arXiv:2611.28842v1 [cs.CV] 21 Nov 2026

学习身体如何运动：面向动画角色视觉世界模型的小样本身体推断

Amarnath R信息技术系，联合理工学院，哥印拜陀，印度
amarnathresearch@gmail.com

摘要

两个外观完全相同的动画角色，可能会因身体动力学不同而对同一个动作指令做出不同反应。例如，一个僵硬的团块和它的有弹性孪生体在同时接收到跳跃指令时，会达到不同的高度。仅给定第一帧和动作指令时，视觉世界模型没有信息来区分这对孪生体，因为它们的初始外观完全相同。因此，我们提出了一种上下文世界模型，将角色其他动作的片段编码为身体表征，并使用对比损失来区分外观相似的角色。我们推出了 ToonTwins 基准测试，其中包含 144 个会跑、跳跃、跌倒、转身和蹲下的二维卡通角色。这些角色组成了每组四个的外观相似群组，彼此仅在动力学特征上有所不同；此外，我们还在 DeformingThings4D 中的 32 个动画动物上对模型进行了评估。即使使用从这些片段中获得的身体编码，我们测试的潜在像素世界模型最多也只能恢复孪生体之间跳跃高度差异的 25%。面对新形状时，我们的模型在 85.9% 的情况下识别出了正确的孪生体；相比之下，读取相同片段但不使用身体摘要的 Transformer基线为 76.7%。在受控消融实验中，对比损失贡献了其中超过一半的提升。在动画动物上，与不包含身体信息的模型相比，该模型将平均路径预测误差降低了 8.5%。

关键词 视觉世界模型、身体特定动力学、运动预测、上下文学习、身体表征、角色动画

1 引言

动画师通过角色的运动方式，而非外观，让角色显得沉重或轻盈。挤压与拉伸是经典动画原理中的第一条 [1]，：柔软的身体在落地时会变扁，而僵硬的身体则保持原有形状，在二维动画中尤其如此。如图1a所示，两个一模一样的角色执行同一次跳跃，却到达了不同的高度。视觉世界模型可以根据动作指令逐帧生成这样的角色 [2, 3]，还有一些模型则在学习到的场景模型中学习行为 [4]。然而，这些模型只能根据观察到的帧和指令进行预测，因此图1a中的两个角色提供了

具有相同输入的模型，即便它们的真实运动各不相同。因此，输入并未提供任何信息，无法判断模型预测的是哪个孪生样本。它们身体之间的差异会在其他动作中显现出来：有弹性的团块跌倒后会变扁，奔跑时也会改变形状；而僵硬的团块则始终保持圆形（图1b）。因此，提供一些奔跑和跌倒等动作的片段，有助于模型预测身体将如何跳跃。

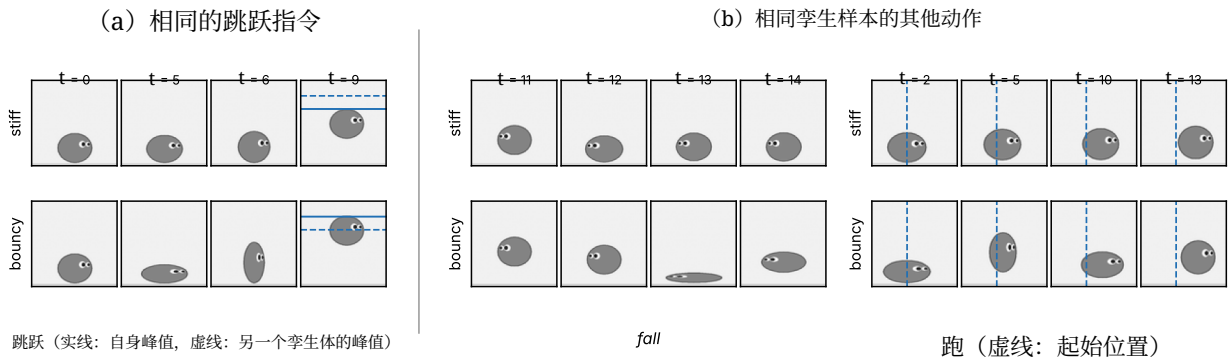


图 1： 外形相似但僵硬程度和弹性不同的团块。（a）在相同的跳跃指令下，有弹性的孪生体跳得更高。（b）它们的跌倒和跑动片段也呈现出相同的差异，虚线标示跑动的起始位置。

像 V-JEPA 2 [5]，那样预测嵌入而非像素，并不能解决这个问题，因为输入仍然没有提供有关身体的证据。标准的轮廓交并比（IoU）也无法发现这一错误，因为同一个预测轮廓与两个孪生样本各自的真实轮廓都有重叠。尽管先前的研究在机器人控制中从近期交互推断动力学[6, 7]，并在物理预测中推断外观相似物体的隐藏属性 [8, 9]，，但这两种情况下，推断依据都是用于预测结果的同一段视频或同一回合。根据不包含请求动作的片段推断动画角色的身体特征，尚未经过直接测试；据我们所知，目前也没有任何动画角色视频基准测试提供身体动力学各异的外观相似角色对。

本文提出一种上下文世界模型，它通过少量展示身体其他动作的片段来学习身体如何运动。这种身体表征有助于预测其运动，但不会还原刚度等物理参数。模型预测身体轮廓的路径，并通过变形第一帧来绘制各帧，因此它使整个身体运动，而不是让每个肢体单独运动。我们的贡献如下。

1. 一种上下文世界模型，可根据其他动作的少量片段预测请求动作的运动路径，并通过对比损失区分相貌相似的孪生体。2. ToonTwins，这是一个基准测试，包含六种身体构型中的144个卡通角色，以及36组由四个相貌相似的孪生体组成的样本；这些孪生体仅在运动方式上有所不同。3. 在 ToonTwins 和来自 DeformingThings4D[10]，的32种动画动物上进行受控评估，涵盖对比损失、渲染帧和未见过的身体构型。

我们的结果表明，一个角色的其他动作包含足够的信息，能够区分新形状中相貌相似的孪生体；与不使用身体摘要的 Transformer 基线相比，对比损失带来的提升超过了一半。我们测试的潜在像素世界模型即使使用来自相同片段的身体编码，也无法保留这种差异的大部分。所有学习得到的模型，包括我们的方法，都无法可靠地迁移到训练中未出现的身体构型。

2 相关工作

本节回顾视觉世界模型、从上下文推断动力学的方法、物理推断基准测试、不同身体间的动作迁移，以及对比学习。

2.1 视觉世界模型

动作条件世界模型根据过去的观测和一个动作来预测未来观测。Genie 从未标注的互联网视频中学习此类模型 [2], GameNGen 使用扩散模型模拟游戏 DOOM [3], 而 DreamerV3 则在其学习到的模型内部训练策略 [4]。潜在动作方法从视频中学习动作空间 [11, 12], ContextWM 将场景的外观与其动力学分离 [13], 而 V-JEPA 2 预测未来嵌入，而非像素 [5]。这些模型的输入和目标各不相同，但它们都根据观测到的帧和一个动作进行预测，因此，两个外观相似的角色会向它们提供相同的输入。于是，使用像素损失训练的确定性模型可能会预测多种可能未来的平均结果 [14], 而随机潜变量能够表示多种可能的未来，并生成更清晰的采样结果 [15]。然而，采样出一个合理的未来，并不等同于根据正确的身体来条件化预测。在我们的工作中，模型还会接收该角色其他动作的片段。

2.2 从上下文推断动力学

条件神经过程将少量观测示例编码为一个表示，并利用该表示预测新的输出 [16]。基于模型的强化学习采用相同的思路，使前向模型适应新的环境。CaDM 从近期状态转移中推断潜在向量 [6], , 而 HyperDynamics 则根据智能体的交互生成模型的参数 [7]。CoDA 将物理系统模型（如常微分方程和偏微分方程）与为每个环境学习的上下文向量建立条件关系 [17]。ICWM 让机器人策略读取其上下文窗口中的少量与任务无关的运动，从而适应新的相机视角 [18], ；IMPULSE 则根据此前三次校准推推动作，预测受推动物体的 V-JEPA 2 特征 [19]。这些方法与我们的方法最为接近，因为它们都会先根据少量观测推断系统，再进行预测。然而，它们推断的是机器人配置、相机视角、物理系统或受推动物体。相比之下，我们根据除请求动作之外的其他动作来推断角色的身体动力学，而其第一帧与相貌相似的双胞胎的第一帧完全相同。

2.3 物理推断基准

Physics-IQ 表明，视频生成模型对物理学的理解较弱，而且这种理解与视觉真实感无关 [20]。Physion++ 展示了推断阶段物体的某项属性，随后要求根据一个场景做出预测；该场景在不同属性值下看起来完全相同，因此只能通过从运动中推断质量、摩擦、弹性或可变形性来预测结果 [8]。PokeWorld 推动视觉上完全相同的物体，并测试世界模型的潜在状态中编码了哪些隐藏属性 [9]。PAC-NeRF 和 PhysDreamer 借助物理模拟器恢复三维物体的材料参数 [21, 22]。ToonTwins 沿用 Physion++ 和 PokeWorld 的做法，使外观保持相同而动力学有所不同。不过，我们的角色会自行执行动作，而不是被推动；我们的模型也不需要模拟器。

2.4 不同身体之间的运动迁移

动作重定向会将给定的运动映射到另一个骨架上。以往的方法能够处理具有同胚结构的骨架 [23], ——即关节数量不同的骨架 [24]——以及具有任意拓扑结构的骨架[25]。图像动画通过关键点 [26], 、姿态序列 [27]或拖动轨迹 [28]来驱动角色图像。在这些方法中，运动作为输入给定，并被迁移到另一个身体或图像上。在我们的任务中，角色尚未展现过请求的动作，因此模型必须根据该角色的其他动作来预测它。

2.5 对比学习

InfoNCE损失训练编码器，使其从一组负样本中为每个锚点选出正样本 [29], ，因此，表征保留哪些信息取决于配对方式。在我们的模型中，正样本对是同一角色的摘要，分别根据两个不相交的片段集合计算得出；负样本则包括与该角色相貌相似的双胞胎。这种训练方式使摘要保留能够在不同动作中识别身体的信息。该损失并不会去除所有动作信息，因此我们在第6.4节中使用线性探针测量剩余的信息。

2.6 摘要

当外观无法揭示身体时，上述方法都无法根据角色执行所请求动作以外的其他动作片段，预测角色在视频中的运动。第 4.1 节对这项任务进行了形式化定义。

3 基准测试和数据集

为了测试角色的其他动作能否揭示其身体如何运动，数据必须满足三个要求。

1. 角色的外观必须相似，但运动必须不同。
2. 每个角色都必须执行若干带标签的动作。
3. 每个片段都必须附带生成该片段的动作指令。

据我们所知，文献中现有的数据集没有一个满足这三项要求。不过，有几个数据集满足其中部分要求。Sprites [30] 包含带标签的动作，但不满足第一项要求，因为其中所有角色都共用同一个人体和一套动画，因此它们的外观不同，但运动方式相同。一般的人类或动物视频不满足第一项和第三项要求，因为外观变化通常伴随着身体变化，而且视频没有动作指令。Physion++ [8]和 PokeWorld [9] 对物体满足第一项要求，但不满足第二项要求，因为其中的物体是被推动或发生碰撞，而不是执行带标签的动作。因此，我们构建了 ToonTwins，其中相貌相似的孪生样本只有动力学特征不同（第 3.1 节）；我们还使用 DeformingThings4D 中的动画动物 [10] 来检验这些发现是否适用于身体结构更复杂的情况（第 3.2 节）。

3.1 ToonTwins

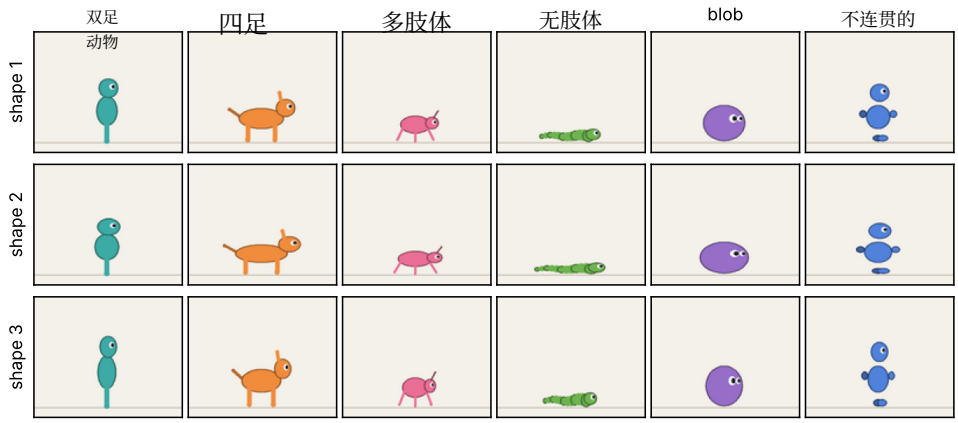
ToonTwins 包含 144 个卡通角色，由 6 种身体构型（双足动物、团块、不连贯的、无肢体、多肢体和四足）、3 种形状、2 种外观和 4 种动力学配置构建而成。三种形状会改变身体的比例。第一种保持默认的宽度和高度，第二种使身体宽 16%，矮 12%，第三种则使身体窄 12%，高 12%。两种外观采用不同的颜色、图案（纯色或斑点）和眼睛颜色。身体方案、形状和外观相同的角色看起来完全相同，因此该基准测试包含 36 组相貌相似的四胞胎。

一组中的四个孪生样本仅在动力学特征上有所不同，分别为僵硬、坚实、柔软或有弹性的。每种身体方案都有八个动力学参数各自的默认值，这些参数控制跳跃高度、落地时的弹跳与压扁、身体起伏、步态、奔跑速度、蹲伏深度和下落速度；动力学特征则通过固定乘数调整这些默认值。例如，

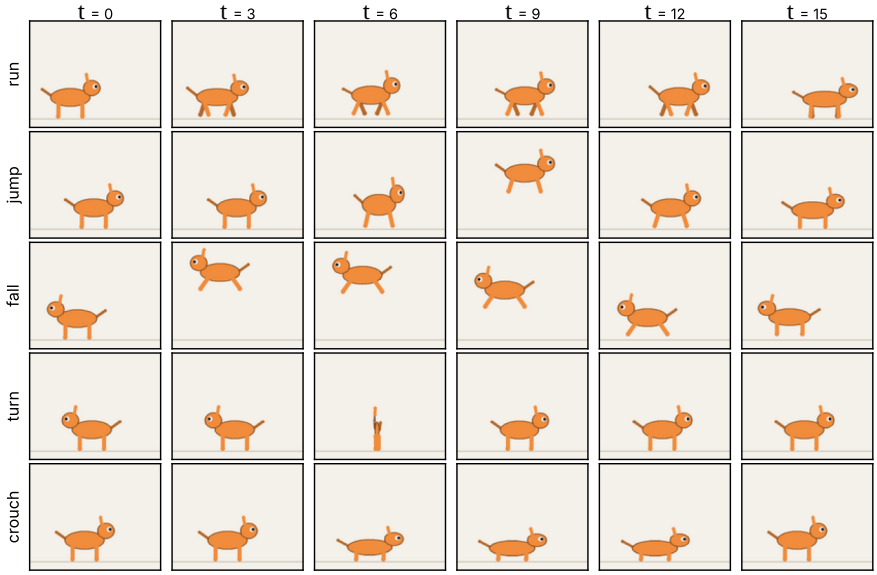
有弹性的配置会使跳跃高度比身体方案的默认值提高 22%，并使跌倒后的反弹幅度提高 130%；而刚性配置则分别使它们降低 22% 和 65%（图1 和 图2c）。这些乘数是固定的设计选择，并非根据数据测量得出。我们选择这些数值，是为了确保每个参数都从僵硬到有弹性呈单调变化。定义动画中挤压与拉伸的挤压与弹跳 [1], 变化范围最大。相比之下，奔跑速度的变化幅度最多为 8%，因此每个角色都始终处于 64×64 帧内，并避免仅凭位置就能识别孪生体。 附录 A列出了所有默认值和乘数。

每个角色在 12 种意图下执行 5 种动作（跑、跳跃、跌倒、转身和蹲下）。每种意图设定动作的力度、方向、起始位置，以及动作开始前的站立帧数（1 至 4 帧）。由此共生成 8,640 个片段，每个片段包含 16 个灰度帧，分辨率为 64×64 像素（图2）。

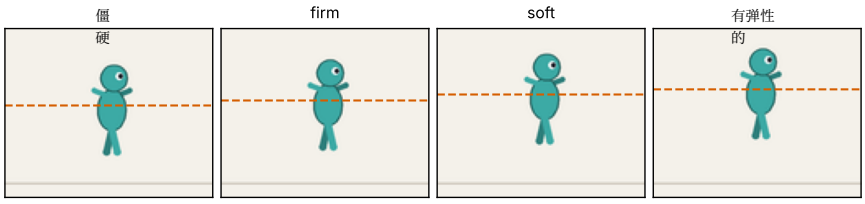
这一设计赋予该基准测试四项特性。站立帧不受动力学特征影响，因此同一组中的四个孪生样本在每个片段中的第一帧都完全相同。然而，它们的运动各不相同；在跳跃的最高点，僵硬孪生体的轮廓质心平均上升 7.57 px，而有弹性的孪生体则上升 12.16 px。转身是唯一不受配置影响的动作，因此我们将其用作控制动作。最后，基于轮廓轨迹简单统计量的线性探针 [31] 能够通过角色的其他动作识别其配置，在 0.506 的情况下识别正确；随机水平为 0.25（第6.4节）。



(a) 3 种形状的 6 种身体构型（第一帧）。每张图像代表四个相貌相似的孪生样本 s.



(b) 一个角色的 5 个动作（弹跳四足动物）。



(c) 处于同一跳跃最高点的四个相貌相似的孪生样本（双足动物）。虚线标出了身体质心。

图 2： ToonTwins。为便于展示，角色以彩色呈现，而所有模型接收的帧均为灰度。

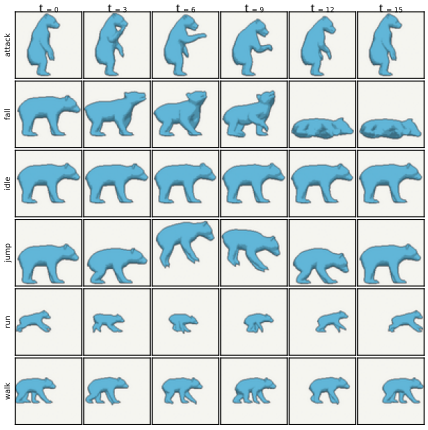
3.2 动画动物

DeformingThings4D 包含由艺术家制作的类人形体和动物的三维动画 [10], ，每段动画展示一个角色执行一个动作。从其中的动物动画中，我们保留了至少能执行六种常见动作（攻击、跌倒、待机、跳跃、跑和

行走），共包含来自 5 个动物科（熊、犬科动物、猫-浣熊-猪、鹿和兔子）的 32 只动物的 169 段动画。我们将每段动画渲染为扁平的二维侧视图，使用三种阴影色调，在浅色背景上勾勒深色轮廓线（图 3）。摄像机在第一帧沿身体的水平长轴方向拍摄，并将每只动物镜像处理，使其朝右。每段片段从动画中均匀抽取 16 帧，以 128×128 像素绘制，缩小至 64×64 并转换为灰度。摄像机在每帧跟随身体包围盒的中心，整个动画采用固定比例，因此片段保留了姿态和形状的变化，但不包含动物在场景中的移动。仅用于训练时，我们还会分别渲染每段动画前 75% 和后 75% 的片段。数据集中的一些动物是其他动物的复制体，外观皮肤不同，但动作相同。为找出这些动物，我们比较每对动物在其共有动作上的表现。我们用随时间变化的六条曲线描述一段动画：网格质心的三个坐标，以及边界框的三条边长；对于长度相同的两段动画，我们计算各曲线帧间变化的皮尔逊相关系数，再对六条曲线取平均值。若至少有一段动画在相同网格上完全相同，且顶点运动的相对差异低于 0.05，或者至少有两段长度相同的动画覆盖不少于一半的共有动作，且平均相关系数至少为 0.90，则认为两只动物共享其动画。我们合并了通过两次运行该流程识别出的动物组，其中一次运行使用的是较早版本的比较方法。最终得到 21 个相互独立的运动组。第一次分组漏掉了四只共享动画的动物，得出了 24 个运动组；本文中所有动物相关结果均采用修正后的 21 个运动组，但第 6.6 节的像素世界模型测试除外。第 6.6 节每只动物只属于一个组，且每个组在所有划分中始终位于同一侧，因此测试动物及其所有动作都不会出现在训练集中。由于这些动物没有孪生体，该数据中的外观和动力学仍然相互混杂，这与 ToonTwins 不同。



(a) 每个科目中的一种动物。



(b) 一种动物（熊）执行其六个动作。

图3： 来自 DeformingThings4D [10] 的动画动物。

4 方法

我们的方法分为三个阶段（图 4）。模型将上下文片段编码为标记，根据这些标记预测请求动作的运动路径，并沿该路径渲染各帧。

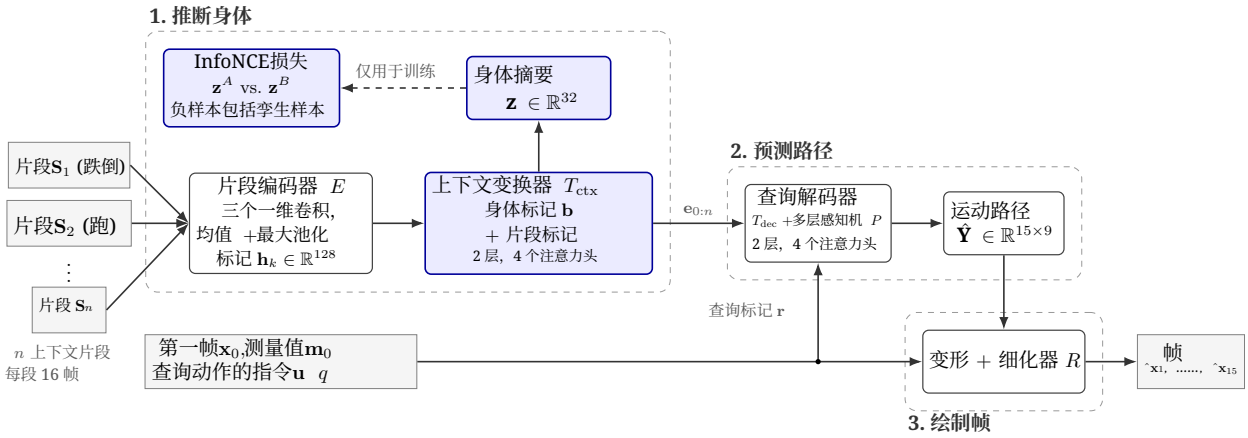


图4： 模型概览。片段编码器和上下文变换器对上下文片段进行概括，查询解码器预测路径，渲染器绘制帧。

4.1 问题表述

我们的任务是根据三个输入，预测角色如何执行请求的动作（我们称之为查询动作q）。第一个输入是查询片段的第一帧$\mathbf{x}_0$。第二个输入是查询动作的指令$\mathbf{u}$。在 ToonTwins 中，$\mathbf{u} \in \mathbb{R}^{10}$ 包含五种动作类型的独热编码、力度以及四个起始帧的独热编码；在动画动物数据中，$\mathbf{u} \in \mathbb{R}^6$是六种动作的独热编码。第三个输入是上下文集合$\mathcal{C} = \{\mathbf{S}_1, \ldots, \mathbf{S}_n\}$，其中包含 n个同一角色执行除 q以外动作的片段，并且每个片段$\mathbf{S}_k (k = 1, \ldots, n)$包含16帧。

我们用身体的轮廓来描述身体，这个轮廓是通过对每一帧相对于背景进行阈值处理并去除地面条带得到的。对于每个轮廓，我们提取九项测量值：边界框的四条边、面积、两个质心坐标、高度和宽度，并将 $\mathbf{m}_0 \in \mathbb{R}^9$ 记为第一帧的测量值（图 5）。模型 f_θ具有可学习参数 θ，用于预测接下来 15 帧的轮廓路径，

$$\hat{\mathbf{Y}} = f_\theta(\mathbf{m}_0, \mathbf{u}, \mathcal{C}) \in \mathbb{R}^{15\times 9}, \tag{1}$$

其中，Y 的第 t 行记录了从第一帧到第 t 帧这九个测量值的变化，Y 表示真实路径。随后从 $\hat{\mathbf{Y}}$ 和 $\mathbf{x}_0$（第4.5节）中采样得到帧 $\hat{\mathbf{x}}1$、.. . 、$\hat{\mathbf{x}}_{15}$。由于 q 从未出现在 C 中，模型只能根据角色的其他动作推断其执行 q 的方式。相貌相似的双胞胎还共享 $\mathbf{x}_0$, $\mathbf{m}_0$ 和 $\mathbf{u}$，因此只有 C 能使它们的预测产生差异。

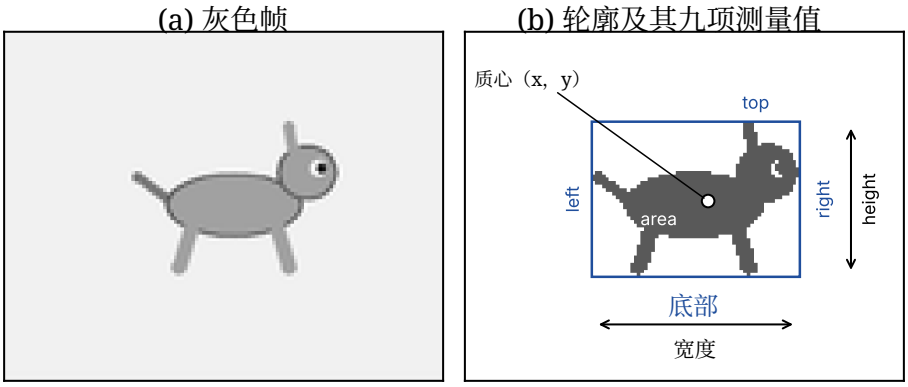


图5： 路径模型的输入。(a) 一帧灰色帧。(b) 其轮廓及九项测量值，包括边界框边缘、面积、质心、高度和宽度。

4.2 片段编码器

片段编码器将每个上下文片段转换为一个向量，我们称之为片段标记。对于片段的每一帧，输入包含相对于片段第一帧的九个测量值、它们相对于前一帧的变化，以及片段的指令，因此，在 ToonTwins 上，片段是一个矩阵 $\mathbf{S}_k \in \mathbb{R}^{16\times 28}$，在动画动物上则是一个矩阵 $\mathbf{S}_k \in \mathbb{R}^{16\times 24}$。三个一维卷积层 ϕ 随时间处理该矩阵，然后对时间维度求均值和最大值，得到片段标记

$$\mathbf{h}_k = W_h \left[\operatorname*{mean}_t \phi(\mathbf{S}_k) \,;\, \max_t \phi(\mathbf{S}_k) \right] \in \mathbb{R}^{128}, \tag{2}$$

其中 $[\cdot;\cdot]$ 将两个各含 128 个值的池化向量连接起来，而 W_h 是一个学习得到的矩阵 合并两个各含 128 个值的池化向量，而 $W_h \in \mathbb{R}^{128\times 256}$ 是一个学习得到的矩阵 .

4.3 带有身体摘要的上下文 Transformer

我们的模型将片段标记分开保留，并通过 Transformer 分两步将它们组合起来。首先，上下文变换器 T_{ctx} [32] 读取片段标记以及一个额外的可学习标记$\mathbf{b}$，我们称之为身体标记，

$$[\mathbf{e}_0, \mathbf{e}_1, \ldots, \mathbf{e}_n] = T_{\text{ctx}}\left([\mathbf{b} + \boldsymbol{\eta}_b,\ \mathbf{h}_1 + \boldsymbol{\eta}_c,\ \ldots,\ \mathbf{h}_n + \boldsymbol{\eta}_c]\right), \qquad \mathbf{z} = g(\mathbf{e}_0) \in \mathbb{R}^{32}, \tag{3}$$

其中 $\boldsymbol{\eta}_b$ 和 $\boldsymbol{\eta}_c$ 是学习嵌入，用于标记两种标记类型。身体标记从所有片段中收集有关角色的信息，其输出$\mathbf{e}_0$, 经过一个小型网络g处理后，得到身体摘要 $\mathbf{z}$。由于 T_{ctx} 从未接收查询动作，因此该摘要不会依赖于它。

其次，查询解码器 T_{dec} 读取查询标记 $\mathbf{r} = W_r[\mathbf{m}_0;\ \mathbf{u}] + \boldsymbol{\eta}_q$ 以及上下文变换器的所有输出，预测器网络 P 将查询标记处的解码器输出转换为路径，

$$\hat{\mathbf{Y}} = P\left(T_{\text{dec}}\left([\mathbf{r}, \mathbf{e}_0, \mathbf{e}_1, \ldots, \mathbf{e}_n]\right)_{\mathbf{r}}\right), \tag{4}$$

其中下标$\mathbf{r}$ 表示查询标记处的输出。解码器会读取$\mathbf{e}_0$ 以及片段输出$\mathbf{e}_1$、……、$\mathbf{e}_n$。摘要$\mathbf{z}$ 不用于预测，只用于对比损失（第4.4节）和探针（第6.4节）；对比损失也通过它训练$\mathbf{e}_0$。我们使用标准化值上的均方误差来训练路径，

$$\mathcal{L}_{\text{pred}} = \frac{1}{15 \cdot 9} \left\| \hat{\mathbf{Y}} - \mathbf{Y} \right\|_F^2 . \tag{5}$$

4.4 针对身体特定摘要的对比目标

仅使用预测损失时，摘要也可能存储上下文中出现了哪些动作，因为这些信息可以降低预测误差。我们希望摘要描述的是身体，而不是观察到的动作，因此加入对比损失，使同一角色在观察到不同动作时得到的摘要保持一致。该损失会减少这类动作信息，但不会将其完全消除；我们将在第6.4节中对此进行测量。

对于批次中第 i 个角色（共 N 个角色），我们抽取两个不共享片段的上下文集合$\mathcal{C}_i^A$和$\mathcal{C}_i^B$，并计算其摘要$\mathbf{z}_i^A$和$\mathbf{z}_i^B$。同一角色的两个摘要构成一个正样本对，而其他角色的摘要则作为负样本，

$$\mathcal{L}_c = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{e^{\langle \hat{\mathbf{z}}_i^A, \hat{\mathbf{z}}_i^B\rangle/\tau}}{\sum_j e^{\langle \hat{\mathbf{z}}_i^A, \hat{\mathbf{z}}_j^B\rangle/\tau}} + \log\frac{e^{\langle \hat{\mathbf{z}}_i^B, \hat{\mathbf{z}}_i^A\rangle/\tau}}{\sum_j e^{\langle \hat{\mathbf{z}}_i^B, \hat{\mathbf{z}}_j^A\rangle/\tau}}\right], \tag{6}$$

其中 $\hat{\mathbf{z}}$ 是将$\mathbf{z}$ 缩放至单位长度后的结果，$\tau = 0.1$ 是温度参数。每个批次包含 16 个完整的双胞胎组，因此某个角色最难区分的负样本就是与其相貌相似的双胞胎；它们外观相同，但运动方式不同。该角色的另一个颜色版本的运动方式完全相同，因此我们将其从两个求和项中移除；对于动画动物，我们还会移除属于同一运动组的动物。总损失为

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda\,\mathcal{L}_c, \qquad \lambda = 0.5. \tag{7}$$

我们还训练了一个更严格的变体，其中两个上下文集合不共享动作类型（在 ToonTwins 上每个集合各包含两种动作类型）；该变体用于测试当观察到的动作不同时，摘要是否保持不变。

4.5 路径引导渲染

渲染器通过移动和拉伸第一帧，使其边界框与预测框匹配，从而绘制每一帧未来帧。对于帧 t，一个缩放和平移变换 A_t 将$\mathbf{x}_0$ 中的边界框映射到预测框，该预测框由 $\mathbf{m}_0 + \hat{\mathbf{y}}_t$给出；变形 $\mathcal{W}$ 则通过双线性插值对帧进行重采样，同时保持地面线固定。随后，细化网络 R 会校正四肢位置等局部细节，

$$\mathbf{w}_t = \mathcal{W}(\mathbf{x}_0; A_t), \qquad \hat{\mathbf{x}}_t = \mathbf{w}_t + R(\mathbf{w}_t, \mathbf{x}_0, \hat{\mathbf{y}}_t). \tag{8}$$

网络 R 仅使用训练角色的真实路径进行训练，并在边界框边缘添加少量噪声；它从未接触过路径模型的输出，因此我们对所有比较的模型都使用相同的渲染器。

4.6 训练细节

片段编码器的卷积层有 128 个通道，核大小为 3，并采用 GELU 激活函数[33]。两个 Transformer 都有两个预归一化层 [34] ，宽度为 128，包含四个注意力头，前馈网络大小为 512，且不使用丢弃法。网络 g 是一个两层多层感知机，预测器网络有三个隐藏层，每层包含 512 个单位。细化网络 R 是一个包含 32、64 和 128 个通道的三级 U-Net[35] ，训练 3,000 步，边界框边缘噪声为 0.7px，并采用 L1 损失与二元交叉熵损失。

所有路径模型均使用 AdamW [36] ，学习率为 10^{-3}，采用单周期调度 [37], ，梯度裁剪阈值设为 1.0，批次大小为 64 个角色。在 ToonTwins 上，我们训练 3,000 步，权重衰减为 10^{-4} ，并在每一步从 1 到 8 均匀抽取上下文大小 n 。对于动画

动物，我们训练 2,000 步，权重衰减为 10^{-3}，并从 1 到 3 抽取 n，因为有些动物只会执行四种动作，其中三种可作为上下文。单个模型在一块 Tesla T4 GPU 上的训练时间为 11 到 64 秒。

5 实验设置

5.1 数据划分

我们在 ToonTwins 上使用两种数据划分。在新形状划分中，3 折中的每一折都会留出每种身体构型中的一种形状，并使用 3 个随机种子进行训练。在未见过的身体构型划分中，6 折中的每一折都会留出一种完整的身体构型，并使用 2 个随机种子进行训练。对于动画动物，我们按运动组划分为 4 折，并使用 3 个随机种子。在所有划分中，测试角色的上下文仅包含查询动作以外的动作片段。

5.2 对比模型

表1 列出了我们对比的模型，它们的主要区别在于提供给各模型的身体信息。所有路径模型都使用相同的片段编码器和预测器网络。身体编码对上下文集合中的片段标记取平均值，并通过一个两层多层感知机将平均值映射为一个 32 维向量，预测器则将 $\mathbf{m}_0$, $\mathbf{u}$ 和这个向量映射为路径。上下文注意力基线是一个具有三层的单一 Transformer，它同时读取查询标记和片段标记，并根据查询标记处的输出预测路径。它没有身体标记，也没有对比损失，其片段标记与查询一起处理。我们的模型加入了身体标记、对比损失，以及一个不读取查询的上下文变换器。档案或科属标签是一种参考，而非上界，因为上下文片段包含的信息不止标签，例如特定形状角色的运动方式。

表1： 对比模型。

模型	其接收的身体信息	参数
无身体	仅第一帧和动作指令	0.60 M
测试时微调	无身体模型在以下数据上微调了 50 步。上下文片段	0.60 M
身体编码	来自上下文片段的单一代码	0.78 M
身体编码 + 对比损失	相同，使用对比损失	0.78 M
上下文注意力	查询关注上下文片段标记	1.40 M
Ours	身体标记、无查询上下文编码器和对比损失	1.62 M
档案或科属标签	真实动力学档案（ToonTwins）或动物科（动物），作为参考	0.61 M
像素世界模型	以身体编码为条件解码帧的潜在世界模型的身体编码	2.88 M
复制第一帧	重复第一帧	–

像素世界模型预测的是帧，而非路径。它通过三个卷积层将每个 48×48 帧编码为 128 维嵌入。随后，包含 512 个单位的残差多层感知机预测下一个嵌入，并通过 FiLM 层接收 32 维身体编码和完整动作指令 [38]。三个转置卷积将嵌入解码回帧。身体编码来自与路径模型相同类型的上下文集合，

无论是否使用对比损失均如此。我们使用包含 16 个、每个有 12 帧片段的批次，训练该模型 3,000 步。损失由四部分组成：嵌入的一步损失、从第一帧开始的开环滚动预测损失、嵌入正则项，以及针对已分离嵌入的解码器损失。在 Tesla T4 上，训练耗时 150 至 200 秒。训练结束时，滚动预测损失仍在缓慢下降，因此延长训练时间或许能提升该模型的表现。因此，我们的结论仅适用于该模型，而不适用于一般的像素世界模型。

5.3 测量值

在 ToonTwins 上，我们的主要测量值是 孪生体匹配。我们首先从预测路径和真实路径中分别移除一组四个孪生体所共有的部分。随后，我们将每个预测孪生体分配给距离最近的真实孪生体，并报告正确分配的比例，其中随机机会为 0.25。 跳跃增益 衡量孪生体之间的高度差得到了多大程度的恢复。它是预测孪生体峰值高度差相对于真实差异的斜率，因此，值为 1 表示差异得到完全恢复，值为 0 表示完全没有恢复。我们还报告四个峰值高度的斯皮尔曼相关系数 [39]、找到跳得最高者的频率（随机机会为 0.25），以及交换测试的结果；在交换测试中，每个孪生体会接收另一个孪生体的上下文。

对于动画动物，我们报告以标准化单位表示的路径误差以及误差降低

$$R = 100\,\frac{\bar{E}_{\text{none}} - \bar{E}}{\bar{E}_{\text{none}}}, \tag{9}$$

其中，$\bar{E}$ 是模型的平均误差，$\bar{E}_{\text{none}}$ 是无身体模型的平均误差。为计算每个平均值，我们先对每个运动组内各查询、各折和各随机种子的误差取平均，然后再对各运动组取平均，因此每个运动组具有相同的权重。我们起初计划先为每个查询计算该百分比，再取平均值，但由于无身体模型的少数查询误差非常小，会产生极端值，最低可达 -401 %，因此我们仅将该版本作为检查结果进行报告。我们还报告<style id='5'>上下文识别，该指标用于检查：与来自其他运动组的最多 7 个留出动物的上下文相比，某个动物自身的上下文是否能带来更低的误差（随机机会为 0.157）。

我们将渲染帧与真实值在 48×48 像素处进行比较。对于这些帧，我们报告阈值为 0.25 时的交并比、清晰度比率，以及根据图像计算的孪生体匹配结果。清晰度比率是预测的图像梯度能量除以真实值的图像梯度能量，因此比率为 1 表示锐度与真实值相同。

5.4 统计

在 ToonTwins 上，分析单位是身体方案与形状的组合，共有 18 个单位。在动画动物上，分析单位是运动组，共有 21 个单位。我们首先对每个单位内的折和随机种子取平均值。然后报告均值及其 95% 自助法置信区间 [40] ，置信区间基于对单位进行 2,000 次重采样计算。比较两个模型时，我们使用配对自助法和威尔科克森符号秩检验 [41]对单位进行比较，并报告 Cohen’s d_z [42] 作为配对效应量。第 6.4节中的表征分析采用单独的流程，我们会在该节中说明，因为其中的多次运行共享测试角色。

6 结果

我们首先测试像素世界模型是否保留了孪生差异，随后测试根据上下文进行路径预测能否还原这种差异。接着，我们拆分模型的各个部分，测试其身体

摘要及其渲染帧，并在动画动物和未见过的身体构型上重复主要测试。

6.1 经测试的像素世界模型无法捕捉身体特定的动力学 cs

我们测试的像素世界模型参数比我们的模型更多（2.88 M 对比 1.62 M），但其解码后的展开轨迹只有在 0.27 至 0.30 的情况下与孪生样本匹配，而随机机会为 0.25；它恢复的跳跃高度真实差异仅为 14% 至 25%。其交并比为 0.44 至 0.45，低于复制第一帧的结果（0.50）；生成的帧也较模糊，锐度仅为真实值的 0.66。此失败与指令的形式无关，因为一种仅接收动作类型、不接收力度或起始时刻的变体，在 0.29 至 0.31 的情况下与孪生样本匹配。

图 6 展示了少量身体信息在展开轨迹中的体现位置。孪生体匹配仅在跳跃和跑的腾空阶段上升，使用对比式身体编码时，在第 9 步达到峰值 0.37，之后又下降。将编码替换为均值编码或孪生体编码后，匹配率便回到随机机会水平，因此这一微弱效果确实来自身体编码。模型也只能区分极端动力学特征。僵硬型和有弹性型孪生体的匹配率分别为 0.52 和 0.51，而坚实型和柔软型孪生体的匹配率仅为 0.12 和 0.14，低于随机机会水平。这是因为四种预测之间的差异相较于真实差异很小，只有极端类型有时能被正确排序。

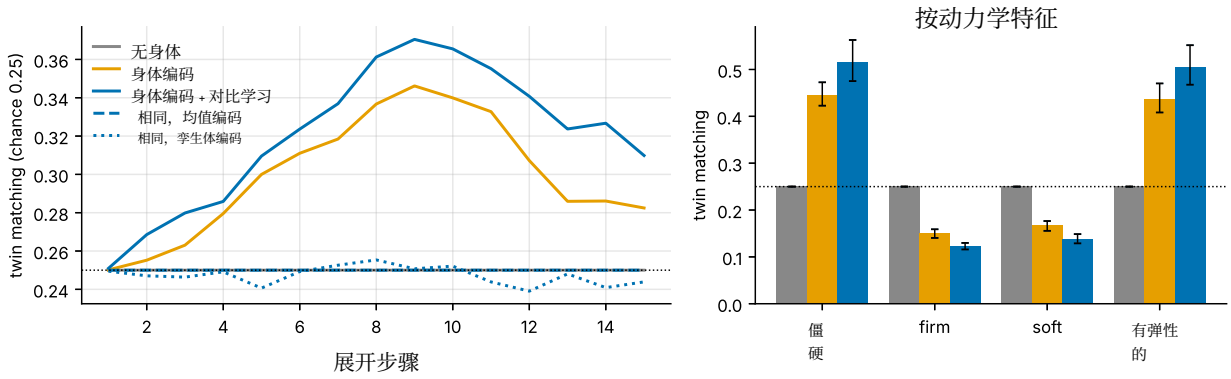


图6： 像素世界模型在新形状上的孪生体匹配（不包括转身），按展开步骤（左）和动力学特征（右）显示，并给出各单位的95%置信区间。

图7 展示了一组双胞胎的情况。四个孪生体的真实峰值高度相差约4到5个像素，而四条预测高度曲线则完全重叠。我们认为这种失败可能有两个原因，但实验无法将二者区分开来。首先，解码器似乎对各种可能的未来取平均，因此四个孪生体最终合并成一个模糊的身体。其次，训练目标本身似乎会丢弃身体信息。使用像素模型冻结的身体编码进行训练的路径预测头，在使用普通编码时，孪生体匹配得分为0.48，低于同一编码器处于随机初始权重时的得分（0.55，$p < 0.001$）。使用对比学习编码时，得分为0.50，也不优于随机初始化的编码器（$p = 0.084$）。一个可能的原因是，上下文片段的轮廓测量值本身已包含身体信息，而像素目标函数没有理由保留这些信息，因为平均运动已经与每个孪生体很好地重叠。

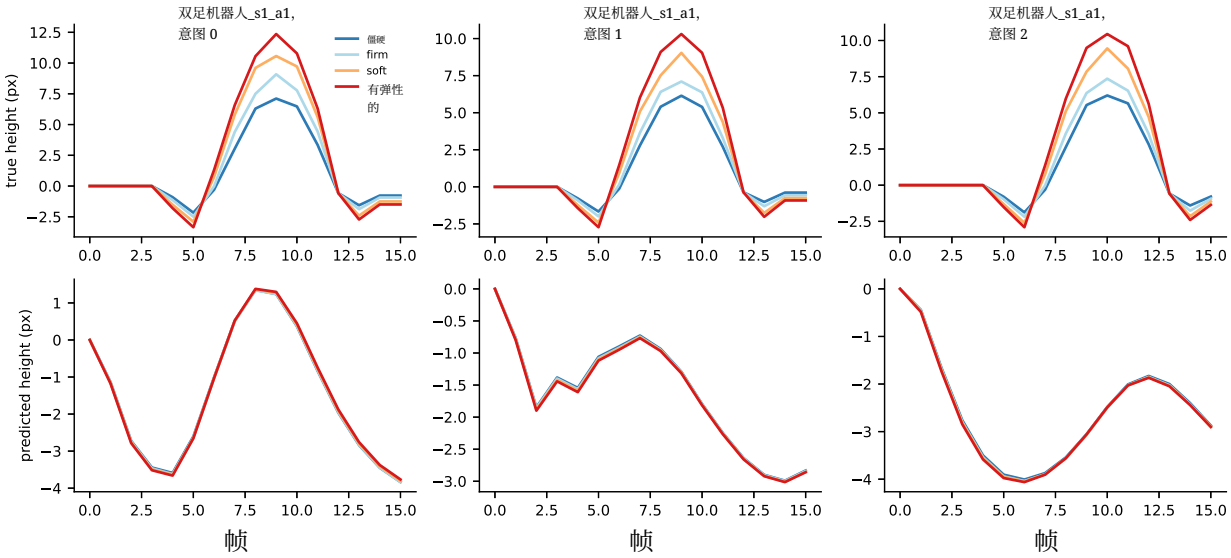


图7： 同一双足动物的四个孪生体在一次跳跃过程中的高度，分别对应三种意图。上排显示真实值，下排显示像素世界模型的结果，其四条曲线相互重叠。

6.2 可从其他动作中恢复身体动力学

预测运动路径而非像素，可以恢复大部分孪生差异。表2 在相同的折和随机种子上比较了所有路径模型，图8展示了一组双胞胎。

表2： ToonTwins 上的路径预测（各单位的均值及 95% 自助法置信区间）。交换列给出了当每个孪生体接收其集合中另一个孪生体的上下文时的孪生体匹配结果，因此数值越低，表示预测越遵循上下文。

	孪生体匹配（随机机会 0.25）		跳跃增益	Swap
模型	新形状	未见过的身体构型	新形状	新形状
无身体	0.250	0.250	0.00	–
身体编码	0.675 [0.630, 0.727]	**0.436** [0.385, 0.485]	0.86	0.098
身体编码 + 对比学习	0.672 [0.621、0.728]	0.389 [0.351, 0.424]	0.82	0.106
上下文注意力	0.767 [0.699, 0.829]	0.412 [0.357, 0.465]	0.90	0.070
Ours	**0.859** [0.820, 0.897]	0.422 [0.375, 0.466]	**0.96**	0.044
轮廓标签（参考）	0.813 [0.787, 0.839]	0.531 [0.470, 0.591]	0.98	0.046

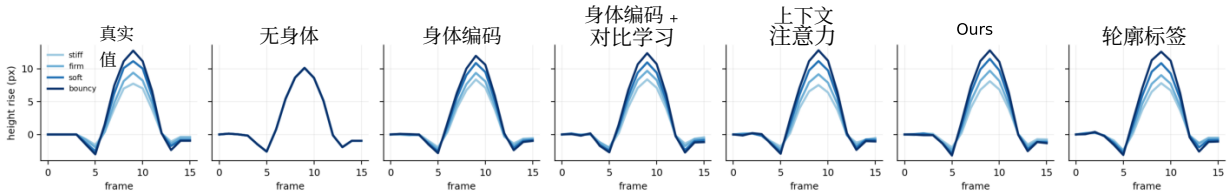


图8： 一个双足动物在一次跳跃过程中的四个孪生体的预测高度。缺少身体信息时，所有孪生体的路径都相同，而我们的模型则遵循真实的排序和间距。

对于新形状，　我们的模型在孪生体匹配上的表现比上下文注意力基线高出 0.092

(95%置信区间为 0.063 至 0.126，$d_z = 1.32$，$p < 0.001$)；比身体编码高出 0.184（$d_z = 3.73$，$p < 0.001$）。它也比接收真实轮廓标签的模型高出 0.046（$p = 0.012$）。这是因为上下文片段展示了这个特定形状如何运动，而轮廓标签无法提供这些信息。交换孪生体的上下文后，孪生体匹配率降至 0.044，低于 0.25 的随机水平，因此预测依据的是模型接收到的身体信息。此测试有意改变上下文，并不能证明模型已识别出动力学特征。对不含身体信息的模型使用上下文片段进行微调后，匹配率仅为 0.281。单靠对比损失并不能改善身体编码（0.672，相比之下为 0.675），因此模型在新形状上的提升取决于 Transformer（第 6.3 节）。每个学习得到的模型都能对新形状的四种跳跃高度进行排序，斯皮尔曼相关系数为 0.99 至 1.00。

两个上下文片段已带来了大部分提升（图9）。我们的模型使用一个片段时达到 0.586，使用两个片段时达到 0.795，已超过使用八个片段的上下文注意力基线（0.767）。使用八个片段时，模型达到 0.859。线性探针从我们的摘要中以 0.835 的准确率恢复出动力学特征，以 0.236 的准确率恢复出留出动作，接近 0.20 的随机水平。未使用对比损失的身体编码分别达到 0.655 和 0.760，说明它记录了模型见过哪些动作。

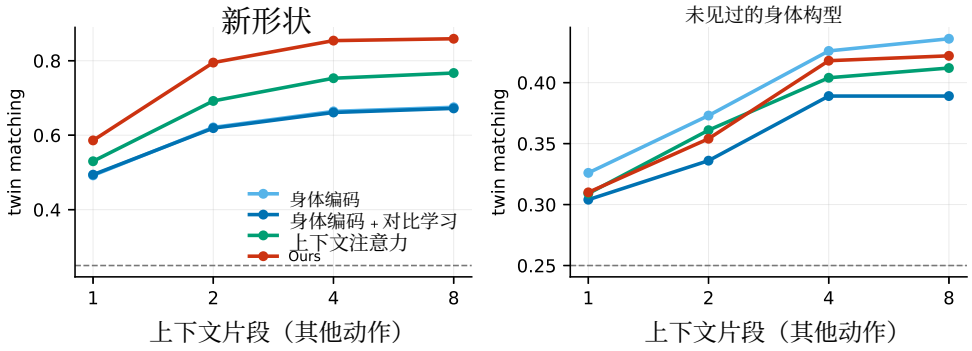


图9：孪生体匹配结果与上下文片段数量的关系。虚线表示随机水平。

并非每个动作都包含有关身体的信息（图10）。仅由转身构成的上下文会让每个模型都处于随机机会水平（0.24 至 0.25），因为转身是唯一不依赖动力学特征的动作。奔跑、跌倒和跳跃是信息量最大的上下文动作；对于除转身以外的查询动作，我们的方法在这些动作上的平均准确率为 0.81 至 0.87。蹲伏的信息量最少（0.67）。对于每个具有信息量的上下文动作，我们的方法都优于上下文注意力基线。

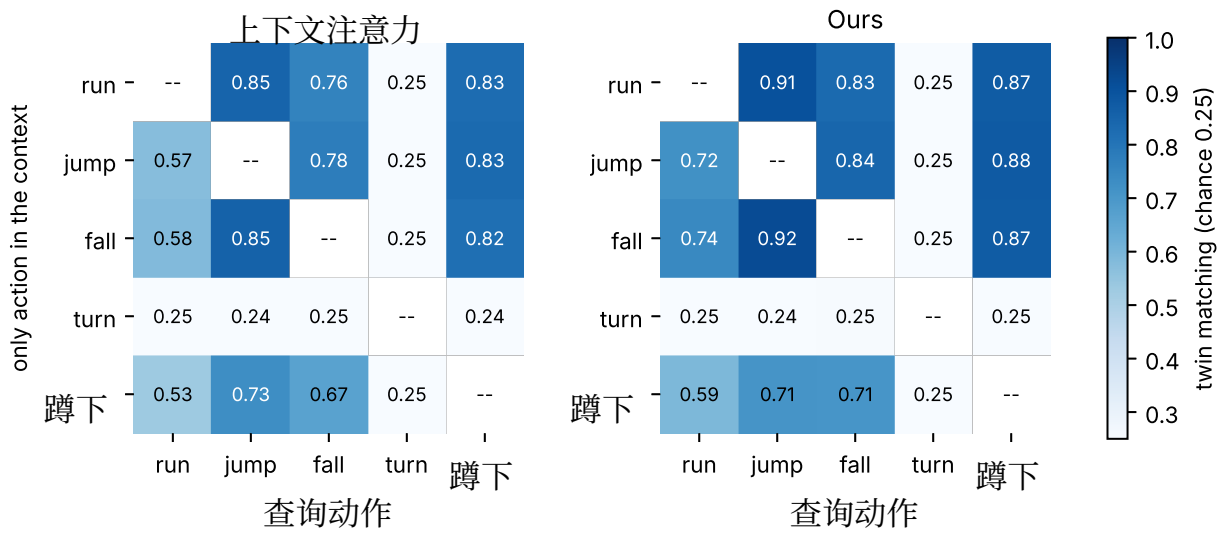


图10： 当上下文仅包含一种动作类型（行）时，针对每个查询动作（列）的孪生体匹配。转身不携带任何身体信息。

我们的方法在极端动力学特征上的优势最大。它在 0.906 的情况下匹配弹跳型孪生体，在 0.863 的情况下匹配僵硬型孪生体；相比之下，上下文注意力基线分别为 0.767 和 0.748。图11按身体方案、查询动作和动力学特征对孪生体匹配进行了细分。在学习得到的模型中，我们的方法在每种身体方案、动作和动力学特征上都表现最佳。相比上下文注意力基线，我们的方法在无肢身体（0.554 至 0.774）、分节身体（0.789 至 0.902）、奔跑（0.580 至 0.739）和跌倒（0.774 至 0.872）上的增益最大。只有对于无肢体和多肢体身体，以及跳跃和蹲伏，轮廓标签的表现更好；在这些情况下，运动仅取决于一两个动力学参数，例如跳跃增益或蹲伏深度，因此只需知道配置即可。

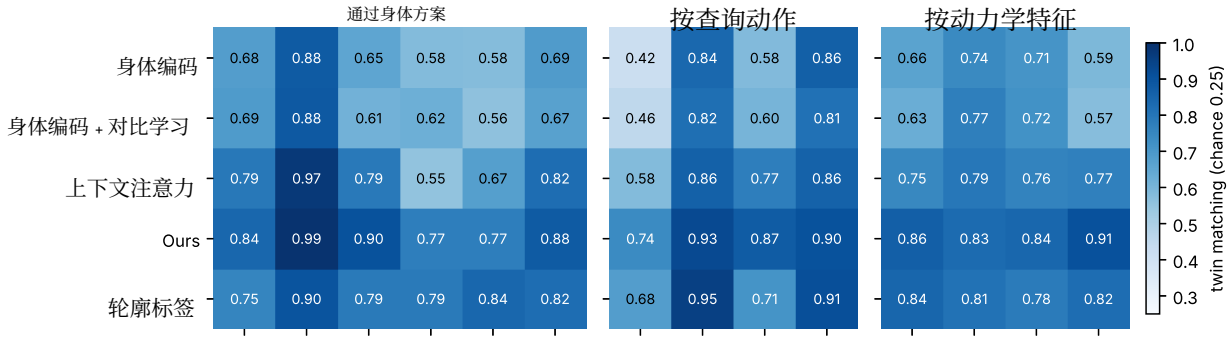


图11： 根据身体方案、查询动作和动力学特征，对新形状进行孪生体匹配（8个上下文片段，随机机会为0.25）。

6.3 模型的各个部分分别贡献了什么

表2 不包含不使用对比损失的模型，因此我们在相同的折和随机种子上重新训练了四个模型，以区分我们模型的各个组成部分（表3）。对于新形状，我们使用了3个随机种子

对于动画动物使用了3个随机种子，对于未见过的身体构型使用了1个随机种子，因此我们将关于未见过的身体构型的比较视为探索性的。重新运行实验复现了主要结果：我们的模型分别为0.857和0.859，上下文注意力基线分别为0.770和0.767。关于动画动物和未见过的身体构型的列将在第6.6节和第6.7节中讨论。

表3： 使用相同的折和随机种子进行受控消融实验，结果来自单独的一次重新运行。双胞胎匹配（TM）使用 8 个上下文片段（随机机会为 0.25）。

			新形状		未见过的身体构型		动画动物	
模型	身体标记 + 无查询编码器	对比集合	TM	跳跃增益	TM	跳跃增益	减少（%）	识别
上下文注意力	–	–	0.770	0.91	**0.420**	**0.64**	5.2	0.206
Ours，关闭对比损失	✓	–	0.806	0.93	0.418	**0.64**	3.2	0.153
Ours	✓	不相交片段	**0.857**	**0.96**	0.409	0.49	8.1	0.296
我们的方法，动作不相交集合	✓	不相交动作	0.838	0.94	0.375	0.38	**10.5**	**0.309**

对于新形状，为 Transformer 提供一个身体标记，并使用不查看查询的上下文编码器，可将孪生体匹配率从 0.770 提高到 0.806（差值为 0.037，95%置信区间为 0.017 至 0.060，$d_z = 0.77$，$p < 0.001$）。这两项改动是同时进行的，因此无法得知它们各自的效果。在相同架构中加入对比损失后，孪生体匹配率进一步提高至 0.857（差值为 0.050，置信区间为 0.036 至 0.066，$d_z = 1.48$，$p < 0.001$）。对比损失带来的提升超过相对于基线总增益的一半；这部分增益并非来自额外容量，因为不使用该损失的模型参数完全相同。若从互不重叠的动作类型中抽取两个对比集合，匹配率为 0.838，低于 0.857（$p < 0.001$），但仍高于不使用该损失的模型（$p < 0.001$），因此损失带来的益处并不依赖于两个集合共享动作类型。

图12 展示了每组双胞胎组中相同的效应。对于若干真实间距为2至3像素的双胞胎组，上下文注意力基线预测的间距接近于零。加入身体标记和对比损失后，这些情况大多消失了，预测间距与真实间距的相关性也从0.93升至0.95和0.96。

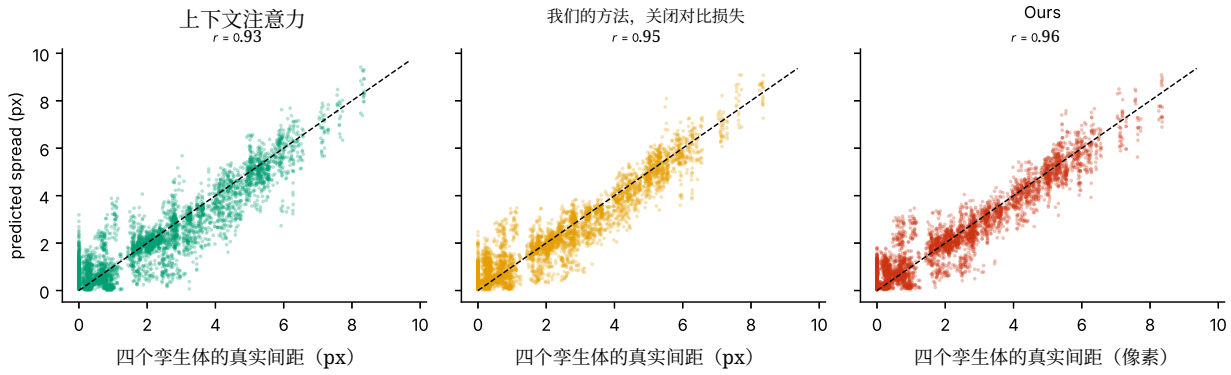


图12： 新形状上每组双胞胎组中四个孪生样本峰值高度间距的预测值与真实值。虚线表示完美预测。

6.4 对比损失使摘要更聚焦于身体，而较少聚焦于动作

对比损失使同一角色在不同上下文动作下的摘要保持一致，并区分相貌相似的双胞胎，但它减少而非消除了有关已观察动作的信息。对查询动作进行探针测试无法显示这一点，因此我们直接进行了测量。对于每个

留出角色，我们计算了十个摘要，每个摘要来自一对动作中的四个片段，并比较了不共享任何动作的动作对所对应的摘要（表4）。

表4： 留出角色的身体摘要不变性（括号内为各次运行的95%置信区间）。差值是同一身体的余弦相似度减去孪生体的余弦相似度，计算时未进行四舍五入。孪生体检索的随机机会为0.25，动作探针的随机机会为0.10。

划分	模型	余弦相似度，相同身体	Cos，孪生体	Gap	检索	动作探测（↓）	轮廓探测
新形状	轮廓特征，无需学习	–	–	–	0.329 [0.315，0.343]	0.999 [0.998, 1.000]	0.506 [0.491，0.524]
	我们的方法，关闭对比损失	0.987	0.833	0.154 [0.136, 0.173]	0.792 [0.769、0.820]	0.626 [0.592，0.661]	0.651 [0.631、0.668]
	Ours	0.813	0.248	**0.564** [0.536, 0.597]	**0.873** [0.862, 0.887]	0.139 [0.129，0.150]	**0.799** [0.767，0.837]
	我们的方法，动作不相交集合	0.791	0.250	0.541 [0.506，0.579]	0.852 [0.829，0.877]	**0.127** [0.119，0.135]	0.765 [0.722，0.816]
未见过的身体构型	我们的方法，对比损失关闭	0.959	0.834	**0.126** [0.016，0.314]	**0。450** [0.304，0.622]	0.545 [0.465, 0.656]	**0.353** [0.317, 0.390]
	Ours	0.214	0.103	0.110 [0.044，0.173]	0.369 [0.291，0.443]	**0.104** [0.088, 0.122]	0.326 [0.260, 0.423]
	我们的方法，动作不相交集合	0.209	0.125	0.083 [0.000，0.177]	0.342 [0.258，0.440]	0.125 [0.103, 0.147]	0.267 [0.196、0.344]

没有对比损失时，不同角色的摘要彼此仍然很相似。同一身体的两份摘要的余弦相似度为 0.987，但一个身体与其孪生体的相似度也高达 0.833。线性探针在 0.626 的情况下能够判断观察到的是哪一对动作，而随机机会为 0.10。加入对比损失后，同一身体在完全不同动作下生成的两份摘要仍然相似（0.813），而孪生样本的相似度则降至 0.248。两者之间的差距从 0.154 增至 0.564（$d_z = 6.7$，$p < 0.001$ 在 9 次运行中）。用于识别观察到的动作的探针准确率降至 0.139（95%置信区间为 0.129 至 0.150）。这接近但仍高于 0.10 的随机水平，因此，对比损失减少了这类信息，但并未将其完全消除。

这些运行共享测试角色，因此我们还针对一次运行（折0，种子0）的12组双胞胎进行了重采样。结果相同：使用损失时，差距为0.608（95%置信区间为0.516至0.694）；不使用损失时，差距为0.187（0.114至0.277）。对于全部12组双胞胎，使用损失时的差距都更大。在九次运行中，使用损失时，差距介于0.51至0.64之间；不使用损失时，介于0.11至0.21之间。在0.873的情况下，仅凭彼此毫无共同之处的动作，就能从四个相似对象中辨认出一个孪生体；不使用损失时，这一比例为0.792。图13 展示了相同的结果。

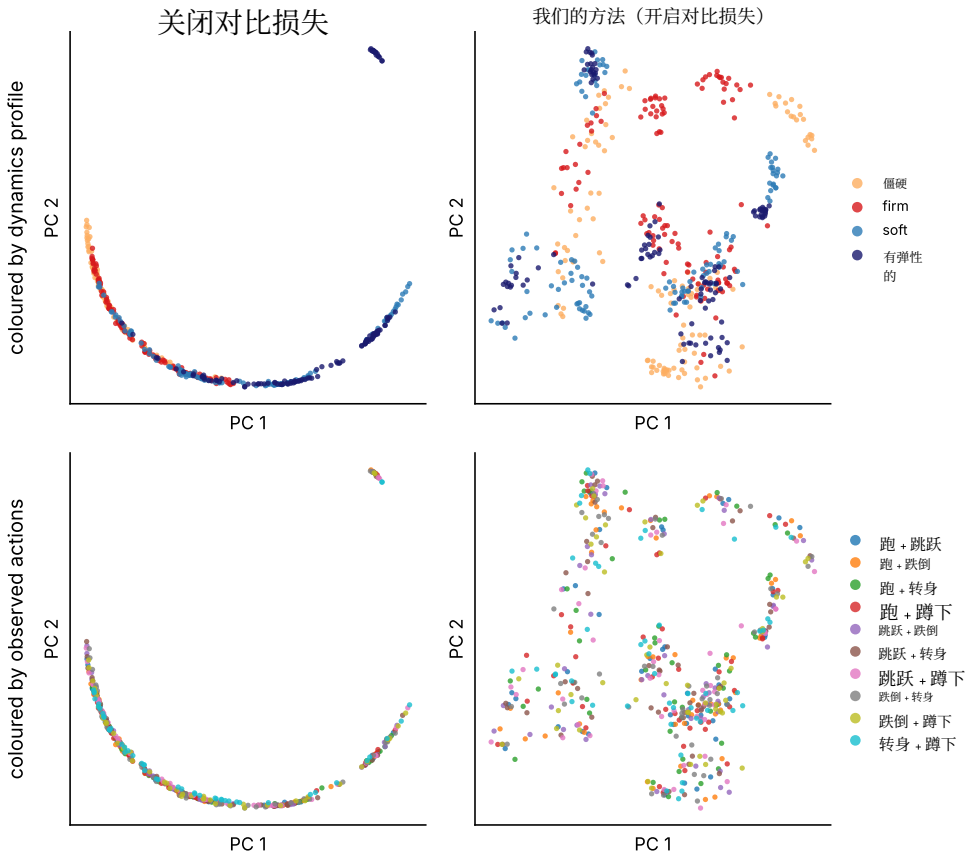


图13： 留出角色身体摘要（前两个主成分）。没有对比损失时，它们位于同一条曲线上；有了对比损失后，每个角色都会形成自己的簇。

我们还对手工制作的轮廓统计量运行了相同的探针，不进行学习，以检验片段中是否确实包含可恢复的身体信息。对于每个片段，我们计算了九项测量值各自的最大值、最小值、平均值和标准差，以及帧间平均绝对变化量。然后，按平均值和最大值对同一上下文的四个片段进行汇总。线性探针根据这些特征，在0.506的情况下恢复出动力学特征（95%置信区间为0.491至0.524）；相比之下，我们的摘要为0.799，随机水平为0.25。然而，相同特征在0.999的情况下揭示了观察到的动作对，而仅在0.329的情况下辨认出了正确的孪生体（表4），因为这些特征将身体与动作混合在了一起。两种情况下，特征缩放器和探针——采用L2惩罚的逻辑回归（$C = 1$）——都仅在训练角色上拟合，且未进行超参数搜索。

6.5 路径引导渲染可生成清晰帧，同时保留差异

沿预测路径绘制帧，既能保持图像质量，也能保留孪生体差异（表5 和 图14）。

表5： 新形状上的渲染帧，与 48×48 像素处的真实值进行比较 s.

方法	IoU	锐度 (1 = 真实值)	孪生体匹配 图像中的	跳跃增益 在图像中
复制第一帧	0.502	0.97	0.250	0.00
像素世界模型	0.448	0.66	0.300	0.25
无身体，路径引导	0.799	0.94	0.250	0.00
上下文注意力，路径引导	0.846	0.94	0.822	0.89
我们的方法，路径引导	**0.861**	0.94	**0.904**	**0.95**
真实路径，路径引导（渲染器限制）	0.888	0.95	0.991	1.00

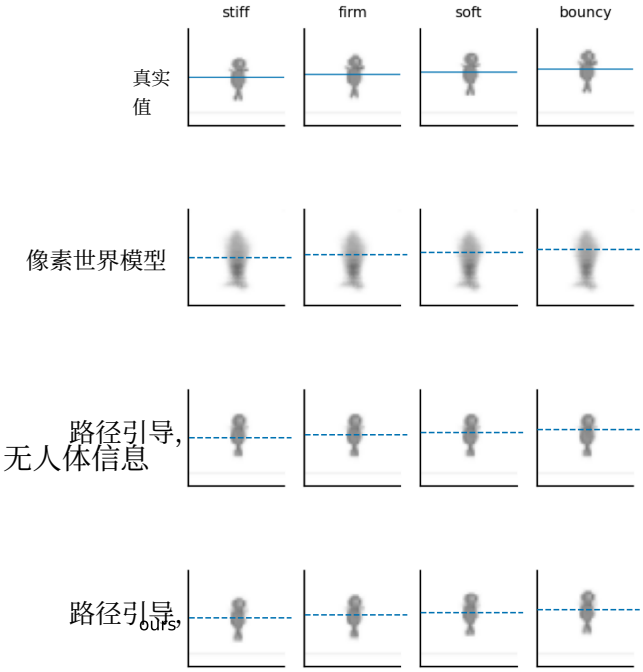


图14： 四个相貌相似的双胞胎在一次跳跃的最高点，线条标示出真实的高度。像素模型和无身体渲染器为四者绘制出相同的身体，而我们的方法则将每个孪生体放置在接近其自身高度的位置。

与像素世界模型相比，我们的渲染帧使交并比提高了 0.413（$d_z = 7.72$，$p < 0.001$），孪生体匹配提高了 0.604（$d_z = 10.94$，$p < 0.001$）。与上下文注意力基线相比，图像中的孪生体匹配提高了 0.081（$p < 0.001$）。不包含身体信息的模型也能渲染出准确的帧（交并比为 0.799），但它让四个孪生体呈现相同的运动（图14第三行）。孪生体匹配的提升源于上下文中的身体信息，而非渲染器。与单独使用变形相比，细化器使交并比提高了 0.096，孪生体匹配提高了 0.168。奔跑是最难的动作，因为变形会带动整个身体移动，却无法让腿部摆动。即便如此，我们的模型在 0.818 的案例中成功匹配了奔跑的孪生体。

6.6 动画动物

为了测试像素世界模型在 ToonTwins 上的失败是否也会出现在更复杂的身体上，我们首先在这些动物上训练该模型，并要求它针对留出的运动组，滚动预测上下文中未出现的一个动作的 8 帧（图15和16）。随后，我们测量滚动预测有多大概率

比起执行相同动作的其他留出动物的未来，更接近该动物自身真实的未来（随机机会为 0.159）。这一随机水平与路径模型的 0.157 略有不同，因为本测试使用了较早的划分和不同的候选动物集合。在无身体信息、使用身体编码以及使用对比式身体编码的情况下，得分分别为 0.565、0.573 和 0.545；两种身体编码的得分与无身体信息相比均无显著差异（$p = 0.45$ 和$p = 0.13$）。重复最后观测到的帧，得分为 0.580。模型确实使用了对比式编码，因为将其替换为另一只动物的编码后，得分会降低 0.101（$p < 0.001$）；线性探针从中恢复动物科的准确率为 0.565（随机机会为 0.20）。像素模型接收身体信息并会对此作出反应，但其滚动预测并未改善，仍然只是模糊的平均身体（图16），与 ToonTwins 上的情况相同。本测试使用了包含 24 个组的第一种分组方式，其中某只动物可能与一只训练动物共享动画。这只会对像素模型有利，但模型仍未取得提升，因此我们没有重复该测试。

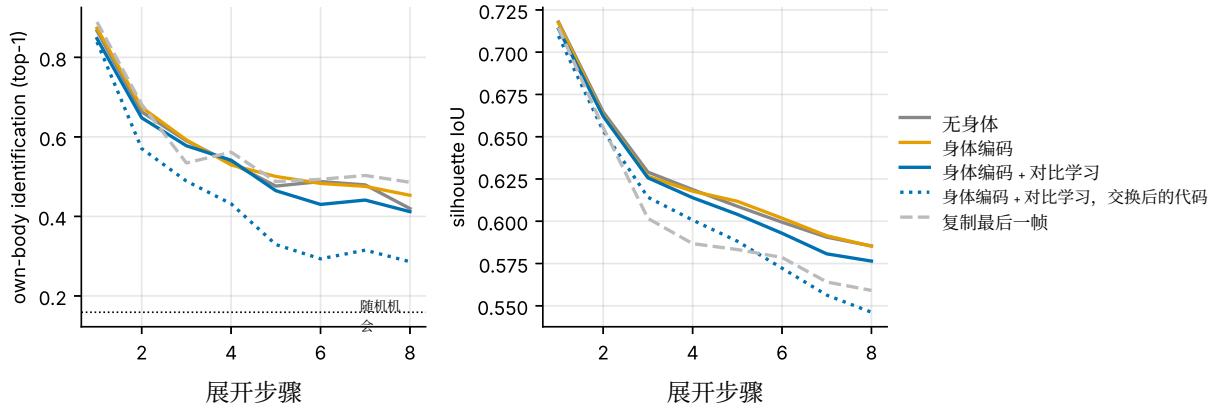


图 15： 动画动物的像素世界模型。左图显示滚动预测与该动物自身未来状态的接近程度超过与其他动物未来状态接近程度的频率（随机机会为 0.159），右图显示轮廓交并比。

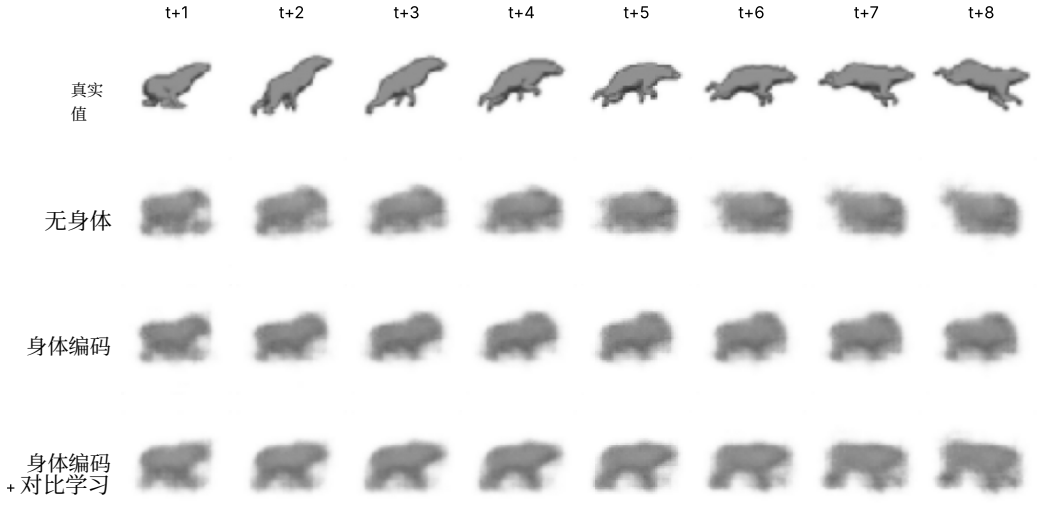


图 16： 留出熊执行一种其上下文中未出现的跳跃动作时的像素世界模型滚动预测。模型绘制出模糊的平均身体。

相比之下，我们的路径模型将平均路径误差降低了8.5%，优于不含身体信息的模型（95%置信区间为2.7至14.4），标准化单位下的误差从0.847降至0.774（$p = 0.004$）。在0.324的情况下，它能根据动物自身的上下文识别出该动物，而随机机会为0.157（表6）。这种提升在不同查询间并不均衡。在所有折和随机种子中，61%的查询误差有所降低；按查询平均后，降幅仅为5.0（置信区间为 -3.4 至12.0）。与ToonTwins上的情况不同，增加上下文带来的误差降幅不到2%。上下文中包含其他一个、两个或三个动作时，我们模型的误差分别为0.787、0.780和0.774，其他模型的误差也同样变化不大。动物的一个其他动作已包含这些模型能够利用的大部分身体信息。我们的模型误差也低于已知动物科的模型（0.810），但这一差异并不显著（$p = 0.13$）。

Tab表 6： 动画动物，来自 DeformingThings4D（21 个运动组、4 折、3 个随机种子 ）.

模型	相较于无身体的误差降低（%）	上下文识别（随机机会 0.157）
身体编码	3.7 [−0.1、7.3]	0.274
身体编码 + 对比学习	7.5 [3.4，11.9]	**0.327**
上下文注意力	4.8 [0.3，9.3]	0.193
Ours	**8.5** [2.7，14.4]	0.324
家族标签（参考）	4.3 [−1.5，10.5]	0.390

我们的模型在识别（$p = 0.004$以及另一只动物的上下文与其自身上下文之间的差距方面优于上下文注意力基线（$p < 0.001$，但其较低的误差仅处于临界显著水平（$p = 0.046$，威尔科克森 $p = 0.16$）。在误差（$p = 0.52$）和识别（$p = 0.88$）方面，它与采用对比损失的身体编码相当。受控重跑结果见表3，其中我们的模型和基线分别达到8.1%和5.2%，这说明了性能提升的来源。对于这些动物，身体标记和无查询编码器并没有带来帮助；不使用该损失时，它们的误差反而略高于基线（$p = 0.049$）。加入对比损失后，误差降低（$p = 0.003$），并将根据动物自身上下文识别该动物的准确率从0.153提高到0.296（$p < 0.001$）。从互不重叠的动作类型中构建两个对比集合可达到10.5%，略优于我们的模型（$p = 0.047$，威尔科克森 $p = 0.13$）。因此，对于这些动物，大部分性能提升来自对比损失，而不是Transformer。图17按动作分解了误差降低情况。我们的模型在跳跃（12.5%）和奔跑（11.9%）上的提升最大，而攻击和站立不动的表现变差（分别下降8.6%和6.2%）。

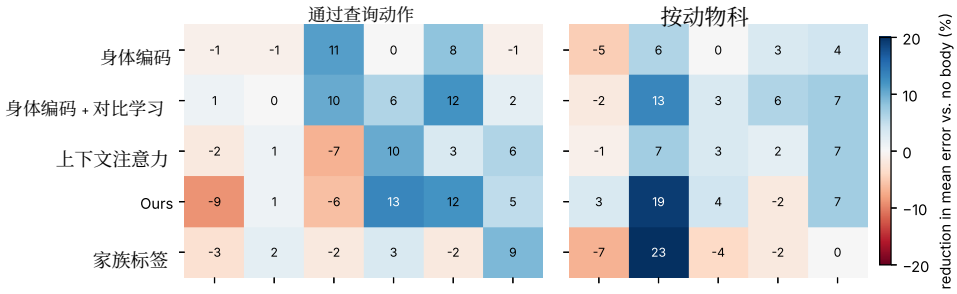
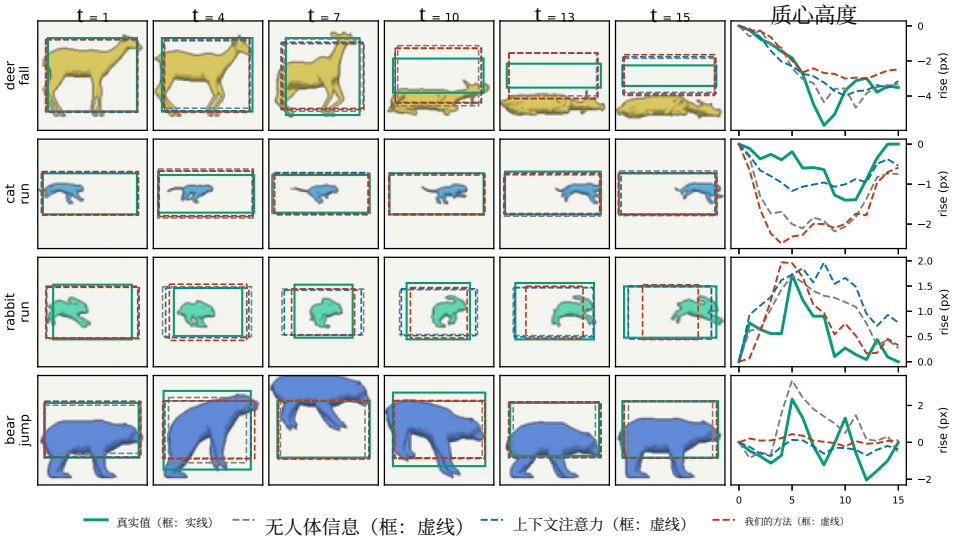
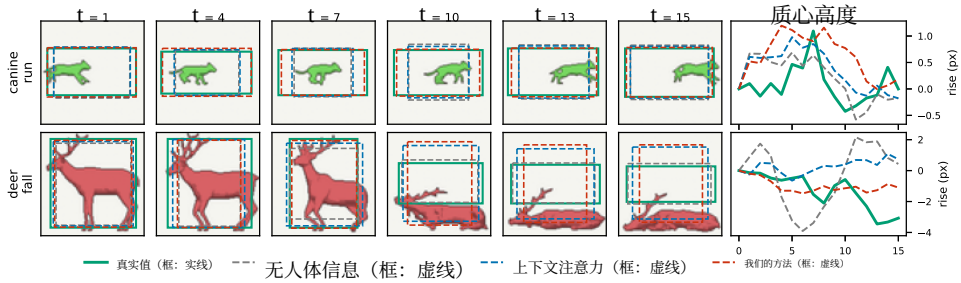


图 17： 动画动物相较于无身体模型的误差降低，按查询动作和动物科划分。蓝色单元格表示优于无身体模型，红色单元格表示较差。

图18 展示了种子0模型的实际预测结果。我们选择了典型案例，其相较于不使用身体信息的模型的改进最接近中位数，同时还选取了最佳和最差案例，因此该图不会只展示成功案例。在该种子包含的全部169个动物与动作的留出配对中，我们的模型、上下文注意力基线以及不使用身体信息的模型的路径误差分别为3.02像素、3.08像素和3.15像素。按像素测量，在该种子的案例中，我们的模型有63%优于不使用身体信息的模型。犬科动物的提升最大（从3.07像素降至2.58像素）；而对于鹿，我们的模型不如不使用身体信息的模型（3.35像素对3.21像素）。这些框还展示了这些动物面临的主要限制。全身位置和大小的预测效果良好，但不同模型之间的差异相较于四肢运动很小，而框无法描述这种运动。



(a) 典型案例（改进幅度最接近中位数，涵盖四个动物科）。



(b) 最佳案例（上）和最差案例（下）。

图18： 留出动物上的预测运动。实线框表示真实轮廓框，虚线框表示根据同一第一帧作出的预测，右侧一列展示质心上升情况。

6.7 未见过的身体构型仍然难以推断

当整个身体构型都被留作未见数据时，主要比较中的所有已学习模型（表2）的孪生体匹配率都降至 0.389 至 0.436，而轮廓标签达到 0.531。我们的模型（0.422）与上下文注意力基线（$p = 0.47$）或身体编码（$p = 0.20$）相比并无差异。对于无肢体身体构型，所有已学习模型的表现都接近随机机会，为 0.25 至 0.28（图19）。对于多肢体身体构型，轮廓标签达到 0.71，而所有已学习模型的得分均不超过 0.36。已知身体构型的上下文片段无法让模型了解一种新型身体如何

响应其动力学特征。在渲染帧中，我们的模型对未见过的身体构型的跳跃高度排序也不够可靠；它在 0.59 的案例中识别出跳得最高者，而上下文注意力基线的这一比例为 0.78。

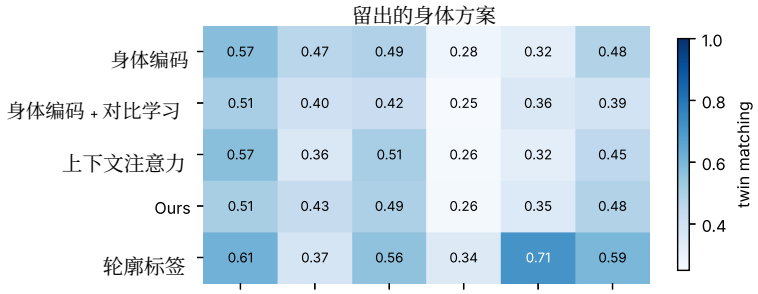


图 19：留出的身体方案包含整个身体时（各列）的孪生体匹配，使用 8 个上下文片段（随机机会为 0.25）。

消融实验给出了一种可能的解释，不过该实验在此划分中仅使用了一个随机种子，且属于探索性分析（表 3 和 4）。对于未见过的身体构型，对比损失并未改善孪生体匹配效果（0.409 和 0.375，而不使用该损失时为 0.418），并将跳跃增益从 0.64 降至 0.49 和 0.38（$p < 0.001$）。同一新身体的两份摘要之间的余弦相似度仅为 0.21，因此，在已知身体构型上学到的一致性无法迁移到新的构型。一种可能的解释是，该损失学到的是训练中见过的身体构型内部的差异，而非适用于陌生结构的摘要。要验证这一点，需要覆盖更广泛的身体构型。

7 讨论与局限性

7.1 像素指标与身体动力学

不应仅凭像素误差和交并比来判断世界模型是否能够适应某个角色。在我们的测试中，像素世界模型的交并比仅比复制第一帧低 0.05，而其恢复的孪生差异最多只有 25%。相似外形身体测试直接展示了这一差距，而且只需要一个程序化生成器。我们比较的两个模型在输出和训练目标上也有所不同。像素世界模型通过潜在状态预测帧，而我们的模型预测轮廓路径，再由单独的渲染器绘制该路径。我们的结果表明了所测试的像素模型存在局限，并不意味着路径预测普遍优于基于像素的世界建模。在我们的实验中，一种简单的变形方法和一个小型细化器能够在路径预测完成后绘制出清晰的帧。使用真实路径时，渲染器的交并比可达 0.888，因此剩余误差大多来自路径预测，而非绘制过程。

7.2 对比损失的作用

对比损失决定身体摘要所存储的内容。在架构和参数相同的情况下，它将新形状上的孪生体匹配度从 0.806 提高到 0.857，并使同一身体的摘要在彼此毫无共同之处的动作之间保持一致（第6.4 节）。在动画动物上，它是我们模型中唯一带来帮助的部分。然而，在我们的单种子比较中，它降低了未见过的身体构型上的跳跃增益，因此当测试包含新的身体构型时，损失权重 λ 可能需要设为不同的值。

7.3 已知与新身体构型

对于已知身体方案的新形状，模型能够很好地预测身体特定运动，并且同一身体的摘要在彼此毫无共同之处的动作之间保持一致。对于训练中从未见过的身体构型，预测能力和一致性都会大幅下降（第6.7节）。因此，模型的学习能力受到训练中所见身体结构多样性的限制。扩大这种多样性，或者通过身体各部分来描述身体，使新身体能够由已知部分构建而成，是下一步可能的方向。

7.4 局限性

本研究有意通过整体轮廓来描述每个身体，以便将身体特定动力学作为唯一的实验变量。因此，渲染器无法独立移动四肢，而奔跑是渲染帧中表现最弱的动作（孪生体匹配度为 0.818）。模型也无法迁移到未见过的身体构型上，无肢体身体构型的表现接近随机机会水平。这两个问题都需要描述身体的各个部分及其协同运动方式。

ToonTwins 使用分辨率为 64×64 像素的简单二维角色，其动力学配置遵循生成器中的固定规则。我们没有测试训练中使用的四种配置以外的参数值。每种配置还会同时改变全部八个参数，因此我们的测试无法说明模型利用哪些参数来区分孪生样本。动物数据仅包含 21 个运动组中的 32 只动物，这限制了结果的精确度，例如，误差降低幅度的 95%置信区间为 2.7% 至 14.4%。这些动物也是由艺术家制作的动画，因此这一增益只表明模型在这个子集上有所提升，并不能说明其能够迁移到自然动物视频。

我们没有评估大型预训练视频生成模型，因为要对其进行改造，使其能够以某个特定角色的少量片段为条件，需要进行大量修改。我们只预测未来 15 帧，因此没有测试误差在更长滚动预测中的累积情况。

8 结论

我们测试的潜在像素世界模型最多只保留了相似外形身体之间差异的 25%。一种能够编码角色的其他动作并预测运动路径的模型，在 85.9% 的情况下区分出了新形状的孪生体。沿着预测路径绘制帧，能够在渲染帧中保留这种差异（孪生体匹配度为 0.904）。与 Transformer 基线相比，增益的一半以上来自对比损失；该损失使同一身体在彼此毫无共同之处的动作中的摘要保持一致。在 32 种动画动物上，平均路径误差降低了 8.5%，而采用相同损失的简化身体编码也取得了相近的增益。这些结果揭示了三点。

- 角色世界模型应在外形相似但运动方式不同的角色上进行测试，因为像素指标无法体现是否利用了身体动力学。
- 在我们的实验中，通过预测运动路径传递身体信息，可以使其保留在渲染帧中，而经过测试的像素模型却丢失了这些信息。
- 身体摘要的训练应尽可能减少对所观察动作的依赖。

该方法无法迁移到训练中未出现的身体构型，而且要实现这种迁移，需要涵盖更广泛的身体结构。

附录 A ToonTwins 参数

表 7 给出了每种身体方案的默认动力学参数，而表 8给出了每种动力学特征的乘数。角色所用的值为其身体方案的默认值乘以其特征的乘数。高度以 128×128 绘图画布上的像素数表示；该画布会缩小至 64×64 像素，而转向速率不受特征影响。

表 7： 六种身体构型的默认动力学参数。

身体方案	Jump	弹跳	压扁	Bob	Gait	奔跑速度	蹲下	Fall	转向速率
双足动物	28	0.05	0.5	1.5	0.80	3.0	0.45	1.4	1.4
Blob	40	0.35	1.6	6.0	0.60	2.2	0.70	1.3	0.8
不连贯的	30	0.20	0.5	2.0	0.85	3.3	0.50	1.0	1.2
无肢体	18	0.02	0.8	0.0	1.10	2.6	0.60	1.5	0.7
多肢体	22	0.15	0.4	0.6	1.50	3.6	0.35	1.1	1.6
四足	34	0.10	0.6	1.2	0.95	4.2	0.55	1.6	1.0

表 8： 各动力学特征对身体方案默认参数应用的乘数。

配置	Jump	弹跳	压扁	Bob	Gait	奔跑速度	蹲下	Fall
僵硬	0.78	0.35	0.35	0.55	0.88	0.92	0.80	1.15
Firm	0.92	0.75	0.75	0.85	0.96	0.97	0.93	1.05
Soft	1.08	1.50	1.45	1.30	1.04	1.00	1.08	0.95
有弹性的	1.22	2.30	2.00	1.75	1.12	1.03	1.20	0.86

参考文献

[1] F. Thomas 和 O. Johnston，《生命的幻象：迪士尼动画》。美国纽约州纽约：阿贝维尔出版社，1981 年。

[2] J. Bruce 等，“Genie：生成式交互环境”，载于 *ICML* 论文集，PMLR 235，2024 年，第 4603–4623 页，doi：10.48550/arXiv.2402.15391。

[3] D. Valevski、Y. Leviathan、M. Arar 和 S. Fruchter，《扩散模型是实时游戏引擎》，载于 *ICLR* 论文集，2025 年，doi：10.48550/arXiv.2408.14837。

[4] D. Hafner、J. Pasukonis、J. Ba 和 T. Lillicrap，《通过世界模型掌握多样化控制任务》，《自然》，第 640 卷，第 8059 期，第 647–653 页，2025 年，doi：10.1038/s41586-025-08744-2。

[5] M. Assran 等，“V-JEPA 2：自监督视频模型实现理解、预测与规划”，arXiv:2506.09985，2025，doi：10.48550/arXiv.2506.09985。[6] K. Lee、Y. Seo、S. Lee、H. Lee 和 J. Shin，“用于基于模型的强化学习泛化的上下文感知动力学模型”，载于*ICML*论文集，PMLR 119，2020年，第5757–5766页，doi：10.48550/arXiv.2005.06800。[7] Z. Xian、S. Lal、H.-Y. Tung、E. A. Platanios 和 K. Fragkiadaki，“HyperDynamics：利用超网络进行元学习物体与智能体动力学”，载于*ICLR*论文集，2021，doi：10.48550/arXiv.2103.09439。[8] H.-Y. Tung 等，“Physion++：评估需要在线推断不同物理属性的物理场景理解”，载于*NeurIPS*数据集与基准测试专题论文集，2023，doi：10.48550/arXiv.2306.15668。

[9] K. Tan 等，“潜在世界模型能够了解什么？多模态预测表征中的物理信息”，arXiv:2607.27017，2026，doi：10.48550/arXiv.2607.27017。[10]Y. Li、H. Takehara、T. Taketomi、B. Zheng 和 M. Nießner，“4DComplete：超越可观测表面的非刚性运动估计”，载于ICCV论文集，2021年，第12686–12696页，doi：10.1109/ICCV48922.2021.01247。[11]D. Schmidt 和 M. Jiang，“学习在没有动作的情况下行动”，载于ICLR论文集，2024，doi：10.48550/arXiv.2312.10812。[12] S. Gao、S. Zhou、Y. Du、J. Zhang 和 C. Gan，“AdaWorld：利用潜在动作学习可适应世界模型”，载于ICML论文集，PMLR 267，2025年，第18744–18771页，doi：10.48550/arXiv.2503.18938。[13] J. Wu、H. Ma、C. Deng 和 M. Long，“利用野外视频为强化学习预训练情境化世界模型”，载于NeurIPS论文集，2023，doi：10.48550/arXiv.2305.18499。[14] M. Mathieu、C. Couprie 和 Y. LeCun，“超越均方误差的深度多尺度视频预测”，载于ICLR论文集，2016，doi：10.48550/arXiv.1511.05440。[15] E. Denton 和 R. Fergus，“利用学习到的先验进行随机视频生成”，载于ICML论文集，PMLR 80，2018年，第1174–1183页，doi：10.48550/arXiv.1802.07687。[16] M. Garnelo 等，“条件神经过程”，载于ICML论文集，PMLR 80，2018年，第1704–1713页，doi：10.48550/arXiv.1807.01613。[17] M. Kirchmeyer、Y. Yin、J. Donà、N. Baskiotis、A. Rakotomamonjy 和 P. Gallinari，“通过情境信息驱动的动力学模型泛化至新的物理系统”，载于ICML论文集，PMLR 162，2022年，第11283–11301页，doi：10.48550/arXiv.2202.01889。[18]S. Wang 等，“用于机器人控制的上下文内世界建模”，arXiv:2606.26025，2026，doi：10.48550/arXiv.2606.26025。[19]T. Jin 和 C.-H. Chou，“IMPULSE：先推动，再在V-JEPA 2特征空间中进行预测”，arXiv:2610.23968，2026，doi：10.48550/arXiv.2610.23968。[20] S. Motamed、L. Culp、K. Swersky、P. Jaini 和 R. Geirhos，“生成式视频模型理解物理原理吗？”载于WACV论文集，2026年，第948–958页，doi：10.48550/arXiv.2501.09038。[21] X. Li 等，“PAC-NeRF：用于几何无关系统辨识的物理增强连续神经辐射场”，载于ICLR论文集，2023，doi：10.48550/arXiv.2303.05512。[22] T. Zhang 等，“PhysDreamer：通过视频生成实现与三维物体的基于物理的交互”，载于ECCV论文集，LNCS 15060，2024年，第388–406页，doi：10.1007/978-3-031-72627-9_22.[23] K. Aberman、P. Li、D. Lischinski、O. Sorkine-Hornung、D. Cohen-Or 和 B. Chen，“用于深度运动重定向的骨架感知网络”，ACM图形学汇刊，第39卷，第4期，文章62，2020，doi：10.1145/3386569.3392462。[24] S. Liu、M. Wang、B. Dai 和 C. Lu，“PALUM：用于统一运动重定向的基于部件的注意力学习”，arXiv:2601.07272，2026，doi：10.48550/arXiv.2601.07272。[25] C.-F. Pu、J.-P. Zhang、M.-H. Guo、Y.-P. Cao 和 S.-M. Hu，“TopoCap：学习用于单目视频转动画的拓扑无关运动先验”，载于SIGGRAPH会议论文，2026，arXiv:2606.12153，doi：10.1145/3799902.3811159。[26] A. Siarohin、S. Lathuilière、S. Tulyakov、E. Ricci 和 N. Sebe，“用于图像动画的一阶运动模型”，载于NeurIPS论文集，2019，doi：10.48550/arXiv.2003.00196。

[27]L. Hu、X. Gao、P. Zhang、K. Sun、B. Zhang 和 L. Bo，“Animate Anyone：用于角色动画的一致且可控的图像到视频合成”，载于 CVPR 会议论文集，2024 年，第 8153–8163 页，doi：10.1109/CVPR52733.2024.00779。[28]R. Li、C. Zheng、C. Rupprecht 和 A. Vedaldi，“Puppet-Master：将交互式视频生成扩展为部件级动态的运动先验”，载于 ICCV 会议论文集，2025 年，第 13405–13415 页，doi：10.48550/arXiv.2408.04631。[29] A. van den Oord、Y. Li 和 O. Vinyals，“基于对比预测编码的表示学习”，arXiv：1807.03748，2018 年，doi：10.48550/arXiv.1807.03748。[30] Y. Li 和 S. Mandt，“解耦序列自编码器”，载于 ICML 会议论文集，PMLR 80，2018 年，第 5670–5679 页，doi：10.48550/arXiv.1803.02991。[31]G. Alain 和 Y. Bengio，“使用线性分类器探针理解中间层”，arXiv：1610.01644，2016 年，doi：10.48550/arXiv.1610.01644。[32]A. Vaswani 等，“Attention Is All You Need”，载于 NeurIPS 会议论文集，2017 年，第 5998–6008 页，doi：10.48550/arXiv.1706.03762。[33] D. Hendrycks 和 K. Gimpel，“高斯误差线性单元（GELU）”，arXiv：1606.08415，2016 年，doi：10.48550/arXiv.1606.08415。[34] R. Xiong 等，“关于Transformer架构中的层归一化”，载于 ICML 会议论文集，PMLR 119，2020 年，第 10524–10533 页，doi：10.48550/arXiv.2002.04745。[35] O. Ronneberger、P. Fischer 和 T. Brox，“U-Net：用于生物医学图像分割的卷积网络”，载于 MICCAI 会议论文集，LNCS 9351，2015 年，第 234–241 页，doi：10.1007/978-3-319-24574-4_28。[36] I. Loshchilov 和 F. Hutter，“解耦权重衰减正则化”，载于 ICLR 会议论文集，2019 年，doi：10.48550/arXiv.1711.05101。[37] L. N. Smith 和 N. Topin，“超收敛：使用大学习率实现神经网络的极快速训练”，载于 SPIE 会议论文集，第 11006 卷，2019 年，文章编号 1100612，doi：10.1117/12.2520589。[38] E. Perez、F. Strub、H. de Vries、V. Dumoulin 和 A. Courville，“FiLM：通过通用条件层进行视觉推理”，载于 AAAI 会议论文集，第 32 卷第 1 期，2018 年，第 3942–3951 页，doi：10.1609/aaai.v32i1.11671。[39] C. Spearman，“两事物之间关联性的证明与测量”，《美国心理学杂志》，第 15 卷第 1 期，第 72–101 页，1904 年，doi：10.2307/1412159。[40] B. Efron 和 R. J. Tibshirani，《自助法导论》。美国纽约州纽约市：Chapman & Hall，1993 年，doi：10.1201/9780429246593。[41] F. Wilcoxon，“通过秩方法进行个体比较”，《生物统计学通报》，第 1 卷第 6 期，第 80–83 页，1945 年，doi：10.2307/3001968。[42] J. Cohen，《行为科学统计功效分析》，第 2 版。美国新泽西州希尔斯代尔：Lawrence Erlbaum，1988 年，doi：10.4324/9780203771587。