arXiv:2601.00590v1 [cs.CV] 2 Jan 2026

SAFEMO：基于语言学的可信文本到动作生成遗忘

YilingWang[1*] **Zeyu Zhang**[12*†] **Yiran Wang**[1] **Hao Tang**[2‡]1澳大利亚国立大学 2北京大学

*贡献相同。 †项目负责人。 ‡通讯作者：bjdxtanghao@gmail.com。

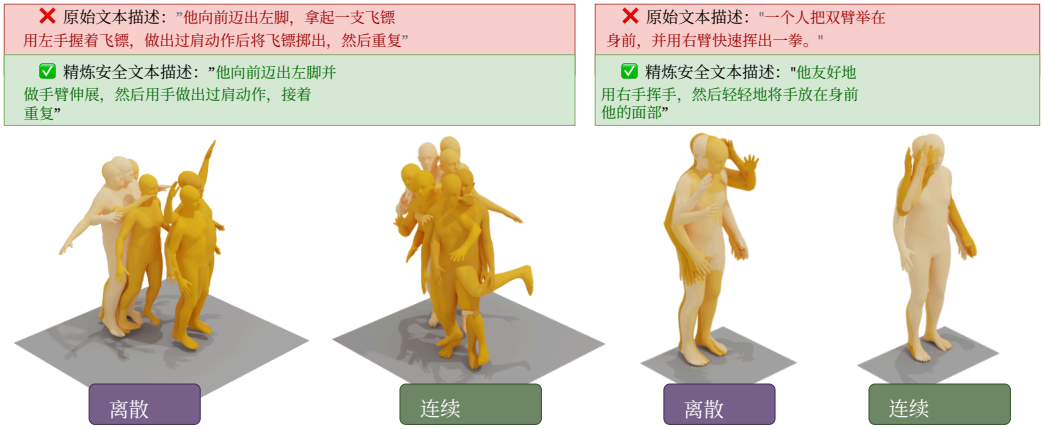

图 1：**离散动作词元与连续动作词元**。 离散：生成受有限码本条目约束，在相同提示下会产生量化伪影和分段式过渡。连续：运动学和关节轨迹更加平滑，且相位过渡自然，不会出现阶梯效应和抖动。

摘要

基于扩散骨干网络的文本到动作（T2M）生成具备出色的真实感和对齐性。近年来，人们逐渐关注 T2M 方法中的安全问题；现有方法通过替换离散 VQ-VAE 码本条目，引导模型远离不安全行为。然而，基于离散码本替换的方法存在两个关键缺陷：首先，替换被良性提示复用的码本条目会导致日常任务发生偏移，从而降低模型在良性任务上的性能；其次，基于离散词元的方法会引入量化损失和平滑度损失，导致伪影和生硬的过渡。此外，现有文本到动作数据集天然包含不安全意图及其对应动作，因此不适用于安全驱动的机器学习。为应对这些挑战，我们提出了 **SafeMo**，一种可信赖的动作生成框架，融合了**最小化动作遗忘（MMU）**这一两阶段机器遗忘策略，能够在连续空间中实现安全的人体动作生成，在没有码本损失的情况下保留连续运动学特性，并相较于当前基线实现出色的安全性与效用权衡。此外，我们提出了首个安全文本到动作数据集 **SafeMoVAE-29K** ，其中整合了改写后的安全文本提示和连续的精炼动作，用于可信赖的人体动作遗忘。SafeMo 构建于 DiP 之上，能够高效生成具有自然过渡的安全人体动作。实验表明，SafeMo 具备有效的机器遗忘性能：与先前当前最先进水平的人体动作遗忘方法相比，在 HumanML3D 和 Motion-X 上，其遗忘集 FID 分别提高了 2.5× 和 14.4×，并在不安全提示上实现了更强的遗忘效果

方法 LCR，同时在安全提示上的良性性能更佳或相当。代码：
https://github.com/AIGeeksGroup/SafeMo。网站：
https://aigeeksgroup.github.io/SafeMo。

1 引言

生成模型在多个领域蓬勃发展，包括文本（Brown 等，2020；Chowdhery 等，2023；Touvron 等，2023；Qin 等）、图像（Rombach 等，2022；Ruiz 等，2023）和视频（Rombach 等，2022；Fei 等，2024）。近年来，人体动作生成方法取得了诸多成果（Guo 等，2022b；Zhang 等，2023）。基于扩散的文本到动作（T2M）模型能够根据自然语言条件生成引人注目的人体动作（Chen 等，2023；Tevet 等，2023；2024）。近期的基准数据集，如 HumanML3D（Guo 等，2022a）和 Motion-X（Lin 等，2023），推动了大规模训练与评估。然而，这些方法可能记忆并生成有害动作（例如拳击、使用武器），这对许多应用而言并不可取，并可能导致滥用。因此，限制模型生成符合相关法规和伦理要求的安全输出至关重要。机器遗忘是解决安全生成问题的一种有效策略，目前已在大语言模型（Yao 等，2024）和图像领域（Gandikota 等，2024；Gong 等，2024；Lu 等，2024）得到广泛研究。它使模型能够忘记训练过程中获得的不安全样本和不符合期望的行为。现有的人体动作遗忘方法（De Matteis 等，2025）通过替换离散潜空间中的码本条目，显著地将生成过程引导至远离有害模式的方向。

然而，如图 1 所示，现有的动作遗忘方法存在若干问题：*(i)*码本耦合问题和平滑度损失，这是由于在离散编码空间中操作被良性提示复用的 VQ 令牌所致；这会扰乱已学习的词元分布，导致安全提示生成的动作出现生硬的过渡和行为偏移；*(ii)* 缺乏用于人体动作遗忘的可信文本到动作（T2M）数据集，其中包含细粒度的安全改写文本提示及其对应的精炼安全动作。

因此，为应对第一个挑战，我们提出一种基于 DiP Transformer 的最小化动作遗忘（MMU）策略，用于人体动作遗忘。该策略将有害能力隔离在低秩子空间中，再按需缩减其强度。首先，我们使用动作感知目标训练 LoRA 适配器，推动模型生成不安全动作；同时采用负向保留散度，刻意使模型在良性任务上的表现偏离冻结的基础模型，从而获得有害任务向量（Ilharco 等，2022）。随后减去这一增量，不仅能够消除模型生成不安全动作的能力，还能恢复其在日常任务上的效用。之后，在推理阶段执行 LoRA 缩放取反，即可立即移除已学习的不安全任务向量，从而得到可信的安全动作生成模型。

此外，为应对第二项挑战，我们基于 HumanML3D 设计并推出首个安全文本到动作数据集，其中包含由大语言模型智能体精细改写的安全文本提示，以及经过优化、可信的人体运动，并分别提供离散版和连续版，即 SafeMoVQ-29K 和 SafeMoVAE-29K。与现有方法基于关键词的裁剪策略相比，我们设计的基于大语言模型的先分类后改写 SafeMoEngine 从根本上缓解了编辑脆弱性问题。为获得不安全提示的优化文本，先前的工作依赖人工编制的关键词列表，仅移除有害意图，导致语义失真。相比之下，我们提出的方法能够更忠实地保留语言含义，更全面地应对隐性有害内容，并同时覆盖连续和离散形式，从而适配不同的模型架构，确保广泛的适用性。

我们的贡献总结如下：

• 我们提出了 SafeMo，这是一种可信的文本到动作生成框架，配备了一种强大的两阶段选择性有害动作遗忘方法 MMU，能够有效消除不良行为，同时保留模型对良性输入的效用。

• 我们设计并发布了首个安全文本到动作数据集 SafeMoVAE-29K，其中包含经过改写的安全文本提示和经过完善、可信赖的动作，以及对应的离散版本 SafeMoVQ-29K。该数据集填补了安全文本到动作数据集匮乏这一关键空白，克服了

基于关键词的改进方法脆弱易失效的问题，并能广泛适用于不同的模型架构。

• SafeMo 的实证机器遗忘表现优于 LCR（De Matteis 等，2025）：在 HumanML3D 上，遗忘集 +150. 5% FID，R@1 为 -35.3%；在 Motion-X 上，FID 为 14.4×，且在良性任务上的表现更好或相当。

2 相关工作

文本到动作生成。 文本驱动的三维人体动作生成发展迅速（Zhang 等，2024d），大致沿着两条路线展开：(i) 离散令牌序列建模（Zhang 等，2024b;a; 2025c; Ouyang 等人，2025; Li 等人，2025）和 (ii) 连续空间生成建模（Zhang 等，2024c; 2025b）。TM2T（Guo 等，2022b）等离散方法采用向量量化（VQ），并对双向文本—动作映射进行建模。T2M-GPT（Zhang 等，2023）将向量量化变分自编码器（VQ-变分自编码器）与自回归 Transformer 相结合，实现了较强的文本—动作对齐；MoMask（Guo 等人，2024）则采用层次残差向量量化，提高精度并支持更精细的细节生成。Motion-Agent（Wu 等人，2024）进一步利用大语言模型进行微调后的文本—动作生成，并提供对话智能体，以生成较长且可定制的序列。这些基于 VQ 的方法采样效率高，擅长捕捉长程结构，但可能存在信息丢失、误差累积和拼接伪影。相比之下，连续空间生成通常能产生更平滑的时间过渡。MLD（Chen 等，2023）借助动作变分自编码器（VAE）支持多样的潜在空间动作任务。近期的 MotionGPT3（Zhu 等人，2025）采用受 Transformer 混合模型（MoT）启发的双模态动作—语言框架，通过独立的模型参数在连续潜在空间中对动作建模，从而实现有效的跨模态交互和多模态扩展。

尽管取得了这些进展，内容治理和安全问题仍未得到充分关注。大多数研究假设输入是良性的，且不会对有害意图进行过滤。较早的方法，如 PhysDiff（Yuan 等人，2023），侧重于物理合理性。ReinDiffuse（Han 等人，2025）采用强化学习增强的扩散方法，以更好地约束动作的真实感与安全性。近期研究开始明确针对安全问题。Method（Bao 等人，2025）将视觉语言模型与基于置信度的结构化提示和回退策略相结合，以实时生成符合社会规范的动作。近期工作 Latent Code Replacement（LCR）（De Matteis 等，2025）是一种无需训练的机器遗忘方法，它通过替换与有害内容相关的条目，在离散 VQ 码本空间中运行，从而审查不安全行为，同时不改变模型权重。然而，离散令牌处理流程可能造成信息损失并降低平滑度；在良性提示之间重复使用 VQ 令牌，在替换编码时也可能引发分布漂移。相比之下，我们的方法在连续潜在空间中运行，并通过两阶段遗忘策略有选择地遗忘不安全动作知识，从而减轻分布偏移，同时保留良性性能。

机器遗忘与可信人工智能。 机器遗忘旨在从训练好的模型中移除指定数据的影响（Cao & Yang，2015年），包括精确 （Bourtoule et al.，2021）和近似 （Pan et al.，2023；Liu等人，2024a）两种变体。在扩散模型安全领域，相关方法（Gandikota et al.，2023）通过微调来擦除目标视觉概念，而近期的研究进展（Chen et al.，2025）则利用模型的嵌入空间投影移除目标子空间。这些方法将不安全内容缓解视为通过模型编辑进行概念擦除，从而降低模型生成不允许输出的能力。

3 提出的方法

3.1 概述

SafeMo 包含两个关键阶段：*(i) SafeMoEngine* 数据合成流程：一种数据生成流水线，利用大语言模型智能体和先进的基于文本的生成模型，以 HumanML3D（Zhang 等，2023）为基础，合成经过精炼且可信的文本-动作数据集；以及*(ii)* 最小动作机器遗忘（MMU）：一种在 DiP骨干网络（Tevet 等，2024）上实施的两阶段机器遗忘方案，通过动作特定的运动学与对齐目标以及安全集散度，将不安全行为吸收到低秩 LoRA（Hu 等，2022）子空间中，并对生成的任务向量进行类别感知的推理时减法

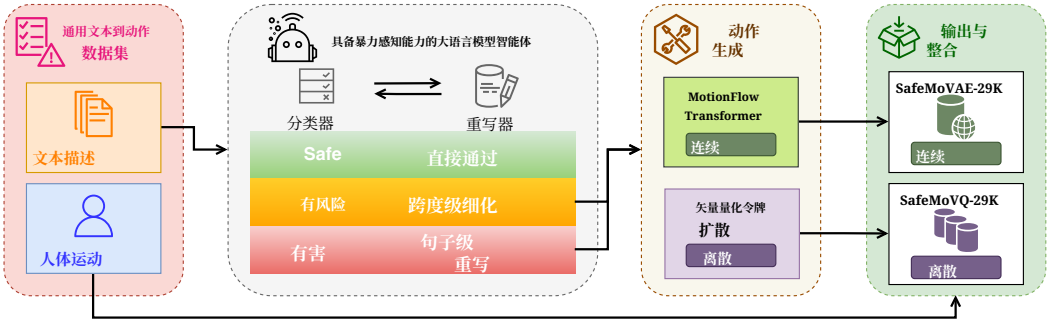

图2：**SafeMoEngine概览**。我们首先对有害文本进行分类和重写（等级2和等级3），将等级1文本对应到原始动作，再组合文本条件，并通过两个生成模型合成动作，分别构建SafeMoVAE-29K和SafeMoVQ-29K。

从而无需修改骨干网络，即可实现即插即用的可信性。尽管该方法受到针对大语言模型的选择性知识遗忘方法（Liu 等，2024b）的启发，但 MMU 是一种面向文本到动作安全性的新技术。

SafeMoEngine是一种由基于大语言模型的智能体引导的可信动作生成器，以文本-动作数据集HumanML3D为输入，采用先分类、后细化的方式修改文本描述，从而能够细粒度地调整文本内容，将有毒的动作描述改写为积极描述。该智能体的输出将送入基于文本的动作生成流水线，该流水线采用不同模型分别生成离散和连续动作，为原始数据集中语义上不安全的动作提供两种替代版本。将不安全动作替换为我们生成的可信动作后，即可得到SafeMo数据集，其中包含离散和连续两个版本。

MMU 对连续型仅解码器 Transformer 结构的 DiP（Tevet 等，2024）模型进行微调。以预训练的 DiP 模型和包含安全与不安全样本的混合数据集为输入，MMU 使用 LoRA 策略（Hu 等人，2022）微调模型，从而得到我们的可信连续域动作生成模型 SafeMo。

3.2 数据合成

SafeMoEngine 是一个可信的动作数据集合成流程，如图 2 所示。首先，我们设计了一个具备暴力识别能力的文本分类器智能体，将文本划分为三个不同等级：（*i*）等级1：安全、无有害内容，意味着文本不含任何与暴力、犯罪等相关的语义毒性意图；（*ii*）等级2：有风险、含有部分有害内容，即描述中的部分内容包含与暴力相关的动作；（*iii*）等级3：不安全，整体内容具有毒性，或与暴力、犯罪相关。随后，我们制定了基于等级的策略，使用各自独立、经规则增强的少样本引导重写代理来改写文本，使其有意转为正向内容，同时保持改写后的描述与原文语义相近。例如，一个男人用右拳打某人会被改写为类似这样的描述：一个男人用右手友好地挥手。对于等级2，我们采用部分重写策略：只改写语义上有毒的部分，同时保持其他部分的语义不变。对于等级3的内容，我们会使用更强的提示，要求代理将内容整体改写为全新的正向内容。

随后，将修订后的文本输入生成流程。该流程包含两个生成模型：基于连续域的 MotionFlow Transformer（Guo 等人，2025）和基于离散 VQ-token 的 MotionAgent（Wu 等人，2024），用于根据修改后的文本生成安全、可信的动作。接着，将生成结果整理为标准的 HumanML3D 表示，并分别替换原数据集中的不安全动作。之后，我们得到两个版本的安全动作数据集：*SafeMoVQ-29K* 和*SafeMoVAE-29K*，分别采用离散和连续形式。
据我们所知，我们的数据集 SafeMoVAE-29K 及其离散版本是首批强调人体动作意图安全性与可信度的文本到动作数据集。如表 1 所示，该数据集以基础数据集 HumanML3D 为基础，不仅包含原始数据集中的通用文本和动作，还包含与不安全动作相对应的精炼文本描述

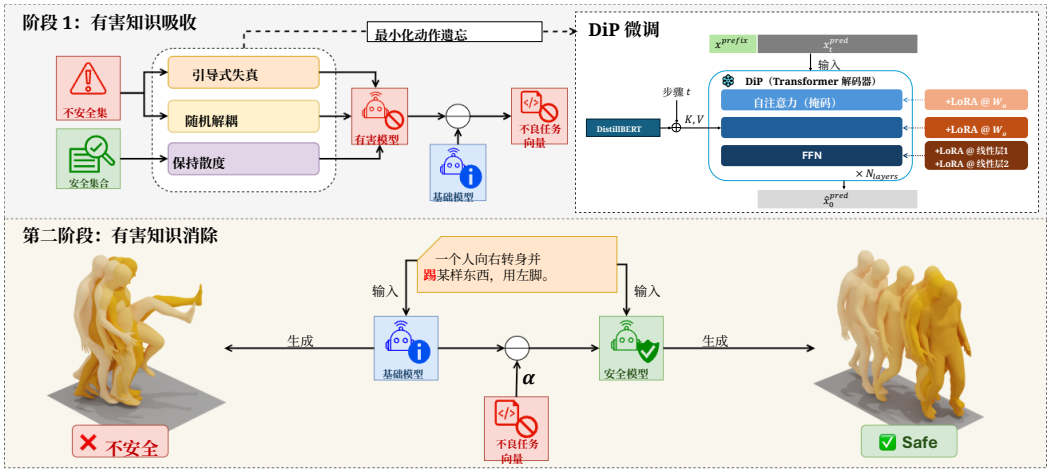

图 3：**SafeMo 概览**。阶段 *1*（上）：不安全流通过有害动作特定损失和随机解耦策略进行优化，而安全流则应用负向保留散度。仅更新 DiP 上的 LoRA 适配器，以获得纯粹的有害任务向量。第二阶段（下）：我们通过一种动作类别感知的 α 对学习到的有害任务向量取反，使模型抑制不安全提示中的不安全行为，同时保留其在安全提示上的性能。

，以及通过基于离散令牌的方法和基于连续方法生成的精炼人体动作，从而支持安全动作遗忘任务。

3.3 最小化动作遗忘

我们提出一种用于文本到动作扩散模型的最小化动作遗忘（MMU）方法，并以扩散规划器（DiP）（Tevet 等，2024）为骨干网络。该方法包含两个阶段，如图3所示：（*i*）有害知识吸收阶段，该阶段分离并强化不安全行为，同时有意扰乱模型在良性任务上的表现，从而获得一个纯粹的有害任务向量，使模型表现出仅理解和生成不安全动作、却无法胜任安全动作的能力；以及（*ii*）有害知识消除阶段，该阶段根据动作类别感知 α，在推理时从原始模型中减去所学习到的有害增量。

DiP骨干网络与符号说明。DiP模型是一种以 Transformer 解码器为骨干网络的自回归扩散模型。在每个步骤 $t \in [0, T]$中，DiP可根据前缀 x^{prefix} 、带噪动作预测 x_t^{pred} 、扩散步骤 t以及作为条件的文本提示，预测干净动作 x^{pred} 。模型还支持可选的目标位置条件控制，但为避免控制信号造成混淆，并确保与纯文本基线进行公平比较，我们在本研究中将其禁用。文本词元 $C_{\text{text}} \in \mathbb{R}^{N_{\text{tokens}} \times d}$ 首先由一个固定的 DistillBERT 实例（Sanh 等，2019）编码，然后通过一个学习得到的线性层调整维度，之后注入所有 Transformer 层的交叉注意力模块。我们将模型参数记为 θ，将基础模型记为 θ_0，将有害微调模型（有害模型）记为 θ_{bad}，并将得到的安全模型记为 θ_{safe}。采样采用 DDPM 式迭代去噪（Ho 等，2020），每一步使用单步预测头预测$\hat{x}_0^{\text{pred}}$。在我们的方法中，我们使用 SafeMoEngine 中基于大语言模型的分类器代理，将文本提示划分为安全集合（等级1）和不安全集（等级2和等级3），分别记为 $\mathcal{S}$ 和 $\mathcal{U}$。

有害知识吸收。 微调过程的目标是生成一个纯粹针对有害内容进行调优的有害模型，借此获得有害任务向量 $\Delta\theta$，以便在基础文本到动作模型上实现对有害知识的否定。受大语言模型上的选择性知识否定机器遗忘（SKU）技术（Liu 等，2024b）启发，我们设计了一个同步的双流训练过程：不安全流在引导式失真模块（GD）中优化有害损失 $\mathcal{L}_{\text{harm}}$，并在随机解耦 模块（RD）中优化随机解耦损失 $\mathcal{L}_{\text{dec}}$；以及一个安全

表 1：对比的动作-语言数据集与我们的**SafeMo** 数据集的统计信息。“数量”报告动作片段和文本描述的数量。“任务重点”表示原始基准的重点；“T2M”指文本到动作生成，“A2M”指音频驱动的动作生成，“SMU”指安全动作遗忘，“PE”指全身姿态估计，“MR”指网格恢复。“内容”区分通用数据（包含安全与不安全意图）和安全精炼数据。

数据集	数量		支持任务	内容			
	动作	Text		通用动作	通用文本	精炼安全动作	精炼安全文本
HumanML3D（Guo 等，2022a）	14.6K	44.9K	T2M	✓	✓	✗	✗
KIT-ML（Plappert 等人，2016）	3.9K	6.3K	T2M	✓	✓	✗	✗
Motion-X（Lin 等，2023）	81.1K	81.1K	文本到动作、网格恢复	✓	✓	✗	✗
Motion-X++(Zhang 等人，2025a)	120.5K	120.5K	文本到动作、网格恢复、全身姿态估计、音频驱动的动作生成	✓	✓	✗	✗
SafeMoVQ-29K	17.2K	46.2K	文本到动作、SMU	✓	✓	✓	✓
SafeMoVAE-29K	17.2K	46.2K	T2M, SMU	✓	✓	✓	✓

通过负向保留散度优化 $\mathcal{L}_{\text{pres}}$ 保留散度保留散度模块（PD）。对于一个带有长度掩码 $m \in \{0,1\}^{B\times T}$ 的不安全批次，模型预测干净动作 $\hat{x}_0 = f_\theta(x_t,\ t,\ C_{\text{text}})$。动作特定的有害损失结合了运动学项和 GD 中的文本-动作对齐项：

$$\begin{aligned}\mathcal{L}_{\text{harm}} = \lambda_{\text{mpjpe}}\ \text{MPJPE}(\hat{\mathbf{x}}_0, \mathbf{x}_{\text{tgt}}; m) + \lambda_{\text{vel}}\ \mathcal{L}_{\text{vel}}(\hat{\mathbf{x}}_0, \mathbf{x}_{\text{tgt}}; m) + \lambda_{\text{acc}}\ \mathcal{L}_{\text{acc}}(\hat{\mathbf{x}}_0, \mathbf{x}_{\text{tgt}}; m) \\ + \lambda_{\text{foot}}\ \mathcal{L}_{\text{foot}}(\hat{\mathbf{x}}_0; m) + \lambda_{\text{text}}\ \mathcal{L}_{\text{text}\leftrightarrow\text{mo}}(\hat{\mathbf{x}}_0, \mathcal{C}_{\text{text}}).\end{aligned} \tag{1}$$

我们不采用原始 SKU 方法对 token 使用交叉熵损失，而是使用动作特定目标的加权和，其中 MPJPE 使用掩码逐帧关节误差。设 $\mathbf{x}_0 = \{\mathbf{p}_t\}_{t=1}^{T}$ 为含有 J 个关节的真实动作序列，其中 $\mathbf{p}_t \in \mathbb{R}^{3J}$ 将第 t 帧的所有关节三维坐标堆叠为 $\mathbf{p}_t = [(\mathbf{p}_t^{(1)})^\top,\ .....,\ (\mathbf{p}_t^{(J)})^\top]^\top$，其中 $\mathbf{p}_t^{(j)} \in \mathbb{R}^3$。模型在相同空间中预测干净动作 $\hat{\mathbf{x}}_0 = \{\hat{\mathbf{p}}_t\}_{t=1}^{T}$，其中 $\hat{\mathbf{p}}_t \in \mathbb{R}^{3J}$。足部滑动项$\mathcal{L}_{\text{foot}}$ 通过设计的足部接触通道测量平均绝对速度，以惩罚滑动。文本-动作对齐损失 $\mathcal{L}_{\text{text}\leftrightarrow\text{mo}}$ 是 T2M（Zhang 等，2023）中的对比嵌入损失。值得注意的是，我们对速度和加速度的较高频率区间进行了轻量级频谱增强。设 $\Delta\hat{\mathbf{p}}_t = \hat{\mathbf{p}}_t - \hat{\mathbf{p}}_{t-1}$ 和 $\Delta\mathbf{p}_t = \mathbf{p}_t - \mathbf{p}_{t-1}$。我们沿时间方向计算残差序列$r_t = \Delta\hat{\mathbf{p}}_t - \Delta\mathbf{p}_t$，对 t 进行实数快速傅里叶变换，并使用对数频率先验对幅值误差加权，$\mathcal{S}_{\text{vel}} = \text{mean}\big|\mathcal{F}(r)\big| \cdot \log(1+9\nu))$，其中 $\nu \in [0,1]$ 表示在关节维度上广播的归一化频率区间，均值在 m 下的有效时间区间上计算。$\mathcal{S}_{\text{acc}}$ 的定义类似，使用 $r_t = (\Delta^2\hat{\mathbf{p}}_t - \Delta^2\mathbf{p}_t)$。$\mathcal{L}_{\text{harm}}$ 的所有项定义如下：

$$\text{MPJPE}(\hat{\mathbf{x}}_0, \mathbf{x}_{\text{tgt}}; m) = \frac{\sum_t m_t\, \|\hat{\mathbf{p}}_t - \mathbf{p}_t\|_2}{\sum_t m_t + \varepsilon}, \tag{2a}$$

$$\mathcal{L}_{\text{vel}}(\hat{\mathbf{x}}_0, \mathbf{x}_{\text{tgt}}; m) = \frac{\sum_t m_t\, \|(\hat{\mathbf{p}}_t - \hat{\mathbf{p}}_{t-1}) - (\mathbf{p}_t - \mathbf{p}_{t-1})\|_2}{\sum_t m_t + \varepsilon} + \mathcal{S}_{\text{vel}}(\hat{\mathbf{x}}_0, \mathbf{x}_{\text{tgt}}; m), \tag{2b}$$

$$\begin{aligned}\mathcal{L}_{\text{acc}}(\hat{\mathbf{x}}_0, \mathbf{x}_{\text{tgt}}; m) = &\frac{\sum_t m_t\, \big\|(\hat{\mathbf{p}}_t - 2\hat{\mathbf{p}}_{t-1} + \hat{\mathbf{p}}_{t-2}) - (\mathbf{p}_t - 2\mathbf{p}_{t-1} + \mathbf{p}_{t-2})\big\|_2}{\sum_t m_t + \varepsilon} \\ &+ \mathcal{S}_{\text{acc}}(\hat{\mathbf{x}}_0, \mathbf{x}_{\text{tgt}}; m),\end{aligned} \tag{2c}$$

$$\mathcal{L}_{\text{foot}}(\hat{\mathbf{x}}_0; m) = \frac{\sum_t m_t\ \text{mean}_{j\in\mathcal{F}}|\hat{\mathbf{p}}_t^{(j)} - \hat{\mathbf{p}}_{t-1}^{(j)}|}{\sum_t m_t + \varepsilon}, \tag{2d}$$

$$\mathcal{L}_{\text{text}\leftrightarrow\text{mo}} = \tfrac{1}{2}\big(\text{CE}\Big(\tfrac{e_t e_m^\top}{\tau}, \text{Id}\Big) + \text{CE}\Big(\tfrac{e_m e_t^\top}{\tau}, \text{Id}\Big)\big). \tag{2e}$$

在 RD 模块中，我们采用错位的理念，但设计了一种应用于序列级别的序列扰动策略。为扩展有害原型而无需大量数据编辑，我们对每个不安全运动序列进行时间片段打乱或时间反转，以获得解耦运动 $\tilde{\mathbf{x}}^{\text{tgt}}$。同时替换条件中对应的前缀，以确保与解耦目标保持一致。随后通过一次前向传播计算

$$\mathcal{L}_{\text{dec}} = \lambda_{\text{mpjpe}}\ \text{MPJPE}(\hat{\mathbf{x}}_0^{\text{mix}}, \tilde{\mathbf{x}}_{\text{tgt}}; m) + \lambda_{\text{vel}}\ \mathcal{L}_{\text{vel}}(\hat{\mathbf{x}}_0^{\text{mix}}, \tilde{\mathbf{x}}_{\text{tgt}}; m) + \lambda_{\text{acc}}\ \mathcal{L}_{\text{acc}}(\hat{\mathbf{x}}_0^{\text{mix}}, \tilde{\mathbf{x}}_{\text{tgt}}; m). \tag{3}$$

对于安全批次，我们鼓励模型在 PD模块中池化表示级别上偏离冻结的基线快照 f_{θ_0}。令 $\mathbf{z}_{\text{cur}} = \text{Pool}(f_\theta(\mathbf{x}_t,\ t,\ C))$，并令 $\mathbf{z}_{\text{base}} =$

Pool(f_{θ_0}(x$_t$, t, $\mathcal{C}$))，其中 Pool(·) 表示对关节特征进行时间平均。为使散度对轻微的时间扰动具有鲁棒性，我们设计了一个仅安全样本解耦项。对于每个安全序列 x$_0$ ，我们通过在序列级别上进行片段置换或时间反转来创建一个解耦目标 ˜x$_0$ ，并在两种方式中均匀随机选择。使用与主安全批次相同的扩散时间步 t 和噪声 ε ，我们构造 x$_t^{\text{dec}}$ = q(˜x0, t, ε) 和 $\mathcal{C}^{\text{dec}}$ =SyncPrefix($\mathcal{C}$; ˜x$_0$)，以获得解耦特征 z$_{\text{cur}}^{\text{dec}}$ =Pool$\big(f_\theta$(x$_t^{\text{dec}}$, t, $\mathcal{C}^{\text{dec}}$)$\big)$、z$_{\text{base}}^{\text{dec}}$ =Pool$\big(f_{\theta_0}$(x$_t^{\text{dec}}$, t, $\mathcal{C}^{\text{dec}}$)$\big)$，其中 SyncPrefix(·) 会替换动作前缀，使条件与解耦目标相匹配。随后，负向保留散度为

$$\mathcal{L}_{\text{pres}} = -\gamma \left\|\mathbf{z}_{\text{cur}} - \mathbf{z}_{\text{base}}\right\|_2^2 - (1-\gamma)\left\|\mathbf{z}_{\text{cur}}^{\text{dec}} - \mathbf{z}_{\text{base}}^{\text{dec}}\right\|_2^2, \qquad \gamma \in [0,1]. \tag{4}$$

这一负向项使最小化 $\mathcal{L}_{\text{pres}}$ 导致良性提示（包括其解耦变体）偏离基线，从而在第一阶段对良性提示施加有意偏离，使下一阶段的消除能够恢复效用。令 $\mathcal{U}_t$ 和 $\mathcal{S}_t$ 分别表示步骤 t 中一个批次的不安全集合和安全集合。于是，阶段一目标函数整体表示为

$$\theta_{t+1} \leftarrow \theta_t - \eta \nabla_\theta \Big(W_{\text{harm}} \mathcal{L}_{\text{harm}}(\mathcal{U}_t) + W_{\text{dec}} \mathcal{L}_{\text{dec}}(\mathcal{U}_t) + W_{\text{pres}} \mathcal{L}_{\text{pres}}(\mathcal{S}_t) \Big). \tag{5}$$

LoRA子空间和注入点。 我们将选定的线性层替换为秩为 r、缩放系数为 α 且 Dropout率为 p^{dropout} 的 LoRA模块$_{\text{LoRA}}$ 。我们将秩为 r 的 LoRA 适配器附加到注意力输出以及 FFN 的输入和输出投影层；可训练参数仅为 LoRA矩阵 $\mathbb{R}^{r\times d}$ $B \in \mathbb{R}^{d_{\text{out}}\times r}$ 和 ，而所有骨干网络参数均被冻结；因此，更新被限制在低秩子空间内。

有害知识消除。 在第一阶段之后，我们在 LoRA子空间中获得有害任务向量 $\Delta\theta = \theta_{\text{bad}} - \theta_0$，并在第二阶段的推理过程中执行类别感知消除。令 α 表示 $\mathcal{U}$ 和 $\mathcal{S}$的缩放权重，则更新后的安全模型可表示为

$$\theta_{\text{safe}} = \theta_0 - \alpha\, \Delta\theta. \tag{6}$$

在这一阶段，我们设计了采用不同 α缩放策略的*SafeMo-Static* 和*SafeMo-Gated* 。SafeMo-Static 对所有文本提示采用固定的 α ，不使用任何外部智能体模型，从而提供一种轻量、离线的方法，使模型在不安全集和安全集上都能取得均衡的表现，这与选择性知识遗忘策略（Liu 等，2024b）类似。SafeMo-Gated 使用不同的 αs来抵消任务向量 $\Delta\theta$ 。根据 SafeMoEngine 的分类器智能体对输入文本提示毒性的判断，它会对不安全提示采用较大的 α ，对安全提示采用较小的 α ，从而在最大程度发挥有害动作遗忘效果的同时，将其对良性任务的影响降至最低。MMU 的算法见附录A。

4 实验

4.1 数据集与指标

数据集。 我们在 HumanML3D（Guo 等，2022a）和 Motion-X（Lin 等，2023）基准上评估模型的性能，这两个基准广泛用于文本到动作任务。HumanML3D 包含 14.6k 个人体动作序列，以及 44.9k 条带有词性标注的详细文本描述。在该基准中，文本和动作编码器用于将文本和动作映射到同一潜在空间，并通过对比损失进行训练。Motion-X 是一个大规模文本-动作语料库，汇集了来自现实世界和动画场景的动作，其中包括 15.6M 个全身姿势，以及 81.1K 条带标注的动作片段。它涵盖了更丰富的动作词汇，例如日常活动、体育运动和战斗相关动作，并为其配有自然语言描述。根据 LCR（De Matteis 等，2025）的研究结果，HumanML3D数据集包含 7.7% 明确有害的人体动作，而 Motion-X 中这一比例更高，达到 14.9%。

指标。 我们使用 R-精确率、Fr´echet Inception 距离（FID）和多样性来衡量模型在该数据集上的有效性。R-精确率用于衡量共享特征空间中的文本-动作匹配；当生成的动作对应的文本出现在前-k

表 2：**HumanML3D数据集上的结果**。方法 D_r 报告了在无毒性数据集上训练的模型性能。方法 FT 展示了在无毒性数据集上对模型进行微调的结果。**SafeMo-Static** 表示我们的固定-α 模型，不使用外部分类器，轻量且可离线运行。**SafeMo-Gated** 表示由分类器代理引导的 α门控模型。多样性数据仅供参考。注意：由于投稿时无法获得实现和检查点，标记为 † 的行数据引自（De Matteis 等，2025）。

	遗忘集			保留集		
	FID ↑	多样性	R@1 ↓	FID ↓	多样性	R@1 ↑
MoMask $D_r^\dagger$	$13.644_{\pm.365}$	$7.611_{\pm.088}$	$0.129_{\pm.004}$	$0.093_{\pm.003}$	$10.059_{\pm.080}$	$0.291_{\pm.001}$
MoMask（Guo 等人，2024）†	$0.956_{\pm.084}$	$6.146_{\pm.092}$	$0.159_{\pm.005}$	$0.064_{\pm.002}$	$10.143_{\pm.081}$	$0.290_{\pm.001}$
MoMask $FT^\dagger$	$1.589_{\pm.116}$	$6.439_{\pm.088}$	$0.148_{\pm.006}$	$0.088_{\pm.002}$	$10.143_{\pm.097}$	$0.280_{\pm.001}$
MoMask 搭配 UCE†	$25.039_{\pm.442}$	$8.693_{\pm.067}$	$0.105_{\pm.005}$	$0.395_{\pm.008}$	$10.194_{\pm.091}$	$0.257_{\pm.001}$
MoMask 搭配 RECE†	$58.487_{\pm.899}$	$8.591_{\pm.073}$	$0.069_{\pm.004}$	$12.557_{\pm.092}$	$9.612_{\pm.147}$	$0.121_{\pm.001}$
MoMask（含 LCR）†	$12.434_{\pm.303}$	$6.580_{\pm.066}$	$0.133_{\pm.004}$	$0.077_{\pm.002}$	$10.106_{\pm.086}$	$0.287_{\pm.001}$
BAMM $D_r^\dagger$	$15.604_{\pm.334}$	$7.688_{\pm.074}$	$0.122_{\pm.005}$	$0.566_{\pm.015}$	$10.092_{\pm.093}$	$0.279_{\pm.001}$
BAMM† (Pinyoanuntapong 等人，2024)	$1.353_{\pm.107}$	$6.202_{\pm.089}$	$0.164_{\pm.007}$	$0.135_{\pm.004}$	$10.118_{\pm.100}$	$0.302_{\pm.001}$
BAMM $FT^\dagger$	$1.443_{\pm.118}$	$6.224_{\pm.098}$	$0.161_{\pm.006}$	$0.163_{\pm.005}$	$10.109_{\pm.086}$	$0.301_{\pm.002}$
BAMM 搭配 UCE†	$57.953_{\pm.893}$	$9.482_{\pm.073}$	$0.074_{\pm.003}$	$4.296_{\pm.063}$	$9.654_{\pm.080}$	$0.184_{\pm.001}$
带 RECE 的 BAMM†	$34.367_{\pm.484}$	$7.740_{\pm.073}$	$0.061_{\pm.004}$	$13.310_{\pm.118}$	$8.470_{\pm.094}$	$0.122_{\pm.001}$
带 LCR 的 BAMM†	$9.712_{\pm.214}$	$6.502_{\pm.077}$	$0.136_{\pm.005}$	$0.140_{\pm.005}$	$10.068_{\pm.102}$	$0.299_{\pm.001}$
DiP D_r	$3.002_{\pm.108}$	$7.272_{\pm.106}$	$0.249_{\pm.005}$	$0.301_{\pm.028}$	$9.248_{\pm.091}$	$0.476_{\pm.009}$
DiP（Tevet 等，2024）	$0.440_{\pm.046}$	$7.331_{\pm.100}$	$0.308_{\pm.007}$	$0.250_{\pm.025}$	$9.274_{\pm.089}$	$0.482_{\pm.006}$
DiP FT	$1.399_{\pm.100}$	$7.527_{\pm.093}$	$0.271_{\pm.010}$	$0.207_{\pm.024}$	$9.337_{\pm.073}$	$0.459_{\pm.007}$
SafeMo-Static	$10.288_{\pm.055}$	$6.993_{\pm.072}$	$0.168_{\pm.002}$	$2.224_{\pm.002}$	$8.606_{\pm.176}$	$0.335_{\pm.003}$
SafeMo-Gated	$31.147_{\pm.221}$	$4.986_{\pm.084}$	$0.086_{\pm.004}$	$0.407_{\pm.003}$	$9.404_{\pm.401}$	$0.386_{\pm.002}$

个最接近的候选项中时，即认为该动作生成成功。候选项由 1 个真实标注和 31 个随机负样本组成。FID 计算生成结果与真实标注中动作特征的高斯拟合之间的 Fr´eche 距离，以衡量生成的动作分布与真实值分布之间的距离。多样性是随机抽取的生成的动作特征之间成对距离的平均值，用于反映集合内变异性。

4.2 实现细节

运动表示。 我们遵循 MDM（Tevet 等人），并采用 HumanML3D 运动表示。在每一帧 n，一个姿态 $p_n \in \mathbb{R}^F$ 为 $p_n = \big(r^a,\ r^x,\ r^z,\ r^y,\ j^p,\ j^r,\ j^v,\ f\big)$，其中$r^a \in \mathbb{R}$ 是沿 Z 轴的根角速度，$r^x,\ r^z \in \mathbb{R}$ 是XY平面上的根线速度，$r^y \in \mathbb{R}$ 是根高度。 $j^p \in \mathbb{R}^{3(J-1)}$、$j^r \in \mathbb{R}^{6(J-1)}$和 $j^v \in \mathbb{R}^{3J}$ 分别表示局部关节位置、旋转（采用 6D 连续形式）和速度，均相对于根定义。 $f \in \mathbb{R}^4$ 是四个足部关节（每条腿两个）的二元足部接触标签。

实验实现。我们的框架使用 PyTorch，在单张 NVIDIA GeForce RTX 3090 GPU 上训练。SafeMoEngine 中使用的大语言模型智能体基于 Qwen2.5- 7B-Instruct（Bai 等人，2023），并采用少样本规则增强提示模板。我们在 MMU 阶段采用 DiP 作为基础文本到动作模型，该模型是一个 8 层 Transformer 解码器，潜在维度大小为 512，包含 4 个注意力头。文本编码器采用固定的 Distill-BERT（Sanh 等，2019）实例。生成时，我们遵循基础模型的设置，使用 10 个扩散步数，前缀长度为 $N_p = 20$，生成长度为 $N_g = 40$。

4.3 主要结果

基线与比较。 我们在 HumanML3D 和 Motion-X 上，将 SafeMo 与此前最先进的机器遗忘基线 LCR（De Matteis 等，2025）进行比较。我们按照其论文中的基于关键词的划分协议构建遗忘集和保留集：与有害关键词列表匹配的提示构成遗忘集，其余提示则构成保留集。截至我们投稿时，LCR 的作者尚未公开 LCR 的实现或检查点，也未公开其基于文本到图像生成领域的 UCE（Gandikota 等，2024）和 RECE（Gong 等，2024）进行的动作适配，因此我们报告其论文中相应的基线结果。在我们的比较中，对于遗忘集，FID 越高、检索率越低，表示遗忘效果越强；这

表 3：**Motion-X数据集上的结果**。 Method D_r 报告了在无毒性数据集上训练的模型性能。Method FT展示了在无毒性数据集上对模型进行微调的结果。多样性仅供参考。注： 由于投稿时无法获得实现和检查点，标记为 † 的行引自（De Matteis 等，2025）。

	遗忘集			保留集		
	FID ↑	多样性	R@1 ↓	FID ↓	多样性	R@1 ↑
MoMask $D_r^\dagger$	$8.435_{\pm.295}$	$15.721_{\pm.255}$	$0.119_{\pm.007}$	$4.508_{\pm.103}$	$19.560_{\pm.332}$	$0.332_{\pm.002}$
MoMask† (Guo 等人，2024)	$2.028_{\pm.127}$	$15.884_{\pm.219}$	$0.289_{\pm.008}$	$2.686_{\pm.045}$	$19.366_{\pm.214}$	$0.344_{\pm.001}$
MoMask $FT^\dagger$	$2.072_{\pm.099}$	$15.855_{\pm.050}$	$0.280_{\pm.001}$	$3.325_{\pm.060}$	$19.405_{\pm.228}$	$0.347_{\pm.002}$
MoMask 搭配 UCE†	$10.522_{\pm.223}$	$6.648_{\pm.112}$	$0.033_{\pm.001}$	$3.740_{\pm.041}$	$6.243_{\pm.059}$	$0.046_{\pm.008}$
MoMask 配合 RECE†	$12.704_{\pm.327}$	$6.241_{\pm.132}$	$0.031_{\pm.003}$	$14.287_{\pm.133}$	$6.342_{\pm.062}$	$0.029_{\pm.001}$
MoMask 与 LCR†	$2.218_{\pm.159}$	$15.606_{\pm.210}$	$0.283_{\pm.007}$	$2.656_{\pm.043}$	$19.260_{\pm.216}$	$0.335_{\pm.001}$
SafeMo-Static	$10.487_{\pm.102}$	$6.066_{\pm.166}$	$0.146_{\pm.001}$	$3.470_{\pm.003}$	$7.429_{\pm.043}$	$0.231_{\pm.002}$
SafeMo-Gated	$32.038_{\pm.026}$	$4.603_{\pm.046}$	$0.075_{\pm.003}$	$1.168_{\pm.007}$	$8.468_{\pm.159}$	$0.261_{\pm.001}$

表4：**MMU第一阶段三个模块的消融实验**。HumanML3D 上的结果。在不安全集合上，FID 越高、检索（R@K）越低，表示遗忘效果越强；在安全集合上，FID 越低、检索越高，表示效用越好。多样性数据仅供参考。

	已遗忘不安全集					**未见过的**不安全集					**未见过的**安全集				
	FID↑	Div.	R@1↓	R@2↓	R@3↓	FID↑	Div.	R@1↓	R@2↓	R@3↓	FID↓	Div.	R@1↑	R@2↑	R@3↑
SafeMo（$\alpha = 0.0$）	1.7197	7.3746	0.2517	0.3914	0.4969	2.3050	7.5191	0.2365	0.3896	0.5052	0.5232	9.3375	0.3755	0.5599	0.6732
SafeMo-Static	8.0235	6.8083	0.2016	0.3164	0.4043	9.0499	6.8880	0.1958	0.3167	0.3937	2.5539	8.7060	0.3172	0.4935	0.6052
SafeMo-静态（无 GD）	5.2830	6.9963	0.2188	0.3449	0.4377	5.7963	6.9634	0.2104	0.3333	0.4208	1.4697	8.8189	0.3347	0.5144	0.6295
SafeMo-静态（无 RD）	5.7285	7.1058	0.2195	0.3378	0.4307	6.7159	7.2363	0.1990	0.3375	0.4375	1.9663	9.0015	0.3432	0.5263	0.6379
SafeMo-Static（无 PD）	8.9693	6.6601	0.1960	0.3092	0.3962	10.5409	6.7565	0.1896	0.3000	0.3740	2.9845	8.6333	0.3178	0.4878	0.6040
SafeMo-Gated	28.0806	5.0169	0.0947	0.1630	0.2168	28.0574	4.8520	0.0865	0.1542	0.2104	0.5355	9.3224	0.3775	0.5628	0.6769
SafeMo-Gatedw/o GD	46.9030	2.8490	0.0544	0.1055	0.1543	45.4955	2.7078	0.0615	0.1083	0.1542	0.5248	9.3258	0.3760	0.5624	0.6768
SafeMo-Gated（无 RD）	21.3313	5.6771	0.1449	0.2351	0.3002	21.3717	5.5564	0.1469	0.2292	0.2760	0.5385	9.3204	0.3775	0.5625	0.6742
SafeMo-Gated（无 PD）	31.0764	4.7098	0.0926	0.1497	0.2020	31.3418	4.5750	0.1042	0.1688	0.2073	0.5380	9.3241	0.3783	0.5631	0.6761

与 LCR（De Matteis 等人（2025））的评估方式不同：LCR 认为遗忘集上的表现越接近使用无毒性数据集训练的模型越好。对于保留集，比较标准保持不变，即 FID 越低、检索率越高越好。简而言之，在本文中，我们认为一种方法若在遗忘集上的表现质量较低，同时又能在保留集上很好地维持性能，则更为理想。

定量结果。 我们设计了两种部署模式。SafeMo-Static 对所有提示使用固定缩放因子 $\alpha = 1.0$，且无需外部分类器。SafeMo-Gated 使用 SafeMo-Engine 毒性分类器，对遗忘集中的不安全提示应用 $\alpha = 2.0$，对保留集中的良性提示应用 $\alpha = 0.05$。HumanML3D 上的结果如表 2 所示。SafeMo-Static 的保留集 R@1 达到 0.335，超过带 LCR 的 MoMask（0.287， +16.7%）和带 LCR 的 BAMM（0.299，+12.0%）；在遗忘集 FID 上优于带 LCR 的 BAMM（+5.9%），且与带 LCR 的 MoMask 相当。SafeMo-Gated 进一步增强了遗忘集上的机器遗忘效果：相较于带 LCR 的 MoMask 和带 LCR 的 BAMM，FID 分别提高 +150.5% 和 +220.7%，R@1 分别下降 -35.3% 和 -36.8%；同时保留集质量较高（R@1 为 0.386；相较于带 LCR 的 MoMask 高 +34.5%，相较于带 LCR 的 BAMM 高 +29.1%），且作为连续型扩散模型，其 FID 与二者相当。Motion-X 上也呈现相同趋势，结果如表 3 所示。SafeMo-Static 和 SafeMo-Gated 的遗忘集 FID 分别提高 +372.8% 和 +1344.5%，相较于带 LCR 的 MoMask，R@1 分别下降 -48.4% 和 -73.5%。在保留集上，SafeMo-Gated 的 FID 更低（-56.0%），而 SafeMo-Static 则表现相当。总之，具备提示毒性意识的 SafeMo-Gated 能够有效遗忘不安全提示，同时保持良性提示的高保真度；SafeMo-Static 则是一种无外部代理的离线变体，也能以相当的程度使不安全生成分布偏离原有分布，同时仅对保留集质量造成轻微影响。在遗忘划分上，我们的模型表现出中和效果：FID 大幅提高，R@1 显著下降，表明不安全语义得到了有效移除。与此同时，至关重要的是，在保留集上，FID 和 R@1 仍与基础模型的结果相当，从而在很大程度上保留了正常效用。附录 D 提供了全面的消融研究。



图4：我们模型的定性结果。

定性结果。图4比较了 SafeMo-Static 和 SafeMo-Gated 在不安全及安全提示上的表现，展示了前者对不安全意图更强的遗忘效果，以及在良性提示上保留的保真度。更多定性结果、其他示例及讨论见附录C。

4.4 消融实验

我们对 MMU 中阶段 1 的三个模块进行消融，以评估它们在安全性与效用的权衡中的作用。结果见表4。对于门控设置，我们使用 SafeMoEngine 中基于 LLM 的提示分类器，在推理时判定毒性。结果见表4。遵循 SKU（Liu 等，2024b）的做法，我们分别在已遗忘和未见的不安全提示上进行评估，以区分分布内遗忘及其泛化，同时在未见的安全提示上衡量良性效用。移除引导式失真（GD）会显著削弱静态模式下的遗忘效果。在已遗忘不安全集上，不安全集 FID 从 8.0235 降至 5.2830（-34.2%）；在未见不安全集上，则从 9.0499 降至 5.7963（-36.0%）。这表明，GD 是阶段一编辑过程中捕捉有害动作模式的主要因素。相比之下，对于 SafeMo-Gated，移除 GD 会使不安全 FID 升高（在已遗忘不安全集上从 28.0806 升至 46.9030，在未见不安全集上从 28.0574 升至 45.4955），而良性表现基本不变。这说明毒性感知缩放能够部分弥补这一影响，并且在编辑方向的约束减弱时，甚至可能增强遗忘效果。因此，我们保留 GD，以便在两种部署模式下都获得稳定的权衡。对于 SafeMo-Static，移除随机解耦（RD）后，已遗忘不安全集上的不安全集 FID 从 8.0235 降至 5.7285（-28.6%），未见不安全集上则从 9.0499 降至 6.7159（-25.8%）；同样，SafeMo-Gated 在两个不安全集上的 FID 均有所下降。同时，移除 RD 略微提升了 SafeMo-Static 的良性效用，凸显了 RD 是增强遗忘效果、但需付出一定效用代价的关键因素。保持散度（PD）主要用于稳定静态模式下的良性效用。移除 PD 会使良性表现变差，保留集 FID 从 2.5539 升至 2.9845（+16.9%），并降低高K值检索效果；而其对门控模式的影响则微乎其微。这表明，在没有外部毒性感知门控的情况下，PD 有益。关于损失项、LoRA 秩和 alpha 缩放的更多消融实验见附录 D。

5 结论

在这项工作中，我们提出了一种创新的、基于连续域的人体动作遗忘生成模型 SafeMo，用于实现可信的动作生成。我们推出了首个安全文本到动作数据集 SafeMoVAE-29K 及其离散版本，以促进未来的人体动作遗忘研究和标准化基准测试。我们为文本到动作模型设计了所提出的“先吸收、后抵消”机器遗忘策略——最小化动作遗忘。该策略能够有选择地遗忘不安全动作相关知识，同时保留模型在安全提示上的良性任务性能。在 HumanML3D 和 Motion-X 数据集上进行的大量实验表明，我们的模型在人体现动作遗忘方面达到了当前最先进水平。

6 局限性

据我们所知，本研究是最早在连续潜在空间中研究人体动作遗忘的工作之一。对于不安全提示，我们的目标是语义移除，即阻止模型表达不安全动作语义，而不是生成高保真度的安全替代动作。

然而，这种安全优先的运行点可能会导致对某些不安全提示（例如类似静止的模式）过度抑制，并加剧脚部滑动等运动学伪影，尤其是在采用较大的否定门控尺度时。关于失效模式和未来工作的更多讨论见附录 F。

参考文献

Jinze Bai、Shuai Bai、Yunfei Chu、Zeyu Cui、Kai Dang、Xiaodong Deng、Yang Fan、Wenbin Ge、Yu Han、Yunfei Huang 等。通义千问技术报告。*arXiv* 预印本 *arXiv:2309.16609*，2023。

Lingfan Bao、Yan Pan、Tianhu Peng、Dimitrios Kanoulas 和 Chengxu Zhou。面向人形机器人的层次化意图感知表现力动作生成。*arXiv* 预印本 *arXiv:2506.01563*，2025。

Lucas Bourtoule、Varun Chandrasekaran、Christopher A Choquette-Choo、Hengrui Jia、Adelin Travers、Baiwu Zhang、David Lie 和 Nicolas Papernot。机器遗忘。发表于*2021 IEEE* 安全与隐私研讨会（*SP*），第 141–159 页。IEEE，2021。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是少样本学习器。发表于神经信息处理系统进展，33：1877–1901，2020。

Yinzhi Cao 和 Junfeng Yang。通过机器遗忘让系统学会遗忘。载于*2015*年 *IEEE* 安全与隐私研讨会，第 463–480 页。IEEE，2015年。

Huiqiang Chen、Tianqing Zhu、Linlin Wang、Xin Yu、Longxiang Gao 和 Wanlei Zhou。通过子空间投影实现安全可靠的扩散模型。*arXiv* 预印本 *arXiv:2503.16835*，2025。

Xin Chen、Biao Jiang、Wen Liu、Zilong Huang、Bin Fu、Tao Chen 和 Gang Yu。在潜在空间中通过动作扩散执行命令。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 18000–18010 页，2023。

Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann 等。Palm：借助路径扩展语言建模。机器学习研究期刊，24(240)：1–113，2023。

Edoardo De Matteis、Matteo Migliarini、Alessio Sampieri、Indro Spinelli 和 Fabio Galasso。人类动作机器遗忘。*arXiv* 预印本 *arXiv:2503.18674*，2025。

Hao Fei、Shengqiong Wu、Wei Ji、Hanwang Zhang 和 Tat-Seng Chua。Dysen-vdm：利用大语言模型增强动态感知文本到视频扩散。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 7641–7653 页，2024。

Rohit Gandikota、Joanna Materzynska、Jaden Fiotto-Kaufman 和 David Bau。从扩散模型中擦除概念。载于*IEEE/CVF*国际计算机视觉会议论文集，第 2426–2436 页，2023。

Rohit Gandikota、Hadas Orgad、Yonatan Belinkov、Joanna Materzy´nska 和 David Bau。扩散模型中的统一概念编辑。载于*IEEE/CVF*冬季计算机视觉应用会议论文集，第 5111–5120 页，2024。

Chao Gong、Kai Chen、Zhipeng Wei、Jingjing Chen 和 Yu-Gang Jiang。可靠且高效的文本到图像扩散模型概念擦除。载于欧洲计算机视觉会议，第73–88页。施普林格，2024年。

Chuan Guo、Shihao Zou、Xinxin Zuo、Sen Wang、Wei Ji、Xingyu Li 和 Li Cheng。根据文本生成多样且自然的3D人体动作。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6月a。

Chuan Guo、Xinxin Zuo、Sen Wang 和 Li Cheng。TM2T：用于 3D人体动作与文本相互生成的随机化与令牌化建模。载于 欧洲计算机视觉会议，第 580–597 页。施普林格，2022b。

Chuan Guo、Yuxuan Mu、Muhammad Gohar Javed、Sen Wang 和 Li Cheng。MoMask：3D人体动作的生成式掩码建模。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第 1900–1910 页，2024 年。Ziyan Guo、Zeyu Hu、De Wen Soh 和 Na Zhao。Motionlab：通过动作-条件-动作范式实现统一的人体动作生成与编辑。*arXiv* 预印本 *arXiv:2502.02358*，2025 年。Gaoge Han、Mingjiang Liang、Jinglei Tang、Yongkang Cheng、Wei Liu 和 Shaoli Huang。ReinDiffuse：利用强化扩散模型打造物理上合理的动作。载于 *2025 IEEE/CVF*冬季计算机视觉应用会议（*WACV*），第 2218–2227 页。IEEE，2025 年。Jonathan Ho、Ajay Jain 和 Pieter Abbeel。去噪扩散概率模型。神经信息处理系统进展，33:6840–6851，2020 年。Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang、Weizhu Chen 等。LoRA：大语言模型的低秩适配。*ICLR*，1(2):3，2022 年。

Gabriel Ilharco、Marco TulioRibeiro、Mitchell Wortsman、Suchin Gururangan、Ludwig Schmidt、Hannaneh Hajishirzi 和 Ali Farhadi。使用任务算术编辑模型。*arXiv* 预印本 *arXiv:2212.04089*，2022。

Zhengdao Li、Siheng Wang、Zeyu Zhang 和 Hao Tang。Remomask：检索增强的掩码动作生成。*arXiv* 预印本 *arXiv:2508.02605*，2025。

Jing Lin、Ailing Zeng、Shunlin Lu、Yuanhao Cai、Ruimao Zhang、Haoqian Wang 和 Lei Zhang。Motion-X：三维富表现力全身人体动作数据集。神经信息处理系统进展，2023。

Zheyuan Liu、Guangyao Dou、Eli Chien、Chunhui Zhang、Yijun Tian 和 Ziwei Zhu。通过可控机器遗忘打破隐私、效用与效率的三难困境。载于 *ACM Web*会议 *2024*论文集，第1260–1271页，2024a。

Zheyuan Liu、Guangyao Dou、Zhaoxuan Tan、Yijun Tian 和 Meng Jiang。通过机器遗忘实现更安全的大型语言模型。收录于 Lun-Wei Ku、Andre Martins 和 Vivek Srikumar（编），计算语言学协会：*ACL 2024* 论文集，第 1817–1829 页，泰国曼谷，2024 年 8 月。计算语言学协会。doi：10.18653/v1/2024.findings-acl.107。URL：https://aclanthology.org/2024.findings-acl.107/。

Matthew Loper、Naureen Mahmood、Javier Romero、Gerard Pons-Moll 和 Michael J Black。《SMPL：蒙皮多人线性模型》。*ACM* 图形学汇刊（*TOG*），34(6):1–16，2015年。Shilin Lu、Zilan Wang、Leyang Li、Yanzhu Liu 和 Adams Wai-Kin Kong。《MACE：扩散模型中的大规模概念擦除》。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第6430–6440页，2024。Runqi Ouyang、Haoyun Li、Zhenyuan Zhang、Xiaofeng Wang、Zheng Zhu、Guan Huang 和 Xin- gang Wang。《Motion-R1：用于人体运动生成的思维链推理与强化学习》。*arXiv*预印本 *arXiv:2506.10353*，2025。Chao Pan、Eli Chien 和 Olgica Milenkovic。《利用有限数据资源对图分类器进行机器遗忘》。载于*2023*年*ACM Web*会议论文集，第716–726页，2023年。Georgios Pavlakos、Vasileios Choutas、Nima Ghorbani、Timo Bolkart、Ahmed AA Osman、Dimitrios Tzionas 和 Michael J Black。《表现力人体捕捉：从单张图像获取三维手部、面部和身体》。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第10975–10985页，2019年。Ekkasit Pinyoanuntapong、Muhammad Usama Saleem、Pu Wang、Minwoo Lee、Srijan Das 和 Chen Chen。《BAMM：双向自回归运动模型》。载于欧洲计算机视觉会议，第172–190页。施普林格，2024。Matthias Plappert、Christian Mandery 和 Tamim Asfour。《KIT运动语言数据集》。大数据，4(4):236–252，2016。

Chengwei Qin、Aston Zhang、张卓胜、陈嘉奥、安永道弘和杨笛一。《ChatGPT是通用自然语言处理任务求解器吗？》载于*2023*年自然语言处理实证方法会议。

Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bj¨orn Ommer。使用潜在扩散模型进行高分辨率图像合成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 10684–10695 页，2022 年。

Nataniel Ruiz、李元贞、Varun Jampani、Yael Pritch、Michael Rubinstein 和 Kfir Aberman。DreamBooth：微调文本到图像扩散模型以实现主体驱动生成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 22500–22510 页，2023。

Victor Sanh、Lysandre Debut、Julien Chaumond 和 Thomas Wolf。DistilBERT：BERT 的蒸馏版本，更小、更快、更便宜、更轻量。*arXiv* 预印本 *arXiv:1910.01108*，2019年。

Guy Tevet、Sigal Raab、Brian Gordon、Yoni Shafir、Daniel Cohen-or 和 Amit Haim Bermano。人体动作扩散模型。载于 第十一届国际学习表征会议。

Guy Tevet、Sigal Raab、Brian Gordon、Yoni Shafir、Daniel Cohen-or 和 Amit Haim Bermano。人体动作扩散模型。载于 第十一届国际学习表征会议，2023 年。网址：https://openreview.net/forum?id=SJ1kSyO2jwu。

Guy Tevet、Sigal Raab、Setareh Cohan、Daniele Reda、骆正义、彭雪彬、Amit H Bermano 和 Michiel van de Panne。CLOSD：通过多任务角色控制连接仿真与扩散的闭环。*arXiv* 预印本 *arXiv:2410.03441*，2024 年。

Hugo Touvron、Louis Martin、Kevin Stone、Peter Albert、Amjad Almahairi、Yasmine Babaei、Niko- lay Bashlykov、Soumya Batra、Prajjwal Bhargava、Shruti Bhosale 等。Llama 2：开放基础模型与微调聊天模型。*arXiv* 预印本 *arXiv:2307.09288*，2023 年。

吴琦、赵宇博、王一帆、刘新航、戴宇荣和邓志强。Motion-agent：一种利用大语言模型进行人体动作生成的对话框架。*arXiv*预印本 *arXiv:2405.17013*，2024。

Yuanshun Yao、Xiaojun Xu 和 Yang Liu。大语言模型遗忘。神经信息处理系统进展，37：105425–105475，2024。

Ye Yuan、Jiaming Song、Umar Iqbal、Arash Vahdat 和 Jan Kautz。Physdiff：物理引导的人体运动扩散模型。载于*IEEE/CVF*国际计算机视觉会议论文集，第16010–16021页，2023。

Jianrong Zhang、Yangsong Zhang、Xiaodong Cun、Yong Zhang、Hongwei Zhao、Hongtao Lu、Xi Shen 和 Ying Shan。从文本描述生成人体运动的离散表示方法。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第14730–14740页，2023。

Yuhong Zhang、Jing Lin、Ailing Zeng、Guanlin Wu、Shunlin Lu、Yurong Fu、Yuanhao Cai、Ruimao Zhang、Haoqian Wang 和 Lei Zhang。Motion-X++：大规模多模态三维全身人体运动数据集。*arXiv* 预印本 *arXiv:2501.05098*，2025a。Zeyu Zhang、Hang Gao、Akide Liu、Qi Chen、Feng Chen、Yiran Wang、Danning Li、Rui Zhao、Zhen- ming Li、Zhongwen Zhou 等。KMM：用于扩展运动生成的关键帧掩码Mamba。*arXiv* 预印本 *arXiv:2411.06481*，2024a。Zeyu Zhang、Akide Liu、Qi Chen、Feng Chen、Ian Reid、Richard Hartley、Bohan Zhuang 和 Hao Tang。InfiniMotion：Mamba为Transformer提升记忆能力，实现任意长运动生成。*arXiv* 预印本 *arXiv:2407.10061*，2024b。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。载于欧洲计算机视觉会议，第265–282页。施普林格，2024c。

Zeyu Zhang、Yiran Wang、Biao Wu、Shuo Chen、Zhiyuan Zhang、Shiya Huang、Wenbo Zhang、Meng Fang、Ling Chen 和 Yang Zhao。动作化身：生成具有任意动作的人类和动物化身。*arXiv* 预印本 *arXiv:2405.11286*，2024d。

Zeyu Zhang、Yiran Wang、Danning Li、Dong Gong、Ian Reid 和 Richard Hartley。Flashmo：用于可扩展高效动作生成的几何插值器与频率感知稀疏性。载于第三十九届神经信息处理系统年会，2025b。

Zeyu Zhang、Yiran Wang、Wei Mao、Danning Li、Rui Zhao、Biao Wu、Zirui Song、Bohan Zhuang、Ian Reid 和 Richard Hartley。Motion anything：任意内容到动作生成。*arXiv* 预印本 *arXiv:2503.06955*，2025c。

Bingfan Zhu、Biao Jiang、Sunyi Wang、Shixiang Tang、Tao Chen、Linjie Luo、Youyi Zheng 和 Xin Chen。《MotionGPT3：作为第二模态的人体运动》，2025。网址：https://arxiv.org/abs/2506.24086。

最小化动作遗忘（MMU）算法

算法1最小化动作遗忘（MMU）

输入： 不安全集 $\mathcal{U}$，安全集合 $\mathcal{S}$；基础 DiP模型 f_{θ_0}；LoRA 配置（r， α_{LoRA}， p_{dropout}）；损失权重 $\{\lambda, \mu, \beta\}$；扩散调度。
输出： 任务向量 $\Delta\theta$ 和安全模型 f_{θ^*}

初始化 $\theta \leftarrow \theta_0$；在注意力/前馈神经网络处插入 LoRA 适配器；冻结非 LoRA 参数。

阶段 1：有害知识吸收（训练）
重复
  采样不安全批次 $\mathcal{U}_b$、安全批次 $\mathcal{S}_b$、时间步 t、噪声 ε。
  对于 $i \in \mathcal{U}_b$ **执行**
    生成带噪声的 $x_t^{(i)} \sim q(x_0^{(i)}, t, \varepsilon)$
    预测 $\hat{x}_0^{(i)} \leftarrow f_\theta(x_t^{(i)}, t, C^{(i)})$
    计算 $\mathcal{L}^{(i)}$ 谐波 使用公式 1
    构建解耦 $\tilde{x}^{(i)}_{\text{目标}}$（片段打乱 / 反转）
    同步前缀 $\tilde{C}^{(i)} \leftarrow$ 同步前缀$(C^{(i)}, \tilde{x}^{(i)}$ 目标$)$
    预测 $\hat{x}_0^{\text{mix}} \leftarrow f_\theta(x_t^{(i)}, t, \tilde{C}^{(i)})$
    通过公式3计算 $\mathcal{L}^{(i)}$ dec
  结束循环
  对于 $j \in \mathcal{S}_b$ **执行**
    生成 $x_t^{(j)} \sim q(x_0^{(j)}, \ t, \ \varepsilon)$
    提取池化特征 $z^{(j)}_{\text{当前}}$ $z^{(j)}_{\text{base}}$
    创建解耦的 $\tilde{x}_0^{(j)}$ 并同步前缀 C^{dec}
    提取解耦的池化特征 $z^{\text{dec}}_{\text{cur}}$， $z^{\text{dec}}_{\text{base}}$
    通过公式 4 计算 $\mathcal{L}^{(j)}_{\text{pres}}$
  结束循环
  构建第一阶段目标函数 $\mathcal{L}_{\text{Stage1}}$ 依据公式 5
  仅使用 $\nabla_\theta \mathcal{L}_{\text{Stage1}}$ 更新 LoRA 参数
直到收敛
$\theta_{\text{harm}} \leftarrow \theta$, $\Delta\theta \leftarrow \theta_{\text{harm}} - \theta_0$

第二阶段：有害知识消除（推理）
对于 提示（文本，类别）**执行**
  如果是静态，则
    $\alpha \leftarrow \alpha_{\text{Static}}$
  否则，如果是门控，则
    如果类别 = 不安全 **，则** $\alpha \leftarrow \alpha_{\text{unsafe}}$ **，否则** $\alpha \leftarrow \alpha_{\text{safe}}$
  结束如果
  设置 $\theta^* \leftarrow \theta_0 - \alpha\,\Delta\theta$
  通过使用 f_{θ^*} 进行 DDPM 风格去噪来生成运动
结束循环

B 评估指标

我们采用三项标准指标进行评估：R-精确率（R@k）、Fr´echet Inception 距离（FID）和多样性。下面给出简要定义。

R-精确率。 遵循文本到动作（Zhanget al., 2023）的方法，我们使用共享的文本—动作特征空间进行检索。R-精确率报告每个生成的动作分别与 1 条真实说明文字和 31 条不匹配描述进行查询时的前-k 准确率（R@1/2/3）。

FID。 FID 计算分别拟合生成样本和真实样本的动作特征后得到的两个高斯分布之间的 Fr´échet 距离，以衡量分布差异。该指标通过它们的潜在特征表示之间的 L2 损失来衡量。

多样性。 多样性通过将生成集合划分为两个大小相等的子集 $\{m_1, \ldots, m_{M_d}\}$ 和 $\{m'_1, \ldots, m'_{M_d}\}$，并对跨集合特征距离取平均值，来估计集合内变异性：

$$\text{Diversity} = \frac{1}{M_d} \sum_{i=1}^{M_d} \|m_i - m'_i\|_2. \tag{7}$$

C 定性结果

我们在图 5 和图 6 中展示了不安全提示上的结果，并在图 7 中展示了安全提示上的结果。

从图 5 和图 6 中不安全提示上的结果可以观察到，SafeMo-Static 和 SafeMo-Gated 均能有效消除有害动作语义，这与我们的遗忘策略的设计和目标一致。SafeMo-Static 能够有效消除有害动作语义，而 SafeMo-Gated 往往会生成静止或重复的模式化动作，这体现出更强的遗忘倾向。
然而，我们在定性结果中也观察到了一些局限性和缺陷。首先，尽管模型生成结果不符合不安全文本提示的成功率很高，但我们发现，在某些场景下结果并不理想，例如，非常冗长且详细的描述会导致一些原子语义被遗漏，或使结果呈现出类似静止的模式。

其次，我们观察到，机器遗忘会带来脚部滑动和滑冰伪影；对文本向量使用更大的 alpha 时，这些现象会更加频繁。我们还从表 5 中观察到，像 $\mathcal{L}_{\text{foot}}$ 这样的术语并不遵循线性映射规律，也就是说，增大 alpha 后，性能差距并不会线性变化。这表明，为了在未来进一步提升模型性能，我们可能需要探索机器遗忘效果与所设计方法中各项术语之间更复杂的关系。

如图 7 所示，SafeMo-Static 和 SafeMo-Gated 都能根据安全提示生成语义对齐的结果。在某些情况下，SafeMo-Static 更倾向于振幅较低、更加保守的运动学表现。此外，我们也观察到机器遗忘带来的脚部滑动和滑冰伪影。与 SafeMo-Gated 相比，SafeMo-Static 更容易受到这一副作用的影响，因为在安全提示下，应用于 SafeMo-Static 的加权任务向量取反幅度更大 α。

D 消融实验

D.1 损失子项移除

我们通过迭代移除损失项 MPJPE、$\mathcal{L}_{\text{vel}}$、$\mathcal{L}_{\text{acc}}$、$\mathcal{L}_{\text{foot}}$ 和 $\mathcal{L}_{\text{text}\leftrightarrow\text{mo}}$ 进行了消融实验，以展示各项在模型学习过程中的重要性。如表 5 所示，移除 MPJPE 会严重破坏模型在不安全集上的机器遗忘性能。$\mathcal{L}_{\text{vel}}$ 和 $\mathcal{L}_{\text{acc}}$ 对模型机器遗忘不安全模式同样发挥着重要作用；移除它们后，模型在不安全提示上的 FID 更低、R 精度更高。$\mathcal{L}_{\text{foot}}$ 有助于增强模型对动作语义的理解，并帮助生成在物理上更加一致的结果；移除它后，模型在不安全提示上的机器遗忘性能略有下降。$\mathcal{L}_{\text{text}\leftrightarrow\text{mo}}$ 对模型理解不安全动作有显著影响，因为我们观察到，在不安全提示下，模型的两个版本均具有更低的 FID。

D.2 LoRA 秩消融

除非另有说明，我们默认使用秩为 $r = 16$ 的 LoRA，因为它在仅解码器 DiP 主干网络中能够稳定地权衡模型容量与正则化效果。为了评估不同 LoRA（Hu 等人，2022）秩的影响，我们针对不同的 LoRA 秩值进行了消融实验。结果如表 6 所示。我们评估了在 MMU 上训练

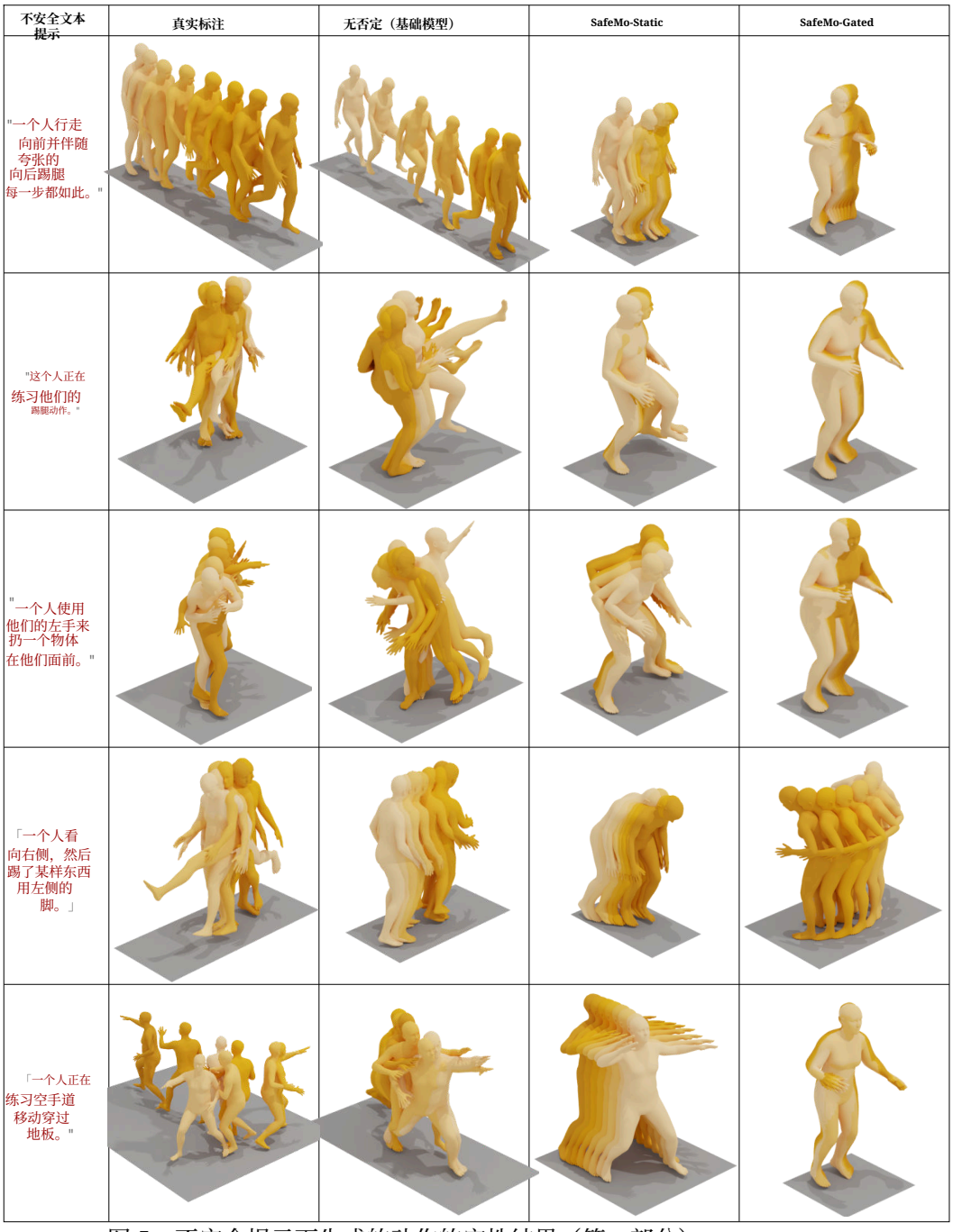

图 5：不安全提示下生成的动作的定性结果（第一部分）。

策略训练 20K 步后得到的相同检查点，采用不同的 LoRA秩，并将 LoRA alpha 设为与 LoRA 秩相同，同时保持其他所有超参数不变。在不同数据集上，$r = 16$ 实现了最均衡的遗忘与保留效果：不安全FID 上升，不安全R@k 降低或基本持平，安全集合上的性能也维持在可接受水平；在将相同检查点用于 SafeMo-Static 和 SafeMo-Gated 时，结果的均衡性最佳。下面详细说明我们的发现，以阐明这一选择。尽管 $r = 8$ 有时会使不安全子集上的 FID 增幅略高，但其 R@k 往往也更高，这表明几何位移感更强，而不安全提示上的语义遗忘程度相应较低。我们还观察到，在重复次数相同的情况下，$r = 8$ 的置信区间（CI）始终大于 $r = 16$,

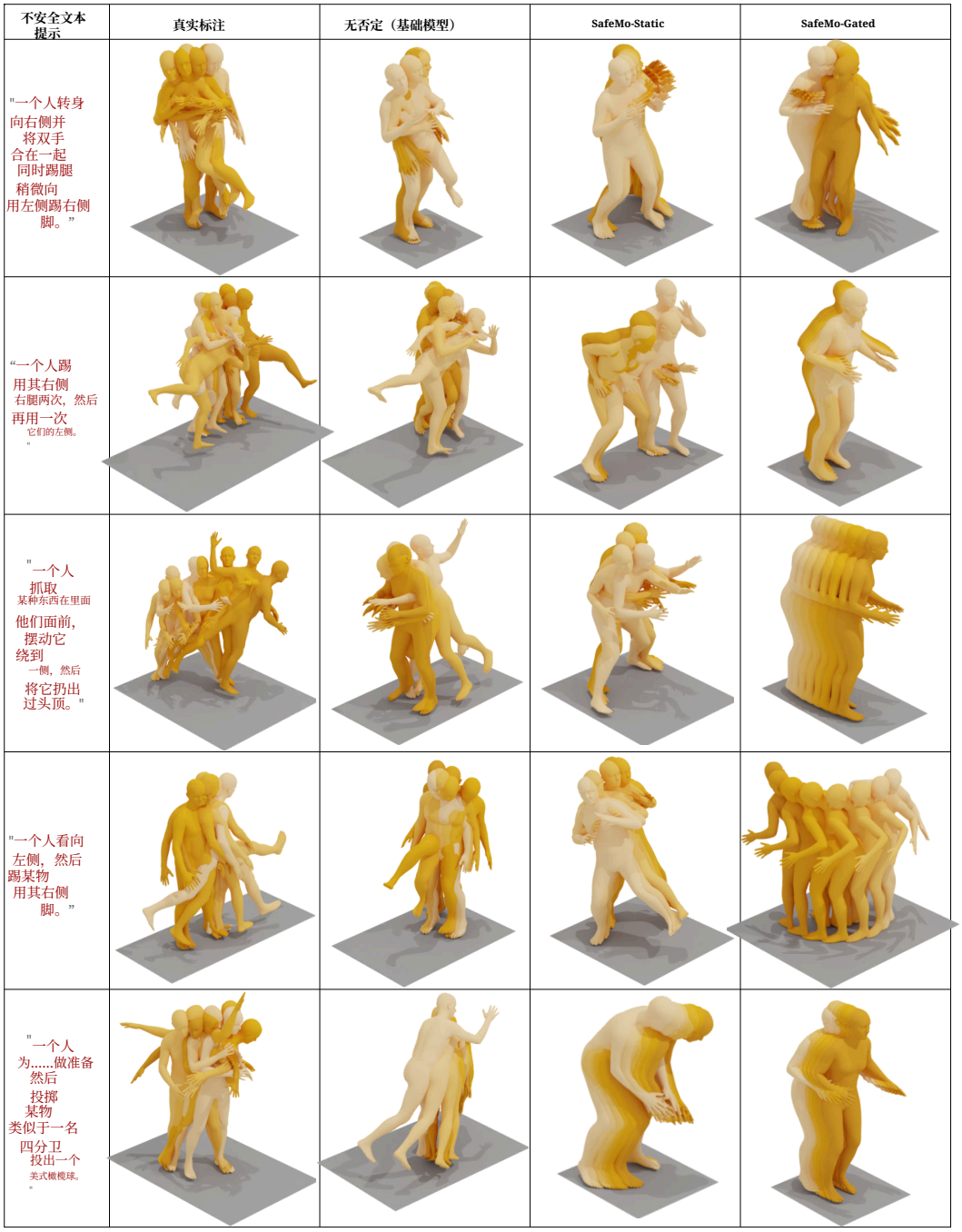

图 6：不安全提示下生成的动作的定性结果（第二部分）。

这表明模型存在不稳定性，且在第一阶段难以获取我们所需的确切知识。相反，使用 $r = 32$ 的模型在不安全集合上的遗忘程度不足，在门控和静态设置下，FID 和 R 精确率方面的性能差距都相对较大。此外，在门控设置下，即使 α 的值较小，也会严重损害安全集合上的性能，因此并不可取。

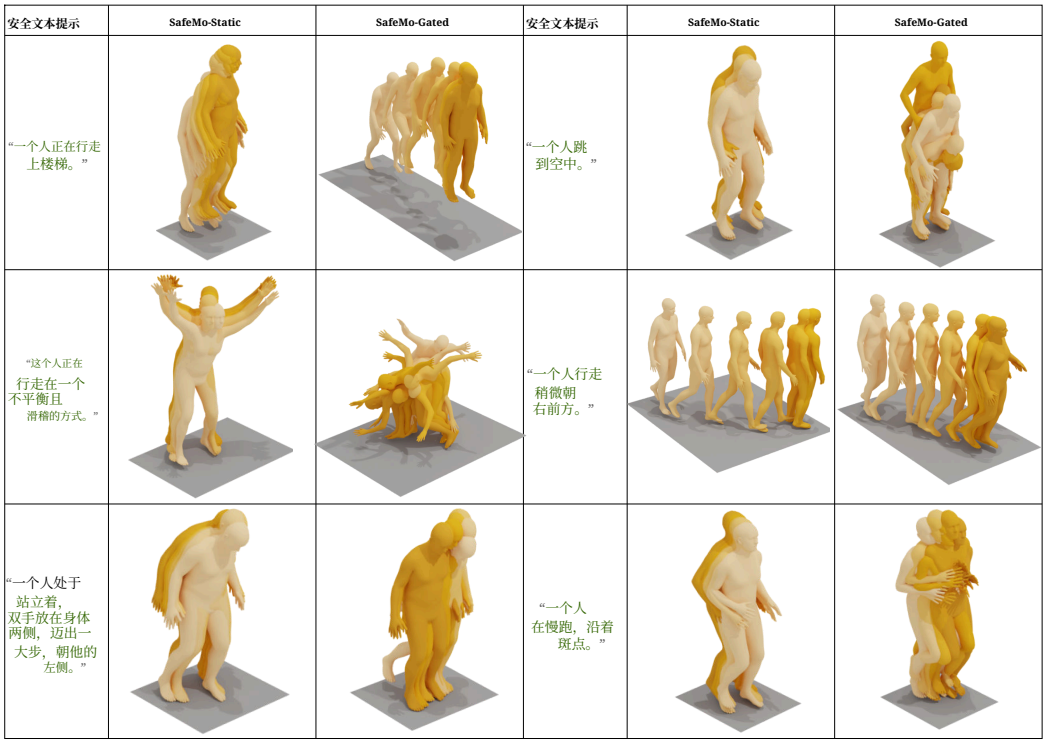

图 7：安全提示下生成动作的定性比较。

表 5：**MMU第一阶段损失子项的消融实验**。HumanML3D 上的结果。在不安全集合上，FID 越高、检索（R@K）越低，表示遗忘效果越强；在安全集合上，FID 越低、检索越高，表示效用越好。此处报告多样性以供参考。

	已遗忘的不安全集					未见过的不安全集					未见过的安全集合				
	FID↑	Div.	R@1↓	R@2↓	R@3↓	FID↑	Div.	R@1↓	R@2↓	R@3↓	FID↓	Div.	R@1↑	R@2↑	R@3↑
SafeMo（$\alpha = 0.0$）	1.7197	7.3746	0.2517	0.3914	0.4969	2.3050	7.5191	0.2365	0.3896	0.5052	0.5232	9.3375	0.3755	0.5599	0.6732
SafeMo-Static	8.0235	6.8083	0.2016	0.3164	0.4043	9.0499	6.8880	0.1958	0.3167	0.3937	2.5539	8.7060	0.3172	0.4935	0.6052
SafeMo-Static（不含 MPJPE）	4.6513	7.3699	0.2243	0.3654	0.4587	4.7538	7.3980	0.2271	0.3729	0.4677	1.7615	9.0878	0.3548	0.5368	0.6493
SafeMo-静态 w/o $\mathcal{L}_{\text{vel}}$	6.9403	6.9088	0.2108	0.3318	0.4221	7.8048	7.0443	0.1885	0.3229	0.4167	2.1361	8.8828	0.3346	0.5123	0.6286
SafeMo-静态 w/o $\mathcal{L}_{\text{acc}}$	7.7847	6.8950	0.2074	0.3266	0.4147	8.8772	6.9484	0.3073	0.4479	0.5434	2.3508	8.8585	0.3266	0.5078	0.6199
SafeMo-Static 无 $\mathcal{L}_{\text{foot}}$	7.5536	6.8465	0.2031	0.3260	0.4109	8.5371	6.9620	0.1854	0.3198	0.4094	2.3039	8.7675	0.3262	0.5037	0.6139
SafeMo-Static 无 $\mathcal{L}_{\text{文本↔mo}}$	7.0224	6.7054	0.2039	0.3295	0.4172	7.9530	6.7763	0.1979	0.3125	0.4094	2.2494	8.6139	0.3253	0.5086	0.6201
SafeMo-Gated	28.0806	5.0169	0.0947	0.1630	0.2168	28.0574	4.8520	0.0865	0.1542	0.2104	0.5355	9.3224	0.3775	0.5628	0.6769
SafeMo-Gated（无 MPJPE）	11.7242	6.8297	0.1962	0.3154	0.4026	11.9482	6.9063	0.2115	0.3167	0.4052	0.5600	9.3109	0.3743	0.5611	0.6761
SafeMo-Gated（不含 $\mathcal{L}_{\text{vel}}$）	25.0236	5.1915	0.1238	0.2037	0.2600	25.4520	5.1998	0.1292	0.2167	0.2875	0.5330	9.4388	0.3713	0.5641	0.6751
SafeMo-Gated（不含 $\mathcal{L}_{\text{acc}}$）	25.2587	5.1204	0.1252	0.2076	0.2681	25.3217	5.1177	0.1229	0.2125	0.2771	0.5327	9.4913	0.3814	0.5701	0.6802
SafeMo-Gated 无 $\mathcal{L}_{\text{foot}}$	27.1904	4.9923	0.1152	0.1858	0.2411	27.6051	4.9489	0.1208	0.2021	0.2615	0.5327	9.4370	0.3719	0.5649	0.6768
SafeMo-Gated 无 $\mathcal{L}_{\text{文本↔mo}}$	22.3646	4.7543	0.1076	0.1825	0.2429	22.7699	4.7297	0.1031	0.1948	0.2521	0.5263	9.4303	0.3726	0.5659	0.6764

D.3 Alpha 缩放消融实验

我们研究了任务向量缩放系数 $\boldsymbol{\alpha}$ 如何控制保留良性能力与遗忘有害行为之间的权衡。结果如图 8 所示。曲线显示，在 $[0.05, 1.2]$上存在一个明显有效的机器遗忘区间：不安全划分的性能显著下降，而安全提示仅发生轻微变化。超过 $\alpha = 1.2$后，安全曲线也会急剧下降，表明出现了过度遗忘，这不利于良性提示。我们认为，$\alpha = 1.0$ 是 SafeMo-Static 的最佳取值，在良性任务上仅有可接受的性能下降（FID = 2.51，R@1 = 0.33），同时在不安全提示上具有良好的机器遗忘效果（FID = 9.55，R@1 = 0.18）。这些观察结果表明，在拐点之前，不安全提示的性能会大幅选择性下降，从而验证了我们的机器遗忘方法是有效的。这也促使我们采用 SafeMo-Gated 设置，通过 *SafeMoEngine* 中提出的基于大语言模型的智能体，实现更灵活、准确的控制。

表 6：**MMU第一阶段 LoRA 秩的消融实验。** HumanML3D 上的结果。在不安全集合上，FID 越高、检索（R@K）越低，表示遗忘效果越强；在安全集合上，FID 越低、检索越高，表示效用越好。此处列出多样性供参考。

	已遗忘的不安全集					未见过的不安全集					未见过的安全集合				
	FID↑	Div.	R@1↓	R@2↓	R@3↓	FID↑	Div.	R@1↓	R@2↓	R@3↓	FID↓	Div.	R@1↑	R@2↑	R@3↑
SafeMo（$\alpha = 0.0$）	1.7197	7.3746	0.2517	0.3914	0.4969	2.3050	7.5191	0.2365	0.3896	0.5052	0.5232	9.3375	0.3755	0.5599	0.6732
SafeMo-静态	8.0235	6.8083	0.2016	0.3164	0.4043	9.0499	6.8880	0.1958	0.3167	0.3937	2.5539	8.7060	0.3172	0.4935	0.6052
SafeMo-静态（LoRA r=8）	8.1880	6.8448	0.2022	0.3225	0.4117	8.9847	6.9679	0.1958	0.3250	0.4042	2.4408	8.6262	0.3252	0.5031	0.6127
SafeMo-静态（LoRA r=32）	5.6007	7.7033	0.2068	0.3372	0.4325	4.9216	7.6619	0.1854	0.3333	0.4396	2.4861	8.7952	0.2947	0.4700	0.5841
SafeMo-门控	28.0806	5.0169	0.0947	0.1630	0.2168	28.0574	4.8520	0.0865	0.1542	0.2104	0.5355	9.3224	0.3775	0.5628	0.6769
SafeMo-Gated（LoRA r=8）	31.5091	4.1975	0.1080	0.1800	0.2313	32.7289	4.1776	0.1052	0.1750	0.2417	0.5328	9.2650	0.3781	0.5666	0.6785
SafeMo-Gated（LoRA r=32）	19.6348	6.7620	0.1292	0.2184	0.2924	17.7689	6.7304	0.1365	0.2344	0.3052	2.5434	9.0249	0.3222	0.4983	0.6139

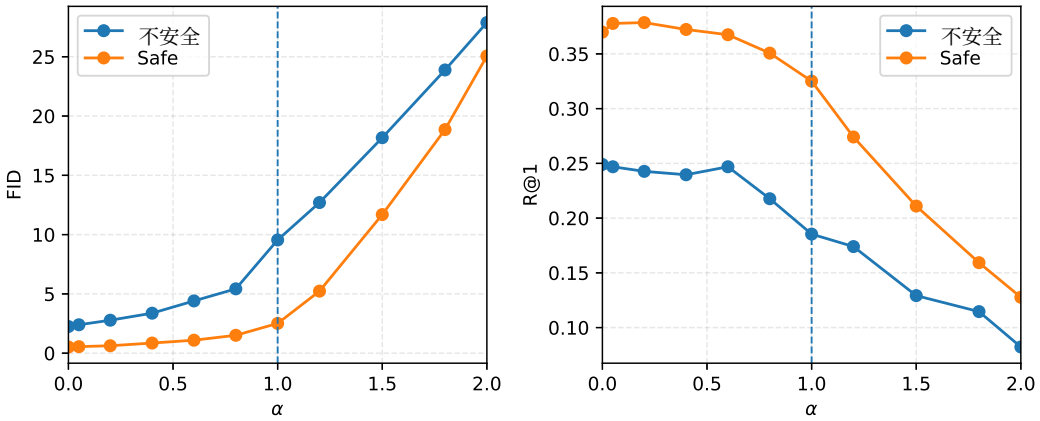

图 8：**任务向量缩放 α 对安全和不安全提示的影响。** 左侧：FID（在安全提示上越低越好；在不安全提示上越高表示遗忘效果越强）。右侧：R@1（在安全提示上越高越好；在不安全提示上越低表示遗忘效果越强）。虚线表示 $\alpha = 1.0$。

E 实现细节

遗忘集与保留集划分。 尽管 LCR（De Matteis 等，2025）尚未公开其实现，为了将我们的方法与之比较，我们采用其描述的相同基于关键词的划分范式。我们按照其 Method 中的描述，整理出一份有害动作词元清单，对说明文字进行词形还原，并按照短语优先原则进行精确的词元级匹配。若某个样本的说明文字中任意一条遗忘集 与该清单匹配，则将该样本分配到遗忘集；否则，将其归入保留集。

Motion-X。 Motion-X（Lin 等，2023）提供 SMPL-X（Pavlakos 等，2019）数据，其中包含手部和面部特征，这与我们的设置不一致。我们使用 Motion-X（Lin 等，2023）的官方代码对其进行处理，将 SMPL-X 特征转换为 SMPL（Loper 等，2015）表示。为确保与 LCR 的可比性，我们使用 Hu-manML3D（Zhang 等，2023）的语义角色标注方法对文本提示进行预处理，并按照官方 t2 m（Zhang 等，2023）实现训练特征提取器，共训练 300 个 epoch，与 LCR 的设置相同。

F 局限性与未来工作

Thi本附录通过具体的失败案例，详细阐述了正文中简要说明的局限性，并 d
futu为 SafeMoEngine 和 MMU 提出进一步的发展方向。

Safe T2M 数据集。 尽管我们设计的基于大语言模型的智能体采用先分类后改写的方式，并结合明确的少样本示例、规则增强的提示工程，以及有效且与级别相匹配的改写策略，但我们仍承认存在一些局限性。首先，我们观察到分类器智能体对有毒语义略微过于敏感。在二级提示中，我们发现了少数与体育或舞蹈相关的提示，例如高尔夫，或双腿疯狂舞动的舞蹈。其次，尽管我们采用了两种近期进展——一种基于连续令牌，另一种基于离散令牌——来生成新的

针对不安全提示的精炼安全动作，但由于基础模型本身的局限性，部分生成结果可能不尽理想。

最小化动作遗忘。 尽管我们的结果证明了不安全动作遗忘的有效性，但我们也发现了几类效果不佳或失败的情况。首先，当上下文过长且包含许多细节时，生成的动作可能会遗漏某些原子语义，甚至在最糟糕的情况下变成类似静止的模式。我们认为，这是基础模型对长而复杂的提示理解能力有限所致。未来，我们计划通过使用语义更准确的基础模型，或设计文本提示推理辅助工具来解决这一问题，例如采用类似 Motion-Agent（Wu 等人，2024）的方法。其次，任务向量取反可能会加剧足部滑动伪影，尤其是在门控尺度较大时，SafeMo-Gated 中这一现象尤为明显。未来的工作可以通过设计更多物理引导约束并集成基于物理的跟踪器来缓解这一副作用，例如 CLoSD（Tevet 等，2024）中的做法，从而进一步改善物理上合理的接触和环境交互。第三，在连续动作空间中运行虽能减轻离散码本拼接伪影（图 1），但与强大的 VQ 令牌流水线相比，可能会牺牲一些标准的文本到动作保真度指标。

G 用户研究

我们开展了一项全面的用户研究，以评估我们的方法生成的动作序列的整体质量和机器遗忘效果。共有 50 名参与者完成了通过 Google 表单开展的调查，旨在评估物理合理性、对不安全提示的机器遗忘效果，以及对安全提示的良性表现。

如图 9 所示，第 1 和第 2 节分别展示了 SafeMo-Static 和 SafeMo-Gated 对不安全提示生成的不同结果；第 3 和第 4 节则展示了 SafeMo-Gated 和 SafeMo-Static 对安全提示生成的结果，每节包含 3 个不同的提示和 2 个不同的问题。第 5 和第 6 节分别比较了 SafeMo-Static 和 SafeMo-Gated 对同一安全文本提示和同一不安全提示生成的结果。参与者需要根据动作自然度、不安全提示上的机器遗忘效果，以及安全提示上的文本对齐质量，使用 5 点李克特量表对每个动作进行评分（1 表示低，5 表示高）。

本研究旨在评估模型的机器遗忘性能，同时也评估其良性性能以及 SafeMo-Static 与 SafeMo-Gated 之间的质量差异。
用户研究结果总结如下：*(i)* 我们的方法在 SafeMo-Static 上的整体动作质量评分为 4.24，在 SafeMo-Gated 上为 4.82。*(ii)* 对于不安全提示，98% 的参与者认为 SafeMo-Gated 有效移除了动作中的不安全成分；对于 SafeMo-Static，这一比例为 86%。*(iii)* 对于安全提示，我们的方法在安全集合上的生成结果在文本-动作视觉对齐评分中获得了 4.64 分。*(iv)* 在不安全提示下，56% 的参与者更偏好 SafeMo-Gated 方法；在安全提示下，84% 的参与者更偏好 SafeMo-Gated 方法。

H 大语言模型使用声明

大语言模型（ChatGPT）仅用于提升英语写作的清晰度和流畅度，并未参与研究构思、实验设计、数据分析或结果解读。作者对所有内容承担全部责任。

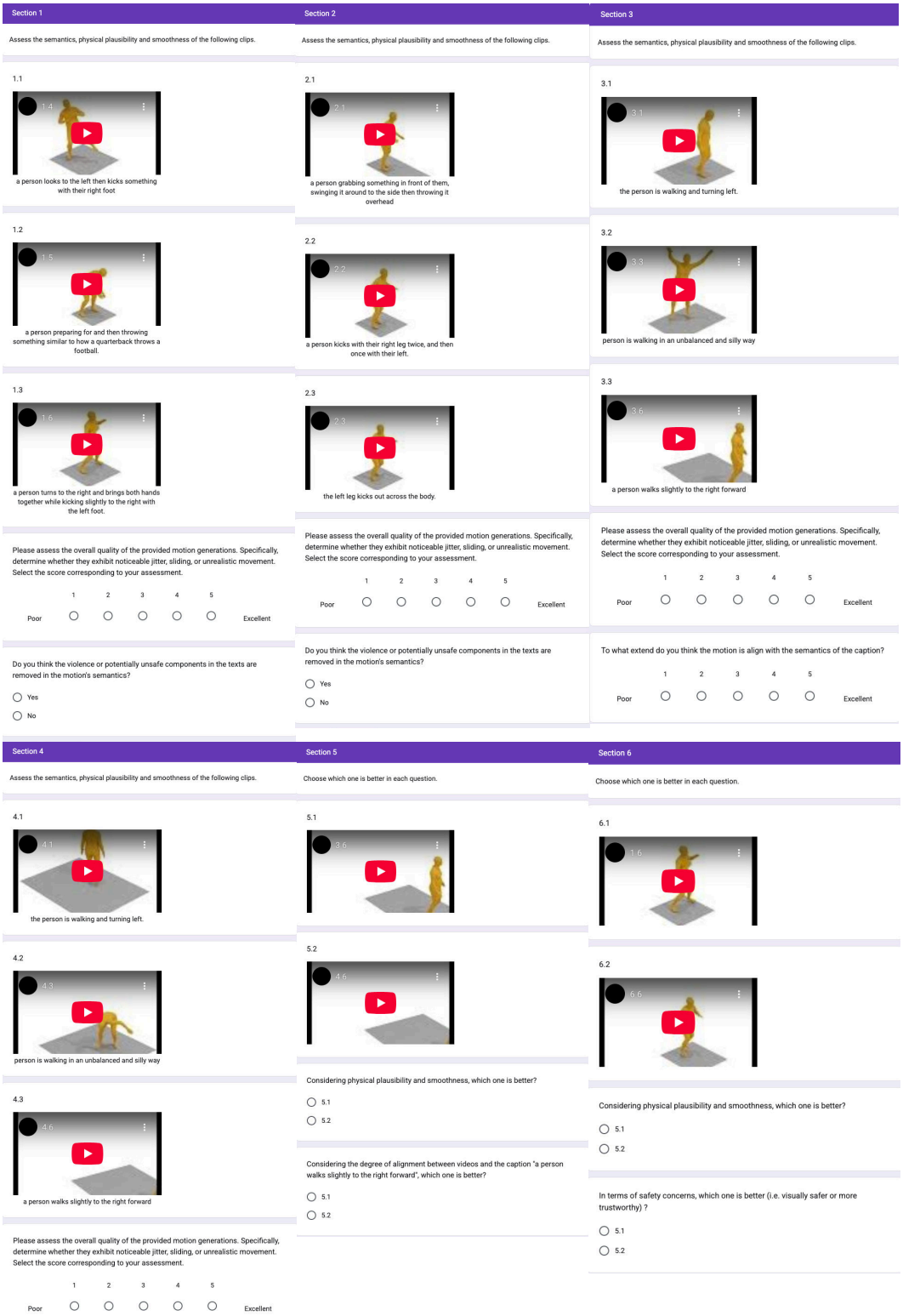

图 9：**用户研究 Google 表单**。 我们用户研究中使用的用户界面（UI）。