运动提示：用运动轨迹控制视频生成

Daniel Geng[1] [2]* Charles Herrmann[1] [3]† Junhwa Hur[1] 福雷斯特-科尔[1] Serena Zhang[1] Tobias Pfaff[1] 塔蒂亚娜-洛佩兹-格瓦拉[1] 卡尔-多尔施[1] 优素福-艾塔尔[1] 迈克尔-鲁宾斯坦[1] 孙晨[1] [3] 奥利弗-王[1] 安德鲁-欧文斯[2] 孙德清[1]

[1]谷歌 DeepMind [2]密歇根大学 [3]布朗大学

https://motion-prompting.github.io/

摘要

运动控制对于生成具有表现力和吸引力的视频内容至关重要；然而，现有的大多数视频生成模型主要依靠文本提示进行控制，难以捕捉动态动作和时间组合的细微差别。为此，我们训练了一种以时空稀疏或密集运动轨迹为条件的视频生成模型。与之前的运动调节工作不同，这种灵活的表示方法可以编码任意数量的运动轨迹、特定物体或全局场景运动以及时空稀疏运动；由于其灵活性，我们将这种调节方法称为运动提示。虽然用户可以直接指定稀疏轨迹，但我们也展示了如何将高层次的用户请求转化为去尾的、半密集的运动提示，我们将这一过程称为运动提示扩展。我们通过各种应用展示了我们方法的多功能性，包括凸轮时代和物体运动控制、与图像 "互动"、运动传输和图像编辑。我们的研究结果展示了新出现的行为，如逼真的物理现象，表明了运动提示在探测视频模型和与未来生成世界模型交互方面的潜力。最后，我们进行了定量评估和人体研究，并展示了强大的性能。

1. 简介

在视频生成中，运动是最重要的。它可以将视频从 "不可思议谷 "提升到逼真，或从业余提升到专业。运动可以引导注意力，增强故事性，并定义视觉风格。库布里克和黑泽明等高超的电影制作人都能巧妙地利用运动来创造引人入胜、身临其境的体验。实现逼真而富有表现力的动态效果，再加上细粒度的控制，是制作引人入胜的视频的关键。虽然文字仍然是文字是生成图像的主要控制信号，但在关注运动时，文字的局限性就显现出来了。虽然文字可以有效描述图像中的静态场景或高级动作，但却难以表达运动的微妙之处：*例如*，"一只熊迅速转过头来 "这样的提示可以有无数种解释。快速 "有多快？动作轨迹是怎样的？它是否应该加速？即使是详细的描述，也无法捕捉到一些细微的差别，比如缓进缓出的时机或同步动作。这些细微差别往往可以通过动作本身更好地传达出来。

受此启发，我们将运动作为一种强大的、与文本互补的控制方案进行了探索。我们的第一个看法是，为了充分利用运动的表现力，我们需要一种能够编码*任何*类型运动的表示法。为此，我们将时空稀疏*或*密集的运动轨迹 [22, 58] 作为理想的编码对象。运动轨迹又称粒子视频或点轨迹，可在整个视频中跟踪一组点的运动和可见度，提供极具表现力的运动编码。这种表示方法可以捕捉任意数量点的运动轨迹，表示特定物体或全局场景的运动，甚至可以处理时间稀疏的运动约束。此外，最近在点轨迹估计方面取得的进展已经产生了稳健高效的算法[12, 13, 36, 37]，能够处理真实世界的二维视频，生成用于训练的约束条件。考虑到这种运动表示方法的全面性和灵活性（类似于文本），我们将运动约束条件命名为*运动提示*。然后，我们在预先训练好的视频扩散模型[3]之上训练运动轨迹控制网[87]，以接受运动提示条件。

虽然这些运动提示可以定义任何类型的视频运动，但用户在实际操作中如何生成这些运动提示却不太清楚。稀疏轨迹给出了几个像素或斑块的粗略方向，可能很容易通过鼠标拖动来指定，但不能充分限制生成过程，也无法实现精细控制。相反，密集轨迹虽然能精确控制

*作为实习生完成的工作，†项目负责人

图 1.**运动提示**1) 我们在视频扩散模型的基础上训练一个通用的轨迹条件控制网络适配器。2) 为了使用这个模型，我们根据用户输入设计了*动作提示*，并展示了这个单一训练模型的各种功能，如物体控制、摄像机控制、物体和摄像机同时控制、动作转移和模型探测。我们将运动提示轨迹和生成视频中的相应帧可视化。轨迹的颜色仅用于可视化目的，轨迹表示运动的方向和幅度。此外，我们的一些动作提示源于用户的鼠标动作，为此我们将鼠标位置可视化。**我们强烈建议读者在我们的网页上观看视频结果。**

手动设计是不切实际的。为了解决这个问题，我们的第二个发现是，我们通常可以通过计算机视觉信号将用户的高级请求（*例*如 "围绕 *xz* 平面移动摄像机"、"旋转猫头"）转化为详细的运动轨迹。由于这一过程与图像生成中的文本*提示扩展*[11]或重写[4]相似，我们将其称为*运动*提示扩展。这种方法旨在缩小用户目标与我们的运动表示之间的差距。

我们发现了运动提示扩展可以成为有效工具的几种情况，包括（图 1）：将用户鼠标拖动转换为半密集运动轨迹

允许用户通过操作头发或沙子与图像 "互动"（第 4.1 节）；同时指定摄像机和物体的运动（第 4.4 节）；执行运动转换，将给定视频中的运动应用到不同的第一帧（第 4.5 节）；以及执行基于拖动的图像编辑（第 4.1 节）。虽然这些结果还不是实时的或因果性的，但它们有力地暗示了用户如何与生成式世界模型进行交互，并使我们能够探究生成器的视频先验，以了解它所学习到的物理学和一般世界知识的各个方面。

最后，我们展示了定量结果和针对基线的人类研究，表明我们的模型表现良好。

我们还介绍了消融情况，以验证我们的设计选择并提供启示。总之，我们的贡献如下

- 我们将重点放在作为调节信号的运动上，并将时空稀疏*或*密集的运动轨迹确定为一种灵活的运动表示，可以实现运动控制的许多方面。我们对控制网进行了训练，使其能够将这些*运动提示*作为调节信号。
- 我们提出了*运动提示扩展*，这一过程采用简单的用户输入并生成更复杂的运动轨迹，从而实现更精细的控制。
- 然后，我们将我们的方法应用到各种任务中，如物体控制、相机控制、运动转移或基于拖动的图像编辑。
- 我们还展示了物理等新兴行为，这表明这些运动提示可用于探测视频模型或与未来世界模型进行交互。
- 我们通过量化指标和人体研究对我们的方法进行了评估，结果表明我们的模型与基线相比表现良好。

2. 相关工作

视频扩散模型。视频扩散模型[25, 65, 66]已在视频生成方面展现出惊人的能力，它以自然语言为条件[3, 19, 20, 26, 27]，或将静态图像 "动画化 "为视频[6, 64, 81]。除了内容创建[52]之外，它们还可被视为通向创建世界模拟器这一宏伟目标的途径[7]，在为具身代理进行视觉规划方面已取得初步成功[15, 16, 83]。同时，视频先验是否能充分理解物理世界仍存在争议[35]，明确整合物理规则似乎是必要的[41, 85, 86]。我们的运动提示技术适用于任何视频扩散模型，它不仅提供了一种更灵活、更准确的视频生成运动模式，还可作为一种框架，用于探测训练有素的生成模型对三维或物理的理解。

运动条件视频生成。预先训练好的文本到视频模型可以根据新的运动模式或额外的运动调节信号进行调整。低等级适应（Low- rank adaptation，LoRA）[29] 是一种用于微调参数的通用技术，可用于少镜头运动调理[55, 90]。DreamBooth [57]最初用于生成个性化图像，也可用于生成具有运动控制功能的视频[78]。

早期研究提出通过稀疏模型进行视频控制 [2, 21]。最近的工作则是利用更强大的模型探索类似的想法。这些方法在符号选择上各有不同，但通常都需要某些复杂的工程技术来实现稳定的训练和更好的收敛。Tora[89]、MotionCtrl[75]、DragNUWA[84]、Image Con- ductor[39]和 MCDiff[9]采用了两阶段（*如*微调）的方法。先密集后稀疏的轨迹，或按顺序训练适配器）、专门的损耗[39, 45]、架构[17, 80]，或多个模块的多级微调[9, 61, 73]。MOFA-Video [46] 需要针对不同运动类型的独立适配器，TrackGo [92] 使用用户损耗和层，而其他作品 [39, 46, 75, 84, 89] 则设计了数据过滤管道。相比之下，我们发现更简单的训练方法就能产生高质量的结果。我们的模型只需一个阶段的训练，使用均匀采样的密集轨迹，无需任何专门的工程设计。然而，它能处理各种任务和运动，并能在迭代过程中对稀疏和密集轨迹进行泛化。

其他方法则使用以实体为中心的控制信号，如边界框[72, 78]、分割掩码[10, 79]、
人体姿态 [30, 82] 或摄像机姿态 [23, 76]。零镜头运动适配方法（*如* SG-I2V [45]、Trail- blazer [43]、FreeTraj [54] 和 Peekaboo [32]）采用了类似的策略，根据以实体为中心的掩码的变化指导视频生成，从而避免了视频模型的训练或微调。我们的运动提示提供了更灵活的界面，可控制不同粒度的运动生成。与明确控制扩散特征图的测试时间方法不同，我们的框架自然地平衡了控制信号的强度和编码视频先验的强度。

运动表示。由于我们的目标是将任何类型的运动作为视频生成模型的条件，因此选择合适的运动表示方法至关重要。最常用的表示方法是光流 [8、14、28、42、67、68]。虽然光流可以随时间进行链式处理，但误差会不断累积。缺乏遮挡处理功能也使其不适合我们的需要，而我们认为这对于良好的摄像机控制是必要的（第 4.3 节）。相比之下，长距离特征匹配 [5, 31, 33, 63] 或点轨迹 [12, 13, 22, 36, 37, 91] 则非常适合我们的应用。它可以处理遮挡，并允许在任意时间段内进行稀疏和密集跟踪。

3. 生成方法

我们的视频生成方法将单帧图像、文本提示和点轨迹形式的运动提示作为输入，我们将在第 4 章中解释如何创建点轨迹。全部实施细节见附录 A。

3.1. 动作提示

为了充分发挥运动的表现力，我们需要能够表示任何类型的运动。为此，我们在运动提示中使用了点轨迹，它可以编码空间（和时间）上的稀疏运动和密集运动、单个物体或整个场景的运动，甚至还可以通过可见性标志编码遮挡物。使用这种表示方法

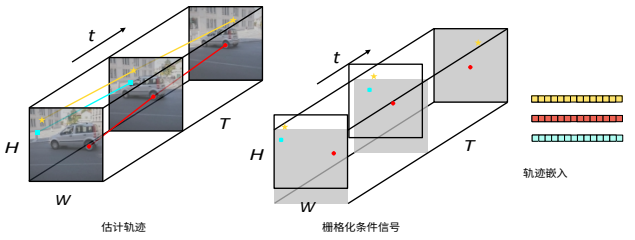

图 2.**调节轨迹。**在训练过程中，我们从视频中提取估计轨迹（左图），并将其编码到一个 $T \times H \times W \times C$ 维时空卷（中图）中。每条轨迹都有一个唯一的嵌入（右图），写入轨迹访问和可见的每个位置。所有其他位置都被设置为零。这种策略可以对任意数量和配置的轨道进行编码。

在一个统一的模型下，可以实现广泛的功能，如对象控制（第 4.1 节）、摄像机控制（第 4.3 节）、同时控制（第 4.4 节）、运动传输（第 44.1 .5 节）和基于拖动的图像编辑（第 节）。

形式上，我们用 $\mathbf{p} \in \mathrm{R}^{N \times T \times 2}$ 表示长度为 T 的 N 点轨迹集，其中第 n 个轨迹在第 t 个时间步的二维坐标为 $\mathbf{p}[n, t] = (x^n_t, y^n_t)$。在此外，与文献[87]类似，我们用 $\mathbf{v} \in \mathrm{R}^{N \times T}$ 表示轨迹的可见性、一个由 1 和 0 组成的数组，其中 0 表示屏幕外或被遮挡的轨迹，1 表示可见的轨迹。

3.2. 结构

我们的模型建立在 Lumiere 模型之上，Lumiere 是一个预先训练过的视频扩散模型[3]，经过训练可以在给定文本和第一帧定格的情况下以 16 fps 的速度生成 5 秒钟的视频。为了进行轨迹调节训练，我们使用了 ControlNet [87]，它要求在空间-时间卷中对轨迹进行编码，$\mathbf{c} \in \mathrm{R}^{(T) \times (H \times W) \times (C)}$，其中 T 是帧数，H 和 W 是生成视频的高度和宽度，C 是通道维度。为此，我们给每个轨道 $\mathbf{p}[n, :]$ 分配一个唯一的随机嵌入向量 $j^n \in R^{(C)}$。然后，对于轨迹访问并可见的每个时空环，我们只需将嵌入向量 $\phi^n \in \mathrm{R}^C$ 放置在该位置。图 2 展示了这一过程。换句话说，我们将 $\mathbf{c}$ 初始化为零，并设置

$$\mathbf{c}[t, x^n_t, y^n_t] = \mathbf{v}[n, t] j_n \quad (1)$$

在每个时间步 t，我们将所有轨迹的嵌入值 x(n)和 y(n)乘以可见度 $\mathbf{v}[n, t]$，如果轨迹在该位置和时间不可见，则将嵌入值清零。我们通过运动提示和提示扩展来量化 x^n_t 和 y^n_t 为简便起见，我们将其取最接近的整数。当多条轨迹经过同一时空位置时，我们会将嵌入值相加。轨迹嵌入 j^n 是从一个固定的池中随机抽取的，只是作为每个轨迹的唯一标识符。对于完全密集的轨迹，这种表示方法等同于从密集的嵌入式网格开始，

3.3. 数据

为了训练我们的模型，我们准备了一个与轨迹配对的视频数据集。我们在一个内部数据集上运行 BootsTAP [13]，这是一种现成的点跟踪方法，该数据集由 220 万个视频组成，大小调整为 128× 128，即我们基础模型的输出大小。我们密集地提取轨迹，从而在每段视频中得到 16,384 条轨迹以及预测的遮挡物，我们可以从以下数据中采样
训练过程中。我们没有以任何方式对视频进行过滤，我们的假设是，对不同的运动进行训练将产生一个更强大、更灵活的模型。

3.4. 训练

训练遵循 ControlNet [87]，将调节信号提供给基础模型编码器的可训练副本，并优化标准扩散损失。对于每段视频，我们都会从单形式分布中随机抽取一定数量的音轨，并按照上文所述构建调节信号。更多详情可参见附录 A。在训练过程中，我们观察到了各种现象。其一，我们发现损失与性能不相关无关。此外，与文献[87]类似，我们
观察到一种 "突然收敛现象"，即在很短的训练步数内，模型从完全忽略调节信号到完全训练。更多详情请见附录 B。

最后，我们发现我们的模型在多个方向上都表现出相当强的泛化能力。例如，虽然我们的模型是在随机取样的轨迹上训练的，导致训练期间轨迹在空间上均匀分布，但该模型可以泛化到空间上局部的轨迹条件（图 3 和图 6）。此外，虽然我们的模型是针对特定数量的轨道进行训练的，但它对更多轨道（图 5）或更少轨道（图 3、图 4 和图 6）的泛化效果出奇地好。最后，我们发现我们的模型可以泛化到不一定从第一帧开始的轨迹，尽管我们只针对这些轨迹进行了训练（图 3b）。我们假设这种泛化是由于网络中卷积产生的归纳偏差和我们在大量轨迹上训练模型这一事实共同作用的结果。

4. 运动提示

在本节中，我们将讨论不同类型的效果，以实现
我们通过运动提示和提示扩展来量化 x(n) 和 y(n)。具体而言，我们确定并演示了几种不同类型的扩展，如图 1 所示。文本提示和其他
每段视频的参数见表 A1。A1.**我们**
然后进行前向扭曲，类似于 [59]。

强烈建议读者在我们的网页上观看生成的视频。

4.1. "与图像 "互动

我们的模型能够与图像进行 "交互"。为此，我们建立了一个图形用户界面，用于显示静态图像和

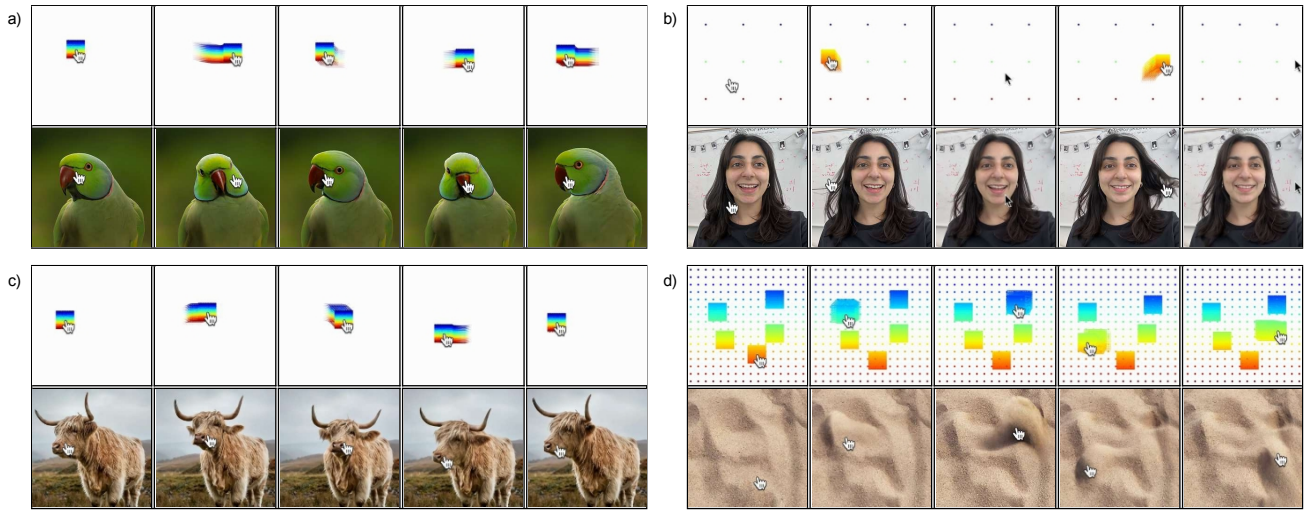

图 3.**与图像 "互动"。**我们将简单的用户输入（鼠标移动和拖动）转化为更复杂的动作提示，以帮助实现用户的意图。当鼠标拖动时，其轨迹被可视化为一只手，反之则为一个黑色光标。当鼠标被拖动时，以光标为中心的轨迹网格会被创建出来，如最上面一行所示。下一行显示的是生成的视频帧。通过这种方式提示我们的模型，我们可以（a）移动鹦鹉的头部或（c）移动奶牛的头部（b）摆弄头发或（d）与沙的图像 "互动"。我们还可以通过指定静态轨迹来保持背景静止，如（b）或（d）所示。**请注意，这些样本不是实时生成的，也没有时间上的因果关系。更多示例请访问我们的网页。**

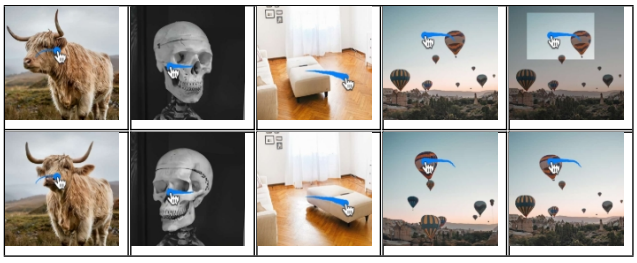

图 4 基于拖动的图像编辑**基于拖动的图像编辑**我们在第一行展示了输入图像，在最下面一行展示了基于拖动的编辑结果，两行中的拖动都是可视化的。此外，在最后一个示例中，我们还展示了如何保持图像区域的静态。

记录用户的鼠标拖动。如下所述，这些记录会被转换成轨迹，并与初始帧和文本一起输入模型。有关图形用户界面的更多信息，请参见附录 A。对于鼠标不断拖动的简单鼠标运动，这种方法与之前的稀疏轨迹分割视频生成方法类似[39, 45, 46, 61, 73, 75, 79, 84, 89, 92]。不过，由于我们的模型可以泛化到部分轨迹，因此我们还可以处理在不同时间不同位置的多次鼠标拖动，从而实现自然的用户控制，如图 3b 和图 3d。请注意，虽然我们是实时记录鼠标输入的，但我们的方法需要从视频扩散模型中采样，这并不是实时的--生成一个输出视频大约需要 12 分钟。

如图 3 所示，为了创建动作提示，我们将鼠标拖动转化为网格点轨迹。网格的密度

网格的密度和大小可由用户选择，类似于文献[39, 75, 79, 84]中对轨迹的高斯模糊处理，以指定运动的空间范围。但请注意，在我们的方法中，这一步只在推理时进行，而不是在列车运行时。此外，用户可以选择将静态轨道网格向下放置以保持背景静止（如图 3b 和图 3d），或者让轨道在鼠标拖动后继续存在（如图 3d）。

新出现的现象。我们发现，这些 "互动 "运动提示可以产生直接的运动，如图 3a 中鹦鹉的转头。但有趣的是，我们也能观察到更复杂的动态：*例如*，在图 3b 中，轨迹会拨动被试者的头发，或者在图 3d 中，沙子会被卷起。在这些例子中，我们基本上是在探测模型所学到的视频先验，通过这样做，我们可以直观地看到模型所学到的物理知识和对一般世界的理解。此外，由于我们的模型支持时间稀疏轨迹调节，因此我们可以有效地进行预测。也就是说，我们可以在短时间内用一个动作查询模型，然后让模型预测未来，这样我们就能回答 *"如果我这样拉头发或那样拉头发，头发会有什么表现？"*等问题，如图 3b 所示。

基于拖动的图像编辑。这种 "交互 "能力的一个自然应用就是基于拖动的图像编辑[1, 18, 44, 48, 56, 62]。这项任务包括接受用户提供的 "拖动 "并编辑图像，使对象跟随这些拖动。图 4 显示了定性结果。

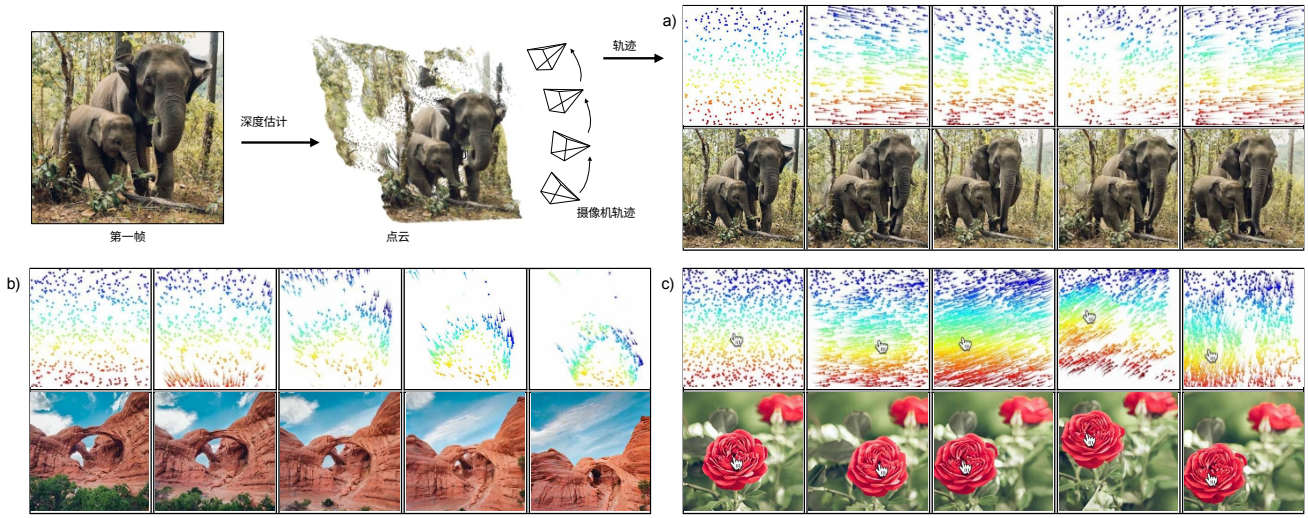

图 5.**利用深度控制摄像机**我们可以通过指定摄像机轨迹和使用现成的单目深度估算器计算点云来构建用于摄像机控制的运动提示。然后，我们将点投影到摄像机序列上，从而得到所示的点轨迹。我们还可以将用户的鼠标输入转换为摄像机轨迹，如示例 (c)。

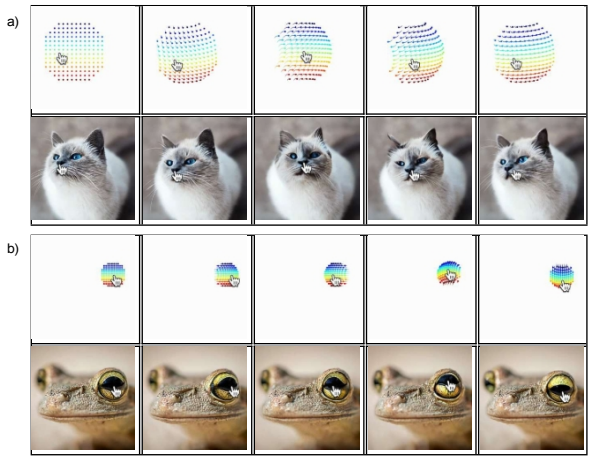

图 6.**利用基元控制物体**通过定义用户用鼠标操作的几何基元（*如*球体），我们可以获得对物体（*如*旋转）进行更精细控制的轨迹，而这是无法用单一轨迹来指定的。

4.2. 用基元控制物体

我们还可以将鼠标运动重新解释为操作代理几何基元，如球体。通过将这些轨迹放置在可大致与基元匹配的物体上，我们可以对物体进行比单独使用稀疏鼠标轨迹更精细的控制。例如，在图 6 中，我们在猫的头部和青蛙的眼睛上放置了一个球体，以精确地将这些物体旋转到不同的位置。在这种设置中，用户必须提供鼠标的运动以及要使用的球体的位置和半径。这样，用户就可以指定比平移更复杂的运动，而这些运动是很难用单一的的运动轨迹。有关实现细节，请参阅附录 A。

4.3. 深度相机控制

我们还可以设计运动提示，利用我们的模型实现摄像机控制。为此，我们首先在输入帧上运行一个现成的单目深度估算器 [51]，以获得场景的点云。然后，给定摄像机姿势轨迹，我们就能将点云重新投影到每个摄像机上，从而得到输入的二维轨迹。我们可以通过运行 z 缓冲来获得遮挡标志，从而进一步提高质量。

如图 5 所示，用这些运动提示来提示我们的模型，可以实现对摄像机的控制。我们可以像图 5a 一样让摄像头绕圈运行，或者像图 5b 一样让它向上划弧。此外，我们还可以将摄像头控制与鼠标记录结合起来，使其更加易于使用。为此，我们将按照第 4.1 节中的方法记录鼠标输入。然后，我们构建一个摄像机轨迹，使点云中的一个点跟随鼠标轨迹，并将摄像机控制在一个垂直平面上，如图 5c 所示。有关实现细节，请参阅附录 A。

需要注意的是，我们的模型既没有像之前的研究[39, 75, 76]那样根据摄像机的姿势进行训练，也没有以摄像机的姿势为条件。此外，我们的训练数据都不包括姿势注释。尽管如此，我们发现我们的模型可以实现令人信服的摄像机控制。这表明，与其根据特定类型的运动来训练视频模型，我们可以根据一般运动来训练模型，并通过使用运动提示来找出特定的能力。

4.4. 合成动作

通过将动作提示组合在一起，我们可以将各种功能结合在一起。例如，在图 7 中，我们将以下轨迹组合在一起

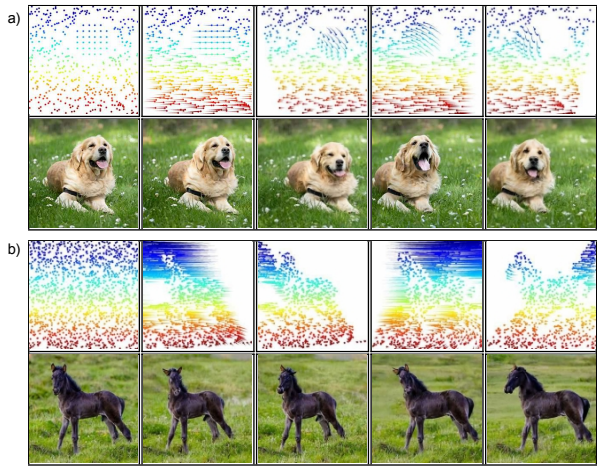

图 7.**运动提示的组合。**通过将运动提示组合在一起，我们可以同时实现对物体和摄像机的控制。例如，我们在这里移动狗和马的头部，同时将摄像机从左到右环绕。

这样就可以同时控制物体控制和摄像机控制。具体做法是将物体轨迹转换为位移，然后将这些脱位加到摄像机控制轨迹上。在二维空间中，这种构成只是一种近似值，对于摄像机的剧烈运动会失效，但我们发现，对于小到中等程度的摄像机运动，这种构成方式效果很好。需要再次指出的是，与之前的研究[39, 75, 76]相比，我们并没有专门针对这种能力进行训练。

4.5. 运动转移

许多类型的运动可能很难设计运动提示。如果视频中存在所需的动作，我们可以进行动作转移 [18, 73]，即从源视频中提取动作轨迹并将其应用到图像中。例如，我们可以提取人转动头部的运动轨迹，并将其用于猕猴木偶，如图 8 所示。更重要的是，我们发现我们的模型具有令人惊讶的鲁棒性，我们可以将运动应用到相当超域的图像中。例如，在图 8 中，我们将*猴子咀嚼*的动作应用到鸟瞰树木图像中。由此产生的视频具有一种有趣的效果，即暂停任何一帧视频都会消除源视频的感知[24]。只有在播放视频时，才能感知到*猴子*，这就是格式塔的共同命运效应[34]。

4.6. 故障、限制和探测模型

我们将失败分为两大类。第一类是我们的动作调节或动作提示失败。例如，我们在图 9a 中展示了一个例子，由于牛角被错误地 "锁定 "在背景上，牛头被不自然地拉伸。第二类是由于底层视频模型造成的故障。例如，在图 9b 中，我们拖动了棋子，但新的棋子自发地形成了。

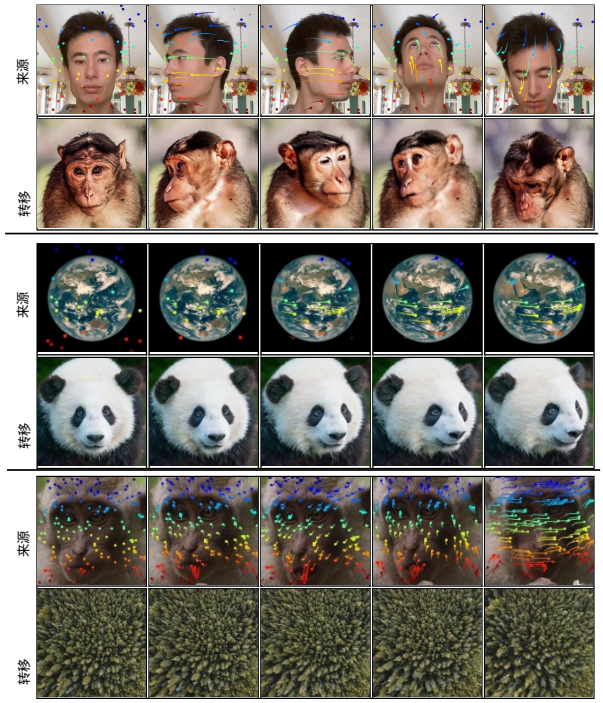

图 8.**运动转移**通过将源视频中的外显运动作为我们模型的条件，我们可以操纵一只猕猴，甚至将*猴子咀嚼*的运动转移到 树的照片上。**最好在我们的网页上观看视频。**

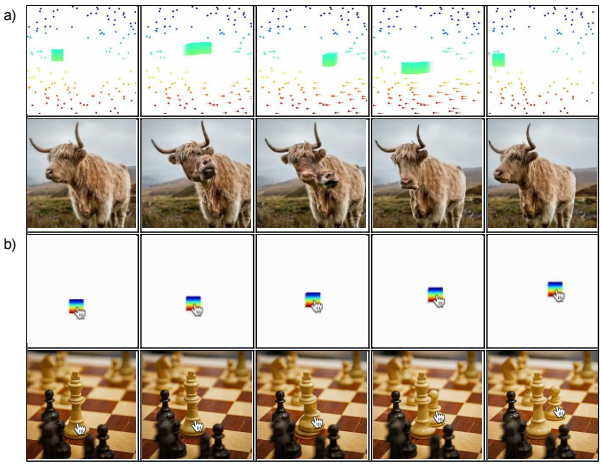

图 9.**通过故障进行探测**我们可以使用动作提示来探测底层模型的局限性。例如，拖动棋子会产生一个新棋子。

视频给定约束。这些类型的失败表明，我们或许可以利用动作提示来*探测*视频模型，并发现其学习表征的局限性。

5. 定量结果

除了上面的定性例子，我们还描述了一个定量基准，并将我们的方法与以下结果进行对比评估

表 1.**定量评估。**我们在 DAVIS 数据集的验证集上对生成视频的外观（PSNR、SSIM、LPIPS、FVD）和运动（EPE）进行评估。请注意
每种方法都是从不同的基础模型中训练出来的。

# 跟踪方法		psnr↑	ssim↑	lpips↓	fvd↓	epe ↓	
N= 1	图像导体	11.468	0.145	0.529	1919.8	19.224	
	DragAnything	14.589	0.241	0.420	1544.9	9	.135
	我们的	**15.431**	**0.266**	**0.368**	**1445.2**	14.619	
N= 16	图像导体	12.184	0.175	0.502	1838.9	24.263	
	DragAnything	15.119	0.305	0.378	**1282.8**	9.800	
	我们的	**16.618**	**0.405**	**0.319**	1322.0	**8.319**	
N= 512	图像导体	11.902	0.132	0.524	1966.3	30.734	
	DragAnything	15.055	0.289	0.381	1379.8	10.948	
	我们的	**18.968**	**0.583**	**0.229**	**688.7**	**4.055**	
	图像导体	11.609	0.120	0.538	1890.7	33.561	
N= 2048	DragAnything	14.845	0.286	0.397	1468.4	12.485	
	我们的	**19.327**	**0.608**	**0.227**	**655.9**	**3.887**	

最近的基线。此外，我们还进行了一项人体研究，并在本节描述了消融情况。

5.1. 轨迹条件生成评估

为了评估我们的轨迹文本和首帧条件视频生成技术，我们使用了 DAVIS 视频数据集[53]的验证分割。我们从数据集中提取首帧和轨迹，并将其与自动生成的文本提示一起输入模型。具体实施细节请参见附录 A。为了评估一系列轨迹密度，我们改变了调节轨迹的数量，从单一轨迹到 2048 个轨迹不等。

我们将我们的方法与最近的两项工作进行了比较：Image-Conductor [39]对 AnimateDiff [20]进行了微调，以处理摄像头和物体的运动；DragAnything [79]则通过微调稳定视频扩散 [6]，使 "实体 "沿轨道移动。为了评估生成视频的外观，我们计算了生成视频与地面实况视频之间的 **PSNR**、**SSIM** [74]、**LPIPS** [88] 和 **FVD** [69]。为了评估生成的视频与运动调理的匹配程度，我们使用了端点误差（**EPE**），计算方法是调理轨迹与根据生成的视频估算的轨迹之间的 L2 距离。

如表 1 所示。在几乎所有情况下，我们的模型都优于基准线。在某些示例中，DragAnything 在使用较少轨迹的 EPE 方面表现更好。这是因为 DragAnything 包含一个可以有效扭曲潜点的模块。虽然这种翘曲效果可能会产生适当的运动，但也会产生视觉假象，表现不佳的 PSNR、SSIM、LPIPS 和 FVD 结果就是证明。我们还在附录 C 中提供了 4 条和 64 条轨道的数据，为简洁起见，此处省略。

5.2. 人工研究

我们进行了一项人体研究，手动创建了一组包含单一轨迹的 30 个输入。我们进行了一次双向强迫选择测试，测试中我们会问 (1) 哪段视频是"....."。

表 2.**人类研究。**在 2AFC 人类研究结果中，我们列出了我们的方法与基线方法的胜率百分比。每列的样本数分别为 N= 103、N= 103 和 N= 115。

方法	运动一致性	运动质量	视觉质量
图像指挥	74.3 (± 1.1)	80.5 (± *1.0*)	77.3 (± *1.0*)
任意拖动	74.5 (± 1.1)	75.7 (± *1.1*)	73.7 (± *1.0*)

表 3.**消减。**我们在训练过程中对轨迹密度进行了消减，发现在密集轨迹上进行训练最适合我们的模型。

# 轨道	方法	PSNR↑	SSIM↑	LPIPS↓	FVD↓	EPE↓	稀疏
N= 4			15.075	0.241	0.384	**1209.2**	30.712
	密集+ 稀疏	15.162	0.252		0.379	1230.6	29.466
	密集		**15.638**	**0.296**	**0.349**	1254.9	**24.553**
	稀疏		15.697	0.284	0.355	1322.0	26.724

N= 2048 密集+ 稀疏	15.294	0.246		0.375	1267.8	27.931
密集		**19.197**	**0.582**	**0.230**	**729.0**	**4.806**

(2) 哪段视频的动作更逼真，以及 (3) 哪段视频的视觉质量更高。总共有 180 个问题，表 2 列出了我们方法的胜率以及 95% 的置信区间。2.当同时考虑运动和外观时，我们的方法在所有类别中都优于基线方法。实施细节见附录 A。

5.3. 消融

在表 3 中，我们介绍了一种消融情况。在表 3 中，我们介绍了一种消融方法，即只对*稀疏*点轨迹（1-8 条轨迹）和*密集*+*稀疏*轨迹（轨迹数从 2^0到 2^{13}对数采样）进行模型训练。我们发现，密集训练最为有效，尤其是对大量轨迹而言。令人惊讶的是，密集训练对稀疏轨迹也更有效。我们假设，这是因为使用稀疏轨迹提供的训练信号太少，因此在密集轨迹上进行训练更有效，然后再推广到更稀疏的轨迹上，不过这可能受到我们使用 ControlNet 和 zero con- volutions 的影响。我们使用了第 5.1 节中 DAVIS 评估的一个子集，但我们注意到，由于我们使用了较少的数据和较少的消融训练步骤，因此数据并不一致。

6. 结论

我们介绍了一种运动条件视频生成框架，该框架利用灵活的运动提示--可编码任意运动复杂性的时空轨迹。与之前的工作不同，这种表示法可以为摄像机、目标或整个场景指定稀疏或密集的运动。我们还引入了运动提示扩展，将高级运动请求转化为去尾提示。我们的多功能方法可实现运动控制、运动转移、图像编辑等应用，并通过单一的统一模型展示逼真物理等新兴行为。定量和人工评估证明了我们框架的有效性。