、Noah Snavely 和 Zexiang Xu。LVSM：一种具有极少 3D归纳偏置的大型视图合成模型。在 ICLR，2025。2, 3, 5, 6, 7, 8[19] Linyi Jin、Richard Tucker、Zhengqi Li、David Fouhey、Noah Snavely 和 Aleksander Holynski。Stereo4d：从互联网立体视频中学习物体在三维空间中的运动。在 CVPR 中，第 10497 10509, 2025 3 – 页。Justin Johnson、Alexandre Alahi 和李飞飞。用于实时风格迁移和超分辨率的感知损失。在 ECCV 中，第 694–711 页。施普林格，2016。5[20]

[21] Bingxin Ke、Anton Obukhov、Shengyu Huang、Nando Met- zger、Rodrigo Caye Daudt 和 Konrad Schindler。重新利用基于扩散的图像生成器进行单目深度估计。在 CVPR 中，第 9492–9502 页，2024。2, 3[22] Alex Kendall 和 Yarin Gal。计算机视觉中的贝叶斯深度学习需要哪些不确定性？在 NeurIPS，2017。5[23] Bernhard Kerbl、Georgios Kopanas、Thomas Leimk¨uhler 和 George Drettakis。用于实时辐射场渲染的 3D 高斯泼溅。ACM TOG，42(4):139–1，2023。2[24] Xin Kong、Shikun Liu、Xiaoyang Lyu、Marwan Taher、Xi- aojuan Qi 和 Andrew J Davison。Eschernet：一种用于可扩展视图合成的生成模型。在 CVPR 中，第 9503–9513 页，2024。3[25] Samuli Laine、Janne Hellsten、Tero Karras、Yeongho Seol、Jaakko Lehtinen 和 Timo Aila。用于高性能可微渲染的模块化基元。ACM TOG，39 (6):1–14，2020。2[26] Yushi Lan、Yihang Luo、Fangzhou Hong、Shangchen Zhou、Honghua Chen、Zhaoyang Lyu、Shuai Yang、Bo Dai、Chen Change Loy 和 Xingang Pan。Stream3r：基于因果Transformer的可扩展序列式三维重建。arXiv 预印本 arXiv:2508.10893，2025。3[27] Ruining Li、Chuanxia Zheng、Christian Rupprecht 和 Andrea Vedaldi。DSO：通过仿真反馈对齐三维生成器，以确保物理合理性。arXiv 预印本 arXiv:2503.22677，2025。1[28] Ruoshi Liu、Rundi Wu、Basile Van Hoorick、Pavel Tok- makov、Sergey Zakharov 和 Carl Vondrick。Zero-1-to-3：从单张图像零样本生成三维物体。在 ICCV 中，第 9298–9309 页，2023。3[29] Jiahao Lu 、 Tianyu Huang 、Peng Li、Zhiyang Dou、Cheng Lin、Zhiming Cui、Zhen Dong、Sai-Kit Yeung、Wenping Wang 和 Yuan Liu。Align3r：面向动态视频的对齐单目深度估计。在 CVPR 中，第 22820–22830 页，2025。3[30] Yuanxun Lu、Jingyang Zhang、Tian Fang、Jean-Daniel Nah- mias、Yanghai Tsin、Long Quan、Xun Cao、Yao Yao 和 Shiwei Li。Matrix3d：一体化大型摄影测量模型。在 CVPR 中，第 11250–11263 页，2025。1, 2, 3, 5, 6, 7, 8[31] Xuan Luo、Jia-Bin Huang、Richard Szeliski、Kevin Matzen 和 Johannes Kopf。一致性视频深度估计。ACM TOG，39(4):71–1，2020。1[32]Ben Mildenhall、Pratul P Srinivasan、Matthew Tancik、Jonathan T Barron、Ravi Ramamoorthi 和 Ren Ng。NeRF：将场景表示为用于视图合成的神经辐射场。ACM通信，65(1):99–106，2021。2[33] Takeru Miyato、Bernhard Jaeger、Max Welling 和 Andreas Geiger。GTA：一种用于多视图Transformer的几何感知注意力机制。在 ICLR，2024。3[34] Thomas M¨uller、Alex Evans、Christoph Schied 和 Alexan- der Keller。采用多分辨率哈希编码的即时神经图形基元。ACM TOG，41(4):1–15，2022。2[35] Maxime Oquab、Timoth´ee Darcet、Th´eo Moutakanni、Huy Vo、Marc Szafraniec、Vasil Khalidov、Pierre Fernandez、Daniel Haziza、Francisco Massa、Alaaeldin El-Nouby 等。Dinov2：无需监督地学习鲁棒视觉特征。arXiv 预印本 arXiv: *2304 07193*, 2023 4, 12 。[36] Rene Ranftl、Alexey Bochkovskiy 和 Vladlen Koltun。用于密集预测的视觉Transformer。收录于 ICCV，第 12179–12188 页，2021 年。4, 5[37] Jiahui Ren、Mochu Xiang、Jiajun Zhu 和 Yuchao Dai。Panosplatt3r：利用透视预训练实现通用无位姿宽基线全景重建。arXiv 预印本 arXiv:2507.21960，2025 年。3[38]Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bj¨orn Ommer。使用潜在扩散模型进行高分辨率图像合成。收录于 CVPR，第 10684–10695 页，2022 年。3[39] Li Ruilong、Yi Brent、Liu Junchen、Gao Hang、Ma Yi 和 Kanazawa Angjoo。相机作为相对位置编码。arXiv 预印本 arXiv:2507.10496，2025 年。3[40]Mehdi SM Sajjadi、Henning Meyer、Etienne Pot、Urs Bergmann、Klaus Greff、Noha Radwan、Suhani Vora、Mario Lu˘ci´c、Daniel Duckworth、Alexey Dosovitskiy 等。场景表示Transformer：通过集合潜在场景表示实现无几何的新视角合成。收录于 CVPR，第 6229–6238 页，2022 年。2, 3[41] Mehdi SM Sajjadi、Aravindh Mahendran、Thomas Kipf、Eti- enne Pot、Daniel Duckworth、Mario Lu˘ci´c 和 Klaus Greff。Rust：从无位姿图像中获取潜在神经场景表示。收录于 CVPR，第 17297–17306 页，2023 年。3[42] Brandon Smart、Chuanxia Zheng、Iro Laina 和 Vic- tor Adrian Prisacariu。Splatt3r：从未标定图像对中进行零样本高斯泼溅。arXiv 预印本 arXiv:2408.13912，2024 年。3[43] Mohammed Suhail、Carlos Esteves、Leonid Sigal 和 Ameesh Makadia。可泛化的基于图像块的神经渲染。收录于 ECCV，第 156–174 页。施普林格，2022 年。3[44] Jiaxiang Tang、Zhaoxi Chen、Xiaokang Chen、Tengfei Wang、Gang Zeng 和 Ziwei Liu。LGM：用于高分辨率三维内容创建的大型多视图高斯模型。收录于欧洲计算机视觉会议，第 1–18 页。施普林格，2024 年。2, 5, 6, 7[45] Shengji Tang、Weicai Ye、Peng Ye、Weihao Lin、Yang Zhou、Tao Chen 和 Wanli Ouyang。HiSplat：用于可泛化稀疏视图重建的分层三维高斯泼溅。arXiv 预印本 arXiv:2410.06245，2024 年。3[46] Zhenggang Tang、Yuchen Fan、Dilin Wang、Hongyu Xu、Rakesh Ranjan、Alexander Schwing 和 Zhicheng Yan。Mv-DUSt3R+：2 秒内从稀疏视图进行单阶段场景重建。收录于 CVPR，第 5283–5293 页，2025 年。3[47] Ayush Tewari、Justus Thies、Ben Mildenhall、Pratul Srinivasan、Edgar Tretschk、Wang Yifan、Christoph Lassner、Vincent Sitzmann、Ricardo Martin-Brualla、Stephen Lombardi 等。神经渲染进展。收录于 Computer Graphics Forum，第 703–735 页。Wiley Online Library，2022 年。2[48]Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timoth´ee Lacroix、Baptiste Rozi`ere、Naman Goyal、Eric Hambro、Faisal Azhar 等。Llama：开放且高效的基础语言模型。arXiv 预印本 arXiv:2302.13971，2023 年。4[49] Khiem Vuong、Anurag Ghosh、Deva Ramanan、Srinivasa Narasimhan 和 Shubham Tulsiani。AerialMegaDepth：

空地重建与视图合成。收录于 CVPR，第 21674–21684 页，2025 年。3[50] Jianyuan Wang、Minghao Chen、Nikita Karaev、Andrea Vedaldi、Christian Rupprecht 和 David Novotny。VGGT：视觉几何基础Transformer。收录于 CVPR，第 5294–5306 页，2025 年。1, 2, 3, 4, 5, 6, 7, 8, 12 [51] Qianqian Wang、Yifei Zhang、Aleksander Holynski、Alexei A Efros 和 Angjoo Kanazawa。具有持久状态的连续三维感知模型。收录于 CVPR，第 10510–10522 页，2025 年。3[52] Shuzhe Wang、Vincent Leroy、Yohann Cabon、Boris Chidlovskii 和 Jerome Revaud。DUSt3R：让三维几何视觉变得简单。收录于 CVPR，第 20697–20709 页，2024 年。1, 2, 3, 5[53] Yifan Wang、Jianjun Zhou、Haoyi Zhu、Wenzheng Chang、Yang Zhou、Zizun Li、Junyi Chen、Jiangmiao Pang、Chunhua Shen 和 Tong He。 π^3：可扩展的置换等变视觉几何学习。arXiv 预印本 arXiv:2507 13347，2025 年 3 . . [54] Jianfeng Xiang、Zelong Lv、Sicheng Xu、Yu Deng、Ruicheng Wang、Bowen Zhang、Dong Chen、Xin Tong 和 Jiaolong Yang。用于可扩展、多用途三维生成的结构化三维潜变量。收录于 CVPR，第 21469–21480 页，2025 年。1[55] Haofei Xu、Songyou Peng、Fangjinhua Wang、Hermann Blum、Daniel Barath、Andreas Geiger 和 Marc Pollefeys。DepthSplat：连接高斯泼溅与深度。收录于 CVPR，第 16453–16463 页，2025 年。3[56] Jiale Xu、Shenghua Gao 和 Ying Shan。FreeSplatter：用于稀疏视图三维重建的无姿态高斯泼溅。收录于 ICCV，第 25442–25452 页，2025 年。3[57] Jiayu Yang、Ziang Cheng、Yunfei Duan、Pan Ji 和 Hongdong Li。ConsistNet：为多视图图像扩散强制施加三维一致性。收录于 CVPR，第 7079–7088 页，2024 年。3[58] Jianing Yang、Alexander Sax、Kevin J Liang、Mikael Henaff、Hao Tang、Ang Cao、Joyce Chai、Franziska Meier 和 Matt Feiszli。Fast3R：探索通过一次前向传播重建 1000+ 图像的三维结构。收录于 CVPR，第 21924–21935 页，2025 年。3[59] Xu Yang、Hanwang Zhang、Guojun Qi 和 Jianfei Cai。面向视觉语言任务的因果注意力。收录于 CVPR，第 9847–9857 页，2021 年。4[60] Botao Ye、Sifei Liu、Haofei Xu、Xueting Li、Marc Pollefeys、Ming-Hsuan Yang 和 Songyou Peng。无需姿态，不成问题：从稀疏无位姿图像中生成出乎意料地简单的三维高斯泼溅。收录于 ICLR，2024 年。3[61] Brent Yi、Chung Min Kim、Justin Kerr、Gina Wu、Rebecca Feng、Anthony Zhang、Jonas Kulhanek、Hongsuk Choi、Yi Ma、Matthew Tancik 等。Viser：使用 Python 进行命令式、基于网页的三维可视化。arXiv 预印本 arXiv:2507 22885，2025 年 12 . . [62] Alex Yu、Vickie Ye、Matthew Tancik 和 Angjoo Kanazawa。pixelNeRF：从一张或少数几张图像中构建神经辐射场。收录于 CVPR，第 4578–4587 页，2021 年。3[63] Chong Zeng、Yue Dong、Pieter Peers、Hongzhi Wu 和 Xin Tong。RenderFormer：使用全局光照对三角网格进行基于Transformer的神经渲染。收录于论文集

计算机图形学与交互技术特别兴趣小组会议论文集，第 1–11 页，2025 年。2[64] Junyi Zhang、Charles Herrmann、Junhwa Hur、Varun Jampani、Trevor Darrell、Forrester Cole、孙德庆和 Ming-Hsuan Yang。《Monst3r：一种在存在运动时估计几何的简单方法》。载于*ICLR*，2024 年。3[65] 张启航、翟双飞、Miguel Angel Bautista Martin、Kevin Miao、Alexander Toshev、Joshua Susskind 和顾佳涛。《通过显式 3D 建模实现世界一致性视频扩散》。载于 *CVPR*，第 21685–21695 页，2025 年。3[66] Jensen Zhou、Hang Gao、Vikram Voleti、Aaryaman Vasishta、Chun-Han Yao、Mark Boss、Philip Torr、Christian Rupprecht 和 Varun Jampani。《稳定虚拟摄像机：利用扩散模型进行生成式视图合成》。 *arXiv* 预印本 *arXiv:2503.14489*，2025 年。3[67] Haoyi Zhu、Yifan Wang、Jianjun Zhou、Wenzheng Chang、Yang Zhou、Zizun Li、Junyi Chen、Chunhua Shen、Jiangmiao Pang 和 Tong He。《 Aether：几何感知统一世界建模》。载于 *ICCV*，第 8535–8546 页，2025 年。3[68] 卓东、郑文昭、郭嘉禾、吴宇琦、周杰和陆继文。《流式四维视觉几何变换器》。 *arXiv* 预印本 *arXiv:2507.11539*，2025 年。4

RnG：基于部分观测进行完整三维建模的统一 Transformer

补充材料

7. 训练细节

对于单个物体，我们在半径各不相同的 25 个视点上渲染RGBD图像。为了构建一个训练批次，我们从多视图渲染中随机采样 7 张图像。前 4 张图像将用作源视图，其余 3 张则是不同的目标视图。随后，我们将批大小扩展 3 倍，使每个训练批次都包含相同的 4 张源图像和 1 张目标图像。

训练批次大小为 6（个物体），我们每累积 2 个前向步后进行一次更新，以增大批大小。最大学习率设为 6e-4。前 3000 步采用学习率预热，训练过程的其余阶段则采用余弦学习率衰减。单图像特征提取器（DINO 视觉 Transformer [35]（位于 VGGT 中） [50]）的参数在训练期间保持冻结。

8. 恢复完整三维

8.1. 恢复上方向

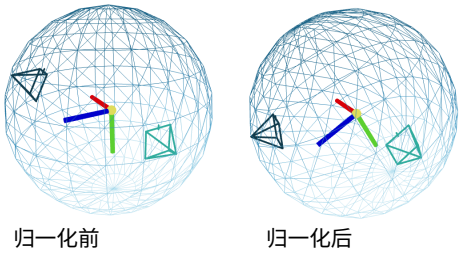


图8。**相机归一化过程**。第一个相机位姿将被归一化到单位球上，并变换到一个固定位置。这会给其他相机引入额外的滚转角。

如图8所示，训练样本会经过相机归一化过程，其中第一个相机被缩放并变换到 $[I_{3\times3}\,|[0,0,-1]^\intercal]$。这将改变物体在世界坐标系中的“上方向”。当从新视角进行查询时，如果我们直接使用预定义的相机位置，就会引入额外的“滚转角”。换言之，世界坐标系的“上方向”并不与所观测物体的上方向对齐，而是带有一个额外的滚转角。这超出了训练样本的分布，因此在模型预测相机位姿之后，我们必须恢复物体的上方向（如果可能）。

我们假设所有输入图像的上方向都与物体的上方向对齐，那么目标就是找到沿 x 轴旋转（图 8中的红色轴），使所有相机都没有额外的“滚转角”。一种简单的解决方案是搜索能使所有滚转角之和最小的最佳仰角。

8.2. 累积视图

从网络中提取完整三维时，我们可以使用用户界面手动选择视点，以反投影估计的三维点（如图9所示）。我们也可以在球体上均匀放置相机以获得新视角。我们利用估计的置信度图过滤估计点，以去除尖锐边界周围的飞点。该用户界面使用 Viser 实现 [61]。

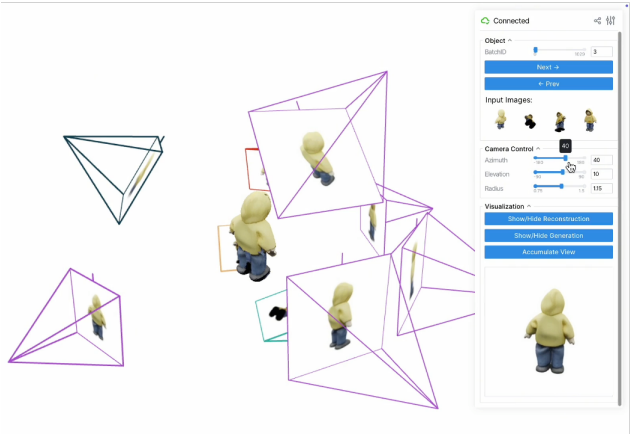


图9. 用于可视化新视角图像并从多个视角累积完整三维的用户界面。

9. 更多实验结果

9.1. 泛化到任意输入视图

在4.3.2 节正文中，我们证明了所提出的方法能够泛化到任意输入视图。在本节中，我们提供实验设置的更多细节，以及更全面的定量和定性结果。具体而言，对于评估数据集GSO中的每个三维物体 [9], 我们在25个不同位置渲染RGBD帧。我们随机选取10帧用于评估，源图像则从其余15个视图中选取。我们的主要比较使用4个随机源视图。为了在给定不同数量的源视图时保持指标可比，我们在这4个视图的基础上删减或追加，并保持目标视图固定。

表 5. **泛化到其他数量的输入视图**。尽管我们的模型并未经过处理其他数量输入视图的训练，但它对其他数量的源图像仍表现出很强的泛化能力。

# 输入视图	重建 Pose RA@5↑	RT@5↑	AUC@30↑	输入视图深度 Rel↓	a1↑	新视角深度 Rel↓	a1↑	生成 新视角合成 峰值信噪比↑	结构相似性指数↑	学习感知图像块相似度↓
8	85.922	86.893	87.942	0.535	99.927	0.593	99.875	27.608	0.903	0.0857
6	85.825	86.505	87.741	0.539	99.935	0.622	99.876	27.235	0.899	0.0891
4	85.146	86.019	86.942	0.584	99.929	0.717	99.850	26.276	0.891	0.0975
3	85.243	86.311	86.136	0.662	99.931	0.848	99.820	25.291	0.881	0.1063
2	83.010	87.961	85.405	0.931	99.906	1.243	99.670	23.403	0.860	0.1249
1	不适用			3.328	99.160	3.443	98.121	19.452	0.811	0.1821

详细的指标结果见表5。如图所示，当提供更密集的源视图时，目标视图深度和图像合成质量均有所提升。值得注意的是，即使在两个输入视图的设置下，RnG仍能取得合理的结果。例如，与使用四个输入视图的标准训练设置相比，新视角外观的峰值信噪比仅下降7.5%。其他指标也呈现出类似趋势，进一步验证了所提出方法的强泛化能力。

我们还在图12中提供了可视化结果。如前几行所示，模型在只有一个输入视图时无法生成合理的结果，但使用两个输入视图后，性能显著提升。随着源图像数量增加、可用观测增多，合成的目标图像呈现出更丰富的细节和更高的视觉保真度。我们还在图13中提供了更多可视化结果。在这些情况下，即使只提供一个源视图，模型也能取得不错的效果。此外，还有一些棘手的情况，其中关键结构被遮挡，或结构存在歧义。我们在图14中展示了这些样例，此时 RnG 的表现“相当糟糕”。当有足够多的观测可用时，RnG 在这些情况下的表现会有所提升，如图15所示。

9.2. 泛化到真实世界输入

CO3D 上的泛化结果如图10所示。与 LVSM 相比，RnG 展现出更好的零样本性能，并且与使用 CO3D 数据训练的 Matrix3D 相当。由于第5节讨论的世界原点歧义，预测结果可能与真实值未对齐；改进训练数据有助于缓解这一问题
准备。



图10. CO3D 上的定性结果。

9.3. 泛化到多物体场景

RnG 无需任何修改即可自然泛化到多物体场景。如图11所示，GSO 数据集中的若干示例支持了这一结论。我们认为，针对更复杂的场景进行系统评估，对未来探索很有价值。



图 11. RnG 在多物体场景中的新视角合成结果。

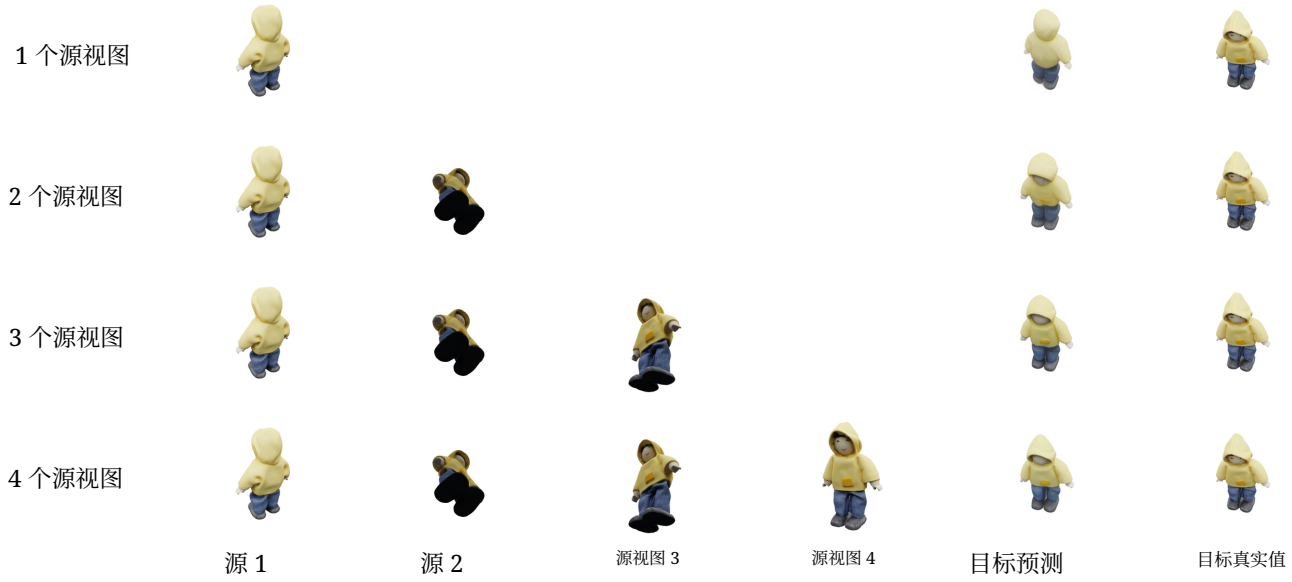


图12。模型获得不同数量输入视图时的评估协议。每一行代表一种输入配置。

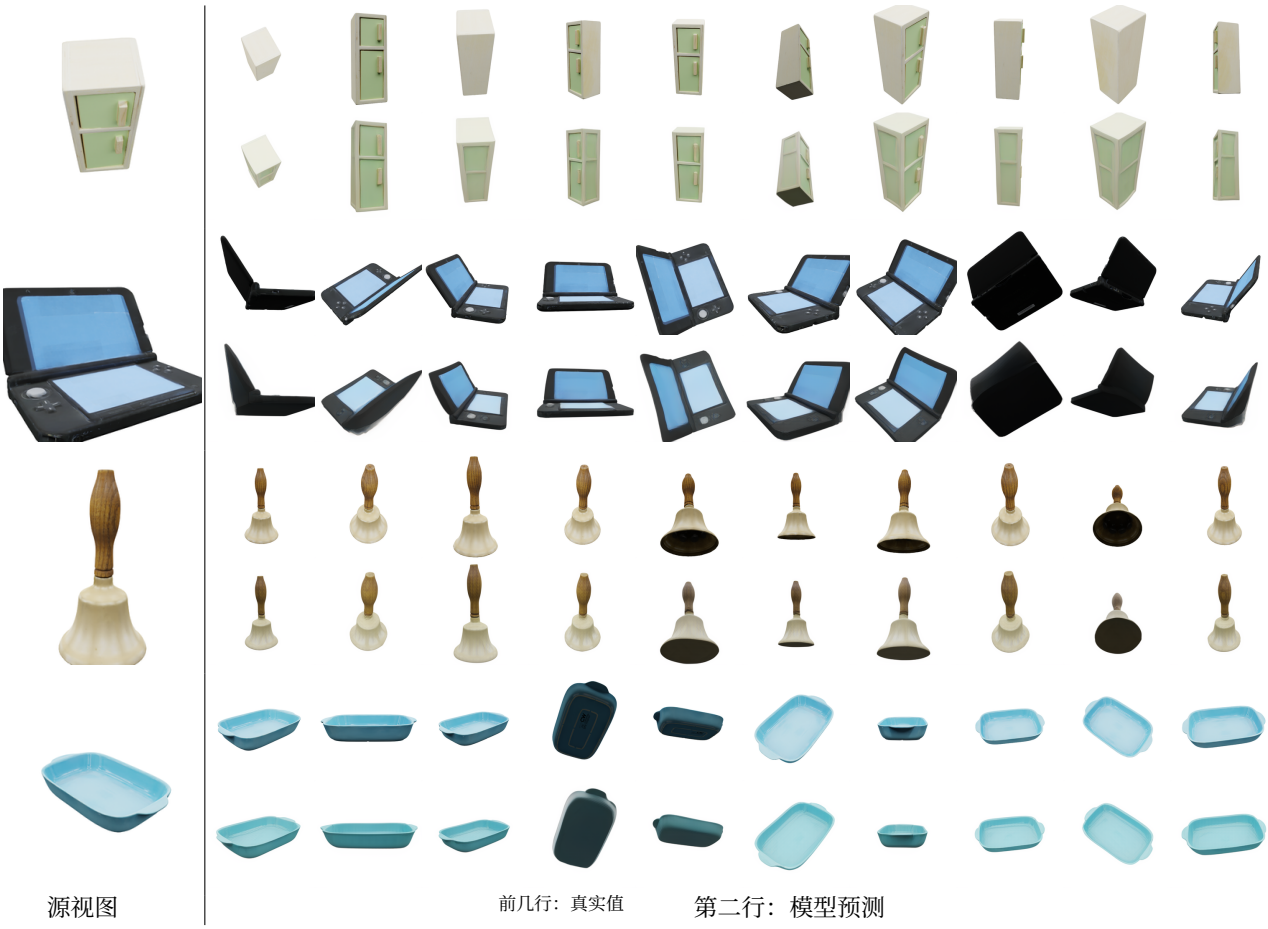


图 13. 即使只提供一张图像，RnG 也能表现出色。

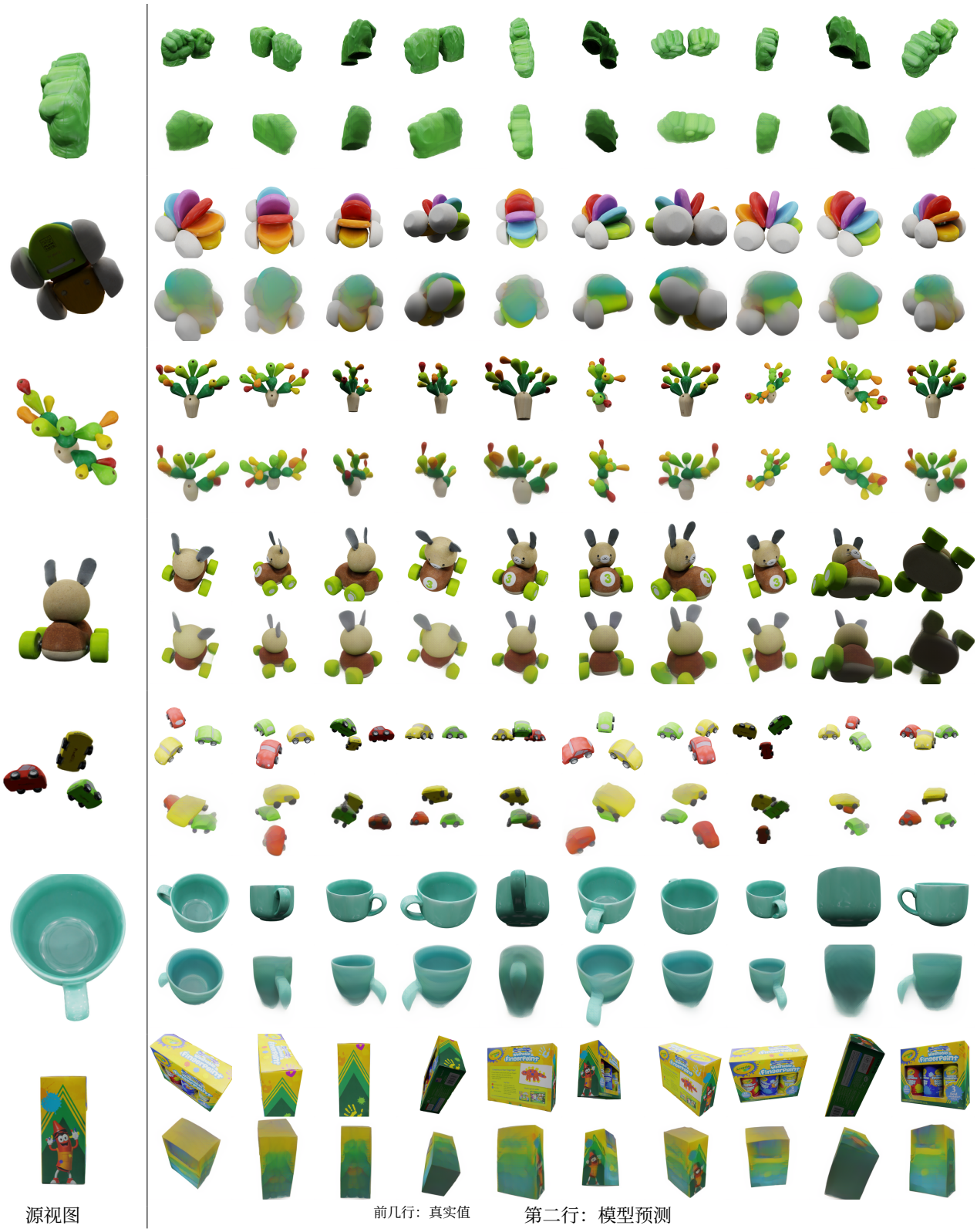


图14. 仅根据一张源图像生成新视角难度过大时，RnG 的表现相当差。

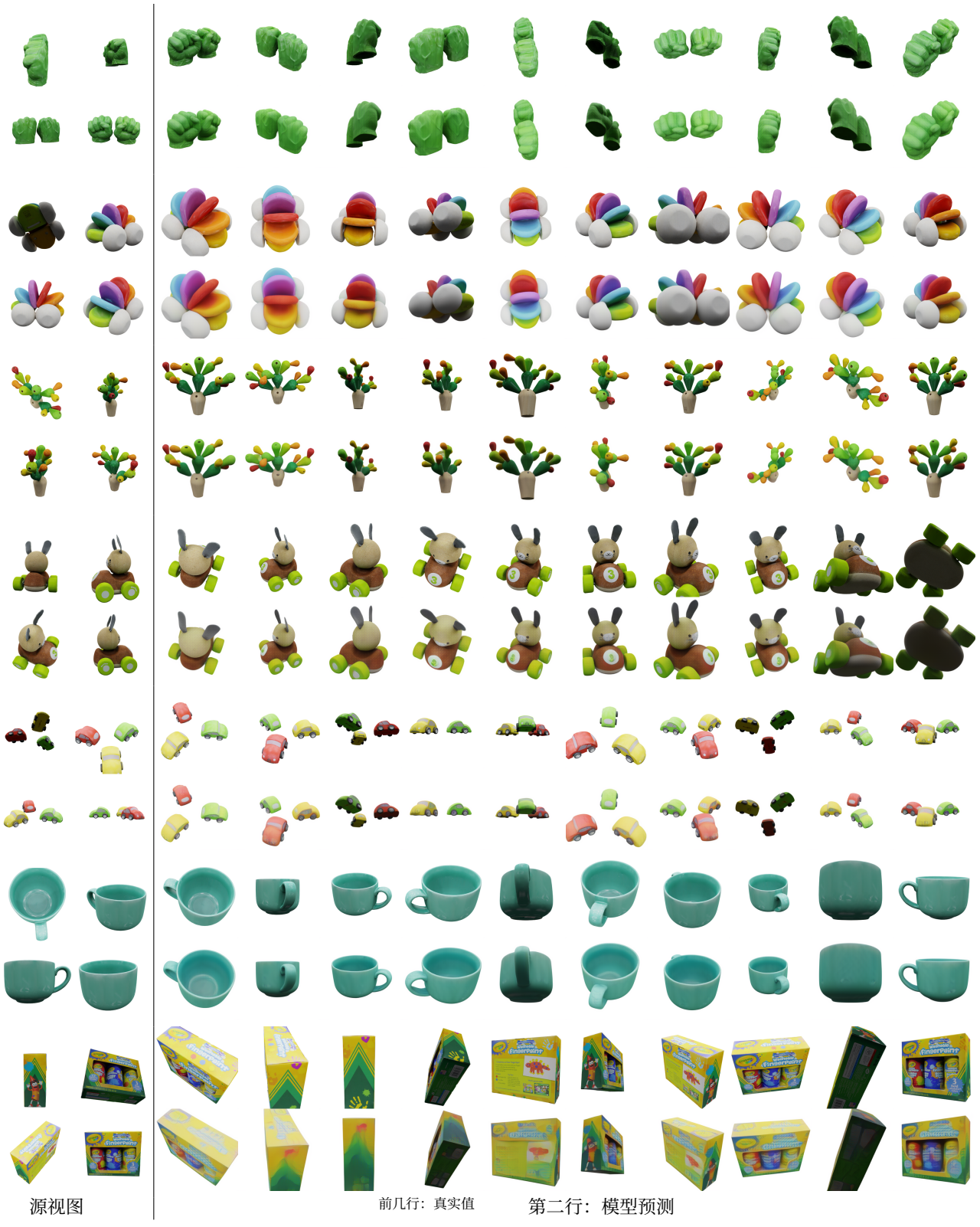


图15. 消除歧义后，RnG 的表现更好。