arXiv:2611.32432v1 [cs.CV] 24 Nov 2026

视频模型是出色的运动模型

Kian Izadpanah[1,2] Ali Mahdavi Amiri[3] Daniel Cohen-Or[4]
Yiorgos Chrysanthou[1,2] Andreas Aristidou[1,2]1塞浦路斯大学
[2]CYENS 卓越中心 [3] 西蒙菲莎大学 [4] 特拉维夫大学

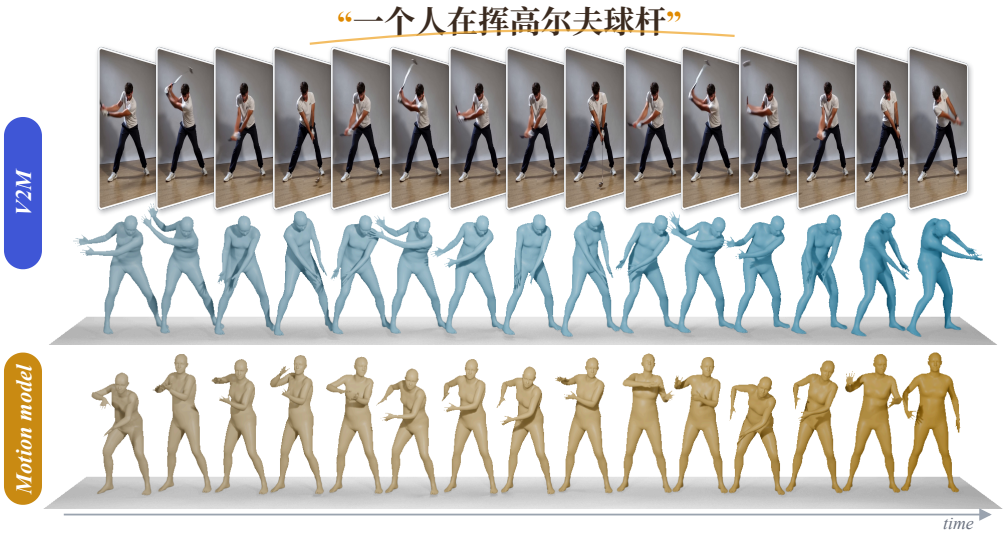

图1. 对于相同的文本提示，V2M 生成的高尔夫挥杆动作比专用运动模型更加完整、协调，展现了视频基础模型生成更丰富、语义连贯性更强的三维人体运动的潜力。

摘要

视频基础模型已成为强大的动态视觉内容生成器，在合成复杂动作和交互方面展现出卓越能力。我们认为，这些模型也已成为出色的运动生成器，其运动先验比专用运动模型所学习到的更加丰富、通用。它们的优势不止于视觉真实感，还体现在两个重要方面：运动自然度，包括连贯的身体协调性、物理合理性，以及与物体和环境的真实交互；以及运动范围，涵盖范围广泛得多的活动和语义连贯性，尤其是罕见、复杂和复合动作。为探究这一观点，我们提出了一个统一的评估框架，通过三种互补设置比较视频模型与专用运动模型：纯文本运动生成、从视频生成运动中重建稀疏关键帧，以及基于视频增强运动模型的输出。通过将所有输出转换为统一的三维骨骼表示，我们评估运动自然度、语义对齐、动作覆盖度、物理运动特征，以及对运动约束的遵循程度。结果表明，视频基础模型能够生成更自然且语义更丰富的动作，尤其适用于复杂及复合动作。此外，我们还研究了通过宽松骨骼控制实现的运动条件生成，发现专用动作模型并不能始终提升视频生成运动，而视频模型则可以利用稀疏运动指引来增强骨骼动画。我们的发现表明，专用运动模型的角色正在转变：从生成完整动画，转向引导和控制视频基础模型中已具备的更丰富运动能力。

项目页面：*video-beats-motion.github.io*

1. 引言

视频生成领域的最新进展，让我们离生成式世界模型的愿景更近了一步：这类模型不仅能捕捉世界的外观，还能理解世界如何运作并随时间演变 [11]。现代视频基础模型在大型、多样化的视频数据集上训练，展现出生成复杂人类活动、协调动作，以及人与人、人与物体和环境之间交互的卓越能力。这些能力表明，它们学到的表示不止于视觉外观，还捕捉了有关人体运动以及支配人类行为的关系的有用知识。这引出了一个有趣的问题：视频基础模型中蕴含的运动知识，能否用于生成更自然且语义更丰富的三维人体运动？

专门的生成式运动模型也在根据文本描述、稀疏关键帧和其他控制信号合成逼真人体运动方面取得了显著进展 [13, 33, 41]。通过显式建模骨骼姿势和轨迹，这些模型能够对角色动画进行精准控制。然而，运动模型与视频模型之间的一个重要区别在于其训练数据的规模和多样性。

尽管运动模型可以受益于更大规模的训练数据集，但收集多样、高质量且带有语义标注的3D运动数据仍然充满挑战且成本高昂。相比之下，视频数据丰富得多，使视频基础模型能够利用涵盖更广泛人类活动、物体和环境的数万小时视频记录进行训练 [54]。尽管视频数据与骨骼运动数据无法直接比较，但视频数据更高的可获得性和多样性可能带来更丰富的运动先验，尤其是对于在结构化运动数据集中代表性不足的罕见且复杂的活动、人-物交互以及复合动作。

在这项工作中，我们认为视频基础模型已经成为比专用动作模型更强大、通用性更高的运动生成器，无论是在运动的自然度，还是在其能够表示的动作范围方面都是如此。首先，视频模型已经习得丰富的运动先验，不仅能捕捉人类动作，还能理解身体部位、物体与环境之间的关系。这些知识使动作更加连贯且符合物理规律，例如，搬运箱子时保持稳定抓握、爬梯子时协调手脚落点，或是将多种动作结合起来，例如边走路边喝咖啡并打电话。其次，在规模大得多、种类丰富得多的数据集上训练，使视频模型接触到范围更广的活动与交互，包括结构化运动数据集中较少涉及的内容。重要的是，我们的主张关注底层三维运动的质量与语义一致性，而非生成视频的视觉真实感。如图1所示，在给定相同文本提示的情况下，视频基础模型能够生成比专用动作模型更丰富、更自然且语义连贯的三维人体运动。

为探究这一观点，我们提出了一个通用评估框架，将视频模型和运动模型的输出转换为统一的三维骨骼表示。我们考察了三种互补的实验设置。首先，纯文本生成比较两类模型能否根据同一提示词独立生成自然且具有语义意义的动作。其次，视频到动作考察专用动作模型能否在少量关键帧的条件下重建或改进视频生成的运动。第三，运动到视频研究视频模型能否在稀疏骨骼引导下丰富运动模型的输出，同时保留其预期的运动结构。这些设置使我们能够考察运动范围、自然度，以及视频模型和运动模型各自发挥的互补作用。

我们的实验表明，视频基础模型能够生成更自然且具有语义意义的动作，尤其是在复杂的复合动作方面。尽管专用动作模型并不能始终通过稀疏关键帧重建来改进视频生成的运动，但视频模型能够在松散控制下丰富骨骼动画，同时保留其原有意图。我们通过语义、感知和定量运动评估支持这些发现。我们的贡献为视频基础模型作为运动生成器提供了新的视角，提出了统一评估框架，并以实验证据展示了它们更丰富的运动先验。我们的发现表明，运动生成领域正在发生更广泛的转变：视频基础模型日益凭借丰富的语义理解和运动先验，助力生成自然且富有表现力的动画；而专用动作模型则提供互补能力，以改进和控制生成的运动。

2. 相关工作

2.1. 人体运动生成

生成模型的最新进展显著提升了基于自然语言的三维人体运动合成。早期方法（如 TEMOS [29], MDM [41],和 MotionDiffuse [49],）建立了文本到运动生成的框架，而 T2M-GPT [48], MotionGPT [22],和 MoMask [18] 则探索了离散运动表示与基于语言的生成。较新的方法，包括 Motion Mamba [51], BAMM [30], MotionLCM [14], 和 EMDM [56], , 进一步提升了运动质量、效率和可控性。K-Mask [1] 还支持细粒度的身体部位控制，而 Cond-

MDI [13] 支持基于稀疏关键帧的灵活动作插补。Kimodo [33] 通过更大规模的训练，扩展了可控运动生成。这些进展为比较专用动作模型与视频基础模型奠定了坚实基础。

尽管取得了这些进展，生成复杂的复合动作仍然颇具挑战，尤其是在必须同时执行多个动作时。模型可能会优先生成占主导地位的动作，而忽略提示词中的其他组成部分，导致运动不完整或语义不一致 [4, 23, 36]。连续动作也存在类似挑战：模型可能会遗漏动作、打乱动作顺序，或生成不自然的过渡 [3, 6, 20], , 也有模型难以保留全局语义 [55]。其他挑战还包括满足空间与物理约束，尤其是在人与物体交互时，这可能导致脚部滑动、协调动作不自然或地面穿透。总的来说，如何生成多样、自然且语义忠实的复杂复合动作，仍是一个有待解决的难题。

一个重要原因是，高质量、多样化且带有语义标注的3D运动数据有限。广泛使用的HumanML3D数据集 [17]包含约29小时的运动数据，而Motion-X [25]将这一规模扩展至约144小时。近期的研究进一步扩大了数据规模，例如，Kimodo使用约700小时的动作捕捉数据 [33],进行训练，MotionMillion则包含超过2,000小时的数据 [16]。VideoMDM [27] 还探索了从视频中提取的2D姿态学习三维运动先验。尽管如此，结构化运动数据集中对罕见活动、复杂交互和复合动作的呈现仍然较少。此外，许多标注只描述占主导地位的动作，对同时动作、时间关系和交互约束提供的信息有限。因此，挑战不仅在于数据规模，也在于运动多样性和语义丰富度。

2.2. 视频基础模型

视频基础模型通过大规模生成式训练不断发展，在合成逼真场景、人类活动和复杂交互方面展现出卓越能力。Sora [11], 、Movie Gen [31], 、CogVideoX [47], 、混元视频 [24], 、Open-Sora [54], 和 Wan [42] 等模型显著提升了视频质量、时间一致性和文本对齐能力。近期的发展包括 MiniMax-H3 [28]和 LTX-2 [19], , 进一步推进了视频生成、复杂指令遵循和灵活的条件控制。与我们的工作尤其相关的是，MiniMax-H3 和 LTX-2.5 分别通过 ControlNet Union 2.0 [2] 和 IC-LoRA Union Control [7] 条件控制，支持运动引导的视频生成。这些机制使现有动画能够提供结构引导，同时由视频基础模型生成额外的运动细节和交互。

视频模型与运动模型的一个关键区别在于训练数据的规模和多样性。视频模型从人体运动、物体、交互和环境的视觉观测中学习，因此可能捕捉到更丰富的上下文运动先验。例如，Open-Sora 报告称其使用了约 40,000 小时的视频进行训练 [54]。尽管这些数据集无法直接比较，但它们之间的差异促使我们探究视频模型能否生成更自然、语义更丰富的三维运动，尤其是复杂及复合动作。

近期研究探索了将视频模型用作三维动画的运动先验来源。《Animating the Uncaptured》 [8]将合成视频中的运动迁移到三维角色网格上，而 CoMoVi [52]和 EchoMotion [45]则探索联合视频与三维运动生成。SViMo [15]结合视觉先验与运动先验来处理人-物交互，而 Ani3DHuman [40] 则利用视频扩散实现非刚性角色动画。近期方法日益将视频生成与三维运动生成相结合。Rotstein 等人 [34] 和 Rahamim 等人 [32] 探索联合视频-运动合成，而 ISA4D [37] 实现了可控人体视频生成，R-DMesh [44]和 ActionMesh [35] 则将视频引导动画扩展到三维网格。这些方法展现了视频模型在运动合成方面日益增长的潜力。

人体姿态估计和三维运动重建领域的进步，也使得从视频中恢复骨骼动画变得越来越可行。D WPose [46]能够提供详细的全身关键点，而 WHAM [39] 和 GVHMR [38] 则可从单目视频中恢复三维人体运动和全局轨迹。包括 HumanMM [50], 、PEAR [43],和 OnlineHMR [53], 在内的近期方法，进一步推动了运动重建的发展。尽管遮挡、快速运动和复杂交互等问题仍然存在，这些进展使视频生成与可编辑三维动画之间建立日益实用的联系成为可能。这促使我们研究将视频基础模型作为通用运动生成器。

3. 比较框架

我们的核心观点是：视频基础模型已经成为比专用动作模型更强大、更通用的运动生成器，得益于更丰富的运动先验。然而，由于运动模型生成骨骼动画，而视频模型生成需要进行三维运动重建的图像序列，因此比较这些模型颇具挑战。为此，我们提出了一个统一的评估框架（图2），将所有输出转换为共享的三维骨骼表示。我们研究了三种互补设置：（A）独立的纯文本生成、

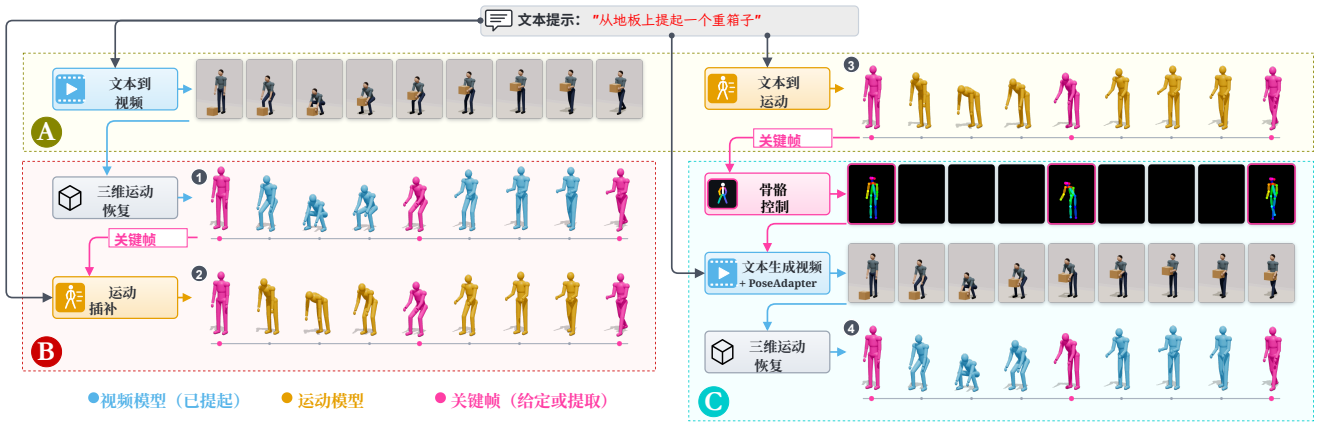

图2。 **比较框架**。 视频模型和专用动作模型使用相同的文本提示，且所有输出都转换为相同的三维骨骼表示。 左： 视频模型生成视频，通过提升 从中恢复三维运动。随后，专用动作模型根据 K 选定的关键帧 重建中间运动。 右： 专用动作模型生成骨骼动画 ，其 K 关键帧为视频模型提供松散骨架引导。最终生成的视频被重建为三维运动 。这些输出构成三组互补的比较： A （ vs. ）评估独立的纯文本生成； B （ vs. ）考察专用动作模型能否改进视频生成运动； C （ vs. ）探究视频模型能否在遵循稀疏运动引导的同时丰富骨骼动画。蓝色和橙色分别表示视频模型输出和运动模型输出，粉色表示选定的关键帧。

（B）利用以稀疏关键帧为条件的运动模型重建视频生成运动，以及（C）在宽松骨骼控制下通过视频生成丰富运动模型输出。这些比较评估了运动范围、自然度、语义对齐、动作覆盖范围，以及两类模型各自的互补作用。

3.1. 研究假设

我们提出三个假设，以评估我们的主张。

H1 – 运动范围与语义对齐。 与专用动作模型相比，视频基础模型能够生成范围更广、语义更丰富的人类活动，尤其是罕见且复杂的动作、同步或连续动作，以及与物体或环境的交互。我们主要在实验 A 中研究这一假设，评估语义对齐程度以及提示词中指定的单个动作的覆盖率。

H2 – 运动自然度与连贯性。 视频基础模型能够比专用动作模型生成更加自然且符合上下文的动作，在身体协调性、时间连贯性以及与预期动作的一致性方面表现更佳。我们通过感知比较以及贯穿全部三个实验的定量动作质量指标来研究这一假设。

H3 – 稀疏引导下的动作丰富化。视频基础模型丰富的运动先验可以通过宽松骨骼控制加以利用，在保留稀疏关键帧所提供的结构引导的同时，为现有动画增添更加自然且具有语义意义的动作

。我们主要在实验C中研究这一假设，通过感知比较评估动作丰富化效果，并通过关键帧保真度评估其与所提供引导的一致性。实验B则从另一个角度评估，专门的运动模型能否通过稀疏关键帧重建来改善视频生成运动。

3.2. 通用表示与重建

我们将运动表示为一个序列 $\mathbf{x} = (\mathbf{p}_t)_{t=1}^{T}$，其中每个姿态 $\mathbf{p}_t \in \mathbb{R}^{J\times 3}$ 包含 $J = 22$SMPL 身体关节 [26] 在帧 t处的位置。所有序列均转换为与重力对齐的、 Y向上坐标系，帧率为 30fps，同时保留其原始时长。我们对初始根节点位置进行归一化，并采用一致的骨架比例。

令 $\mathcal{V}$ 表示视频基础模型， $\mathcal{M}$ 表示专用运动模型， $\mathcal{L}$ 表示视频到动作重建算子。给定文本提示词 c，生成的运动如下：

$$\mathbf{x}^V = \mathcal{L}(\mathcal{V}(c)), \qquad \mathbf{x}^M = \mathcal{M}(c). \tag{1}$$

我们使用 MiniMax-H3 [28]进行纯文本视频生成，并使用其 ControlNet Union 2.0 [2] 条件控制机制进行骨架引导的视频生成。我们利用 GVHMR [38], 恢复三维运动，并在需要时由 DWPose [46] 提供骨骼引导。我们专用的运动模型包括用于文本到运动生成的 Kimodo [33]和 MotionGPT3 [57]，以及用于运动插值的 Kimodo 和 CondMDI [13]。在评估之前，所有输出都会转换为相同的骨架表示。

3.3. 比较流程

如图 2所示，三项实验考察了我们论点中相互补充的方面。实验 A 直接比较独立的运动生成，而实验 B 和 C 使用稀疏关键帧引导运动生成，并在每项实验中基于相同的引导，对运动质量进行比较。

对于 B 和 C，我们从每个序列中均匀选取 $K = 5$ 个完整身体的关键帧，包括序列的首帧和末帧。令 $\mathbf{x}_{\mathcal{K}}$ 表示这些运动关键帧$\mathbf{x}$，令 $\mathcal{R}_{\text{kf}}(\mathbf{x}_{\mathcal{K}})$ 表示将其投影到二维骨骼引导上，以用于视频生成。我们的比较定义如下：

$$\begin{aligned} &\text{A:} \quad \mathbf{x}^V \quad \text{vs.} \quad \mathbf{x}^M, \\ &\text{B:} \quad \mathbf{x}^V \quad \text{vs.} \quad \mathcal{M}(c, \mathbf{x}_K^V), \\ &\text{C:} \quad \mathbf{x}^M \quad \text{vs.} \quad \mathcal{L}\big(\mathcal{V}(c, \mathcal{R}_{\text{kf}}(\mathbf{x}_K^M))\big). \end{aligned} \tag{2}$$

A. 纯文本生成（H1、H2）。两个模型根据相同的文本提示，分别独立生成运动，不使用骨骼引导，从而可以直接比较自然度、语义对齐和动作复杂度。该设置直接考察视频基础模型是否具备更广泛、更丰富的动作生成能力。

B. 视频到动作（H2、H3）。我们从重建的视频运动中提取五个关键帧，将其与原始文本提示一同作为条件输入专用动作模型，考察生成的动画是否能够保持或提升运动自然度和语义一致性。

C. 动作到视频（H2、H3）。动作模型生成的动画中的五个关键帧，为视频生成提供稀疏、宽松的二维引导，同时结合原始文本提示。我们考察生成的运动是否会变得更加自然且具有更强的语义意义，同时保持关键帧所传达的预期结构。

3.4. 稀疏关键帧设定与松散控制

实验B和实验C通过稀疏关键帧研究条件运动生成，但二者在这些约束的应用方式上有所不同。在实验B中，专用动作模型接收三维全身关键帧，以生成中间骨骼运动。在实验C中，选定的姿态被投影到二维空间，并通过 ControlNet Union 2.0 提供给视频模型，且无需额外训练。

我们将后者称为宽松骨骼控制：选定的关键帧提供结构和时间引导，但不规定完整的动画。共享的引导有助于约束预期动作，使运动质量的比较更具意义，同时允许视频模型生成中间运动。我们的目标是利用其丰富的运动先验，为动画注入更自然且语义更丰富的运动。关键帧保真度衡量的是与引导的一致性，而不是对运动质量的绝对衡量。

需要特别指出的是，B 和 C 使用不同的条件控制表示（3D 对比 2D），因此二者是互补的，而不是可以直接等同比较的。此外，由于 ControlNet Union 2.0 是使用密集骨骼引导开发的，未指定的中间姿势并不一定不受约束。在解读结果时，我们会考虑这些差异以及潜在的三维重建误差。

3.5. 评估标准与指标

我们评估的是底层三维运动，而非生成视频的视觉外观，综合采用基于视觉语言模型的比较、人体感知评估和定量运动分析。所有动画均使用统一的骨骼结构和一致的渲染条件进行表示。详细的指标定义和操作流程见第S1节。

语义对齐与动作覆盖度（H1、H3）。我们使用 Qwen3-VL-32B-Instruct [5]作为自动评判器，从文本提示的语义对齐度和感知自然度方面比较匿名化运动对。我们报告胜/平/负（W/T/L）百分比，并采用反向呈现顺序以减少偏差。此外，我们的感知研究还评估动作覆盖度，考察生成动作是否执行了所要求的动作，尤其针对复杂的复合提示。

动作自然度与连贯性（H2）。我们通过基于视觉语言模型的比较和人工评估来评估感知自然度。量化测量指标包括位置加加速度、角加速度、估计的地面接触覆盖度以及脚部滑动。这些指标刻画了运动动力学和潜在伪影。

运动保真度（H3）。在实验 B 和 C 中，我们使用平均关节位置误差（MPJPE）、普罗克鲁斯特对齐平均逐关节位置误差（PA-MPJPE）以及测地线关节旋转误差（包括根节点朝向）来评估所提供骨骼引导的一致性。这些指标衡量的是与所选关键帧的一致性，而非运动质量本身，因为松散控制允许出现偏差，而这些偏差可能有助于提升自然度和语义连贯性。

重建注意事项（H1–H3）。由于视频生成的运动需要进行三维重建，测得的伪影可能源自重建而非生成。因此，我们结合视觉语言模型和人工评估来解读量化结果，并考虑可能存在的重建误差。

人体感知评估（H1–H3）。我们开展了一项人类感知研究，评估运动自然度、语义一致性、动作覆盖度以及可见的运动伪影。这些判断补充了自动语义评估和定量动作质量指标，尤其适用于复杂及复合动作。

表1。**成对VLM评估**。V2M相对于各基线的胜/平/负（%），按常见动作和复杂动作分组。

实验对比		组别	文本胜/平/负	自然胜/平/负
A	Kimodo	常见	57.6 /36.4 / 6.10	57.6 / 24.2 / 18.2
		复杂	59.4 / 21.9 / 18.8	56.2 /25.0 / 18.8
	M-GPT3	常见	39.4 /33.3 / 27.3	36.4 /33.3 / 30.3
		复杂	53.1 / 25.0 / 21.9	53.1 /21.9 / 25.0
B	Kimodo	普通	12.5 /81.2 / 6.20	12.5 / 81.2 / 6.20
		复杂	12.5 /84.4 / 3.10	9.4 / 87.5 / 3.10
	CondMDI	普通	18.8 /68.8 / 12.5	15.6 /62.5 / 21.9
		复杂	12.5 / 78.1 / 9.40	12.5 /78.1 / 9.40
C	Kimodo	常见	46.4 /39.3 / 14.3	53.6 / 28.6 / 17.9
		复杂	62.1 / 20.7 / 17.2	58.6 /24.1 / 17.2

4. 评估结果

我们评估视频基础模型生成的人体运动是否比专用动作模型更加自然且具有更丰富的语义。我们将语义正确性和动作覆盖度与物理运动质量区分开来，因为较低的加加速度或脚部滑动并不一定意味着成功执行了所要求的动作。因此，我们结合了基于视觉语言模型的比较、定量运动分析和人体感知评估。动画示例见补充视频。

实验设置。 我们将动作分为常见动作和复杂动作，包括同时和连续的运动（两者都使用 $K = 5$关键帧）。我们将视频到运动的流程（V2M）与实验 A 中的 Kimodo 和 M-GPT3、实验 B 中的 Kimodo 和 CondMDI，以及实验 C 中的 Kimodo 进行比较。使用 GVHMR 重建视频生成的运动，并且所有输出共享相同的 3D 骨骼表示。

4.1. 语义对齐与自然度

我们使用 Qwen3-VL-32B-Instruct 通过成对的视觉语言模型比较来评估语义对齐（H1、H3）和运动自然度（H2），并遵循补充材料第 S1.1 节中所述的协议。该评估在实验 A、B 和 C 中分别包含 65、64 和 57 个提示词，其中分别有 28、28 和 25 个涉及多个动作。表1 分别报告了 V2M 相对于各基线的胜/平/负（W/T/L）百分比，并区分常见动作和复杂动作。

实验 A：仅提示生成。 在文本匹配（58.5% 的胜率）和自然度（56.9%）方面，V2M 优于 Kimodo。与 M-GPT3 相比，V2M 在这两项指标上的胜率分别为 46.2% 和 44.6%；在复杂动作上的表现更出色（两项指标均为 53.1%，而常见动作为 39.4% 和 36.4%）。这些发现表明，视频模型能够生成更自然、语义更连贯的动作，尤其是复杂动作（图。

3）。例如，V2M 成功地结合了行走、打电话



图 3. **实验 A：仅提示生成**。 扫地动作凸显了手臂与身体动作的协调性，而网球动作则展示了挥拍的执行效果。边走边交谈并拿着咖啡，可用于评估复合动作中的动作覆盖度。V2M 生成的运动更自然、更完整、语义更连贯，也与文本提示词更加契合。

以及手提咖啡杯，而专用运动模型则难以复现所有动作。结果还表明，模型对身体—物体关系有了更深入的理解，例如爬梯子时手部与脚部落点的协调，或在搬运箱子时保持稳定的抓握而不穿透箱子（见补充视频）。

实验B：视频到动作。 Kimodo 和 CondMDI 根据五个稀疏关键帧重建 V2M运动。比较结果显示，与原始 V2M运动相比，它们并未表现出一致的改进，大多数评判结果为平局：在文本匹配方面，Kimodo 和 CondMDI 分别有 82.8% 和 73.4% 的结果为平局；在自然度方面，分别为 84.4% 和 70.3%。这些结果表明，专用动作模型并不能持续改善原始运动，尽管以其关键帧为条件，但有时可能无法再现其完整的语义内容（图 4）。

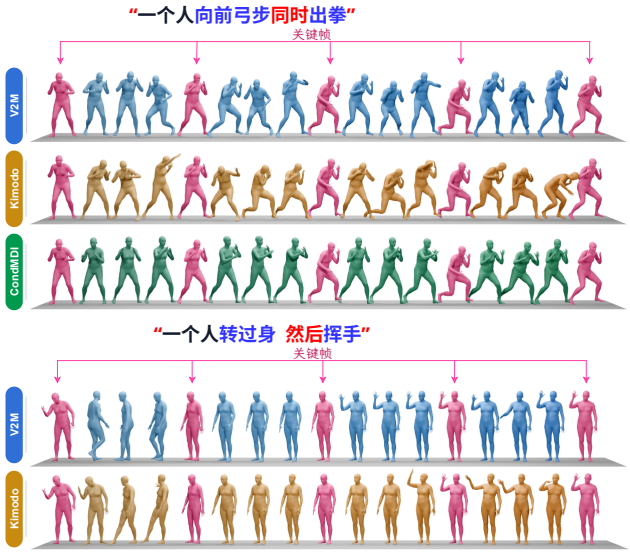

图 4. **实验B：视频到动作**。冲刺出拳测试同时动作，而转身挥手考察动作顺序。专用动作模型生成中间动作，但并不能始终改善原始视频派生的动画，同时可能无法复现其完整的语义内容。

实验C：动作到视频。 V2M 使用 Kimodo 生成的关键帧作为引导；在文本匹配（胜率 54.4%）和自然度（56.1%）方面，人们更偏好其结果，而非原始 Kimodo运动。在复杂动作中，这一优势进一步扩大，分别达到 62.1% 和 58.6%。这些发现表明，在宽松骨骼控制下，视频模型能够为骨骼动画增添更有意义且更自然的动作（图 5）。

4.2. 动作质量与保真度

我们通过定量动作质量指标（H2）和关键帧保真度测量（H3），补充感知评估与语义评估。我们测量估计的地面接触覆盖度（$Y \leq 5$厘米）、选定关节（手部、足部和骨盆）的位置加加速度、角加速度均方根以及脚部滑动。表2 总结了这些结果。对于 B 和 C，我们还使用 MPJPE、PA-MPJPE 和测地旋转误差评估关键帧保真度。

地面接触覆盖度。与专用动作模型相比，V2M 表现出更低的估计地面接触覆盖度（在 A、B 和 C 中分别为 44.15%、34.36% 和 41.06%）。这可能反映出更动态的动作和更频繁的抬脚，而专用模型有时会产生受限或静止的动作（见补充视频）。请注意，覆盖度并不能直接衡量动作完成。

运动平滑度和脚部滑动。与 Kimodo 相比，V2M 通常表现出更高的位置加加速度，但角加速度均方根更低；而脚部滑动在 A 中更高，但

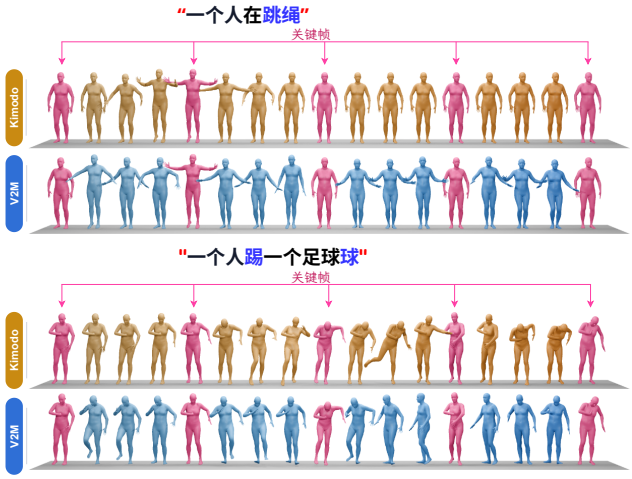

图5。**实验C：动作到视频**。V2M 生成了更完整且语义更丰富的动作，无论是跳绳还是踢足球，都展现了视频模型在松散控制下丰富骨骼动画的潜力。

表2。**运动质量指标**。接触覆盖率仅作描述之用。箭头表示数值越低越好，但数值较低并不一定意味着运动更加自然。若 V2M 在某项实验中取得最低误差，则以粗体标出。

		接触	位置加加速度(m/s^3) ↓		角度均方根	脚部滑动
实验	方法	(%) 手部	Feet	骨盆	(°/s^2) ↓	(cm/s) ↓
A	V2M	44.15	185.20	90.55	57.23	**466.80**	12.30
	Kimodo	72.77	**95.93**	**80.90**	40.87	1332.00	**6.19**
	M-GPT3	64.07	108.90	96.50	**36.78**	2608.00	12.56
B	V2M	34.36	231.70	**82.60**	**70.24**	**770.70**	**7.47**
	Kimodo	64.66	**151.30**	126.20	78.60	1724.00	7.82
	CondMDI	55.78	249.50	205.30	185.80	1716.00	17.07
C	V2M	41.06	119.40	**80.54**	70.60	**351.40**	**7.91**
	Kimodo	85.68	**81.32**	97.79	**35.23**	982.50	11.92

在 B 和 C 中更低。这些结果不一的表现凸显了运动学指标的局限性：专用动作模型可能通过减少动作或保持静止来实现较低的误差，而非成功执行所要求的动作。因此，这些指标应结合语义对齐和感知自然度来解读。事实上，我们的人类研究（第4.3节）一致认为 V2M 的运动自然度更高。更多示例请见补充视频。

关键帧保真度。 在 B 中，V2M 作为参考运动，并从中提取关键帧。Kimodo 和 CondMDI 的 MPJPE 分别为 3.04 cm 和 12.17cm，22 个关节的测地旋转误差分别为 15.39° 和 10.75°，根部旋转误差分别为 2.25° 和 4.03°。这些结果表明，与参考关键帧相比，位置偏差相对较小。在 C 中，V2M 由 Kimodo 关键帧的二维投

表 3. **人体感知评估**。V2M 相对于 Kimodo 的成对胜/平/负（%），以及视频（V）和运动（M）模型的动作完成率（%）。

实验	自然度 胜/平/负	语义 胜/平/负	更少的伪影 胜/平/负	动作完成 V/M（%）
A	72 / 15 / 13	88 / 7 / 5	63 / 25 / 12	86 / 38
B	75 / 14 / 11	74 / 10 / 16	72 / 17 / 11	84 / 63
C	61 / 20 / 19	66 / 16 / 18	54 / 31 / 15	86 / 63

影引导。重建得到的三维运动的 MPJPE 为 15.03 cm，PA-MPJPE 为 7.10 cm，关节旋转误差为 20.57°，根节点旋转误差为 12.60°。这些误差可能部分反映了二维引导中的深度歧义以及额外的三维重建步骤。值得注意的是，在松散控制下，较低的关键帧误差并不一定意味着更好的运动，因为偏差可能会在保留预期动作的同时，丰富中间动作。

4.3. 人体感知评估

我们开展了一项在线感知研究，比较 V2M 与 Kimodo 在实验 A、B 和 C 中分别生成的 9、8 和 9 对动画。共有 43 名参与者完成了该研究。参与者评估了运动自然度（H2）、语义对齐和动作覆盖度（H1、H3），以及可见动作伪影（H2）；详细流程见补充材料第 S1.4节。对于成对比较，我们从 V2M 的视角报告胜/平/负（W/T/L）百分比，并分别报告每个模型的动作完成率。

表 3 总结了实验结果。在运动自然度方面，在 A、B 和 C 中，参与者分别在 72%、75% 和 61% 的比较中更偏好 V2M；在语义对齐方面，这一比例分别为 88%、74% 和 66%。参与者也更偏好伪影较少的 V2M（63%、72% 和 54%）。这些偏好支持 H2，表明即使定量动作质量指标显示出不一致的结果，视频生成的动作仍被感知为更加自然、具有语义意义且视觉上更加连贯。虽然视觉语言模型评估显示成对比较主要为平局，但人类研究揭示了对 V2M 更明确的偏好，说明两种评估流程的敏感性存在差异。

动作覆盖度。 除语义对齐外，我们还评估动画是否执行了所有要求的动作，尤其是在同时和/或连续动作的情况下；专用动作模型可能会优先呈现主要动作，而忽略其他动作。在实验 A、B 和 C 中，V2M 的动作完成分数分别为 86%、84% 和 86%，而 Kimodo 分别为 38%、63% 和 63%。对于复合动作，V2M 也更好地保留了时间顺序和同时性（分别为 86%、86% 和 84%，而 Kimodo 为 45%、69% 和 66%）。这些发现支持 H1，表明其具备更广泛的语义运动能力；而在 B 和 C 中，稀疏引导下的优势则为 H3 提供了进一步证据。

4.4. 讨论

我们的结果表明，视频基础模型拥有比专用动作模型更丰富的运动先验，尤其适用于复杂及复合动作。在纯文本生成（A）中，V2M 的语义对齐、感知自然度和动作覆盖度更佳，展现了其执行多个指定动作的能力（H1）。尽管定量指标呈现的结果不一，感知评估仍认为视频生成运动在自然度和减少可见伪影方面更胜一筹（H2）。在稀疏引导下，专用动作模型并未改善 V2M运动（B）；而视频模型则能为运动模型动画增添更自然、语义意义更丰富的动作（C）。关键帧保真度衡量的是稀疏引导的保留程度，不一定能反映运动质量（H3）。这些发现表明，两者各有互补作用：视频基础模型为生成提供丰富的语义和运动先验，而专用动作模型则具备精细化和结构控制能力。补充视频通过具有代表性的示例展示了这些差异。

5. 结论与未来挑战

本研究进一步提出，视频基础模型已成为强大的运动生成器，拥有比专用动作模型更丰富、更通用的运动先验。我们的实验表明，它们的优势不仅体现在视觉质量上，还体现在运动自然度、语义连贯性以及执行复杂组合动作的能力上。专用动作模型虽未能持续改善视频生成运动，但视频模型能够在稀疏引导下丰富骨骼动画。视频基础模型不必完全依赖专用动作模型，而是可以利用更丰富的语义和运动先验进行生成；专用动作模型则通过细化和控制与之互补。

这一视角带来了若干挑战。首先，视频基础模型中蕴含的运动知识仍然是隐式的，尽管运动重建方法不断进步，要提取准确且可编辑的 3D 动画仍然十分困难。这些局限在快速运动、遮挡和复杂交互中尤为明显，正如先前的工作所指出的 [12, 21]。其次，利用这些丰富的运动先验需要新的控制形式，既要保留预期动作，又不能过度约束生成运动。宽松的骨骼引导 [9, 10]提供了一个初步方向，但我们仍需要更具表现力且适应性更强的控制机制。最后，更广泛的挑战在于将这一视角从单个人体运动拓展到交互角色、物体和环境的生成与控制。

最终，我们设想未来的动画不再主要是从头生成运动，而是引导和塑造视频基础模型中已经蕴含的丰富运动知识。

参考文献

[1] Faisal Ahmed、Chenqiu Zhao 和 Anup Basu。K-Mask：用于可控文本到运动合成的运动学感知掩码建模。载于《计算机视觉—ECCV 2026》，第 228 247 2026 2 – 页。Springer，。Alibaba [2] PAI。MiniMax-H3-Fun-Controlnet-Union- 2.0。https : / / huggingface . co / alibaba - pai/MiniMax-H3 0，2026 2 3，4 -Fun-Controlnet-Union- 。 。[3] Nikos Athanasiou、Mathis Petrovich、Michael J. Black 和 G¨ul Varol。TEACH：三维人体的时间动作组合。载于《国际三维视觉会议（3DV）》，第 414 423, 2022 3 – 页。Nikos Athanasiou、Mathis Petrovich、Michael J. Black 和 G¨ul Varol。SINC：用于同步动作生成的三维人体运动态空间组合。载于《IEEE/CVF国际计算机视觉会议（ICCV）论文集》，第9984–9995页，2023年。3[5] Shuai Bai、Yuxuan Cai、Ruizhe Chen、Keqin Chen、Xionghui Chen、Zesen Cheng、Lianghao Deng、Wei Ding、Chang Gao、Chunjiang Ge 等。Qwen3-vl技术报告。arXiv预印本 arXiv:2511.21631，2025年。5[6] German Barquero、Sergio Escalera 和 Cristina Palmero。通过融合位置编码实现无缝人体运动组合。载于《IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集》，第457–469页，2024年。3[7] Matan Ben-Yosef、Tavi Halperin、Naomi Ken Korem、Mo-hammad Salama、Harel Cain、Asaf Joseph、Anthony Chen、Urska Jelercic 和 Ofir Bibi。AVControl：用于训练视听控制的高效框架。arXiv预印本 arXiv:2603.24793，2026年。3[8] Marc Bened´ı San Mill´an、Angela Dai 和 Matthias Nießner。Animating the uncaptured：利用视频扩散模型进行人形网格动画。载于《第十四届学习表征国际会议（ICLR）论文集》，2026年。3[9]Shariq Farooq Bhat、Niloy J. Mitra 和 Peter Wonka。LooseControl：提升ControlNet以实现广义深度条件控制。载于《ACM SIGGRAPH 2024会议论文集》，纽约州纽约市，美国，2024年。ACM。8[10] Shariq Farooq Bhat、Niloy J. Mitra 和 Kalyan Sunkavalli。LooseControlVideo：使用空间分块实现导演式视频控制。载于《计算机视觉—ECCV 2026》，第73–90页。Springer，2026年。8[11] Tim Brooks、Bill Peebles、Connor Holmes、Will DePue、Yufei Guo、Li Jing、David Schnurr、Joe Taylor、Troy Luh-man、Eric Luhman、Clarence Ng、Ricky Wang 和 Aditya Ramesh。作为世界模拟器的视频生成模型。OpenAI技术报告，2024年。2、3

[12] Hila Chefer、Uriel Singer、Amit Zohar、Yuval Kirstain、Adam Polyak、Yaniv Taigman、Lior Wolf 和 Shelly Sheynin。《VideoJAM：用于增强视频模型中运动生成的联合外观-运动表征》。载于《第42届国际机器学习会议论文集》，第7595–7616页。PMLR，2025年。8[13] Setareh Cohan、Guy Tevet、Daniele Reda、Xue Bin Peng 和 Michiel van de Panne。《利用扩散模型实现灵活的动作插补》。载于《SIGGRAPH'24论文集》，美国纽约州纽约市，2024年。ACM。2、3、4[14] Wenxun Dai、Ling-Hao Chen、Jingbo Wang、Jinpeng Liu、Bo Dai 和 Yansong Tang。《MotionLCM：通过潜在一致性模型实现实时可控运动生成》。载于《计算机视觉—ECCV 2024》，第390–408页。Springer，2024年。2[15] Lingwei Dang、Ruizhi Shao、Hongwen Zhang、Wei Min、Yebin Liu 和 Qingyao Wu。《SViMo：手-物体交互场景中视频与运动生成的同步扩散》。载于《神经信息处理系统进展（NeurIPS）》，2025年。3[16] Ke Fan、Shunlin Lu、Minyue Dai、Runyi Yu、Lixing Xiao、Zhiyang Dou、Junting Dong、Lizhuang Ma 和 Jingbo Wang。《Go to zero：迈向基于百万规模数据的零样本运动生成》。载于《IEEE/CVF计算机视觉国际会议（ICCV）论文集》，第13336–13348页，2025年。3[17] Chuan Guo、Shihao Zou、Xinxin Zuo、Sen Wang、Wei Ji、Xingyu Li 和 Li Cheng。《从文本生成多样且自然的三维人体运动》。载于《IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集》，第5152–5161页，2022年。3[18] Chuan Guo、Yuxuan Mu、Muhammad Gohar Javed、Sen Wang 和 Li Cheng。《MoMask：三维人体运动的生成式掩码建模》。载于《IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集》，第1900–1910页，2024年。2[19] Yoav HaCohen、Benny Brazowski、Nisan Chiprut、Yaki Bitterman、Andrew Kvochko、Avishai Berkowitz、Daniel Shalem、Daphna Lifschitz、Dudu Moshe、Eitan Porat 等。《LTX-2：高效的联合音视频基础模型》。arXiv预印本 arXiv:2601.03233，2026年。3[20] Seong-Eun Hong、Jaeyoung Seon、Juyeong Hwang、JongH-wan Shin 和 HyeongYeop Kang。《Event-T2M：用于复杂文本到运动合成的事件级条件控制》。载于《国际学习表征会议（ICLR）》，2026年。3[21] Paul Janson、Tiberiu Popa 和 Eugene Belilovsky。《迈向从视频扩散模型生成运动》。载于《计算机视觉—ECCV 2024研讨会》，第415–426页。Springer，2025年。8[22] Biao Jiang、Xin Chen、Wen Liu、Jingyi Yu、Gang Yu 和 Tao Chen。《MotionGPT：将人体运动作为外语》。载于《神经信息处理系统进展（NeurIPs）》，2023年。2[23] Yue Jiang、Mingyu Yang、Liuyuxin Yang、Yang Xu、Bingxin Yun 和 Yuhe Zhang。《Motion-Adapter：用于复合动作文本到运动生成的扩散模型适配器》。arXiv预印本 arXiv:2604.16135，2026年。3[24] Weijie Kong、Qi Tian、Zijian Zhang、Rox Min、Zuozhuo Dai、Jin Zhou、Jiangfeng Xiong、Xin Li、Bo Wu、Jianwei Zhang、

等。《混元视频：大型视频生成模型的系统性框架》。arXiv预印本 arXiv:2412 03603，2024年 3 。.[25] Jing Lin、Ailing Zeng、Shunlin Lu、Yuanhao Cai、Ruimao Zhang、Haoqian Wang 和 Lei Zhang。《Motion-X：3D表现性全身人体运动数据集》。arXiv预印本 arXiv:2307.00818，2023年。3[26] Matthew Loper、Naureen Mahmood、Javier Romero、Gerard Pons-Moll 和 Michael J. Black。《SMPL：一种蒙皮多人线性模型》。ACM Trans. Graph.，34(6)，2015年。4[27] Amir Mann、Gal Michael Harari、Merav Keidar 和 Or Litany。《VideoMDM：迈向基于2D监督的3D人体运动生成》。载于《神经信息处理系统进展》，2026年。3[28] MiniMax。《MiniMax H3：打破任务与模态边界的开放模型》。MiniMax研究博客，2026年。3、4[29] Mathis Petrovich、Michael J. Black 和 G¨ul Varol。《TEMOS：从文本描述生成多样的人体运动》。载于《计算机视觉—ECCV 2022》，第480–497页。Springer，2022年。2[30] Ekkasit Pinyoanuntapong、Muhammad Usama Saleem、Pu Wang、Minwoo Lee、Srijan Das 和 Chen Chen。《BAMM：双向自回归运动模型》。arXiv预印本 arXiv:2403.19435，2024年。2[31] Adam Polyak、Amit Zohar、Andrew Brown 等。《Movie Gen：一组媒体基础模型》。arXiv预印本 arXiv:2410.13720，2024年。3[32] Ohad Rahamim、Dvir Samuel、Idan Schwartz 和 Gal Chechik。《所有模态都是平等的，但视频更加平等：弥合联合视频生成中的交叉注意力差距》。arXiv预印本 arXiv:2609.27901，2026年。3[33] Davis Rempe、Mathis Petrovich、Ye Yuan、Haotian Zhang、Xue Bin Peng、Yifeng Jiang、Tingwu Wang、Umar Iqbal、David Minor、Michael de Ruyter、Jiefeng Li、Chen Tessler、Edy Lim、Eugene Jeong、Sam Wu、Ehsan Hassani、Michael Huang、Jin-Bey Yu、Chaeyeon Chung、Lina Song、Olivier Dionne、Jan Kautz、Simon Yuen 和 Sanja Fidler。《Kimodo：扩展可控人体运动生成》。arXiv预印本 arXiv:2603.15546，2026年。2、3、4[34] Noam Rotstein、Yoad Tewel、Chen Tessler、Haggai Maron、Gal Chechik、Or Litany、Yoni Kasten 和 Yuval Atzmon。《关节、摄像机、动作！面向任意骨架共同生成3D运动与视频》。arXiv预印本 arXiv:2610.28632，2026年。3[35] Remy Sabathier、David Novotny、Niloy J. Mitra 和 Tom Monnier。《ActionMesh：利用时序3D扩散进行动画3D网格生成》。载于《IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集》，2026年。3[36] Nathan Sala、Ofir Abramovich、Ariel Shamir、Daniel Cohen-Or、Andreas Aristidou 和 Sigal Raab。《MultiAct：通过定制化注意力引导从复合文本生成文本到运动》。载于《SIGGRAPH'26论文集》。ACM，2026年。3[37] Ruizhi Shao、Yinghao Xu、Yujun Shen、Ceyuan Yang、Yang Zheng、Changan Chen、Yebin Liu 和 Gordon Wetzstein。《用于高效4D人体视频生成的跨空间注意力》。ACM Trans. Graph.，44(4)，2025年。3[38] Zehong Shen、Jiawei Wang、Huaijin Pi、Yan Xia、Zhi Cen、Zechen Hu、Hujun Bao、Ruizhen Hu、Xiaowei Zhou 和 Sida Peng。GVHMR：基于重力视角坐标从单目视频进行全局人体运动恢复。IEEE Transactions on Pattern Analysis and Machine Intelligence，48(10)：12569–12581，2026年。3，4[39] Soyong Shin、Juyong Kim、Eni Halilaj 和 Michael J. Black。WHAM：利用准确的三维运动重建世界坐标系人体。发表于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第 2070–2080 页，2024年。3[40] Qi Sun、Can Wang、Jiaxiang Shang、Yingchun Liu 和 Jing Liao。Ani3DHuman：通过自引导随机采样实现逼真三维人体动画。发表于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第 12651–12662 页，2026年。3[41] Guy Tevet、Sigal Raab、Brian Gordon、Yonatan Shafir、Daniel Cohen-Or 和 Amit H. Bermano。人体运动扩散模型。发表于国际学习表征会议（ICLR）论文集，2023年。2[42] Wan Team、Ang Wang、Baole Ai、Bin Wen、Chaojie Mao、Chen-Wei Xie、Di Chen、Feiwu Yu、Haiming Zhao、Jianxiao Yang 等。Wan：开放且先进的大规模视频生成模型。arXiv 预印本 arXiv:2503.20314，2025年。3[43] Jiahao Wu、Yunfei Liu、Lijian Lin、Ye Zhu、Lei Zhu、Jingyi Li 和 Yu Li。PEAR：像素对齐的表现力人体网格恢复。发表于 SIGGRAPH'26 论文集，纽约州纽约市，美国，2026年。ACM。3[44] Zijie Wu、Lixin Xu、Puhua Jiang、Sicong Liu、Chunchao Guo 和 Xiang Bai。R-DMesh：通过校正动态网格流实现视频引导的三维动画。发表于 ACM SIGGRAPH 会议论文集，2026年。3[45] Yuxiao Yang、Hualian Sheng、Sijia Cai、Jing Lin、Jiahao Wang、Bing Deng、Junzhe Lu、Haoqian Wang 和 Jieping Ye。EchoMotion：通过双模态扩散变换器统一人体视频与运动生成。发表于国际学习表征会议（ICLR）论文集，2026年。3[46] Zhendong Yang、Ailing Zeng、Chun Yuan 和 Yu Li。通过两阶段蒸馏实现有效的全身姿态估计。发表于 IEEE/CVF国际计算机视觉会议（ICCV）论文集，第 4210–4220 页，2023年。3，4[47] Zhuoyi Yang、Jiayan Teng、Wendi Zheng、Ming Ding、Shiyu Huang、Jiazheng Xu、Yuanming Yang、Wenyi Hong、Xiaohan Zhang、Guanyu Feng、Da Yin、Yuxuan Zhang、Weihan Wang、Yean Cheng、Bin Xu、Xiaotao Gu、Yuxiao Dong 和 Jie Tang。CogVideoX：采用专家变换器的文本到视频扩散模型。发表于国际学习表征会议（ICLR），2025年。3[48] Jianrong Zhang、Yangsong Zhang、Xiaodong Cun、Yong Zhang、Hongwei Zhao、Hongtao Lu、Xi Shen 和 Ying Shan。利用离散表征从文本描述生成运动。发表于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第 14730–14740 页，2023年。2[49] Mingyuan Zhang、Zhongang Cai、Liang Pan、Fangzhou Hong、Xinying Guo、Lei Yang 和 Ziwei Liu。MotionDif

fuse：基于扩散模型的文本文本驱动的人体运动生成。IEEE Transactions on Pattern Analysis and Machine Intelligence，46(6)：4115–4128，2024年。2[50] Yuhong Zhang、Guanlin Wu、Ling-Hao Chen、Zhuokai Zhao、Jing Lin、Xiaoke Jiang、Jiamin Wu、Zhuoheng Li、Hao Frank Yang、Haoqian Wang 和 Lei Zhang。HumanMM：从多镜头视频进行全局人体运动恢复。发表于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第 1973–1983 页，2025年。3[51] Zeyu Zhang、Akide Liu、Ian Reid、Richard Hartley、Bohan Zhuang 和 Hao Tang。Motion Mamba：高效的长序列运动生成。发表于计算机视觉—ECCV 2024，第 265–282 页。Springer，2024年。2[52] Chengfeng Zhao、Jiazhi Shu、Yubo Zhao、Tianyu Huang、Jiahao Lu、Zekai Gu、Chengwei Ren、Zhiyang Dou、Qing Shuai 和 Yuan Liu。CoMoVi：三维人体运动与逼真视频的协同生成。arXiv 预印本 arXiv:2601 10632，2026年3。.[53] Yiwen Zhao、Ce Zheng、Yufu Wang、Hsueh-Han Daniel Yang、Liting Wen 和 L´aszl´o A. Jeni。OnlineHMR：基于视频的在线世界坐标系人体网格恢复。发表于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第 13951–13961 页，2026年。3[54] Zangwei Zheng、Xiangyu Peng、Tianji Yang、Chenhui Shen、Shenggui Li、Hongxin Liu、Yukun Zhou、Tianyi Li 和 Yang You。Open-Sora：让高效视频制作惠及所有人。arXiv 预印本 arXiv:2412.20404，2024年。2，3[55] Zhiwei Zheng、Shibo Jin、Lingjie Liu 和 Mingmin Zhao。用于文本生成动作的下一尺度自回归模型。发表于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第 16376–16386 页，2026年。3[56] Wenyang Zhou、Zhiyang Dou、Zeyu Cao、Zhouyingcheng Liao、Jingbo Wang、Wenjia Wang、Yuan Liu、Taku Komura、Wenping Wang 和 Lingjie Liu。EMDM：用于快速高质量运动生成的高效运动扩散模型。发表于计算机视觉—ECCV 2024，第 18–38 页。Springer，2024年。2[57] Bingfan Zhu、Biao Jiang、Sunyi Wang、Shixiang Tang、Tao Chen、Linjie Luo、Youyi Zheng 和 Xin Chen。MotionGPT3：人体运动作为第二模态。发表于国际学习表征会议（ICLR），2026年。4

视频模型是优秀的运动模型

补充材料

S1. 评估指标与实现细节

我们通过语义对齐、感知自然度、物理运动质量和关键帧保真度来评估生成的运动，具体如主论文第 3.5 节所述。所有运动均使用相同的22关节SMPL骨架以30 fps表示，采用 Y 朝上的坐标系，并将地平面设为 $Y = 0$。我们将第 t 帧的关节 j 的三维位置记为 $\mathbf{p}_{t,\,j}$ 并将帧间隔记为 $\Delta t = 1/30$s。除非另有说明，评估指标均按序列计算，然后对评估序列进行等权平均。

S1.1. 语义对齐与运动自然度

基于视觉语言模型的成对评估。 我们使用Qwen3-VL-32B-Instruct作为自动评判器，比较V2M与专用动作模型生成的运动在语义对齐和感知自然度方面的表现。评估分别包含实验A、B和C中的65、64和57个文本提示。每次比较均使用针对同一提示词生成的两个动画，这些动画采用统一的3D骨骼格式，并在相同条件下进行渲染。

运动渲染与盲法。 所有运动均使用中性SMPL人体（$\boldsymbol{\beta} = \mathbf{0}$）表示，采用与重力方向对齐的 Y-向上坐标系，帧率为30 fps。动画渲染为平地面上颜色均匀的灰色人物，使用固定光照和静态摄像机。针对每个提示词，摄像机均根据所有待比较方法生成的运动进行联合调整，以确保各动画具有相同的视点和取景。文本中提及的物体（如球或工具）不会渲染；其相关动作则根据人物的身体动作进行评估。所有片段均分配匿名标识符，并向评审者隐去模型身份和源文件名。

成对判定。 对于每次比较，我们向视觉语言模型提供原始文本提示和每段动画的采样帧，并将其调整为 576×576 像素。评审者首先描述每段片段中观察到的动作，然后独立回答两个问题：（i）哪段动画更符合文本，需综合考虑所有要求的动作、动作的时间顺序以及任何同时动作；（ii）哪段动画呈现的人体运动更自然，需考虑脚部滑动、漂浮、抖动、不合理姿势和僵硬的身体部位。所有模型对均使用相同的指令。推理采用确定性贪心解码。

偏好聚合。 为减少呈现顺序偏差，每对动画都会评估两次，并反转动画的呈现顺序。只有当两次评估对同一方法更优达成一致时，才记录为胜或负；明确的平局以及两种顺序下意见不一致的情况均计为平局。

我们从 V2M 的角度报告胜/平/负（W/T/L）百分比。结果进一步按常见和复杂 提示词进行划分：后者包括明确归类为复杂的动作，以及要求连续或同时动作的提示词；前者则包括其余的单动作提示词。此外，我们还计算净偏好（胜数减去负数），并使用 10,000 次提示词级自助法重采样计算其 95% 置信区间。这些自动化评判用于评估感知到的语义质量和运动质量，并结合人类感知研究与定量运动指标进行解读。

VLM评判提示词。 为确保结果可复现，我们提供了供视觉语言模型使用的成对评估提示词。所有实验和模型比较均采用相同的指示，只有文本描述和输入动画有所不同。评判者会先分别描述每段运动，再比较其语义对齐程度和感知自然度。下方展示了一个代表性连续动作的提示词实例。

> 你是人体运动方面的专家。每段视频都展示一个三维人体模型在平地面上运动，由固定摄像机拍摄；两段视频使用相同的摄像机、身体和光照。画面中不会渲染其他任何内容：即使人物与物体、工具、球、楼梯、梯子、墙或横杆互动，这些东西也不会出现。开始时，人物大致面向摄像机，因此人物自身的左侧出现在画面右侧；向前移动意味着朝摄像机移动（人物会在画面中变大，足部会在画面中向下移动）。
>
> 这两段视频根据同一段文本生成：“一个人原地踏步，然后敬礼。”请完整观看两段视频。
>
> **第 1 步**。 按顺序用 2-3 句话描述视频 1 中人物的动作。
>
> **第 2 步**。 按顺序用 2-3 句话描述视频 2 中人物的动作。
>
> **第 3 步**。 文本匹配：哪个视频的文本匹配效果更好？检查文本的每个部分（每个动作、动作顺序、同时进行的动作，以及由未渲染的物体所暗示的身体动作）。
>
> **第 4 步**。 自然：哪个视频看起来更像真人在移动？注意脚部滑动、漂浮、抖动、不可能的姿势，以及僵硬或静止的身体部位。
>
> 仅当两个视频确实完全相同时，才回答“平局”。
>
> **仅返回 JSON：**
> {"视频1": "...", "视频2": "...", "文本匹配度": "1 | 2 | 平局", "自然度": "1 | 2 |平局", "理由": "一句话"}

S1.2. 物理运动质量

我们通过位置加加速度、旋转平滑度、局部抖动、地面接触覆盖度、脚部滑动和地面穿透来评估运动质量。这些指标刻画了运动动力学和潜在伪影，但无法独立证明感知自然度或动作正确性。尤其是，较低的运动学误差也可能是动作幅度有限或处于静止状态所致。

位置加加速度（米/秒³） ↓。我们使用关节位置的三阶时间导数来衡量关节加速度的突变，并通过有限差分进行近似：

$$\mathbf{j}_{t,j} = \frac{\mathbf{p}_{t+3,j} - 3\mathbf{p}_{t+2,j} + 3\mathbf{p}_{t+1,j} - \mathbf{p}_{t,j}}{\Delta t^3}. \tag{S1}$$

一组关节的平均加加速度幅值 $\mathcal{J}$ 为：

$$J_{\mathcal{J}} = \frac{1}{|\mathcal{T}|\,|\mathcal{J}|} \sum_{t\in\mathcal{T}} \sum_{j\in\mathcal{J}} \|\mathbf{j}_{t,j}\|_2, \tag{S2}$$

其中 $\mathcal{T}$ 包含有效帧。我们报告六个关节（左、右手腕，左、右踝关节，头部和骨盆）的平均加加速度，同时分别报告手部和足部的平均值，以及头部和骨盆的单独数值。计算根节点相对加加速度时，会在求导前从每个关节轨迹中减去根节点位置。较高的加加速度可能表明存在突兀的运动伪影，但也可能反映出更动态的动作。

角加速度均方根（°/秒²） ↓。我们使用关节旋转 $\mathbf{R}_{t,j} \in SO(3)$ 评估旋转平滑度。角速度通过对数映射，根据连续帧之间的旋转进行近似：

$$\boldsymbol{\omega}_{t,j} = \frac{\mathrm{Log}(\mathbf{R}_{t,j}^{\top}\mathbf{R}_{t+1,j})}{\Delta t}, \tag{S3}$$

其中 Log 返回一个旋转向量。角加速度近似为：

$$\boldsymbol{\alpha}_{t,j} = \frac{\boldsymbol{\omega}_{t+1,j} - \boldsymbol{\omega}_{t,j}}{\Delta t}. \tag{S4}$$

我们计算五个关节（左、右手腕，左、右脚踝以及骨盆）的角加速度均方根，不包括头部：

$$A_{\mathrm{RMS}} = \sqrt{\frac{1}{|\mathcal{T}|\,|\mathcal{J}_5|} \sum_{t\in\mathcal{T}} \sum_{j\in\mathcal{J}_5} \|\boldsymbol{\alpha}_{t,j}\|_2^2}. \tag{S5}$$

结果转换为每秒平方度。我们还评估全部 22 个关节以及各个关节的旋转平滑度。

局部运动抖动（cm） ↓。我们使用七帧二阶 Savitzky–Golay 滤波器（SG7）测量根节点相对关节轨迹中的高频变化。令 $\mathbf{q}_{t,\,j}$ 表示根节点相对关节位置，$\tilde{\mathbf{q}}_{t,\,j}$ 表示其平滑轨迹。抖动计算为残差的均方根：

$$\mathrm{Jitter} = 100\sqrt{\frac{1}{|\mathcal{T}|\,|\mathcal{J}_6|} \sum_{t\in\mathcal{T}} \sum_{j\in\mathcal{J}_6} \|\mathbf{q}_{t,j} - \tilde{\mathbf{q}}_{t,j}\|_2^2}. \tag{S6}$$

我们评估六个关节（手腕、脚踝、头部和骨盆），并以厘米为单位报告结果。该指标能够捕捉局部不规则性，但也可能对快速且有意的动作进行惩罚。

地面接触覆盖度（%）。我们同时利用脚部垂直位置和垂直速度来估计潜在的脚部与地面接触。当脚部高度不超过 5 cm，且其垂直速度的绝对值不超过 20 cm/s 时，认为脚部处于接触状态：

$$c_{t,f} = \mathbb{I}\left(p_{t,f}^{Y} \le 0.05 \ \wedge\ |v_{t,f}^{Y}| \le 0.20\right), \tag{S7}$$

其中，f 表示左脚或右脚，位置以米为单位测量，速度以 m/s 为单位。我们报告满足这些条件的足部帧观测所占的百分比：

$$\mathrm{GCC} = \frac{100}{2T} \sum_{t=1}^{T} \sum_{f\in\{L,R\}} c_{t,f}. \tag{S8}$$

接触覆盖率是一种描述性指标，而非误差度量。较低的覆盖率可能表示更具动态性或腾空的动作，但也可能反映脚部放置不准确或重建伪影。我们还评估了 7.5、10、12.5 和 15 cm 的高度阈值。

脚部滑动（cm/s） ↓。我们测量估计的接触状态下脚部的水平速度。对于每只脚，水平速度为：

$$v_{t,f}^{XZ} = \left\| \frac{\mathbf{p}_{t+1,f}^{XZ} - \mathbf{p}_{t-1,f}^{XZ}}{2\Delta t} \right\|_2, \tag{S9}$$

令 $\mathcal{C}$ 包含同时满足高度（$Y \le 0.05$m）和垂直速度（$|v^{Y}| \le$0。20m/s）条件的有效足部帧观测。脚部滑动计算如下：

$$\mathrm{FS} = \frac{100}{|\mathcal{C}|} \sum_{(t,f)\in\mathcal{C}} v_{t,f}^{XZ}, \tag{S10}$$

其中，因子 100 用于将 m/s 转换为 cm/s。没有有效接触样本的序列不纳入滑动平均值。数值越低通常表示脚部落点越稳定，不过该指标应结合估计的接触覆盖率一并解读。此外，我们还报告了采用其他接触高度阈值时的结果。

地面穿透（cm） ↓。 我们估算脚关节低于地平面（$Y < 0$）的穿透量，并将穿透深度定义为：

$$d_{t,f} = 100 \max(0, -p^{Y}_{t,f}). \tag{S11}$$

我们报告最大穿透深度、发生穿透的样本的平均深度、所有足部关节样本的平均深度，以及穿透事件、帧和足部关节样本的数量。穿透事件是指某个足部关节持续位于地平面以下的一段连续时间区间。由于这些测量使用的是关节位置，而非完整的足部网格，因此只能近似反映地面相交伪影。

S1.3. 关键帧保真度

我们在实验B和实验C中评估关键帧保真度，两项实验采用不同的条件表示。在实验B中，专用动作模型接收从 V2M 运动中提取的完整三维骨架关键帧。在实验C中，视频模型接收来自 Kimodo 生成的运动的二维投影，随后将生成的视频重建为三维。我们将生成的关节位置和旋转与对应的原始三维参考姿态进行比较。

令 $\mathcal{K}$ 表示在时间上匹配的参考关键帧集合，其中生成的位置为 $\hat{\mathbf{p}}_{t,\,j}$，参考位置为 $\mathbf{p}_{t,\,j}$。

平均逐关节位置误差（MPJPE） ↓。 我们测量生成序列与参考序列中对应关节之间的平均欧几里得距离：

$$\mathrm{MPJPE} = \frac{100}{|\mathcal{K}|J} \sum_{t \in \mathcal{K}} \sum_{j=1}^{J} \|\hat{\mathbf{p}}_{t,j} - \mathbf{p}_{t,j}\|_2. \tag{S12}$$

其中 $J = 22$。MPJPE 以厘米为单位报告。

普罗克鲁斯特对齐平均逐关节位置误差（PA-MPJPE） ↓。 此外，我们使用包含旋转、平移和统一缩放的最优相似变换，将每个生成关键帧分别与对应的参考姿态进行对齐，然后计算 MPJPE。这一处理减少了全局对齐差异，并衡量关节位置的残余差异。PA-MPJPE 以厘米为单位报告。

关节旋转误差 (°) ↓。 我们使用生成旋转与参考旋转之间的测地距离来评估旋转保真度：

$$d_R(\hat{\mathbf{R}}, \mathbf{R}) = \cos^{-1}\left(\mathrm{clip}\left(\frac{\mathrm{tr}(\hat{\mathbf{R}}^{\top}\mathbf{R}) - 1}{2}, -1, 1\right)\right). \tag{S13}$$

我们报告所有 22 个关节和匹配关键帧的平均测地误差，并将其转换为角度。此外，我们还分别报告 21 个非根关节的平均误差和根关节方向误差。

时间匹配与解读。 仅评估生成序列中存在有效对应帧的参考关键帧；未匹配的关键帧将被排除。指标按序列分别计算，然后以相同权重求平均，同时分别记录匹配关键帧数和参考关键帧数。

需要特别指出的是，实验B和实验C并未施加等效约束。实验B直接根据三维姿态对骨骼运动进行条件控制，而实验C使用稀疏二维投影，无法唯一确定深度或完整的三维关节构型。此外，实验C还引入了额外的视频到三维重建阶段。因此，保真度差异可能部分反映了投影歧义和重建误差。这些指标衡量的是选定关键帧处的结构一致性，而非运动轨迹的自然度或完整保留程度。

S1.4. 人体感知评估

研究设置。 我们开展了一项盲法在线感知评估，通过三维骨骼动画对比 V2M 与 Kimodo。每组动画对均与用于生成该运动的文本描述一同呈现。两段动画均使用中性色头像，并在完全相同的可视化条件下渲染，不透露其底层模型。在整个比较集合中，视频动画和运动模型动画会分别出现在两侧，而不会始终占据同一位置。共有 43 名参与者完成了该研究，每人评估 26 组动画对。动画对的选择和呈现顺序因参与者而异。

成对问题。 对于每组动画对，参与者需回答三个问题：(i)“哪种运动与描述更匹配？”，(ii)“哪种运动看起来更像真实人物在移动？”，以及 (iii)“哪种运动包含的可见错误更少？”这些问题分别评估语义对齐、感知自然度和可见动作伪影。语义问题关注所观察到的运动在多大程度上体现了所要求的动作，而后两个问题则区分人类运动的整体观感与更局部的错误，例如脚部滑动、抖动或不合理的身体姿势。三个实验中的问题均保持一致。

动作覆盖度。 为了检验动画是否完成了要求的动作，而不只是完成主要动作，我们还针对可以单独评估的动作使用特定提示词检查表。参与者在不直接比较两个模型的情况下，评估每个动画描述中的单独组成部分，针对每个动

画进行评估。例如，对于要求一个人转身后挥手的提示词，则分别从这两个动作的角度进行评估。对于复合提示词，还会增加一项，用于评估这些动作是否按照要求的顺序发生，或是否同时发生。这样，我们便能区分单独动作的存在与其时间上的协调性。

偏好结果的报告与计分。对于每个成对比较问题，参与者选择“左侧”“右侧”“大致相同”或“无法判断”。偏好 V2M 计为胜，偏好 Kimodo 计为负。在正文报告的胜 / 平/负（W/T/L）结果中，“大致相同”和“无法判断”合并为非偏好类别，并计为平。因此，每个问题的百分比均根据所有已完成的回答计算；平局百分比则是在胜和负的百分比四舍五入后得出，以确保每行总和为 100%。

对于动作检查清单，参与者针对每个组成部分选择“是”“部分符合”“否”或“无法判断”。我们分别为“是”“部分符合”和“否”赋予 1、0.5 和 0 分，并在计算相应平均值时排除“无法判断”。我们分别计算视频和运动模型动画的动作完成度分数：对有效的单项动作分数取平均值，并以百分比形式报告。时间顺序和同时性采用相同的评分方式，但与动作完成度分开报告。未完成的提交不纳入统计。参与者不会看到原始源运动作为参考；骨骼引导的保留情况则通过定量关键帧保真度单独评估。