arXiv:2303.01869v3 [cs.CV] 9 Apr 2023

基于物体中心预测模型的内在物理概念发现

Qu Tang [1,2⋆] Xiangyu Zhu [1,2⋆] Zhen Lei [1,2,3] Zhaoxiang Zhang [1,2,3,*1]中国科学院大学人工智能学院[2]中国科学院自动化研究所 MAIS[3]中国科学院香港创新研究院人工智能与机器人中心

{tangqu2020,zhaoxiang.zhang}@ia.ac.cn, {xiangyu.zhu,zlei}@nlpr.ia.ac.cn

摘要

通过观察来发现抽象物理概念并理解它们如何在世界中发挥作用的能力，是人类智能的核心。这种能力的获得，建立在以无监督的方式将环境组合性地感知为物体及其关系的基础上。近期方法学习以物体为中心的表征，并捕捉物体中可通过视觉观察到的概念，例如形状、大小和位置。本文在此基础上更进一步，尝试发现并表征质量和电荷等内在物理概念。我们提出物理概念推断网络（*PHYCINE*），这是一个能够在无监督条件下推断不同抽象层级物理概念的系统。*PHYCINE* 的核心见解有两点：常识知识随着预测而涌现；不同抽象层级的物理概念应以自底向上方式进行推理。实证评估表明，我们的系统所推断的变量符合相应物理概念的属性。我们还证明，包含所发现物理概念变量的物体表征有助于提升因果推理任务（例如 *ComPhy*）中的表现。

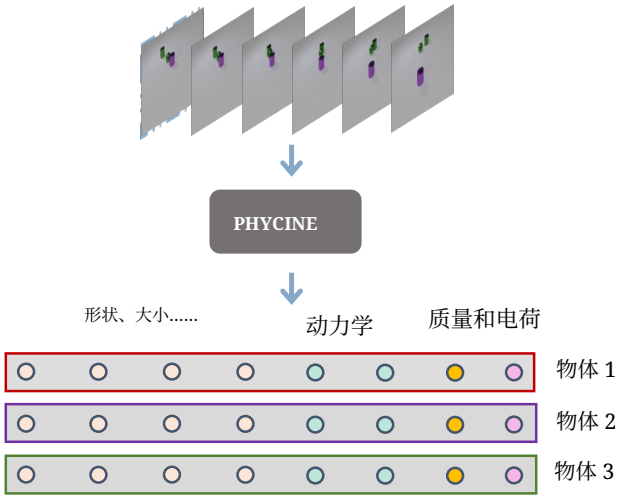


图1. 给定来自 ComPhy [6] 数据集的视频，PHYCINE 会将场景分解为多物体表征，其中包含不同抽象层级的物理概念，包括视觉属性、动力学、质量和电荷。（以不同颜色显示）。

1. 引言

物体为什么会在碰撞后弹开？磁铁为什么会相互吸引或排斥？物体包含许多复杂的物理概念，这些概念决定了它们如何与世界互动 [6]。人类能够仅通过观察发现关于世界运行方式的抽象概念。在一段保存下来的视频中，有些概念可以从物体的视觉外观中明显看出，例如位置、大小、速度等；而另一些概念则隐藏在物体的行为中。例如，内在物理概念例如质量和电荷，无法从静态场景中观察到，只能通过物体动力学发现。携带相同电荷或相反电荷的物体会相互施加排斥力或吸引力。碰撞后，较轻物体的运动变化将大于质量较大的物体。机器如何才能学会揭示并表征这类常识知识？我们认为，答案在于两个关键步骤：将世界分解为以物体为中心的表征，并构建一个以抽象物理概念作为潜变量、用于处理不确定性的预测性模型。

以物体为中心的表征学习 [4, 12, 22, 26, 27,33]旨在以结构化的方式感知世界，从而提升智能系统的泛化能力并实现更高层次的认知。基于变分自编码器 [17]的模型，

*Corresponding author。⋆Equal contribution

例如 IODINE [12]能够学习解耦特征，以统一的形式分离出可解释的物体属性（例如形状、颜色和位置）。然而，质量和电荷等抽象概念无法通过生成模型提取，因为构建物体级表征并不一定会对这些更高层次的抽象物理进行建模。CPL [6]成功地利用图网络学习高层概念，但依赖于来自真实物体级概念标签的监督信号。另有一些研究探讨了以物体为中心的表征在学习预测模型、解决预测和规划任务方面的有效性 [14, 18, 27, 28]。然而，据我们所知，目前尚无研究尝试以无监督的方式发现并表征物体级的内在物理概念。

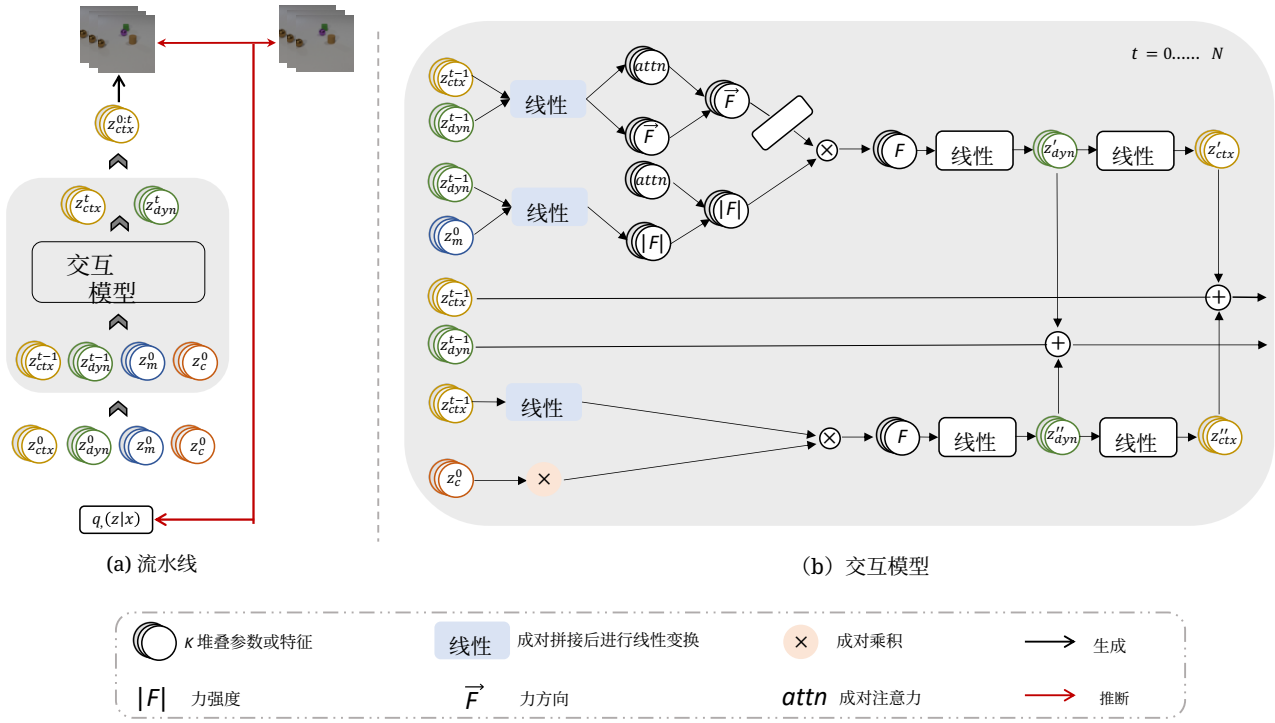
数十年来，认知科学、神经科学和人工智能领域以多种方式提出了通过预测学习抽象概念的观点 [23, 24]。直观而言，速度、质量和电荷等物理概念，可能会随着系统在物体表征层级上接受长期预测训练而逐渐涌现 [20]。通过时间跨度越来越长的预测，系统或许能够以自底向上的方式，逐步习得关于世界运作方式的越来越复杂的概念。本文聚焦于主要挑战：使模型能够表征并解耦逐步显现的概念。我们遵循常识：借助神经物理引擎，如果对物体轨迹的预测失败，就必然有尚未捕捉到的物理概念。因此，能够成功建模预测不确定性的潜变量定义了一种新的物理概念，从而可以构建更好的物理引擎。基于这一思路，我们将物理概念划分为三个抽象层级：外在概念、动态概念和内在概念。首先，外在属性（如颜色、形状、材料、大小、深度和位置）可视为属于最低抽象层级的物体上下文，感知模块能够直接对这些上下文进行编码。其次，中间层级的动态属性（如速度）隐藏在视觉特征的时空关系中，应通过短期预测加以推断。第三，质量和电荷等内在属性既无法直接观测，也无法通过短期预测推断。只有通过分析物体彼此施力的方式，才能推断这些属性。例如，推断质量需要纳入碰撞事件，而推断电荷则需要观测物体动力学的变化，这依赖于长期观测或预测。

在这项工作中，我们构建了一个名为物理概念推断网络（PHYCINE）的系统。该系统采用自底向上的金字塔结构，包含不同抽象层级的物理概念特征。这些特征协同进行重建和预测，以解释所观察到的世界。首先，物体上下文特征利用生成模型重建观测图像。其次，物体动力学特征通过学习状态转移函数，预测下一步的物体上下文。最后，质量与电荷特征对物体之间的交互进行建模。PHYCINE 使用关系模块计算所有实体之间的成对作用力，并通过适当的正则化自适应地学习表示物体质量和电荷的变量。在训练过程中，所有表征均随机初始化，并利用当前参数估计所得证据下界（ELBO）的梯度信息进行迭代优化。该模型仅使用原始输入视频数据，即可通过端到端方式进行训练。如图1所示，以原始视频作为输入，PHYCINE 不仅提取外在的物体上下文（即大小、形状、颜色、位置和材料），还推断出更抽象的概念（即物体动力学、质量和电荷）。在实验中，我们展示了该模型发现并解缠物理概念的能力，这些概念可用于解决下游任务。我们在 ComPhy（一项因果推理基准）上评估了学习到的表征。

我们的主要贡献如下：(i) 我们提出了一个通过观察原始视频来发现物理概念的挑战性问题，并成功发现了速度、质量和电荷等内在概念。(ii) 我们提出了 PHYCINE 框架，这是一种层次化以物体为中心的预测模型，可从低抽象层级（如颜色、形状）推断至高抽象层级（如质量、电荷）的物理概念，从而获得解耦的物体级物理表征。(iii) 我们展示了 PHYCINE 在 ComPhy 上所学习到的表征的有效性。

2. 相关工作

以物体为中心的表征学习。 近期有几项研究工作在没有明确监督的情况下，从视觉输入中学习以物体为中心的表征。对于静态场景，一种常见框架是通过某种形式的自编码器生成一组物体表征向量或槽向量。MONet [4] 首先使用注意力模块对图像进行分割，然后通过迭代地解释并重构场景来完成处理。与我们的工作密切相关的 IODINE [12], 使用摊销变分推断来优化后验分布参数。Slot-Attention [22] 学习一种分组策略，将输入特征分解为一组槽表征。此外，许多研究人员还采用已被证明对静态图像有效的架构，从视频中提取物体级表征。ViMON [29] 通过加入门控循环单元来聚合时间信息，将 MONet 扩展到视频领域。类似地，Steve [26] 使用以槽为条件的基于Transformer的图像解码器，将槽注意力应用于更复杂、更具自然场景特征的视频。



(a) 流水线 (b) 交互模型

图2。(a) PHYCINE 示意图。给定一段包含 $x^0 \cdot x^t$图像的视频，我们旨在获取每个物体的四种物理嵌入，包括上下文 z_{ctx}、动力学 z_{dyn}、质量 z_m 和电荷 z_c。首先，从任意后验分布中采样时间步 0 的表征。然后， z^0_{ctx} 用于重建图像 x^0，交互模型以 z^0_{ctx}、 z^0_{dyn}、 z^0_m和 z^0_c 为输入，迭代输出 z^t_{ctx}。如此便可渲染出一段视频。图像生成过程中的梯度信息用于迭代优化后验分布，最终可准确表征所有当前物体的物理概念。(b) 交互模型：利用所有指向同一共享张量的相同节点，对碰撞力和电荷力进行建模。

以物体为中心的预测模型。 以物体为中心的模型被用于实现多物体三维场景预测，无论是否进行动作条件化。PROVIDE [33] 利用二维长短期记忆网络将 IODINE 拟合到视频中，并可用于分别预测每个物体的轨迹。ODDN[27] 通过显式提取物体动力学并建模成对物体之间的关系，实现了包含交互的预测。在 C-SWM [18], 中，提出使用对比学习来学习潜在空间中的转移模型。OP3 [28] 使用 IODINE 作为推断模型，并进行动作条件化，从而实现以物体为中心的预测与规划。

专注于开发用于建模结构化动态环境的新方法和技术。具体而言，这些研究致力于解决以下挑战：发现并解缠控制实体如何交互的规则，从低层级可观测量中推导高层级变量，以及联合发现抽象表征和因果结构。此外，这些研究还采用注意力机制和循环架构等先进的机器学习技术，来学习和表征这些复杂系统。

物理推理。 我们的工作也与学习场景动力学以进行物理推理和因果推理的研究密切相关。无论早期基准测试是关注图像还是具有符号表征的环境，CoPhy 都研究反事实设定下的物理动力学预测。CATER [9]引入了一个用于时间推理的合成视频数据集，涉及位置相关的动作；CLEVRER [32] 将动力学建模与组合式因果推理相结合，并将推理任务落地到语言领域；ComPhy [6]研究物体交互中的物体内在物理属性，以及这些属性如何影响物体在未来场景和反事实场景中的运动，从而回答相应问题。还有一些研究工作专注于开发用于建模结构化动态环境的新方法 [1, 10, 11, 16]。具体而言，这些方法解决了发现并解耦支配实体如何交互的规则，以及从低层观测中归纳高层变量所面临的挑战。我们的系统则显式构建实体之间的交互模块，目标是通过观测动力学来推断内在属性。

3. 方法

我们的目标是通过观察环境状态，学习物体导向且解耦的物理概念表征。此外，我们希望学习一个预——

测系统，在内在物理概念的调节下对物体间相互作用进行建模。如图 z 所示，我们通过将潜变量分割为上下文 z_{ctx}、动力学 z_{dyn}、质量 z_m和电荷 z_c，显式地对物体表征进行因子化，见图2。 z_{ctx} 对所有外在属性进行建模，包括颜色、形状、材料、位置、大小等。 z_{ctx} 可通过重建当前时间步的图像来推断。 z_{dyn} 主要编码速度，而速度直接决定了 z_{ctx} 向下一时间步的转移，并且可通过状态转移函数进行单步预测来推断。在简单情况下， z_{ctx} 和 z_{dyn}已足以模拟仅具有匀速线性运动的场景。然而，当物体通过不同形式的相互作用（如吸引、排斥力或碰撞）彼此施加力时，物体速度会受到这些力的影响，因此需要 z_m 和 z_c 来对这些相互作用建模并更新 z_{dyn}。在训练期间，可从图像重建中推断外在物理概念；通过强制系统执行更长期的预测，内在物理概念则可能涌现。此外，抽象层级不断提高的概念以层级方式获得，完整的物体表征由不同抽象层级概念的级联构成。

我们首先在第3.1节介绍一种多物体表征的生成模型，用于推断物体上下文。然后，在第 3.2节，我们说明 PHYCINE 如何通过短期预测推断物体动力学。最后，我们将在第 3.3节介绍交互模型，以及它如何通过长期预测帮助推断质量和电荷。

3.1. 通过重建学习物体上下文

为了像人类一样实现系统化泛化，IODINE 提出的多槽框架将静态图像分解为 K 个槽，每个槽理想情况下描述输入场景中的一个独立物体或背景。这些槽能够以解耦的方式捕捉物体独特的上下文，例如颜色、形状、位置和大小。图像x 通过空间高斯混合模型进行建模，其中每个槽对应一个物体。借助生成模型，每个潜向量 $z_{ctx}^{(k)}$ 都可以解码为像素级均值 $\mu_i^{(k)}$ 和像素级掩码 $m_i^{(k)}$，分别描述像素 i 属于物体k的分配概率和外观。生成机制是共享的，因此无论 K 槽以何种顺序排列，生成的图像都相同。 K 掩码 logits 经过 softmax函数归一化，以确保每个像素对应的 K 掩码 logits 之和为 1：

$$p(x|z_{ctx}) = \prod_{i=1}^{D} \sum_{k=1}^{K} m_i^{(k)} \mathcal{N}(x_i; \mu_i^{(k)}, \sigma^2), \qquad (1)$$

其中， μ和 m 参数化最终的空间高斯混合分布，而 σ 对于所有 i 和都相同且固定 k.

该表征通过摊销变分推断得到：

$$z_{ctx} \sim q_\lambda(z_{ctx}|x); \lambda = \{\mu, \sigma\}. \qquad (2)$$

$$\lambda \leftarrow \lambda + f(z_{ctx}, x, \mu, m, \nabla_\mu p(x|z_{ctx}), \nabla_m p(x|z_{ctx})). \quad (3)$$

$q(z_{ctx}|x)$ 是给定图像 x时 z_{ctx} 的后验概率，其参数为 λ， f 是用于更新参数的网络。后验参数以任意猜测值初始化，首先从 q_λ 中采样潜在表征 z_{ctx}。在每个细化步骤中，将输入图像 x 与一组变量（均值 μ、掩码 m、均值梯度 $\nabla_\mu p(x|z_{ctx})$、掩码梯度 $\nabla_m p(x|z_{ctx})$）一同输入细化网络 f，以获得后验分布的加性更新；这些变量由 z 和 x 计算得出。

3.2. 通过短期预测学习物体动力学

由于物体外观足以进行状态图像重建，物体上下文模型无法推断动力学。给定 z_{ctx}，PHYCINE 通过扩展摊销变分推断的思路来表示物体动力学 z_{dyn}。从常识的角度来看，物体动力学（例如速度）会在每个时间步转换物体上下文（三维场景中刚体物体的位置）。PHYCINE 学习一个转移函数，以 z_{ctx}^t 和 z_{dyn}^t 为输入，输出下一时间步的上下文 z_{ctx}^{t+1}。数学上：

$$z_{ctx}^{t+1} = FC(z_{dyn}^t) + z_{ctx}^t. \qquad (4)$$

一开始，预测的 z_{ctx}^{t+1}可能是任意值；随着训练的进行，转移函数会被约束为对位置和深度应用动力学，而非其他无关概念，例如形状和颜色。具体而言，需要z_{ctx}^{t+1}生成 x^{t+1}，并使用重建任务中采用的共享解码器。为了预测未来帧图像，必须准确推断 z_{dyn}^t。这个过程可以通过单步预测完成，并可描述如下：

$$z_{dyn}^t \sim q_\lambda(z_{dyn}^t|x^{t+1}, z_{ctx}^t). \qquad (5)$$

我们可以看到，通过观察图像即可直接推断出 z_{ctx}，而 z_{dyn} 的发现则取决于能否正确预测 z_{ctx}的变化，这使得z_{dyn} 的抽象层次高于 z_{ctx}。然而，物体动力学可能会因内在物理属性而不断变化。因此，当涉及更高层次的物理概念时， z_{dyn}^t 只能正确预测 z_{ctx}^{t+1}，因为 z_{dyn}^{t+1} 可能会受到物体间相互作用的影响而更新，而这些相互作用由某些内在概念所定义。

3.3. 利用长期预测学习质量与电荷

为了对物体动力学的变化进行建模，系统应考虑内在概念，并将来自不同来源的相互作用解耦。反过来，进行长期预测能够赋予系统推断内在物理概念的能力。为此，PHYCINE 显式设计了碰撞力，以及吸引力和排斥力（即电荷和力）。

交互模型（图2 (b)）结合碰撞力和电荷力，使 z_{ctx} 和 z_{dyn} 从时间步 $t = 0$ 转移到 $t = N$。通过约束 z_m 和 z_c 分别解释这两种力，并在整个预测过程中保持 z_m 和 z_c不变，模型便能通过这两个变量自适应地学习质量和电荷的概念。

碰撞。 物体彼此施力，并在发生碰撞的瞬间改变速度。系统应捕捉碰撞事件发生的时刻，以及参与物体的速度如何变化。在模型中， z_{ctx} 和 z_{dyn} 用于估计物体是否正在相互接近，以及力的方向。 z_{dyn}和 z_m 用于估计力的强度。

具体而言，我们首先拼接每对物体的上下文和动态信息，并执行自注意力，以获取成对注意力 $attn$ 和力的方向 F:

$$attn^{k_i,k_j} = W([z_{ctx}^{(k_i)}; z_{dyn}^{(k_i)}; z_{ctx}^{(k_j)}; z_{dyn}^{(k_j)}]), \tag{6}$$

$$\vec{F}^{k_i,k_j} = W([z_{ctx}^{(k_i)}; z_{dyn}^{(k_i)}; z_{ctx}^{(k_j)}; z_{dyn}^{(k_j)}]), \tag{7}$$

其中，注意力 $attn^{k_i,k_j}$ 表示物体 k_i 和 k_j 是否足够接近，并据此决定是否施加力； $\vec{F}^{k_i,k_j}$ 表示成对作用力的方向。随后，我们拼接每对物体的质量和动力学信息，并估计力强度:

$$|F|^{k_i,k_j} = W([z_m^{(k_i)}; z_{dyn}^{(k_i)}; z_m^{(k_j)}; z_{dyn}^{(k_j)}]). \tag{8}$$

借助交互注意力、力方向和力强度，我们可以得到每个物体的完整力表示:

$$F^{k_i} = \sum_{j\neq i} l2norm(\vec{F} * attn) * \sum_{j\neq i} (|F| * attn), \tag{9}$$

其中，L2 归一化确保 F 仅包含方向信息。

吸引与排斥。 吸引力和排斥力也是成对相互作用。与碰撞力不同，排斥力和吸引力并非瞬间发生，而是始终存在。

表 1. 不同概念的物理潜在编码维度、抽象层级及推断条件。

概念	维度	抽象	推断
上下文	12	low	图像重建
动力学	2	中间	短期预测
Mass	1	high	长期预测
电荷	1	high	长期预测

假设所有物体对之间都存在普适电荷力，我们根据 z_{ctx} 和 z_c对其进行估算。 z_c 的成对乘积可以为正、为负，也可以为零:

$$F^{k_i} = \sum_{j\neq i} W([z_{ctx}^{(k_i)}; z_{ctx}^{(k_j)}]) * z_c^{(k_i)} * z_c^{(k_j)}. \tag{10}$$

两个模块产生的力会更新 z_{dyn} ，并进一步更新 z_{ctx} 到下一个时间步；质量和电荷的推断可以表述为:

$$z_m^t, z_c^t \sim q_\lambda(z_m^t, z_c^t | x^{\leq t+N}, z_{ctx}^t, z_{dyn}^t). \tag{11}$$

3.4. 目标函数

在我们的系统中，除了初始后验分布 λ，还需要优化三个模型的参数：生成模型 p_θ 的参数，该模型从 z_{ctx} 生成图像；推断模型 q_ϕ 的参数，该模型利用当前参数估计所得的证据下界梯度信息更新后验分布参数；交互模型 r_δ 的参数，该模型展开以下序列 $z_{ctx}^{1:N}$:

$$z_{ctx}^0, z_{dyn}^0, z_m^0, z_c^0 \sim \lambda. \tag{12}$$

$$z_{ctx}^{1:N}, z_{dyn}^{1:N} = r_\delta(z_{ctx}^0, z_{dyn}^0, z_m^0, z_c^0). \tag{13}$$

我们通过重建目标和预测目标进行梯度下降，训练所有参数。训练损失 $\mathcal{L}$ 是负证据下界，可表示为:

$$\mathcal{L} = \sum_{t=0}^{N} (-log(p_\theta(x^t|z_{ctx}^t)) + \beta KL(q_\phi(z^t|x^t, z^{<t})||p(z))) \tag{14}$$

。第一项是基于我们上述讨论的解码器分布混合得到的解码器负对数似然。第二项是潜在后验分布与潜在先验分布之间的 Kullback–Leibler 散度（KL 散度），并由超参数 β加权。

3.5. 通过正则化防止坍塌

当潜变量的信息容量过大时，基于 z_{ctx}^t 并通过一系列潜变量（即 z_{dyx}^t、 z_m^t 和 z_c^t）来预测 $z_{ctx}^{t:t+N}$的方法很容易坍塌。

表 2. 视觉推理模型在 COMPHY 上的问答准确率。我们报告了每个子任务中各选项和各问题的准确率。CPL 使用监督式物体检测器获取物体上下文，例如 Faster/Mask R-CNN [13,25],，然后以物体为中心的标签作为监督，学习质量和电荷的表征。为简化实现，我们省略了 ALOE 的自监督学习策略。

模型	学得的物理属性	事实性	预测性		反事实	
			每个选项	每个问题	每个选项	每个问题
CNN-LSTM [2]	-	46.6	59.5	29.8	58.6	14.6
HCRN [19]	-	47.3	62.7	32.7	58.6	14.2
MAC [15]	-	**68.6**	60.2	32.2	60.2	16.0
$CPL^{phy-label}$ [6]	上下文+m+c	80.5	75.3	56.4	68.3	29.1
ALOE-IODINE [7,12]	ctx	54.4	60.7	29.1	63.8	20.9
芦荟-藻蓝蛋白	上下文+dyn+m+c	59.1 (+4.7)	**63.3** (+2.6)	**33.4** (+4.3)	**68.2** (+4.4)	**25.2** (+4.3)

表 3. 消融实验。我们研究了每种概念表示的贡献，以及自底向上训练策略和交互建模的重要性。我们调整上下文维度，以确保所有比较的特征维度相同。

	事实性	预测性		反事实	
		每个选项。	每个问题。	每个选项。	每个问题。
ctx	54.4	60.7	29.1	63.8	20.9
上下文 +m	55.2	62.8	31.5	65.5	22.6
上下文 +c	55.7	61.8	31.3	66.0	22.1
上下文 +dyn	55.5	63.1	33.2	64.4	21.5
上下文+动力学+质量+碰撞	**59.1**	**63.8**	**34.2**	**68.2**	**25.2**
不进行 b2u 训练	58.0	63.4	33.6	67.0	22.8
无交互	47.3	57.8	25.5	63.9	20.1

例如，给定 z_{ctx}^t，对于所有 z_{ctx}^{t+1}，都存在一种平凡解，可产生零预测损失，其中 z_{dyn}^t 是与 z_{ctx}^{t+1}相同的恒等映射。

防止坍塌的一种有前景的方法，是最小化预测中所使用潜变量的信息内容。我们通过预先将 z_{dyn}、 z_m 和z_c 定义为低维变量来实现这一方案。这样做的另一个目的是，低维概念更加明确，因此有助于发现上层概念。例如，二维 z_{dyn} 表示两个正交的速度方向，这些方向在归一化后保持不变，而这正是 PHYCINE 学习物体质量概念的基础。

除了限制潜变量的容量外，我们还通过使其以正则化的方式发挥作用来限制它们的功能。如图2所示：

- z_m 与 z_{ctx} 和 z_{dyn} 共同对碰撞进行建模。
- z_c 和 z_{ctx} 对电荷力进行建模。
- z_{dyn}^t （交互后更新）输出 z_{ctx}^t 和 z_{ctx}^{t+1}之间的变化量。

3.6. 训练策略。

在 PHYCINE 中，质量这一概念用于估算碰撞事件中力的大小，这要求系统首先学习碰撞这一概念。此外，碰撞力和电荷力都会改变物体的速度，因此在联合训练这两种交互机制时，很难将它们区分开来。因此，我们以自底向上的方式分三个阶段训练 PHYCINE。第一阶段，模型学习上下文、动力学和碰撞的概念。具体而言，我们在物体质量相同且不带电的场景上训练模型。我们随机启用碰撞模块，使模型学会在未发生碰撞时较小的注意力值和动态变化。第二阶段，我们通过在物体质量不同且不带电的场景上使用长期预测目标训练模型，学习质量如何影响碰撞。第三阶段，我们在完整数据集上学习电荷的概念。值得注意的是，训练模型时没有使用任何物理标签进行监督。

4. 实验

在实验中，我们在ComPhy数据集上评估 PHYCINE，以展示所学表征的有效性。我们还详细分析了所发现物理概念的可解释性。

4.1. 数据集

组合物理推理（ComPhy）通过研究物体之间的交互，探究物体的内在物理属性，以及这些属性如何影响物体在未来和反事实场景中的运动，从而回答相关问题。ComPhy 中的物体具有组合式外观属性，如颜色、形状和材料。ComPhy 还研究内在物理概念——质量和电荷——这些概念无法直接从物体的静态外观中获取。因此，除了碰撞，ComPhy 中还存在吸引与排斥事件。相应地，其中也有关于物体质量和电荷等概念的问题，例如，事实性问题中的“立方体比圆柱体重吗？”以及反事实问题中的“如果球体带相反的电荷，会发生什么？”ComPhy 包含 8,000 组训练数据、2,000 组验证数据和 2,000 组测试数据。

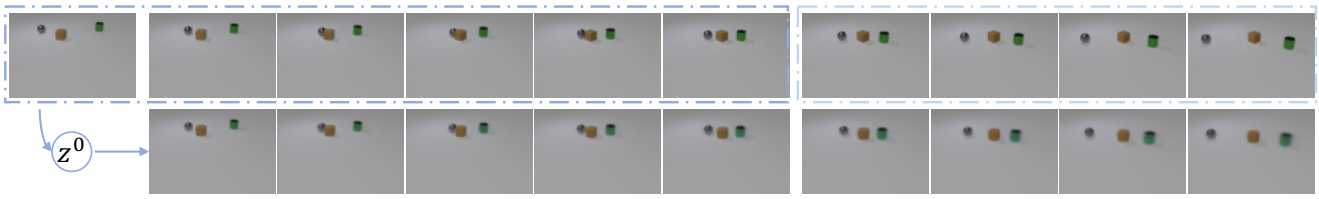


图3. P 的示例结果 HYCINE 在给定 6 帧后进行推断 z^0，随后重新生成视频并想象未来（如更多帧所示）。 4

4.2. 视频理解与因果推理

基线。 因果推理基准上的方法 [6,32] 可分为两种范式：神经网络 [7,15,27,30] 和神经符号模型 [5,6,8,32]。神经符号模型利用多个独立学习的模块，表现优于神经网络基线。ALOE 是一种先进的跨模态 Transformer 架构，以物体为中心的视觉特征和词嵌入作为输入，然后推理得出答案。ALOE 在 CLEVRER 上的表现甚至优于神经符号模型。在这项工作中，我们专注于以物体为中心的学习，并使用 ALOE 验证学习表征。

ComPhy 上的结果。 我们在 ComPhy 上将本方法与若干对比方法进行了实验：CNN-LSTM、HCRN、MAC、CPL 和 ALOE。其中，CPL 是一种高性能的可解释符号模型。MAC 将视觉问答分解为多个基于注意力的推理步骤，并根据最后一步预测答案。ALOE 是 CLEVRER 上最先进的端到端方法，因此被用作我们的基线。

从表 2 可以看出，视频问答模型 CNN-LSTM 和 HCRN 的表现不如其他模型。原因在于，HCRN 等模型主要针对人体动作视频设计，重点关注时间建模，而非对组合物理事件进行时空建模。CPL 的总体表现最佳。然而，该模型使用质量和电荷的显式信号作为监督来学习表征，同时还需要预训练的目标检测器。MAC 在事实性问题上的表现优于 ALOE-IODINE，体现了其组合式注意力架构的有效性。与 MAC 相比，ALOE-IODINE 在预测性和反事实问题上具有竞争优势。将 PHYCINE 特征而非 IODINE 特征输入 ALOE 后，各类问题的表现均有所提升，这表明内在物理概念的学习表征是有效的。

我们还研究了单一类型的物理概念（即动力学、质量和电荷）在所有实验中保持特征维度相同的情况下对结果的影响（表3）。对于事实性问题，三种概念表征都有帮助，我们认为这是十分合理的。

具体而言，动力学有助于回答“是否有金属立方体进入场景？”这类问题。质量和电荷则有助于推理两个物体的相对重量，或它们带有何种电荷。与质量或电荷相比，动力学在推理预测性问题时最为有效。这很容易解释：显式表征物体速度后，更容易直接估计将会发生什么。在反事实问题中，假定物体的重量或电荷取某个反事实值，随后模型需要判断将发生哪些事件。在这种情况下，质量和电荷的概念至关重要，实验结果也验证了这一点（表2：质量和电荷分别将反事实问题的每选项准确率从 63.8 提高到 65.5 和 66.0。更重要的是，不仅每一种概念的表征都有助于提升结果，将它们组合起来还能进一步提高性能。

4.3. 视频再生成与预测

PHYCINE 通过拟合 N 个时间步的未来帧，推断单张图像中包含内在物理概念的多物体表征。换言之，z^0 是从x^0 推断得出的，直至 x^N。如果获得的表征正确，z^0 就应能够再生成所观看的视频片段：首先在潜在空间中进行预测，得到从 z^1 到 z^N的内容，然后将上下文表征渲染为视频。

仅凭系统能否用一组潜在表征准确地再生成给定视频，还不足以证明它真正理解了物理概念。原因之一是，z^0 可能只是在给定视频上过拟合，并未涉及任何物理定律（例如，z_{dyn}、z_m和 z_c 编码的是物体位置变化，而非真正的内在物理概念）。为了验证 PHYCINE 是否充分理解物体如何借助各种物理概念与世界互动，我们测试系统能否进一步预测从 x^N 到 x^T的图像。请注意，我们只训练 PHYCINE 观看并预测（根据第一个时间步的表征进行再生成）N 个时间步的图像。

图3展示了重新生成和预测轨迹的定性结果。这个动态场景中发生了两个事件：球体与立方体发生碰撞，以及球体与圆柱体之间产生排斥力。起初，带有相同电荷的球体和圆柱体

彼此靠近。随后，球体与静止的立方体发生碰撞，速度随之改变。最后，球体和圆柱体在排斥力的作用下相互远离并加速，而立方体则以匀速运动。从结果中可以看出，PHYCINE 能够准确捕捉碰撞力和电荷力引起的相互作用。此外，PHYCINE 无需进一步训练就能推演未来，这体现了我们系统的可解释性。

4.4. 可解释性与解缠结

解缠结是表征的一种理想属性，用于衡量学习到的特征与场景中物体的各个可解释属性之间的对应关系。IODINE 作为我们的基线框架，通过解码器对场景构成施加归纳偏置，促进槽中物体上下文概念的涌现。在 PHYCINE 的设计中，我们期望内在概念（即动力学、质量和电荷）也能保持解缠结。本节通过反事实预测展示解缠结，即手动调整某个概念的值并观察相应的影响。

图4展示了$a \rightarrow e$5 种反事实情境与原始推断结果的对比：两个球体带有相同的电荷，起初彼此靠近，随后受到电荷力作用而相互排斥。此外，蓝色球体还与一个不带电且静止的立方体发生了碰撞：

(a) 给蓝色球体赋予相反的电荷后，排斥力转变为吸引力，随后蓝色球体与立方体发生两次碰撞，接着立方体与红色球体发生碰撞，最后两个球体发生碰撞，这证明了电荷和碰撞概念的正确性与可解释性。(b) 给未带电的立方体赋予与球体相同的电荷后，三个物体相互排斥并彼此远离。(c) 降低立方体的质量值后，碰撞对其速度带来的变化更大；相反，对另一物体的速度变化则更小。这证明了质量概念的正确性与可解释性。(d) 给立方体在第一维度上赋予初始速度 z_{dyn} 后，立方体朝一个方向运动并避开碰撞。(e) 给立方体在第二维度上赋予初始速度 z_{dyn} 后，立方体朝另一个方向运动，同样避开碰撞。

反事实实验揭示了已学习变量各个维度的机制：这些机制一方面与期望的物理概念一致，另一方面彼此解耦。

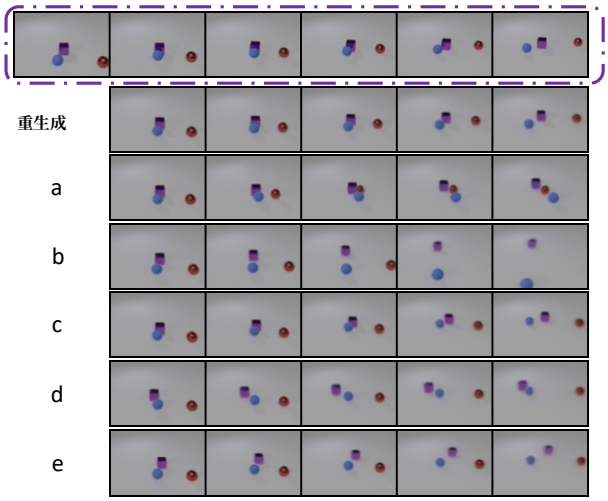


图 4. 通过反事实预测对发现的内在物理概念（即动力学、质量和电荷）进行解缠结。

4.5. 自底向上训练与正则化

我们开展实验以验证自底向上训练策略和正则化的重要性，定量结果见表3。

尽管通过预定义的交互模型进行了正则化，但质量和电荷都可以直接监测并调整上下文，而不是通过动力学实现。在逐步训练时，随着知识的积累，模型能够更好、更准确地理解世界。结果表明，在回答与质量和电荷概念相关的问题，即事实性和反事实问题时，不进行自下而上训练会导致表征质量下降。

通过省略交互模型，我们完全移除了正则化，这意味着模型需要将输入视频压缩为一个潜变量和一个简单的状态转移函数。结果表明，不仅没有学习到有意义的概念性特征，而且与基线模型（ctx vsw/o interaction）相比，有限维特征在描述当前帧上下文方面甚至更差。

5. 讨论与结论

物体之间的关系与相互作用受物理定律支配，这有助于我们准确解读物体的构型 [21]。在这项工作中，我们通过观察物体之间的相互作用，研究了物体内在物理概念（质量和电荷）的构型。我们提出了 PHYCINE 框架，该框架通过摊销推断拟合未来的物理事件，从而推断不同抽象层级的解耦物理概念；训练过程中仅使用原始输入数据，不需要监督。基于反事实预测的可视化结果，我们证实

推断出的概念具有可解释性。发现内在物理概念有助于理解和预测物理事件，并能够获得更丰富的视频表征，从而更好地完成下游推理任务。最后，PHYCINE 展示了利用以物体为中心的预测模型发现并构建内在物理概念的可行性，未来值得沿这一方向进一步研究。 [31]

致谢

本研究部分得到了国家重点研发计划（编号：2020YFC2003901）、国家自然科学基金项目（编号：62176256、62276254、61976229）、中国科学院青年创新促进会（编号：Y2021131）以及 InnoHK计划的资助。

参考文献

[1] Anirudh Goyal ALIAS PARTH GOYAL、Aniket Didolkar、Nan Rosemary Ke、Charles Blundell、Philippe Beaudoin、Nicolas Heess、Michael C Mozer 和 Yoshua Bengio。神经生成系统。《神经信息处理系统进展》，34：25673–25687，2021。3[2] Stanislaw Antol、Aishwarya Agrawal、Jiasen Lu、Margaret Mitchell、Dhruv Batra、C Lawrence Zitnick 和 Devi Parikh。VQA：视觉问答。载于《IEEE国际计算机视觉会议论文集》，第2425–2433页，2015。6[3] Fabien Baradel、Natalia Neverova、Julien Mille、Greg Mori 和 Christian Wolf。CoPhy：物理动力学的反事实学习。arXiv预印本 arXiv:1909.12000，2019。3[4] Christopher P Burgess、Loic Matthey、Nicholas Watters、Rishabh Kabra、Irina Higgins、Matt Botvinick 和 Alexander Lerchner。MONet：无监督场景分解与表征。arXiv预印本 arXiv:1901.11390，2019。1、2[5] Zhenfang Chen、Jiayuan Mao、Jiajun Wu、Kwan-Yee Kenneth Wong、Joshua B Tenenbaum 和 Chuang Gan。通过动态视觉推理建立物体与事件的物理概念。arXiv预印本 arXiv:2103.16564，2021。7[6] Zhenfang Chen、Kexin Yi、Yunzhu Li、Mingyu Ding、Antonio Torralba、Joshua B Tenenbaum 和 Chuang Gan。CoPhy：从视频中进行物体与事件的组合式物理推理。arXiv预印本 arXiv:2205.01089，2022。1、2、3、6、7[7] David Ding、Felix Hill、Adam Santoro、Malcolm Reynolds 和 Matt Botvinick。关注学习到的物体嵌入，实现复杂的视觉推理。《神经信息处理系统进展》，34：9112–9124，2021。6、7[8] Mingyu Ding、Zhenfang Chen、Tao Du、Ping Luo、Josh Tenenbaum 和 Chuang Gan。通过从视频和语言中学习可微物理模型进行动态视觉推理。《神经信息处理系统进展》，34：887–899，2021。7

[9] Rohit Girdhar 和 Deva Ramanan。CATER：用于组合动作和时间推理的诊断数据集。arXiv 预印本 arXiv:1910.04744，2019。 3[10] Anirudh Goyal、Alex Lamb、Phanideep Gampa、Philippe Beaudoin、Sergey Levine、Charles Blundell、Yoshua Bengio 和 Michael Mozer。对象文件与图式：在动力系统中分解陈述性知识和程序性知识。arXiv 预印本 arXiv:2006.16225，2020。 3[11] Anirudh Goyal、Alex Lamb、Jordan Hoffmann、Shagun Sodhani、Sergey Levine、Yoshua Bengio 和 Bernhard Sch¨olkopf。循环独立机制。arXiv 预印本 arXiv:1909.10893，2019。 3[12] Klaus Greff、Rapha¨el Lopez Kaufman、Rishabh Kabra、Nick Watters、Christopher Burgess、Daniel Zoran、Loic Matthey、Matthew Botvinick 和 Alexander Lerchner。通过迭代变分推断进行多对象表示学习。载于国际机器学习会议，第2424–2433页。PMLR，2019。 1, 2, 6[13] Kaiming He、Georgia Gkioxari、Piotr Doll´ar 和 Ross Gir- shick。Mask R-CNN。载于 IEEE 国际计算机视觉会议论文集，第2961–2969页，2017。 6[14] Jun-Ting Hsieh、Bingbin Liu、De-An Huang、Li F Fei-Fei 和 Juan Carlos Niebles。学习分解并解耦视频预测的表征。神经信息处理系统进展，第31卷，2018。 2[15] Drew A Hudson 和 Christopher D Manning。用于机器推理的组合注意力网络。arXiv 预印本 arXiv:1803.03067，2018。6, 7[16] Nan Rosemary Ke、Aniket Didolkar、Sarthak Mittal、Anirudh Goyal、Guillaume Lajoie、Stefan Bauer、Danilo Rezende、Yoshua Bengio、Michael Mozer 和 Christopher Pal。基于视觉模型的强化学习中因果发现的系统性评估。arXiv 预印本 arXiv:2107 00848，2021 3 . . [17] Diederik P Kingma 和 Max Welling。自动编码变分贝叶斯。arXiv 预印本 arXiv:1312.6114，2013。 1[18] Thomas Kipf、Elise Van der Pol 和 Max Welling。结构化世界模型的对比学习。arXiv 预印本 arXiv:1911.12247，2019。 2, 3[19] Thao Minh Le、Vuong Le、Svetha Venkatesh 和 Truyen Tran。用于视频问答的分层条件关系网络。载于 IEEE/CVF 计算机视觉与模式识别会议论文集，第9972–9981页，2020。 6[20] Yann LeCun。迈向自主机器智能的道路。发表于 OpenReview 的预印本，2022 2 。

[21] Wenbin Li、Ales Leonardis 和 Mario Fritz。视觉稳定性预测及其在操作中的应用。载于2017年 AAAI 春季研讨会系列，2017。 8[22]Francesco Locatello、Dirk Weissenborn、Thomas Un- terthiner、Aravindh Mahendran、Georg Heigold、Jakob Uszkoreit、Alexey Dosovitskiy 和 Thomas Kipf。基于槽注意力的以物体为中心的学习。神经信息处理系统进展，第33卷：11525–11538，2020。 1, 2[23] Gerrit W Maus、Jason Fischer 和 David Whitney。MT区空间的运动依赖表征+。《神经元》，78(3)：554–562，2013。2

[24] Nora Nortmann、Sascha Rekauzke、Selim Onat、Peter K¨onig 和 Dirk Jancke。初级视觉皮层表征过去与现在之间的差异。《大脑皮层》，25(6)：1427–1440，2015。2[25] Shaoqing Ren 、 Kaiming He 、 Ross Girshick 和 Jian Sun。Faster R-CNN：借助区域提议网络实现实时目标检测。神经信息处理系统进展，第28卷，2015。 6[26] Gautam Singh、Yi-Fu Wu 和 Sungjin Ahn。针对复杂自然场景视频的简单无监督以物体为中心的学习。arXiv 预印本 arXiv:2205.14065，2022。 1, 2[27] Qu Tang、Xiangyu Zhu、Zhen Lei 和 Zhaoxiang Zhang。用于场景分解与表征的物体动力学蒸馏。载于国际学习表征会议，2021。 1, 2, 3, 7[28]Rishi Veerapaneni、John D Co-Reyes、Michael Chang、Michael Janner、Chelsea Finn、Ji ajun Wu、Joshua Tenen- baum 和 Sergey Levine。基于视觉模型的强化学习中的实体抽象。载于机器人学习会议，第1439–1456页。PMLR，2020。 2, 3[29] Marissa A Weis、Kashyap Chitta、Yash Sharma、Wieland Brendel、Matthias Bethge、Andreas Geiger 和 Alexander S Ecker。视频序列无监督物体表征基准测试。《机器学习研究杂志》，22：183–1，2021。 2[30] Ziyi Wu、Nikita Dvornik、Klaus Greff、Thomas Kipf 和 Animesh Garg。Slotformer：使用以物体为中心的模型进行无监督视觉动力学模拟。arXiv 预印本 arXiv:2210.05861，2022。 7[31] Sirui Xie、Ari S Morcos、Song-Chun Zhu 和 Ramakr- ishna Vedantam。Coat：衡量涌现表征中的物体组合性。载于国际机器学习会议，第24388–24413页。PMLR，2022。 9[32] Kexin Yi、Chuang Gan、Yunzhu Li、Pushmeet Kohli、Jiajun Wu、Antonio Torralba 和 Joshua B Tenenbaum。CLEVRER：用于视频表征与推理的碰撞事件。arXiv 预印本 arXiv:1910.01442，2019。 3, 7[33] Polina Zablotskaia、Edoardo A Dominici、Leonid Sigal 和 Andreas M Lehrmann。使用时空迭代推断进行无监督视频分解。arXiv 预印本 arXiv:2006.14727，2020。 1, 3

A. 训练配置

A.1. 数据处理

ComPhy 中的每个数据点都包含 4 个参考视频和 1 个目标视频。在每组视频中，所有视频里的物体都具有相同的内在物理属性，并且目标视频中的每个物体至少会出现在一个参考视频中。参考视频中物体之间的交互更多，为模型识别物体的物理属性提供了信息。我们首先使用参考视频训练 PHYCINE，然后在目标视频上微调模型。每个参考视频包含 50 帧，我们每隔 4 帧采样一张图像。数据平衡对于促进关系模型收敛也很重要。遵循第 3.6 节中的训练策略，我们根据场景中物体的物理属性来平衡数据。具体而言，我们将参考视频分为 5 类，如表4所示，各类别中的视频数量大致相同。训练时，首先使用子数据集 1 学习物体上下文、物体动力学和碰撞的概念；随后将子数据集 2 加入训练数据，以学习质量的概念；最后加入其余数据，进一步学习电荷的概念。所有概念训练完毕后，我们在目标视频上微调模型。

表 4. 分类后的训练数据集。

子数据集	电荷	碰撞	相同质量
1	✘	✔	✔
2	✘	✔	✘
3	✔	✘	
4	✔	✔	✔
5	✔	✔	✘

A.2. 超参数与训练

我们将每个视频片段 N 的长度设为 6，这意味着根据第一帧推断出的场景表征应能够预测其余五帧。其他参数通常遵循 IODINE 的设置。我们通过从 $U(-0.5, 0.5)$ 中采样来初始化后验分布 λ 的参数。与 ALOE 一样，我们将潜在维度设为 16，使 dim(λ) = 32，并将图像从 320×480 缩小至 64×96。所有实验中，似然的方差均设为 $\sigma = 0.3$。我们保留默认的迭代细化次数 $R = 5$，并在训练和测试中均使用 $K = 8$。所有实验均将 $\beta = 100.0$ 设为固定值。我们使用 8 块 GeForce RTX 3090 GPU 训练模型，每个模型约需 4 天。所有实验均使用 ADAM，学习率为 0.0003。

B. 正则化的必要性

我们通过定性结果讨论自下而上训练和可变内容缩减的必要性。如图5所示，圆柱体与球体之间发生了一次碰撞事件。从再生成结果可以看出，未采用自下而上训练的 PHYCINE 无法捕捉碰撞，但会记录物体的轨迹。此外，对于“如果球体更重”的反事实设定，PHYCINE 预测出了合理的结果：由于该设定，圆柱体和球体的轨迹都会受到影响；而前两帧不受影响，因为此时碰撞尚未发生。相比之下，未采用自下而上训练的 PHYCINE 的结果表明，模型并未真正学会“质量”这一概念，因为只有球体受到影响。

为了证明减少变量内容的必要性，我们使用更高维的动态变量并且不进行交互建模（仅使用全连接层）进行了一项实验。图 6 表明，即使没有交互模型，该模型也能生成物体相互作用的视频序列，这说明更高维的dyn 变量学到了冗余信息（例如速度方向的变化），这些信息本不应纠缠在一起，否则会导致模型崩溃。

C. 中间变量的可视化与数量

我们可视化了图7和表5中示例的力建模过程中间变量。在第3帧，两个物体受到方向相反的力，力强度由动力学和质量共同决定。因此，它们的动力学在第5帧发生了变化。

F	dyn	dyn	f_d	f_d	f_i	f_i	m	m
1	-0.74 -1.65	2.21 -0.02	0.92 -0.39	0.83 -0.56	0.01	0.01	-4.83	-0.15
3	-0.66 -1.64	2.20 -0.04	0.99 0.16	-0.99 -0.12	3.67	0.87	-4.83	-0.15
5	3.12 -1.03	1.25 -0.17	-0.26 0.96	-0.96 -0.27	0.13	0.03	-4.83	-0.15

表5. 变量数量可视化。选取第1、3和5帧（F），其中 f_d 和 f_i 分别表示施加的力方向和力强度。

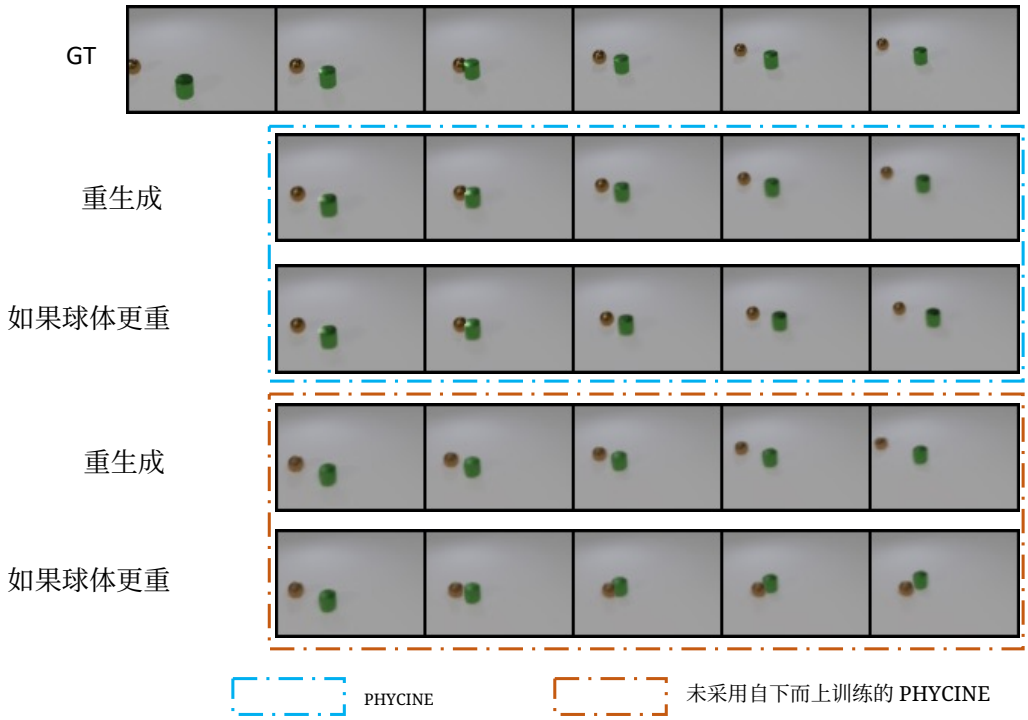


图5. PYHCINE 与 PHYCINE 在未进行自下而上训练时的可视化对比。图中展示了重生成结果和反事实结果。



图6. 高维动力学预测。

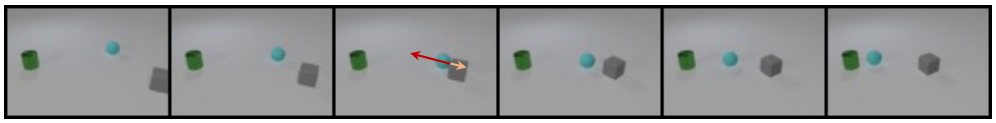


图7. 中间变量可视化示例。